HIERARCHICAL OBJECT REPRESENTATIONS IN THE VISUAL CORTEX AND COMPUTER VISION

EDITED BY: Antonio Rodríguez-Sánchez, Mazyar Fallah and Ales Leonardis
PUBLISHED IN: Frontiers in Computational Neuroscience



frontiers Research Topics



frontiers

Frontiers Copyright Statement

© Copyright 2007-2016 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714
ISBN 978-2-88919-798-9
DOI 10.3389/978-2-88919-798-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

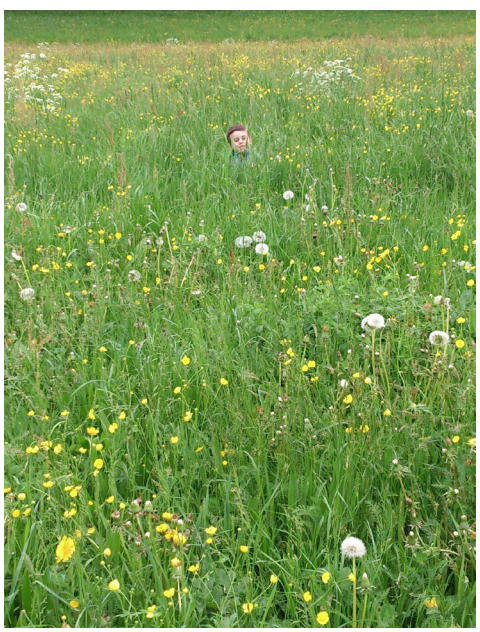
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

HIERARCHICAL OBJECT REPRESENTATIONS IN THE VISUAL CORTEX AND COMPUTER VISION

Topic Editors:
Antonio Rodríguez-Sánchez, University of Innsbruck, Austria
Mazyar Fallah, York University, Canada,
Ales Leonardis, University of Birmingham, UK



Attention is an important mechanism that filters irrelevant information in order to focus on important areas of the visual scene.
Image by Antonio Rodriguez-Sanchez.

Cover: On/Off neurons: Sparse representation: Different stimuli elicit only a small fraction of neurons from a population to be active. Image by Antonio Rodriguez-Sanchez.

Over the past 40 years, neurobiology and computational neuroscience has proved that deeper understanding of visual processes in humans and non-human primates can lead to important advancements in computational perception theories and systems. One of the main difficulties that arises when designing automatic vision systems is developing a mechanism that can recognize - or simply find - an object when faced with all the possible variations that may occur in a natural scene, with the ease of the primate visual system. The area of the brain in primates that is dedicated at analyzing visual information is the visual cortex. The visual cortex performs a

wide variety of complex tasks by means of simple operations. These seemingly simple operations are applied to several layers of neurons organized into a hierarchy, the layers representing increasingly complex, abstract intermediate processing stages.

In this Research Topic we propose to bring together current efforts in neurophysiology and computer vision in order 1) To understand how the visual cortex encodes an object from a starting point where neurons respond to lines, bars or edges to the representation of an object at the top of the hierarchy that is invariant to illumination, size, location, viewpoint, rotation and robust to occlusions and clutter; and 2) How the design of automatic vision systems benefit from that knowledge to get closer to human accuracy, efficiency and robustness to variations.

Citation: Rodríguez-Sánchez, A., Fallah, M., Leonardis, A., eds. (2016). Hierarchical Object Representations in the Visual Cortex and Computer Vision. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88919-798-9

Table of Contents

06 ***Editorial: Hierarchical Object Representations in the Visual Cortex and Computer Vision***
Antonio J. Rodríguez-Sánchez, Mazyar Fallah and Aleš Leonardis

Chapter: Visual architectures and their mechanisms

10 ***Feedforward object-vision models only tolerate small image variations compared to human***
Masoud Ghodrati, Amirhossein Farzmahdi, Karim Rajaei, Reza Ebrahimpour and Seyed-Mahdi Khaligh-Razavi

27 ***Why vision is not both hierarchical* and *feedforward***
Michael H. Herzog and Aaron M. Clarke

32 ***The proactive brain and the fate of dead hypotheses***
Amir Tal and Moshe Bar

38 ***Combining segmentation and attention: a new foveal attention model***
Rebeca Marfil, Antonio J. Palomino and Antonio Bandera

55 ***Object recognition with hierarchical discriminant saliency networks***
Sunhyoung Han and Nuno Vasconcelos

Chapter: The role of ventral and dorsal streams

75 ***A conceptual framework of computations in mid-level vision***
Jonas Kubilius, Johan Wagemans and Hans P. Op de Beeck

94 ***Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation***
Stephan Tschechne and Heiko Neumann

114 ***Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective COSFIRE models***
George Azzopardi and Nicolai Petkov

123 ***Visual dictionaries as intermediate features in the human brain***
Kandan Ramakrishnan, H. Steven Scholte, Iris I. A. Groen, Arnold W. M. Smeulders and Sennay Ghebreab

133 ***Complex cells decrease errors for the Müller-Lyer illusion in a model of the visual ventral stream***
Astrid Zeman, Oliver Obst and Kevin R. Brooks

142 ***Exploration of complex visual feature spaces for object perception***
Daniel D. Leeds, John A. Pyles and Michael J. Tarr

159 ***A neuromorphic system for video object recognition***
Deepak Khosla, Yang Chen and Kyungnam Kim

171 ***Shape representations in the primate dorsal visual stream***
Tom Theys, Maria C. Romero, Johannes van Loon and Peter Janssen

180 ***Modeling the shape hierarchy for visually guided grasping***
Omid Rezai, Ashley Kleinhans, Eduardo Matallanas, Ben Selby and Bryan P. Tripp

193 ***Feature integration and object representations along the dorsal stream visual hierarchy***
Carolyn Jeane Perry and Mazyar Fallah

Chapter: Learning and sparse coding

210 ***Sparsey™: event recognition via deep hierarchical sparse distributed codes***
Gerard J. Rinkus

254 ***Deformation-specific and deformation-invariant visual object recognition: pose vs. identity recognition of people and deforming objects***
Tristan J. Webb and Edmund T. Rolls

270 ***Competition improves robustness against loss of information***
Arash Kermani Kolankeh, Michael Teichmann and Fred H. Hamker

282 ***From image processing to computational neuroscience: a neural model based on histogram equalization***
Marcelo Bertalmío

frontiers in Computational Neuroscience

EDITORIAL
published: 20 November 2015
doi: 10.3389/fncom.2015.00142

Editorial: Hierarchical Object Representations in the Visual Cortex and Computer Vision

Antonio J. Rodríguez-Sánchez[1]*, *Mazyar Fallah*[2] *and Aleš Leonardis*[3]

[1] *Intelligent and Interactive Systems, Department of Computer Science, University of Innsbruck, Innsbruck, Austria,* [2] *Visual Perception and Attention Laboratory, Centre for Vision Research, School of Kinesiology and Health Science, York University, Toronto, ON, Canada,* [3] *School of Computer Science, University of Birmingham, Birmingham, UK*

Keywords: computer model, neurophysiology, computer vision, visual cortex, computational neurosciences

Over the past 40 years, Neurobiology and Computational Neuroscience have proved that deeper understanding of visual processes in humans and non-human primates can lead to important advancements in computational perception theories and systems. One of the main difficulties that arises when designing automatic vision systems is developing a mechanism that can recognize—or simply find—an object when faced with all the possible variations that may occur in a natural scene, and with the ease of the primate visual system. The area of the brain in primates that is dedicated to analyzing visual information is the visual cortex. The visual cortex performs a wide variety of complex tasks by means of seemingly simple operations. These operations are applied to several layers of neurons organized into a hierarchy, the layers representing increasingly complex, abstract intermediate processing stages.

In this research topic we propose to bring together current efforts in Neurophysiology and Computer Vision in order to better understand (1) How the visual cortex encodes an object from a starting point where neurons respond to lines, bars or edges to the representation of an object at the top of the hierarchy that is invariant to illumination, size, location, viewpoint, rotation and robust to occlusions and clutter; and (2) How the design of automatic vision systems benefits from that knowledge to get closer to human accuracy, efficiency and robustness to variations. In fact, the primate visual system has influenced computer vision systems for decades now since Hubel and Wiesel (1968) simple and complex cells inspired the Neocognitron (Fukushima, 1980). Since then, studies about the primate and human visual systems led the way to many more works on biologically-inspired computational vision, such as Tsotsos et al. (1995); Olshausen and Field (1996); Booth and Rolls (1998); Riesenhuber and Poggio (1999); Rodríguez-Sánchez and Tsotsos (2011), to name a few.

The answers to these issues bring hypotheses that are partially addressed in this research topic, raising additional new questions:

1. What are the mechanisms involved in these visual architectures? What are the limitations of feedforward connections? When is feedback and top-down priming necessary? The classical way of seeing feedback connections is for the enhancement of neural responses through top-down attentive processes (Moran and Desimone, 1985; Rodríguez-Sánchez et al., 2006; Perry et al., 2015). But lately, other studies support a role of feedback connections related to cell selectivity through recurrent networks (Neumann and Sepp, 1999; Angelucci and Bressloff, 2006).
2. The ventral stream areas (V1, V2, V4, inferotemporal cortex) have usually been considered to be the ones involved in object recogntion and the subject of several existing models (Serre et al., 2006; Rodríguez-Sánchez and Tsotsos, 2012). But, also recently, there are new findings that relate the dorsal stream with that same task (Konen and Kastner, 2008; Perry and Fallah, 2012). What are the differences between how objects are processed in the ventral and the dorsal streams? Which areas are involved in recognition and which in localization?

OPEN ACCESS

Edited by:
Si Wu,
Beijing Normal University, China

Reviewed by:
Da-Hui Wang,
Beijing Normal University, China

***Correspondence:**
Antonio J. Rodríguez-Sánchez
antonio.rodriguez-sanchez@uibk.ac.at

Received: *21 August 2015*
Accepted: *06 November 2015*
Published: *20 November 2015*

Citation:
Rodríguez-Sánchez AJ, Fallah M and Leonardis A (2015) Editorial: Hierarchical Object Representations in the Visual Cortex and Computer Vision.
Front. Comput. Neurosci. 9:142.
doi: 10.3389/fncom.2015.00142

3. And finally, how much is learned and how much is genetically implemented (Rodríguez-Sánchez and Piater, 2014)? Even more, what is the relation between learning, sparse coding, selectivity and diversity (Olshausen and Field, 1996; Xiong et al., 2015) and how different learning strategies compare?

We present a total of 19 papers related to those questions. The following five papers deal with the questions related to visual architectures and their mechanisms. Ghodrati et al. (2014) studied whether recent relative successes in object recognition on various image datasets based on sparse representations applied in a feedforward fashion represented a breakthrough in invariant object recognition. In their study they showed, using a carefully designed parametrically controlled image database consisting of several object categories, that these approaches fail when the complexity of image variations is high and that their performance is still poor compared to humans. This suggests that learning sparse informative visual features may be one of the necessary components but definitely not a complete solution for a human-like object recognition system. A classical feedforward filtering approach is also challenged in the paper by Herzog and Clarke (2014), where the authors provided ample evidence, stemming from experiments from crowding research, to support their arguments that the computations are not purely local and feedforward, but rather global and iterative. On the same topic, Tal and Bar (2014) explored the role of top-down mechanisms which bias the processing of the incoming visual information and facilitate fast and robust recognition. This work specifically addresses the question of what happens to initial predictions that eventually get rejected in a competitive selection process. The work by Marfil et al. (2014) brings into focus another important aspect of biological visual sytems, namely attention. The authors studied a bidirectional relationship between segmentation and attention processes. They presented a bottom-up foveal attention model that demonstrates how the attention process influences the selection of the next position of the fovea and how segmentation, in turn, guides the extraction of units of attention. In Han and Vasconcelos (2014) the authors also researched the role of attention models, but this time in connection to object recognition. Using their recognition model, hierarchical discriminant saliency network (HDSN), they clearly demonstrated the benefits of integrating attention and recognition.

We provide an interesting discussion on the role of ventral and dorsal streams with a total of 10 articles. Kubilius et al. (2014) discusses the importance of surface representation and reviews recent work on mid-level visual areas in the ventral stream. We include here two models of shape related to those intermediate visual areas. The first approach is a recurrent network that achieves figure-ground segregation by assigning border ownership through the interaction between feedforward and feedback inputs (Tschechne and Neumann, 2014). The second approach is a trainable set of shape detectors that can be applied as a filter bank to recognize letters and keywords as well finding objects in complex scenes (Azzopardi and Petkov, 2014). The question that arises regarding computational models is of course, how faithful they are? This is what Ramakrishnan et al. (2015) answers by comparing the fMRI responses from 20 subjects to two different types of computer vision models: the classical bag of words and the biologically-inspired HMAX. HMAX is also the subject of study in Zeman et al. (2014), here the authors use that model to compare the robustness of complex cells to simple cells in the Müller-Lyer illusion. The final stage in the object recognition pathway is the inferotemporal cortex (IT), Leeds et al. (2014) present an fMRI study that tries to answers the problem of how starting from simple edge-like features in V1 we obtain neurons at the top of the hierarchy that respond to complex features as parts, textures or shapes. Using feed-forward object detection and classification modeling, Khosla et al. (2014) developed a neuromorphic system that also efficiently produces automated video object recognition. However, the visual system is not limited to only detecting objects, but can also detect the spatial relationships between objects and even between parts of the same object. The dorsal stream areas are thus also important for object representation with a focus on action via effectors such as the eyes or the hand. Theys et al. (2014) reviews how 3D shape for grasping is processed along the dorsal stream, focusing on the representations in the anterior intraparietal area (AIP) and ventral premotor cortex (PMv). Rezai et al. (2014) advances this by modeling the curvature and gradient input from the caudal intraparietal area (CIP) to visual neurons in AIP, using superquadric fits—used in robotics for grasp planning—or Isomap dimension reductions of object surface distances. They found that both models fit responses from primate AIP neurons. However, Isomaps better approximated the feedforward input from CIP making it the more promising model of how the dorsal stream produces shape representations for grasping. Yet the features used for grasping are only a subset of an object's features. While the integration of features along the ventral stream to form object representations is well-known, Perry and Fallah (2014) review recent findings supporting dorsal stream object representations and propose a framework for the integration of features along the dorsal stream.

Finally, four papers address the problem of learning and sparse coding. Rinkus (2014) shows that a hierarchical sparse distributed code network provides the foundation for the storage and retrieval of associative memory on top of building up an object representation. The end point of object processing is recognition, which the human visual system is very efficient at and many computational models are based upon. Webb and Rolls (2014) investigated how recognition of the identity of individuals and their poses can be separated. They showed that a model of the ventral visual system using temporal continuity, VisNet, can through learning develop pose-specific and identity-specific representations that are invariant to the other factor. In their biologically inspired study, Kermani Kolankeh et al. (2015) researched different computational principles (sparse coding, biased competition, Hebbian learning) capable of developing receptive fields comparable to those of V1 simple-cells and discovered that methods which employ competitive mechanisms achieve higher levels of robustness against loss of information which may be important to achieve better performance on classification tasks. While these studies have focused on using

biologically-inspired visual processing in computational models, Bertalmío (2014) worked in reverse by taking an image processing technique used for local histogram equalization and applying it to a neural activity model. The resultant model predicts spectrum whitening, contrast enhancement and lightness induction, all behavioral aspects of visual processing. Time will tell if neuronal studies bear out this process.

We are bringing together two seemingly different disciplines: Neuroscience and Computer Vision. We show in this research topic that each one can benefit from the other. The latter can aid Neuroscience for testing hypotheses regarding the visual cortex in a non-invasive way, or otherwise when we reach technical limitations, e.g., how the information flows along the visual architectures (see Rodríguez-Sánchez, 2010 for a recent example). On the other hand, Computer Vision can benefit from Neuroscience in order to develop better, more robust, efficient and general systems than the ones present to date (Krüger et al., 2013).

Due to the complexity of vision (Tsotsos, 1987), objects/locations are considered to *compete* for the visual system's resources. The studies presented here show that—among other aspects—feedforward hierarchies are insufficient, supporting the need for top-down priming or attention. The interaction between feedforward and feedback inputs have an impact in neural encoding as shown in the models presented in this research topic. Not only competition, sparsity is another important mechanism. The aim is achieving efficient codes that represent and store object classes efficiently into memory since not every possible combination of features/parameters is feasible to be stored. Finally, a number of studies stress on the importance of the dorsal stream in shape and identity-object representation in order to interact with specific objects, e.g., grasping.

REFERENCES

Angelucci, A., and Bressloff, P. C. (2006). Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog. Brain Res.* 154, 93–120. doi: 10.1016/S0079-6123(06)54005-1

Azzopardi, G., and Petkov, N. (2014). Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective cosfire models. *Front. Comput. Neurosci.* 8:80. doi: 10.3389/fncom.2014.00080

Bertalmío, M. (2014). From image processing to computational neuroscience: a neural model based on histogram equalization. *Front. Comput. Neurosci* 8:71. doi: 10.3389/fncom.2014.00071

Booth, M., and Rolls, E. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523. doi: 10.1093/cercor/8.6.510

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36, 193–202. doi: 10.1007/BF00344251

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., and Khaligh-Razavi, S. M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Front. Comput. Neurosci.* 8:74. doi: 10.3389/fncom.2014.00074

Han, S., and Vasconcelos, N. (2014). Object recognition with hierarchical discriminant saliency networks. *Front. Comput. Neurosci.* 8:109. doi: 10.3389/fncom.2014.00109

Herzog, M. H., and Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Front. Comput. Neurosci.* 8:135. doi: 10.3389/fncom.2014.00135

Hubel, D., and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243. doi: 10.1113/jphysiol.1968.sp008455

Kermani Kolankeh, A., Teichmann, M., and Hamker, F. H. (2015). Competition improves robustness against loss of information. *Front. Comput. Neurosci.* 9:35. doi: 10.3389/fncom.2015.00035

Khosla, D., Chen, Y., and Kyungnam, K. (2014). A neuromorphic system for video object recognition. *Front. Comput. Neurosci.* 8:147. doi: 10.3389/fncom.2014.00147

Konen, C. S., and Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nat. Neurosci.* 11, 224–231. doi: 10.1038/nn2036

Krüger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., et al. (2013). Deep hierarchies in the primate visual cortex: what can we learn for computer vision? *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1847–1871. doi: 10.1109/TPAMI.2012.272

Kubilius, J., Wagemans, J., and Op de Beeck, H. P. (2014). A conceptual framework of computations in mid-level vision. *Front. Comput. Neurosci* 8:158. doi: 10.3389/fncom.2014.00158

Leeds, D. D., Pyles, J. A., and Tarr, M. J. (2014). Exploration of complex visual feature spaces for object perception. *Front. Comput. Neurosci.* 8:106. doi: 10.3389/fncom.2014.00106

Marfil, R., Palomino, A. J., and Bandera, A. (2014). Combining segmentation and attention: a new foveal attention model. *Front. Comput. Neurosci.* 8:96. doi: 10.3389/fncom.2014.00096

Moran, J., and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784. doi: 10.1126/science.4023713

Neumann, H., and Sepp, W. (1999). Recurrent V1–V2 interaction in early visual boundary processing. *Biol. Cybernet.* 81, 425–444. doi: 10.1007/s004220050573

Olshausen, B., and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Perry, C. J., and Fallah, M. (2012). Color improves speed of processing but not perception in a motion illusion. *Front. Psychol.* 3:92. doi: 10.3389/fpsyg.2012.00092

Perry, C. J., and Fallah, M. (2014). Feature integration and object representations along the dorsal stream visual hierarchy. *Front. Comput. Neurosci.* 8:84. doi: 10.3389/fncom.2014.00084

Perry, C. J., Sergio, L. E., Crawford, J. D., and Fallah, M. (2015). Hand placement near the visual stimulus improves orientation selectivity in V2 neurons. *J. Neurophysiol.* 113, 2859–2870. doi: 10.1152/jn.00919.2013

Ramakrishnan, K., Scholte, H. S., Groen, I. I. A., Smeulders, A. W., and Ghebreab, S. (2015). Visual dictionaries as intermediate features in the human brain. *Front. Comput. Neurosci.* 8:168. doi: 10.3389/fncom.2014.00168

Rezai, O., Kleinhans, A., Matallanas, E., Selby, B., and Tripp, B. P. (2014). Modeling the shape hierarchy for visually guided grasping. *Front. Comput. Neurosci.* 8:132. doi: 10.3389/fncom.2014.00132

Riesenhuber, M., and Poggio, T. (1999). Are cortical models really bound by the "binding problem"? *Neuron* 24, 87–93. doi: 10.1016/S0896-6273(00)80824-7

Rinkus, G. J. (2014). Sparsey^{TM}: event recognition via deep hierarchical sparse distributed codes. *Front. Comput. Neurosci.* 8:160. doi: 10.3389/fncom.2014.00160

Rodríguez-Sánchez, A. (2010). *Intermediate Visual Representations for Attentive Recognition Systems*. PhD thesis, York University, Department of Computer Science and Engineering.

Rodríguez-Sánchez, A., and Tsotsos, J. (2011). "The importance of intermediate representations for the modeling of 2D shape detection: endstopping and curvature tuned computations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, CO), 4321–4326. doi: 10.1109/cvpr.2011.5995671

Rodríguez-Sánchez, A. J., and Piater, J. (2014). "Models of the visual cortex for object representation: learning and wired approaches," in

Brain-Inspired Computing, Vol. 8603 of *Lecture Notes in Computer Science*, eds L. Grandinetti, T. Lippert, and N. Petkov (Springer International Publishing), 51–62.

Rodríguez-Sánchez, A. J., Simine, E., and Tsotsos, J. K. (2006). "Feature conjunctions in visual search," in *Artificial Neural Networks (ICANN)*, eds S. Kollias, A. Stafylopatis, W. Duch, and E. Oja (Athens: Springer), 498–507. doi: 10.1007/11840930_52

Rodríguez-Sánchez, A. J., and Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. *PLoS ONE* 7:e42058. doi: 10.1371/journal.pone.0042058

Serre, T., Oliva, A., and Poggio, T. (2006). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Tal, A., and Bar, M. (2014). The proactive brain and the fate of dead hypotheses. *Front. Comput. Neurosci.* 8:138. doi: 10.3389/fncom.2014.00138

Theys, T., Romero, M. C., van Loon, J., and Janssen, P. (2014). Shape representations in the primate dorsal visual stream. *Front. Comput. Neurosci.* 8:43. doi: 10.3389/fncom.2015.00043

Tschechne, S., and Neumann, H. (2014). Hierarchical representation of shapes in visual cortex - from localized features to figural shape segregation. *Front. Comput. Neurosci.* 8:93. doi: 10.3389/fncom.2014.00093

Tsotsos, J. K. (1987). A complexity level analysis of immediate vision. *Int. J. Comput. Vis.* 1, 303–320.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., and Nuflo, F. (1995). Modeling visual-attention via selective tuning. *Artif. Intell.* 78, 507–545. doi: 10.1007/BF00133569

Webb, T. J., and Rolls, E. T. (2014). Deformation-specific and deformation-invariant visual object recognition: pose vs. identity recognition of people and deforming objects. *Front. Comput. Neurosci.* 8:37. doi: 10.3389/fncom.2014.00037

Xiong, H., Rodríguez-Sánchez, A. J., Szedmak, S., and Piater, J. (2015). Diversity priors for learning early visual features. *Front. Comput. Neurosci.* 9:104. doi: 10.3389/fncom.2015.00104

Zeman, A., Obst, O., and Brooks, K. R. (2014). Complex cells decrease errors for the Müller-Lyer illusion in a model of the visual ventral stream. *Front. Comput. Neurosci.* 8:112. doi: 10.3389/fncom.2014.00112

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Rodríguez-Sánchez, Fallah and Leonardis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 18 July 2014
doi: 10.3389/fncom.2014.00074

Feedforward object-vision models only tolerate small image variations compared to human

Masoud Ghodrati[1,2,3]*, Amirhossein Farzmahdi[1,2,4], Karim Rajaei[1,2], Reza Ebrahimpour[1,2] and Seyed-Mahdi Khaligh-Razavi[5]*

[1] *Brain and Intelligent Systems Research Laboratory, Department of Electrical and Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran*
[2] *School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[3] *Department of Physiology, Monash University, Melbourne, VIC, Australia*
[4] *Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran*
[5] *MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK*

Edited by:
Mazyar Fallah, York University, Canada

Reviewed by:
John Lisman, Brandeis University, USA
Alexander G. Dimitrov, Washington State University Vancouver, USA

***Correspondence:**
Masoud Ghodrati, Department of Physiology, Monash University, Clayton, Melbourne, VIC 3800, Australia
e-mail: masoud.ghodrati@monash.edu;
Seyed-Mahdi Khaligh-Razavi, MRC Cognition and Brain Sciences Unit, Cambridge University, 15 Chaucer Road, Cambridge CB2 7EF, UK
e-mail: seyed.kalighrazavi@mrc-cbu.cam.ac.uk

Invariant object recognition is a remarkable ability of primates' visual system that its underlying mechanism has constantly been under intense investigations. Computational modeling is a valuable tool toward understanding the processes involved in invariant object recognition. Although recent computational models have shown outstanding performances on challenging image databases, they fail to perform well in image categorization under more complex image variations. Studies have shown that making sparse representation of objects by extracting more informative visual features through a feedforward sweep can lead to higher recognition performances. Here, however, we show that when the complexity of image variations is high, even this approach results in poor performance compared to humans. To assess the performance of models and humans in invariant object recognition tasks, we built a parametrically controlled image database consisting of several object categories varied in different dimensions and levels, rendered from 3D planes. Comparing the performance of several object recognition models with human observers shows that only in low-level image variations the models perform similar to humans in categorization tasks. Furthermore, the results of our behavioral experiments demonstrate that, even under difficult experimental conditions (i.e., briefly presented masked stimuli with complex image variations), human observers performed outstandingly well, suggesting that the models are still far from resembling humans in invariant object recognition. Taken together, we suggest that learning sparse informative visual features, although desirable, is not a complete solution for future progresses in object-vision modeling. We show that this approach is not of significant help in solving the computational crux of object recognition (i.e., invariant object recognition) when the identity-preserving image variations become more complex.

Keywords: computational model, invariant object recognition, reaction time, object variation, visual system, feedforward models

INTRODUCTION

The beams of light reflecting from visual objects in the three-dimensional natural environment provide two-dimensional images onto the retinal photoreceptors. While the object is the same, an infinite number of light patterns can be mirrored in the retinal photoreceptors depending on object's distance (size), position, lightening condition, viewing angle (in-depth or in plane), and background. Therefore, the probability of having the same image on retina generated by an identical object in two different times, even in successive frames that are temporally close, is quite close to zero (DiCarlo and Cox, 2007; Cox, 2014). However, the visual system outstandingly performs object recognition, accurately and swiftly, despite substantial transformations.

The human brain can recognize the identity and category membership of objects within a fraction of a second (~100 ms) after stimulus onset (Thorpe et al., 1996; Carlson et al., 2011; Baldassi et al., 2013; Isik et al., 2013; Cichy et al., 2014). The mechanism of this remarkable performance in the unremitting changes of visual conditions in the natural world has constantly been under intense investigations, both experimentally and computationally (reviewed in Peissig and Tarr, 2007; DiCarlo et al., 2012; Cox, 2014). Our visual system can discriminate two highly similar objects within the same category (e.g., face identification) in various viewing conditions (e.g., changes in size, pose, clutter, etc.—invariance). However, this task is a very complex computational problem (Poggio and Ullman, 2013).

It is thought that the trade-off between selectivity and invariance is evolved through hierarchical ventral visual stages starting from the retinal to the lateral geniculate nucleus (LGN), then through V1, V2, V4, and finally IT cortex (Kreiman et al., 2006;

Zoccolan et al., 2007; Rust and DiCarlo, 2010, 2012; Sharpee et al., 2013). Decades of investigations on the visual hierarchy have shed light on several fundamental properties of neurons in the ventral visual stream (Felleman and Van Essen, 1991; Logothetis and Sheinberg, 1996; Tanaka, 1996; Cox, 2014; Markov et al., 2014). We now know that neurons in the higher level visual areas, such as IT, have larger receptive fields (RFs) compared to the lower levels in the hierarchy (e.g., V1). Each higher level neuron receives inputs from several neurons in the lower layer. Therefore, upstream neurons in the hierarchy are expected to respond to more complex patterns such as curvature for V4 neurons (reviewed in Roe et al., 2012) and objects for IT neurons compared to the early visual areas, which are responsive to bars and edges (Carandini et al., 2005; Freeman et al., 2013).

Using a linear read-out method, Hung et al. (2005) were able to decode the identity of objects from neural activities in primate IT cortex while the size and position of objects varied. This shows that representations of objects in IT are invariant to changes in size and position. Moreover, recent studies have reported intriguing results about object recognition in various stages and times in the ventral visual stream using different recording modalities in different species (e.g., Haxby et al., 2001; Hung et al., 2005; Kiani et al., 2007; Kriegeskorte et al., 2008b; Freiwald and Tsao, 2010; Cichy et al., 2014). Nevertheless, the mechanism of invariant object recognition has remained unknown to a certain extent. Most studies that have attempted to address invariant object recognition have used objects with gray backgrounds while either frontal views of objects were presented or only simple objects with limited variations were used (e.g., Alemi-Neissi et al., 2013; Isik et al., 2013; Wood, 2013). Studying the underlying computational principles of invariant object recognition is a very complicated problem with many confounding factors such as complex variations in real-world objects that makes it even more abstruse. This may explain why in most studies more attention is paid to understanding object recognition under restricted conditions by disregarding these complex variations from the stimulus set.

Recent recording studies have evidenced that representations of objects in IT are more invariant to changes in object appearance than intermediate levels of the visual ventral stream, such as V4 (Yamins et al., 2014). This shows that invariant representations are evolving across the visual hierarchy. Modeling results, inspired by biology, have also demonstrated that a great level of invariance is achievable using several processing modules built upon one another in a hierarchy from simple to complex units (e.g., Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999; Rolls, 2012; Anselmi et al., 2013; Liao et al., 2013).

Computational modeling is a valuable tool for understanding the processes involved in biological object vision. Although recent computational models have shown outstanding performances on challenging natural image databases (e.g., Mutch and Lowe, 2006; Serre et al., 2007b; Ghodrati et al., 2012; Rajaei et al., 2012) and compared to human (Serre et al., 2007a), they fail to perform well when they are presented with object images under more complex variations (Pinto et al., 2008). It has also been shown that the representations of object categories in object-vision models are weakly correlated with human and monkey IT cortex (Kriegeskorte, 2009; Kriegeskorte and Mur, 2012; Khaligh-Razavi and Kriegeskorte, 2013). This may explain why models do not yet achieve human level of categorization performance. Some studies have suggested that instead of a random sampling of visual features (Serre et al., 2007a), extracting a handful of informative features can lead to higher recognition performances (Ullman et al., 2002; Ghodrati et al., 2012; Rajaei et al., 2012). Having said that, we show in this study that when image variations are high, yet this approach results in poor performances compared to humans. Furthermore, we also show that the models do not form a strong categorical representation when the image variation exceeds a threshold (i.e., objects in the same category do not form a cluster in higher levels of variations).

Here we compare the performance of several object recognition models (Mutch and Lowe, 2006; Serre et al., 2007a; Pinto et al., 2008; Ghodrati et al., 2012; Rajaei et al., 2012) in invariant object recognition. Using psychophysical experiments, we also compare the performance of the models to human observers. All models are based on the theory of feedforward hierarchical processing in the visual system. Therefore, to account for the feedforward visual processing, images in our psychophysical experiments were rapidly presented to human observers (25 ms) followed by a mask image. As a benchmark test we also evaluated the performance of one of the best known feedforward object recognition models (Krizhevsky et al., 2012) against humans to see how far the best performing object-vision models go in explaining profiles of human categorization performance.

We employed representational similarity analysis (RSA), which provides a useful framework for measuring the dissimilarity distance between two representational spaces independent of their modalities (e.g., human fMRI activities and models' internal representations—see Kriegeskorte et al., 2008a; Kriegeskorte, 2009). In this study we used RSA to compare the representational geometry of the models with that of the human observers in invariant object recognition tasks.

To evaluate the categorization performance of the models and humans we built a parametrically controlled image database consisting of different object categories, considering various object variations, rendered from 3D planes (O'Reilly et al., 2013). Generating such controlled variations in object images helps us to gain better insights about the ability of models and humans in invariant object recognition. It also helps experimentalists to study invariant object recognition in human and monkey by taking advantage of having controlled variations over several identity-preserving changes of an object.

Our results show that human observers have remarkable performances over different levels of image variations while the performances of the models were only comparable to humans in the very first levels of image variations. We further show that although learning informative visual features improves categorization performance in less complex images (i.e., images with fewer confounding variations), it does not help when the level of confounding variations (e.g., variations in size, position, and view) increases. The results of our behavioral experiments also demonstrate that models are still far from resembling humans in invariant object recognition. Moreover, as the complexity level of object variations increases (from low to intermediate and high levels of variations), models' internal representation become

worse in disentangling the representation of objects that fall in different categories.

MATERIALS AND METHODS

IMAGE GENERATION PROCESS

One of the foremost aspects of the evaluation procedure described in this study is the utilization of controlled variations applied to naturalistic objects. To construct various two-dimensional object images with controlled variations, we used three-dimensional meshes (O'Reilly et al., 2013). It allowed us to parametrically control different variations, background, number of objects in each class, etc. Therefore, we were able to parametrically introduce real-world variations in objects.

For each object category (car, motorcycle, animal, ship, airplane), we had on average sixteen 3D meshes (showing different exemplars for each category) in which 2D object images were rendered using rendering software with a uniform gray background for all images. Throughout the paper we call them objects on plain backgrounds. These images were superimposed on randomly selected backgrounds from a set of more than 4000 images (see **Figure S1** for image samples with natural backgrounds). The set included images from natural environments (e.g., forest, mountain, desert, etc.) as well as man-made environments (e.g., urban areas, streets, buildings, etc.). To preserve a high variability in our background images, we obtained all background images using the internet.

Naturalistic object images were varied in four different dimensions: position (across x and y axes), scale, in-depth rotation, and in-plane rotation (**Figure 1**). To alter the difficulty of the images and tasks, we used seven levels of variation to span a broad range of diversity in the image dataset (starting from no particular variations, **Figure 1**-left, to the intermediate and complex image variations, **Figure 1**-right). The amount of object transformations in each level and dimension was selected by random sampling from a uniform distribution. For example, to generate images with second level of variation (i.e., Level 1), we randomly sampled different degrees for in-depth rotation (or in-plane rotation) from a range of 0–15° using a uniform random distribution. The same sampling procedure was applied to other dimensions (e.g., size and position). Then, these values were applied to a 3D mesh and a 2D image was subsequently generated from the 3D mesh. A similar approach was taken for generating images in other levels of variation.

PSYCHOPHYSICAL EXPERIMENT

Two experiments were designed to investigate the performance of human subjects in invariant object recognition: tow- and multiclass invariant object categorization task.

Two-class invariant object categorization

In total, 41 subjects (24 male, age between 21–32, mean age 26) participated in the first experiment. We used 560 object images (300 × 400 pixels, grayscale images) selected from seven levels of variation and two different object categories (80 images for each level with 40 images from each category) for each session. Images were presented on a 21″ CRT monitor with a resolution of 1024 × 724 pixels and a frame rate of 80 Hz. We used Matlab with the Psychophysics Toolbox to present the images (Brainard, 1997; Pelli, 1997). The viewing distance was 60 cm.

Following a fixation cross, which was presented for 500 ms, an image was randomly selected from the dataset (considering levels and categories) and presented at the center of the screen for the duration of 25 ms. Subsequently, a blank screen was presented for the duration of 20 ± 2 ms (interstimulus interval-ISI) and a mask image was presented after the blank screen and stayed on for 100 ms (**Figure 2**). The mask image was a (1/f) random noise.

Subjects were instructed to complete four sessions (cars vs. animals, cars vs. motors, with plain and natural background).

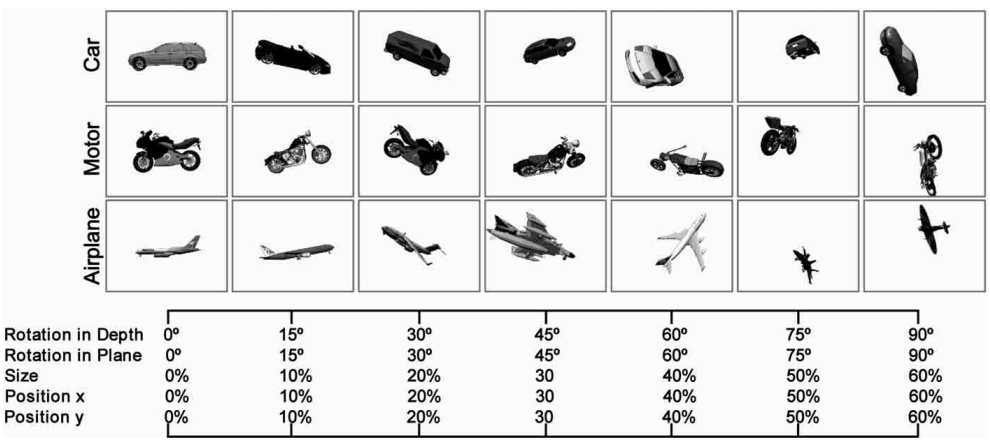


FIGURE 1 | Sample images in different levels of variation with Plain Background. The Object images, rendered from 3D planes, vary in four dimensions: size, position (x, y), rotation in-depth, and rotation in plane. To alter the complexity of the images, we constructed images in seven levels of variations starting from zero level variation, which no variation is applied to 3D object planes (first column at left), to seventh level of variation, which substantial variations are applied to images (last column at right). In each level of variation, we randomly sample different values for each dimension (e.g., size, rotation, and position) from a uniform distribution and finally selected values are applied to a 3D plane. As the level of variation increases, the range of values increases. There are several sample object images with natural background in the supplementary materials (**Figure S1**).

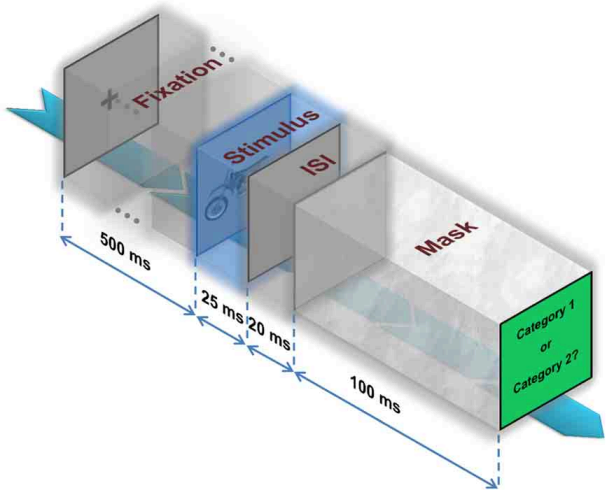


FIGURE 2 | Psychophysical experiment. A fixation cross is presented for 500 ms. Then, an image is randomly selected from the dataset and presented at the center of the screen for the duration of 25ms. Subsequently, a blank screen is presented on the screen for the duration of 20 ± 2 ms (interstimulus interval—ISI) followed by a mask image that stays on the screen for 100ms. Finally, subjects have to decide whether the image belongs to category 1 or 2.

Some subjects completed all four sessions and some only finished some sessions. In each session, 560 images (e.g., 280 cars and 280 motors) were presented in a random order and were divided into 4 blocks of 140 images each. There was a time interval of 5 min between blocks for each subject to take a rest. The reaction times (RTs) of participants were recorded to investigate whether there is any time difference in categorization between levels and categories.

The subjects' task was to determine whether the presented image was a car or a/an motor/animal by pressing "C" or "M" on a computer keyboard, respectively. Keys were labeled on the keyboard with the name of corresponding categories. Subjects performed several training trials, with different images, to become familiar with the task prior to the actual experiment. In training trials (30 images), a sentence was presented on the monitor showing whether the answers were correct or not. During the main procedure, the participants had to declare their decision by pressing the keys; but no feedback was given to them regarding the correctness or incorrectness of the choices. The next trial was instantly started after getting subject's response. Subjects were instructed to respond as fast and accurate as possible to the presented image. All subjects voluntarily accepted to participate in the task and gave their written consent.

Multiclass invariant object categorization

In total, 26 subjects participated in the second behavioral experiment (17 male, age between 21–32, mean age of 26 years). Object images were selected from five categories (i.e., car, animal, motorcycle, ship, and airplane) in seven levels of variation. The procedure was the same as the first experiment: an image was randomly selected and presented on the center of the screen for 25 ms after a fixation cross (500 ms). Subsequently, a blank screen (ISI) of 20 ± 2 ms was presented followed by a mask image, which stayed on for 100 ms (**Figure 2**). Subjects were instructed to indicate the image category by pressing one of the five keys on the computer keyboard, each labeled with a name representing a specific category ("C," "Z," "M," "N," and "/" for car, animal, motorcycle, ship, and airplane, respectively). The next trial was started by pressing the space-bar. The RTs of subjects were not evaluated in this task, so subjects had time to state their decisions. However, subjects were instructed to respond as fast and accurately as possible.

This task was designed to have two sessions (images with plain and natural background). In each session, 700 images (100 images per level, 20 images from each object class in each level) were presented in a random order, divided into 4 blocks of 175 images each. There was a gap of 5 min between blocks for subjects to take a rest. Some subjects completed all sessions and some only finished some of them. Subjects performed a few example trials before starting the actual experiment (none of the images in these trials were presented in the main experiment). In training trials (30 images), a sentence was presented on the monitor as a feedback showing the correctness/incorrectness of the answers. In the main procedure, participants had to declare their decision by pressing one of the keys; but no feedback was given to them regarding the correctness of choices. All subjects voluntarily accepted to participate in the task and gave their written consent.

HUMAN REPRESENTATIONAL DISSIMILARITY MATRIX (RDM)

In the multiclass psychophysical experiment, subjects' responses to the presented stimuli were recorded. Subjects had five choices for each presented stimulus: 1–5 for five categories. We constructed a matrix, R, based on the subjects' responses. The rows of R were labels assigned to an image by different subjects (each row corresponds to one image) and each column contained responses of one subject to all images in the task. Therefore, the size of this matrix was: images × subjects (e.g., for the multiclass experiment the size was 700 × 17 for each task, plain and natural background). Afterwards, we calculated the categorization score for each row of the matrix. To do this, for example, out of 17 participants (e.g., responses in row one), 11 selected category one for the presented image, five responses showed category two, and one classified the image as category three, and no subject classified the image as category four and five. This gives us a response pattern ($R_{1,1:5}$) for the first image (e.g., the image in the first row):

$$R_{1,1:5} = [11\ 5\ 1\ 0\ 0]$$

Finally, we normalized each row by dividing it to the number of responses:

$$R_{1,1:\ 5} = \frac{[11\ 5\ 1\ 0\ 0]}{17} = [0.6471\ \ 0.2941\ \ 0.0588\ \ 0\ \ 0]$$

To calculate the RDMs, we used the RSA toolbox developed by Nili et al. (2014). Each element in a given RDM shows the pairwise dissimilarity between the response patterns elicited by two images. RDM is a useful tool to visualize patterns of dissimilarities between all images in a representational space (e.g., brain

or model). The dissimilarity between two response patterns is measured by correlation distance (i.e., 1-correlation—here we used Spearman's rank correlation). RDMs are directly comparable to each other and they provide a useful framework for comparing the representational geometry of the models with that of the human independent of the type of modalities and represented features (e.g., human behavioral scores and models' internal representations).

COMPUTATIONAL MODELS

V1-like

This model is a population of simple and complex cells fed by luminance images as input. We used Gabor filters at four different orientations (0, 45, 90, and −45°) and 12 sizes (7–29 pixels with steps of two pixels) to model simple cell RFs. Complex cells were made by performing the MAX operation on the neighboring simple cells with similar orientations. The outputs of all simple and complex cells were concatenated in a vector as the V1 representational pattern of each image.

HMAX

The HMAX model, developed by Serre et al. (2007a), has a hierarchical architecture inspired by the well-known simple to complex cells model of Hubel and Wiesel (1962, 1968). The HMAX model that is used here adds two more layers (S2, C2) on the top of the complex cell outputs of the V1 model described above. The model has alternating S and C layers. S layers perform a Gaussian-like operation on their inputs, and C layers perform a max-like operation, which makes the output invariant to small shifts in scale and position. We used the freely available version of the HMAX model (http://cbcl.mit.edu/software-datasets/pnas07/index.html). The HMAX C2 features were used as the HMAX representation.

GMAX

GMAX is an extension of the HMAX model for which in the training phase, instead of selecting a pool of random patches, patches that are more informative for the classification task are selected. The model uses an optimization algorithm (i.e., genetic algorithm) to select informative patches from a very large pool of random patches (Ghodrati et al., 2012). In the training phase of the GMAX model the classification performance is used as the fitness function for the genetic algorithm. A linear SVM classifier was used to measure the classification performance. To run this model we used the same set of model parameters suggested in Ghodrati et al. (2012).

Stable

Stable model is a bio-inspired model with a hierarchy of simple to complex cells. The model uses the adaptive resonance theory (ART-Grossberg, 1976) for extracting informative intermediate level visual features. This has made the model stable against forgetting previously learned patterns (Rajaei et al., 2012). Similar to the HMAX model it extracts C2-like features, except that in the training phase it only selects the highest active C2 units as prototypes that represent the input image. This is done using top-down connections from C2 layer to C1 layer. The connections match the C1-like features of the input image to the prototypes of the C2 layer. The matching degree is controlled by a vigilance parameter that is fixed separately on a validation set. We set the model parameters the same as were suggested in Rajaei et al. (2012).

SLF

This is a bio-inspired model based on the HMAX C2-features. The model introduces sparsified and localized intermediate-level visual features (Mutch and Lowe, 2008). We used the Matlab code freely available for these feature (http://www.mit.edu/∼jmutch/fhlib); and the default model parameters were used.

Pixel

The pixel representation is simply a feature vector containing all pixels of an input image. Each image was converted to grayscale and then unrolled as a feature vector. We used pixel representation as our baseline model.

Convolutional neural networks

Convolutional neural networks (CNNs) are bio-inspired hierarchical models of object-vision that are made of several convolutional layers (Jarrett et al., 2009). Convolutional layers scan the input image inside their RFs. RFs of convolutional layers get their input from various places in the input image, and RFs with identical weights make a unit. The outputs of each unit make a feature map. Convolutional layers are usually followed by subsampling layers that perform a local averaging and subsampling, which make the feature maps invariant to small shifts (LeCun and Bengio, 1998). In this study we used the deep supervised convolutional network by Krizhevsky et al. (2012; Donahue et al., 2013). The network is trained with 1.2 million labeled images from ImageNet (1000 category labels), and has eight layers: five convolutional layers, followed by three fully connected layers. The output of the last layer is a distribution over the 1000 class labels. This is the result of applying a 1000-way softmax on the output of the last fully connected layer. The model has 60 million parameters and 650,000 neurons. The parameters are learnt with stochastic gradient descent. The results for the deep ConvNet are discussed in Supplementary Material.

MODEL EVALUATION

To evaluate the performance of the models, we first randomly selected 300 images from each object category and level (e.g., 300 car images with level one variation). Images were then randomly divided to test and train images. We selected 150 images for the training set and 150 for the test set. All images were converted into grayscale and resized to 200 pixels in height while aspect ratio was preserved. For the case of natural background, we randomly selected equal number of natural images (i.e., 300 images) and superimposed the objects images on these backgrounds. We then fed each model with the images and the performance of each model was obtained for various levels of variation separately. The feature vectors of each model were fed to a linear SVM classifier. The reported results are the average of 15 independent random runs and the error bars are standard deviation of the mean (SD-**Figures 3**,**4**,**6**).

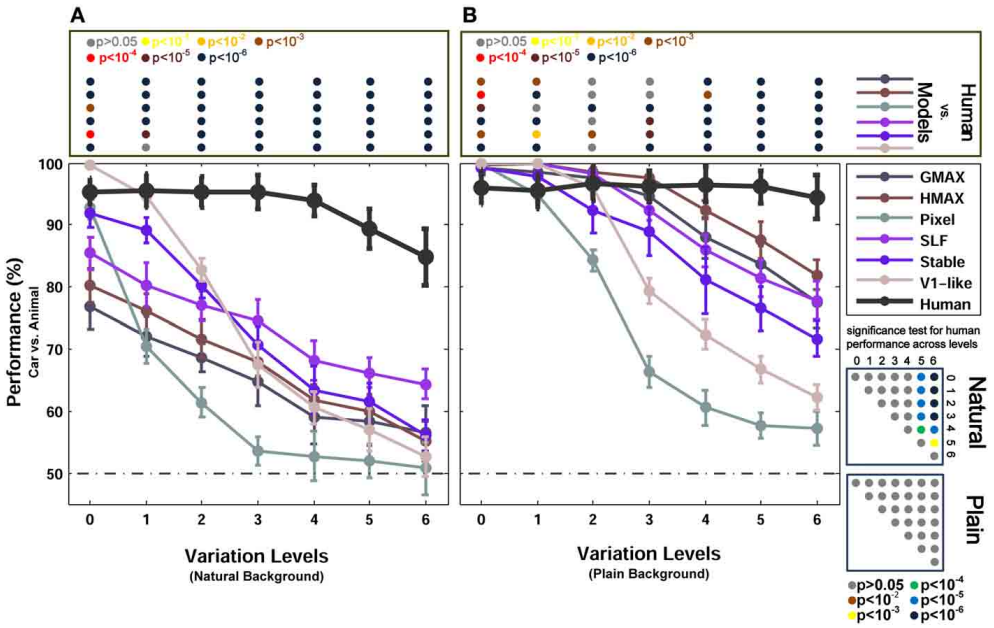


FIGURE 3 | Performances of models and human in two-class (car/animal) invariant object categorization task. (A) Performance comparison when objects were presented on natural backgrounds (performances are the ratio of correct responses). The dashed, horizontal line shows the chance level and each curve represents the performance of a model in different levels of variation, specified with different colors at the right inset. The bold, black curve illustrates human performance. The color-coded circle points at the top of each plot, inside the rectangular box, exhibits the *p*-values for comparisons between human and each model obtained by Wilcoxon signed-rank test (for example the performance of the HMAX model was compared to the human in each level of variation separately. The result of comparison for each model in each level provides us with a *p*-value. *P*-values are reported with different colors). The color-coded circle points at the right insets, inside the square boxes, show the *p*-values for all possible comparisons between human responses in different levels of variation (with plain and natural background). Here, the *p*-values show whether human categorization performances are significantly different at different levels of variation. For example, we compared the performance of human in Level 0, with Level 1, Level 2, and so on and reported a *p*-value for each comparison. These comparisons resulted in a symmetric *p*-value matrix with the size of 7*7 (i.e., 7 levels of variations). **(B)** Performance comparison when objects were presented on plain backgrounds. In both panels **(A,B)**, the results are the average of 15 independent random runs and the error bars show the standard deviation of the mean.

Furthermore, the confusion matrices for all models as well as humans were computed in all levels for both plain and natural backgrounds (for multiclass object classification). To obtain a confusion matrix, we first trained a classifier for each category. Then, using these trained classifiers, we computed multiclass performances as well as errors made in classification. To construct a confusion matrix for a given level, we calculated the percentage of classification performance (predicted labels) obtained by each classifier which was trained on a particular category. Confusion matrices can help us to examine which categories are more mistakenly classified. We can also see whether errors increase in high levels of variation.

RESULTS

TWO-CLASS INVARIANT OBJECT CATEGORIZATION

In this experiment, we compared the categorization performance of different models in invariant object recognition tasks with each other and with the categorization performance of human observers. The categorization performance of human observers was measured in psychophysical experiments where subjects were presented with images in different levels of variation. To evaluate the performance of models, we ran similar categorization tasks in which two groups of object categories were selected to perform a two-class object categorization. In the first group, motorcycle and car images were selected, which are both vehicles. For the second group, we selected more dissimilar categories, car and animal images. There were two different types of animal images in this category (i.e., elephant and dinosaur) with variety of examples for each type. We selected 150 images for the training set and 150 for the testing set (see Materials and Methods). The categorization performance of each model was obtained for all levels of variation separately (i.e., seven levels of variation). **Figures 3**, **4** show the performances of different models as well as human observers in the seven levels of object variation. The results for the deep ConvNet are shown in **Figure S3**, and are explained in Supplementary Material.

Figure 3 shows the results of animal vs. car classification with natural (**Figure 3A**) and plain (**Figure 3B**) backgrounds. In the case of plain background, models performed as accurate as humans in the first two or three levels of variation. Even the Pixel model, which gray values of images were directly fed into the classifier, performed very close to humans in the first two levels of variation. From the level three onward, the performance of the two null models (i.e., V1-like and Pixel) decreased sharply down to 60% in the last level of variation (note that chance level is 50%). Likewise, from the third level up to the sixth level, the

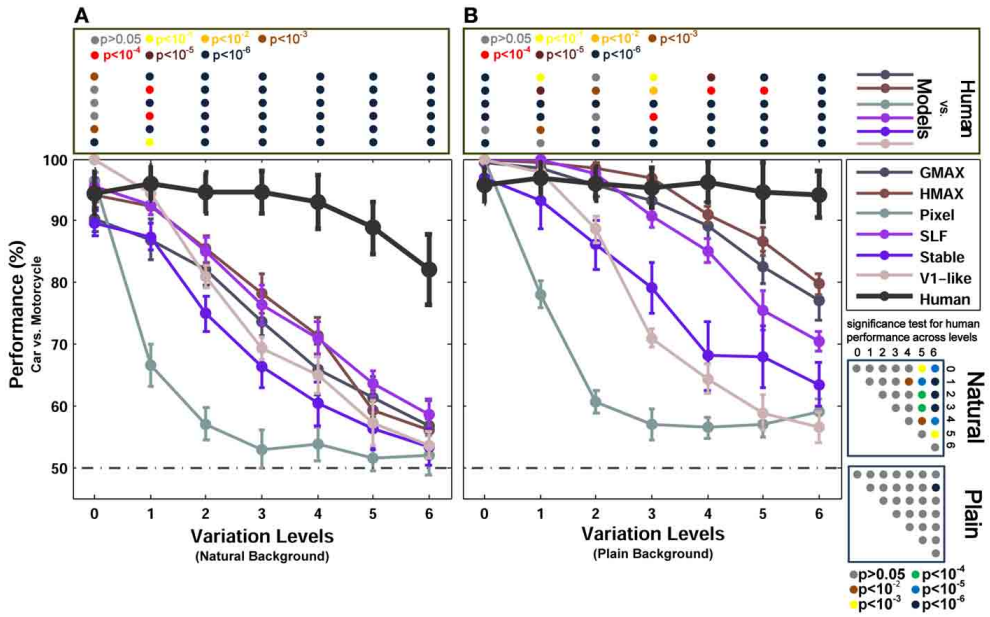


FIGURE 4 | Performances of models and human in two-class (motorcycle/car) invariant object categorization task. (A) Performance comparison when objects were presented on natural backgrounds (performances are the ratio of correct responses). The dashed, horizontal line shows the chance level and each curve represents the performance of a model at different levels of variation, specified with different colors in the right inset. The bold black curve illustrates human performance. The color-coded circle points at the top of each plot, inside the rectangular box, exhibits the *p*-values for comparisons between human and each model obtained by Wilcoxon signed-rank test (for example the performance of the HMAX model was compared to the human in each level of variation separately. The result of comparison for each model in each level provides us with a *p*-value. *P*-values are reported with different colors). The color-coded circle points at the right insets, inside the square boxes, show the *p*-values for all possible comparisons for human responses in different levels of variation (with plain and natural background). Here, the *p*-values show whether human categorization performances are significantly different at different levels of variation. For example, we compared the performance of human in Level 0, with Level 1, Level 2, and so on and reported a *p*-value for each comparison. These comparisons resulted in a symmetric *p*-value matrix with the size of 7*7 (i.e., 7 levels of variations). **(B)** Performance comparison when objects were presented on plain backgrounds. In both panels **(A,B)**, the results are the average of 15 independent random runs and the error bars show the standard deviation of the mean.

performances of other models diminished significantly compared to humans. This shows that the models fail to solve the problem of invariant object recognition when the level of variation grows up. Comparing the performances of the V1-like model and the Pixel model shows that the V1-like model has slightly better invariant responses than the Pixel model. In more complex variations, four other hierarchical models, which implement the hierarchical processing from V1 to V4 and aIT, exhibited higher performances, compared to the null models. Nevertheless, in high levels of variation, even the cortex-like hierarchical models performed significantly lower than human subjects.

Interestingly, when objects are presented on plain backgrounds, the categorization performance of humans in any level of image variation is not significantly different from other levels (see *p*-values in **Figure 3** bottom right inset). This means that human observers, as opposed to the models, were able to produce equally well invariant representations in response to objects under different levels of image variation. Indeed, the models are still far below the performance of humans in solving the problem of invariant object recognition (see *p*-values for all comparisons between the models and human observers at the top inset in **Figure 3**, specified with color-coded circle points inside the rectangular box).

We also compared the performance of the models with humans in a more difficult task, in which objects were presented on randomly selected natural backgrounds instead of plain backgrounds (**Figure 3A**). A natural background makes the task more difficult for models as well as for humans. In this case, overall, there is a significant difference between the categorization performance of the models and human, even in zero level variation (i.e., no variation, Level 0). In the last three levels of variation (i.e., Levels 4–6), we can see a decrease in human categorization performance (see the *p*-values at the bottom right inset in **Figure 3**). Although adding natural backgrounds diminished the performance of human in invariant object recognition, the human responses are still robust to different levels of variations and still significantly higher than the models (see *p*-values for all comparisons between the models and human at the top inset in **Figure 3**).

The lower performances of models in the case of natural backgrounds in comparison to the plain backgrounds show that the feedforward models have difficulties in distinguishing a target object from a natural background. Natural backgrounds impose more complexity to object images and the process of figure-ground segregation becomes more difficult. Studies have suggested that recurrent processing is involved in figure-ground

segregation (Roelfsema et al., 2002; Raudies and Neumann, 2010). This may explain why we observe a dramatic decrease in the categorization performance of feedforward models in the natural background condition. They lack a figure-ground segregation step that seems to arise from feedback signals.

Figure 3 shows the categorization performances for car vs. animal images, which are two dissimilar categories, across different levels of variations. To evaluate the performances of human and models in categorizing two similar categories, we used car and motorcycle images, which are both vehicles with similar properties (e.g., wheels). The results are shown in **Figure 4A** (with natural background) and **Figure 4B** (with plain background). Overall, the results in both experiments are similar, except that the performances are lower in car vs. motorcycle categorization task.

As the level of variation increases the complexity of images grows in both plain and natural backgrounds and the performance decreases. We asked whether the complexity of images affects human RTs in high level of variations. RT is considered as a measure of uncertainty that seems to be associated with the amount of accumulated information required for making a decision about an image in the brain. **Figure 5** reports the average RTs across subjects in all seven levels of image variation and the two rapid categorization tasks (animal/car and motorcycle/car) for both plain and natural background conditions. In the case of plain background (green curves), the mean RTs are approximately the same for low and middle levels of variations. On the other hand, when objects are presented with natural backgrounds, human RTs increases more sharply as the complexity of object variations increases. This indicates that the visual system requires more time, in higher levels of variation, to accumulate enough information to reach a reliable decision. This suggests that the brain responds differently to different levels of object variation and the time course of responses depends on the strength of variation. Furthermore, having higher RTs in the natural background condition compared with the plain background condition, suggests that some further processes are going on in the first condition, probably to separate the target object from a distracting natural background.

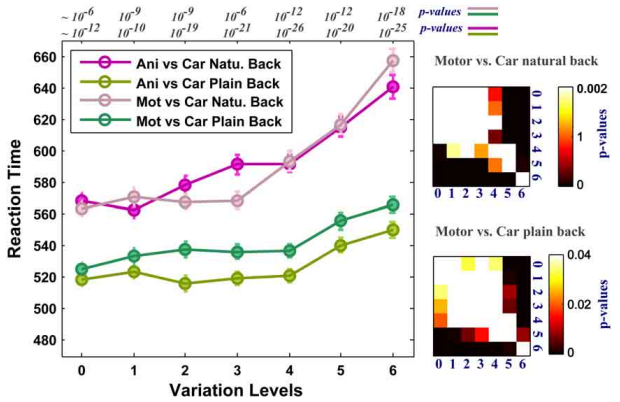


FIGURE 5 | Human reaction times (RTs) for different levels of variation in two-class invariant object categorization tasks with plain and natural backgrounds. The RTs were almost equal across all levels of variation when objects were presented on plain backgrounds (except for the higher levels of variation, see *p*-values for all comparisons at the right insets. We made all possible comparisons between RTs across different levels to find out whether the differences between the RTs are statistically significant. Here we only showed matrices for motorcycle vs. car. Animal vs. car gives similar *p*-value matrices). In contrast, when objects were presented on natural backgrounds, the RTs in all levels of variation increased significantly compared to the plain background condition. Error bars are s.e.m. See *p*-values on the top of the figure show comparisons between natural and plain background conditions.

MULTICLASS INVARIANT OBJECT CATEGORIZATION

We also compared the models with each other and with human observers in multiclass invariant object categorization tasks (five classes of objects). The confusion matrices for all models as well as humans were computed in all seven levels of object variation in both plain and natural background conditions. Overall, the confusion matrices show that the null models make many more errors while categorizing object classes with intermediate and high level of variations compared to the hierarchical cortex-like models. Moreover, they show that humans accurately categorized object images with only a handful of errors even in higher levels of variation in which the complexity of image variation is higher and it is more likely to perceive two different object images as similar.

Figure 6 reports the performances of multiclass object categorization for plain and natural background conditions in all seven levels of object variation. As shown in **Figure 6B**, when objects were presented on plain backgrounds, all models performed as accurate as humans in zero level variation (no variation-Level 0).

In the next level, the performance of the V1-like model was still similar to humans, but it sharply decreased when object images had stronger variations. The performance of the Pixel model dropped dramatically after the zero level variation. This shows that the actual values of pixels do not exhibit an invariant representation. The performances of other models also decreased as the level of image variation increased (from the first level to the last level). In the last level, the performances of the Pixel and V1-like model were very close to the chance level. However, biologically inspired hierarchical models converged on performances higher than chance, although the performances were still much lower than the human performance. Human performances did not significantly differ across different levels of variations, indicating the remarkable ability of human brain in generating invariant representation despite the increasing level of the difficulty in image variations (see *p*-values at the bottom right inset in **Figure 6** for all possible comparisons, specified with color-coded circle points inside the square box).

In the case of natural backgrounds (**Figure 6A**), the performance of the models, even in zero level variation, is significantly lower than the human performance. Interestingly, the V1-like and the Pixel model performed better than other models in zero level variation. This is almost similar to the results reported in Pinto et al. (2008), in which a V1-like model that does not contain any special machinery for tolerating difficult image variations performs better than state-of-the-art models when images have no or very small variations. On the other hand, the representation of these two null models was not informative enough in higher levels of variation and the performance of these models rapidly falls off as the variation gets more difficult (**Figure 6A**).

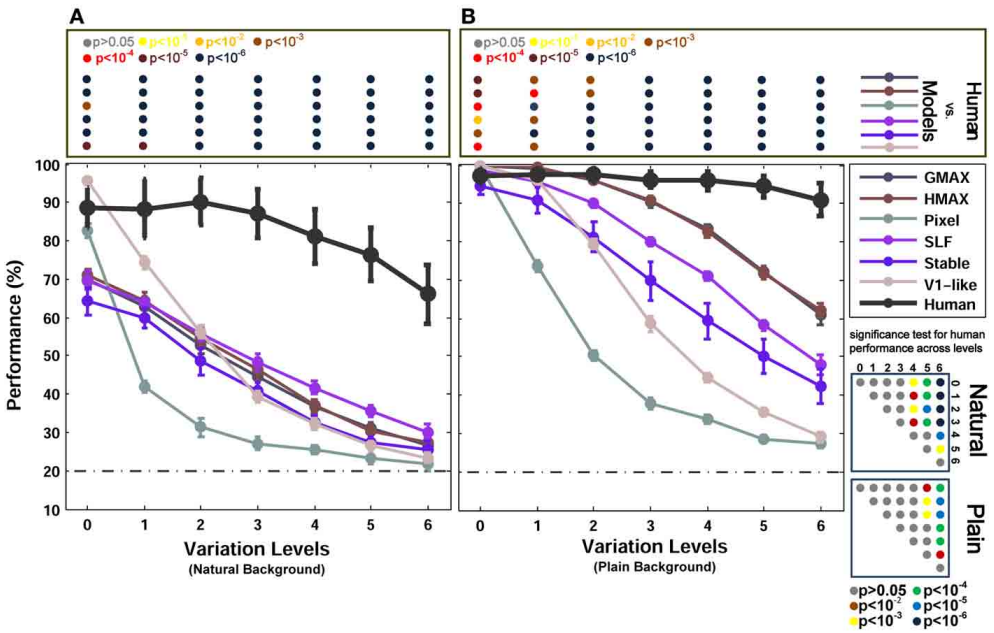


FIGURE 6 | Performance comparisons between different models and human in multiclass invariant object categorization task. (A) Performance comparison when objects were presented on natural backgrounds. The dashed, horizontal line shows the chance level (20%) and each curve represents the performance of a model in different levels of variation, specified with different colors in the top right inset. The bold black curve illustrates human performance. The color-coded circle points at the top of each plot, inside the rectangular box, exhibits the *p*-values for comparisons between human and each of the models obtained by Wilcoxon signed-rank test. The color-coded circle points at the right insets, inside the square boxes, show the *p*-values for all possible comparisons for human responses in different levels of variation (with plain and natural background). Here, the *p*-values show whether human categorization performances are significantly different at different levels of variation. For example, we compared the performance of human in Level 0, with Level 1, Level 2, and so on and reported a *p*-value for each comparison. These comparisons resulted in a symmetric *p*-value matrix with the size of 7*7 (i.e., 7 levels of variations). **(B)** Performance comparisons when objects were presented on plain backgrounds. In both panels **(A,B)** error bars are STD and the performances are the average of 15 runs.

To have a closer look at the performance of humans and models in categorizing each object category and complexity level, we used confusion matrices. **Figures 7**, **8** show confusion matrices for plain and natural backgrounds, respectively. In the plain background condition, confusion matrices for humans in all levels are completely diagonal that shows the ability of humans in discriminating objects without difficulty, even in higher levels of image variation. The confusion matrices of models are also diagonal in the first two levels of variation. However, models made more errors in higher levels of variation. The Pixel and V1-like models, for example, made many errors in classification of different objects in last levels of variations. This shows that the internal representation of these null models does not tolerate identity-preserving variations beyond a very limited extent. Furthermore, we do not expect responses of V1 neurons to be clustered based on semantic categories (e.g., Kriegeskorte et al., 2008b; Cichy et al., 2014). So a linear readout would not be able to readily decode from V1 responses. This is similar to what we see in the V1 model. Although the representation of V1 neurons are not clustered according to object categories, during recurrent interactions between higher and lower visual areas, early visual areas contribute in categorization and perception happening in higher levels of visual hierarchy (Koivisto et al., 2011). Feedback signals, from higher visual areas toward early visual areas, such as V1, have also been shown to play a role in figure-ground segregation (Heinen et al., 2005; Scholte et al., 2008), which is a useful mechanism in discriminating target objects from cluttered background.

Models made more errors when objects were presented on natural backgrounds (**Figure 8**). Incorporating object images with randomly selected natural scenes have made the task more difficult for human observers as well. However, the human observers only made a few errors in the last two levels of variation and the confusion matrices for all levels are still close to diagonal. In the models, there are more errors in high and even moderate levels of image variation. As can be seen, the confusion matrices for models are not strongly diagonal in the last two levels of variation. This indicates that models were unable to discriminate objects in higher variations.

In zero level variation, the Pixel and V1-like models achieved performances comparable to human in both cases, plain and natural background (**Figures 6A,B**). Comparing the internal representation of models gives us further insights about the ability of models in generating identity-preserving invariant representations. To this end, we used RSA (Kriegeskorte et al., 2008a,b) and compared the dissimilarity-patterns of models with each other and with human observers. **Figure 9** represents RDMs for different models, calculated directly from feature vectors of each model in seven levels of variation when objects were presented on plain backgrounds. The RDMs for humans are based on the behavioral

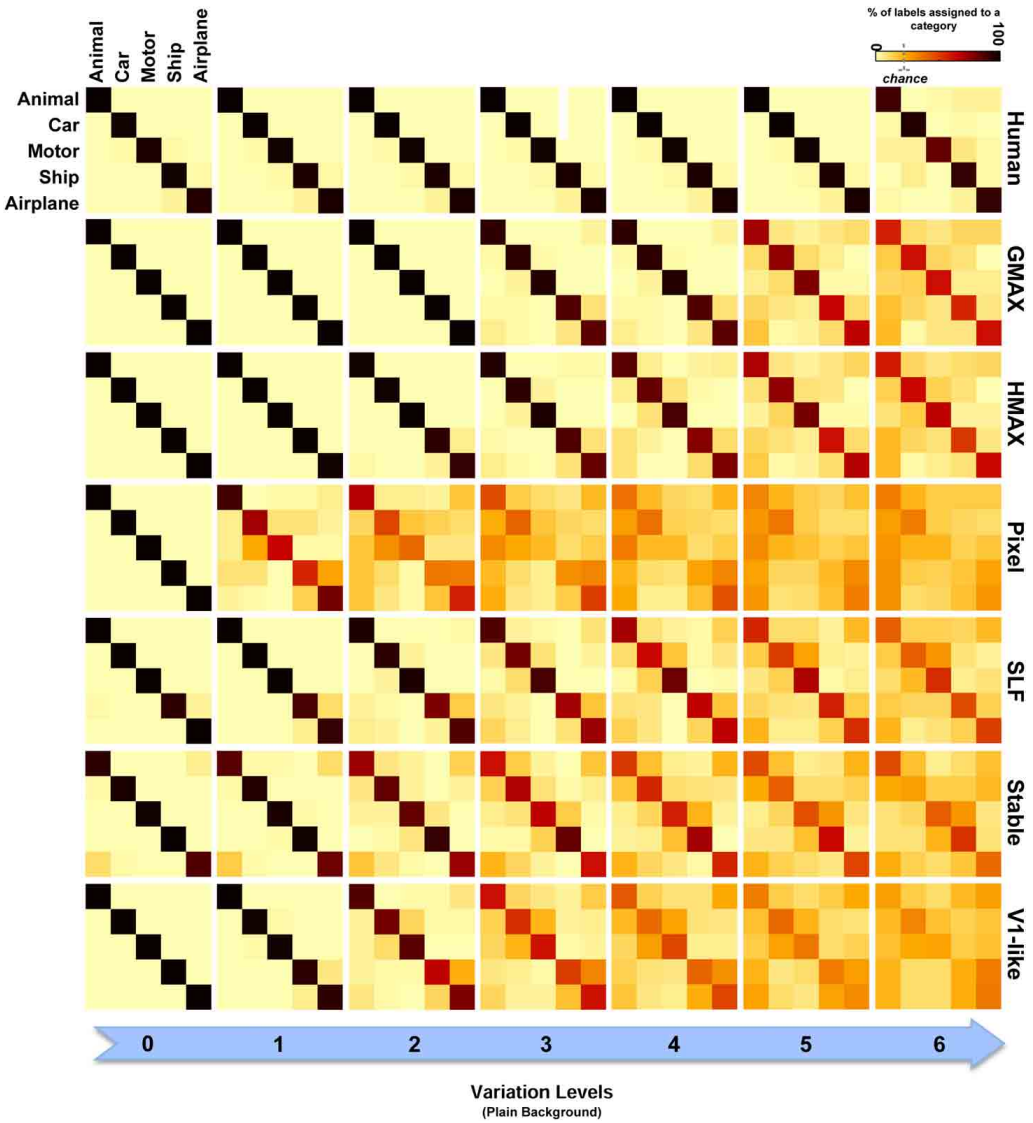


FIGURE 7 | Confusion matrices for the multiclass invariant object categorization task with plain background. Each color-coded matrix shows the performance of a model in categorizing different object categories, as specified in the first matrix at the top-left corner. Matrices in each column show confusion matrices for a particular level of variation (from 0 to 6) and each row shows confusion matrices for one model (model name is written at the right end of each row). The first row illustrates the performance of humans in psychophysical experiments. The color bar at the top-right color codes the percentage of the subject responses (labels) assigned to each category. The chance level is specified with a dashed line on the color bar.

results, using the labels assigned to each image by human subjects (see Materials and Methods). As can be seen, the dissimilarity representation of models, even in the first levels of variation, does not provide a strong categorical representation for different object classes. However, the RDMs of human show clear clustered representations for different object categories across all levels (first row in **Figure 9**).

As described earlier, the human RDMs were built based on the labels given to the presented images while the RDMs of the models calculated using model features. For further comparisons and to make human RDMs more comparable to models' RDMs, we similarly constructed RDMs for models based on the classifier outputs. **Figure 10** illustrates the RDMs of the models based on the SVM responses for the case of objects presented on plain backgrounds. Visual inspection shows that the representations of several models are comparable to humans in different levels of variation. This simply indicates that classifier performs well in categorizing different object categories with high and intermediate levels of variation. However, this similarity structure significantly reduces when models were presented with objects on natural backgrounds (**Figure S2** in Supplementary Materials).

As can be seen from RDMs in **Figures 9**, **10**, some object categories (i.e., ship and airplane) have more similar representations in the model space compared to other categories. Interestingly, this can also be seen in the confusion matrices of the models as well as the confusion matrices of human

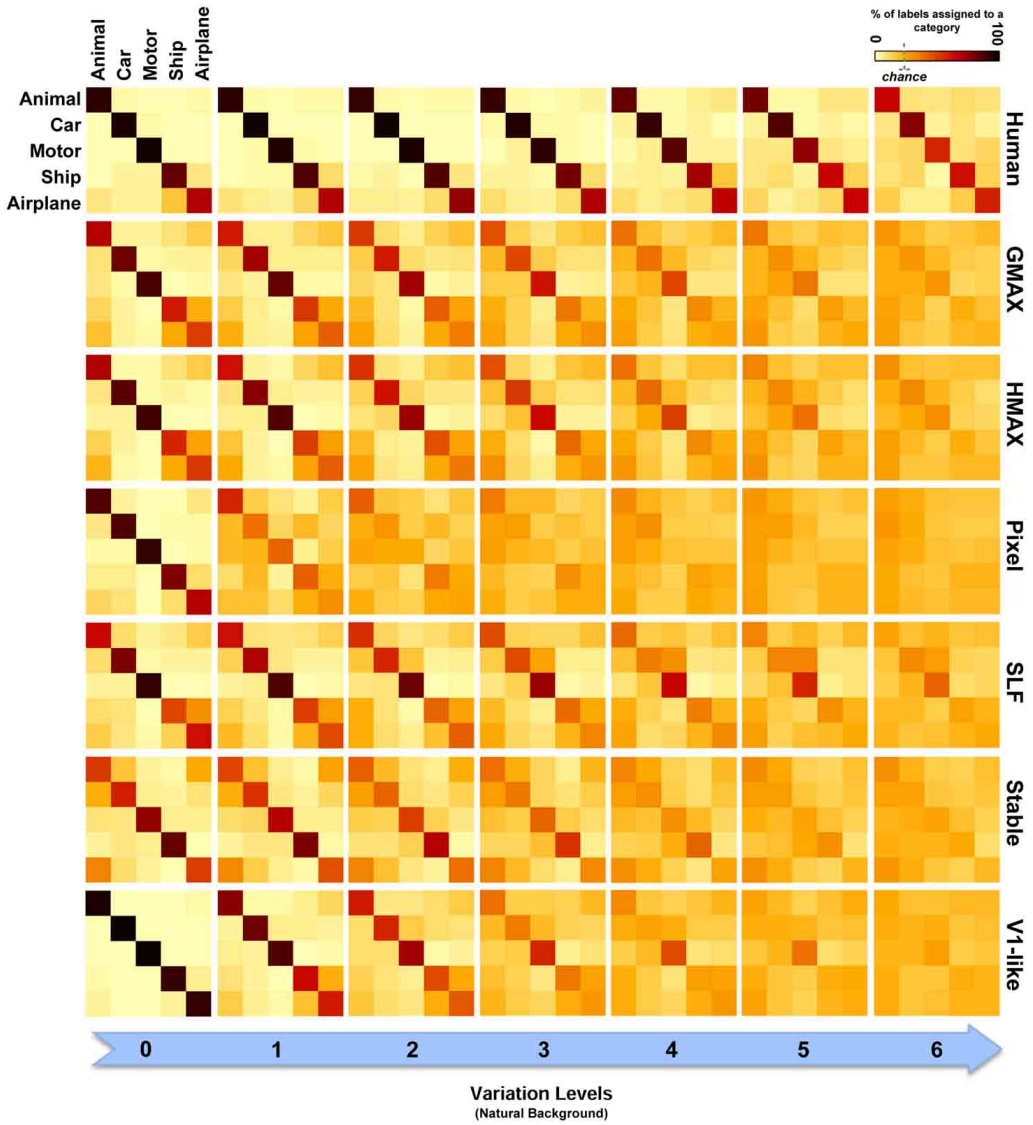


FIGURE 8 | Confusion matrices for the multiclass invariant object categorization task with natural background. Each color-coded matrix shows the confusion matrix for a model or for humans in categorizing different object categories presented on natural backgrounds. Matrices in each column show confusion matrices for a particular level of variation (from 0 to 6) and each column shows confusion matrices for one model (model name is written at the right end of each row). The first row illustrates the performance of humans in psychophysical experiments. The color bar at the top-right color codes the percentage of the subject responses (labels) assigned to each category. The chance level is specified with a dashed line on the color bar.

observers (**Figures 7**, **8**). This effect is clearer in **Figure 8**. These results suggest that the observed similarities are mainly driven by the shape similarly of objects (both ship and airplane share similar shape properties such as body, sail, and wing, etc.). This result was expected for the models since the models were all unsupervised models, and therefore by definition the extracted features were only aware of the shape similarity between the objects and had no additional cue about their category labels. But, human observers similarly made more errors in categorization of these two categories indicating the role of shape similarity in object recognition (Baldassi et al., 2013).

To provide a quantitative measure for better comparisons between human and models, we computed the correlation between each model RDM and human RDM in different levels of variation (Kendall tau-a rank correlation). **Figure 11** shows the correlation between the models and human in different complexity levels and conditions (i.e., plain and natural background). The highest correlation among all is close to 0.5. The correlation between the human RDMs and model RDMs, calculated based on model features, is lower compared to RDMs obtained based on the classification responses (**Figure 11C**). After classification, the responses of several models in different levels are more correlated with human responses, **Figures 11A,B**.

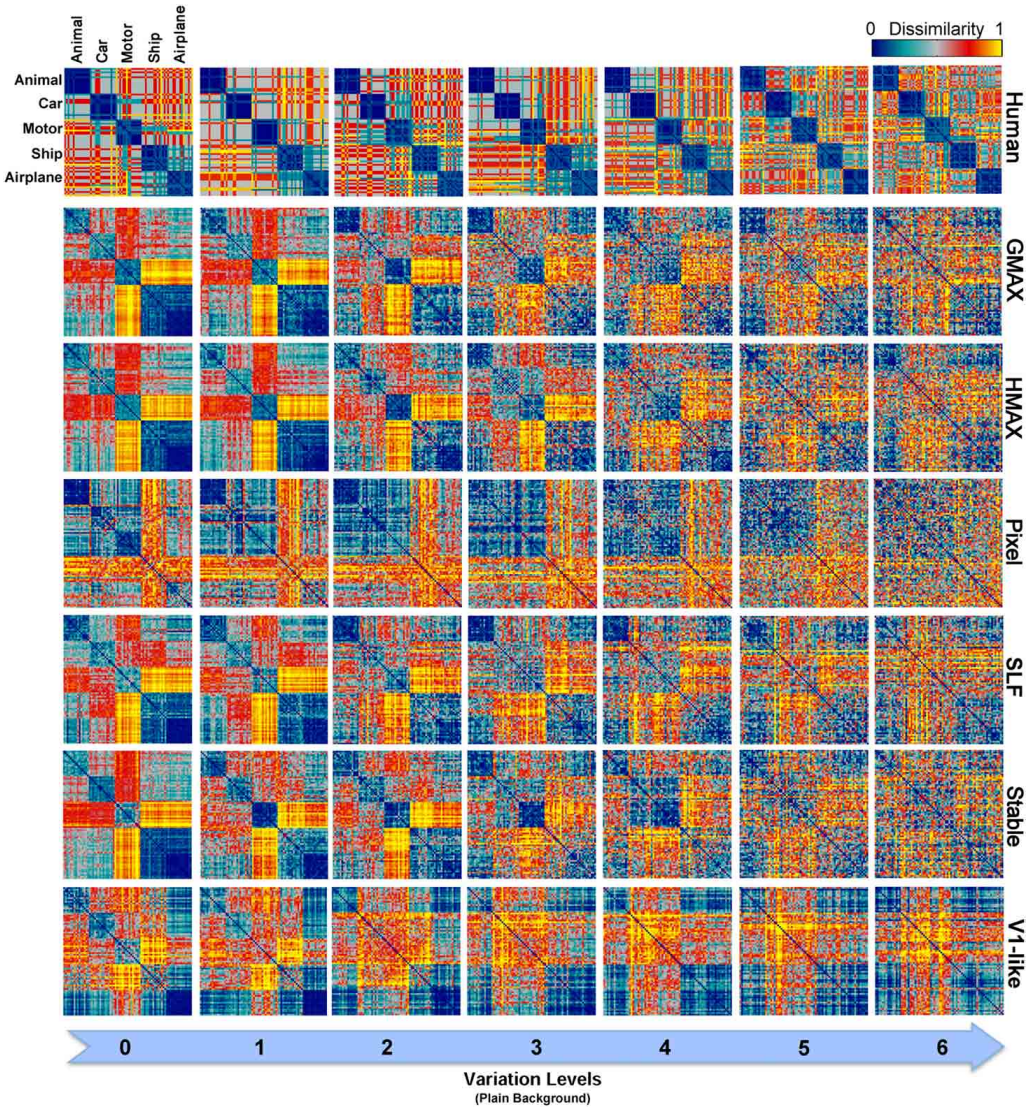


FIGURE 9 | Representational Dissimilarity Matrices (RDM) for multiclass invariant object categorization with plain background across different levels of variation, calculated based on models' features vector. Each element in a matrix shows the pairwise dissimilarities between the internal representations of a model for pairs of objects (see Materials and Methods). Each column in the figure shows the RDMs for a particular level of variation (from 0 to 6) and each row shows the RDMs of a model in different levels of variation. The first row illustrates the RDMs for human calculated based on responses in psychophysical experiments. The color bar at the top-right corner shows the degree of dissimilarity (measured as: 1-correlation— Spearman's rank correlation). The size of each matrix is 75*75. For visualization, we selected a subset of responses to images in each category (15 images from each category).

DISCUSSION

HUMANS PERFORM SIGNIFICANTLY BETTER THAN MODELS IN DISCRIMINATING OBJECTS WITH HIGH LEVEL OF VARIATIONS

Humans are very fast in categorizing natural images and different object categories (e.g., Potter and Levy, 1969; Thorpe et al., 1996; Vanrullen and Thorpe, 2001; Fabre-Thorpe, 2011). Behavioral studies have demonstrated that humans are able to identify ultra-rapidly presented images from different object categories (Kirchner and Thorpe, 2006; Mack and Palmeri, 2011; Potter et al., 2014). These studies indicate that feedforward visual processing is able to perform a great deal of different visual tasks, although limited to a certain extent (Kreiman et al., 2007; Fabre-Thorpe, 2011). Using psychophysical experiments, we showed that humans are able to remarkably perform invariant object recognition with high performance and minimum time. Although the similarity between two different views of the same object is much lower than the similarity between two different objects (Cox, 2014), human observers could accurately and quickly discriminate different objects categories in different complexity levels (both in two- and multiclass rapid categorization tasks). This task is of immense difficulty for models with many false alarms due to lack of selectivity-invariance trade-off and

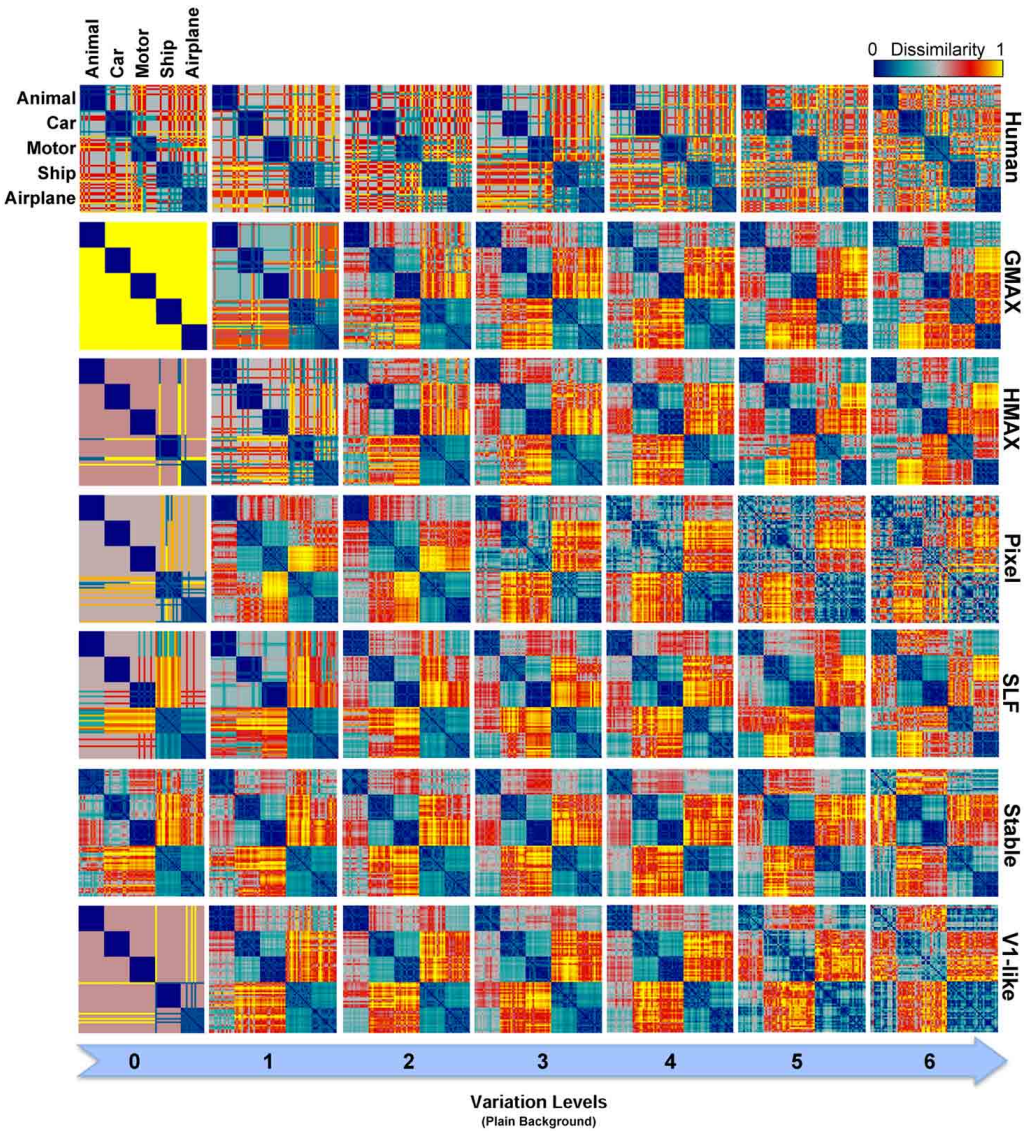


FIGURE 10 | Representational Dissimilarity Matrices (RDM) for multiclass invariant object recognition with plain background across different levels of variation, obtained based on classifier responses. Each element in a matrix shows the pairwise dissimilarities between the internal representations of a model for pairs of objects (see Materials and Methods). Each column in the figure shows the RDMs for a particular level of variation (from 0 to 6) and each row shows the RDMs of a model in different levels of variation. The first row illustrates the RDMs for human calculated based on responses in psychophysical experiments. The color bar at the top-right corner shows the degree of dissimilarity (measured as: 1-correlation—Spearman's rank correlation). The size of each matrix is 75*75. For visualization, we selected a subset of responses to images in each category (15 images from each category).

some other mechanisms, such as figure-ground segregation in cluttered images. Considering the RTs and categorization performances of human observers in the two-class rapid object categorization experiments, we saw that humans were able to respond accurately and swiftly to rapidly presented images with different levels of complexity either when objects were presented on plain backgrounds or on natural backgrounds. This contrasts with the categorization performance of models where they performed weakly in high and intermediate levels of image variation. Further explorations of the errors made in multiclass invariant object recognition, analyzed using confusion matrices, demonstrated that the error rate of the models in categorization was increased in accordance with the complexity of image variations. However, human accuracy remained high even in complex image variations; and humans performed significantly better than the models in categorizing different objects in all seven levels of image variation while objects were only presented for 25 ms.

NOT ALL IMAGE VARIATIONS YIELD THE SAME DIFFICULTY FOR THE VISUAL SYSTEM

Brain responds differently to different types of object variations. For example, size invariant representation appears earlier than

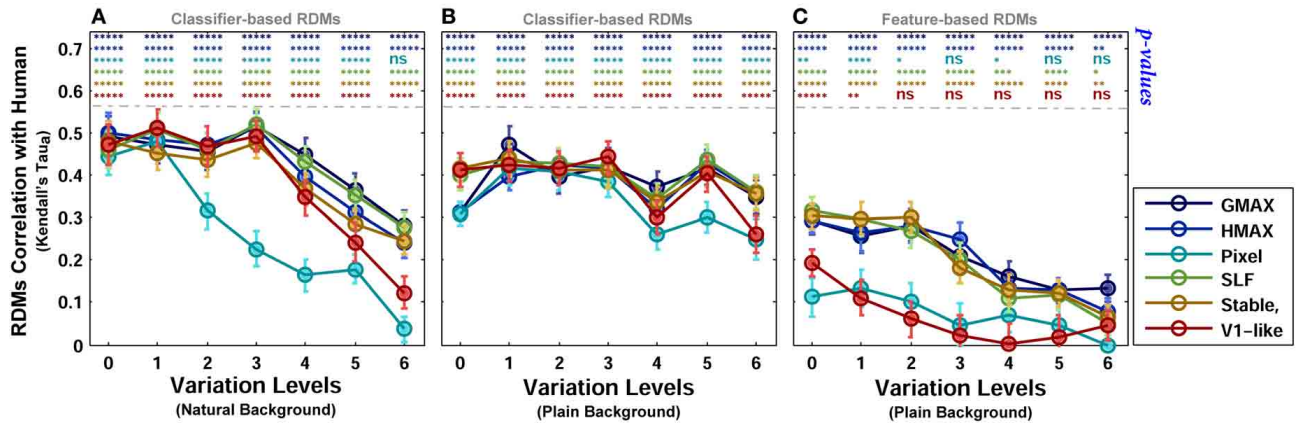


FIGURE 11 | Correlation between human and model RDMs in different background conditions and complexity levels. (A) Correlation between human RDMs and model RDMs across different levels of variation, calculated based on classifier responses, when objects were presented on natural backgrounds. **(B)** Correlation between human RDMs and model RDMs across different levels of variation, obtained based on classifier responses, when objects were presented on plain backgrounds. **(C)** Correlation between human RDMs and model RDMs across different levels of variation, obtained based on models' feature vector, when objects were presented on plain backgrounds. The *p*-values for correlation between human and each of the models are shown at the top of each plot, specified with different colors for different models (ns, means not significant; $^{*}p < 0.05$; $^{**}p < 0.005$; $^{****}p < 10^{-4}$; and $^{*****}p < 10^{-6}$). Error bars are standard deviations of the mean. Correlation results are the average over 10,000 bootstrap resamples (we used Kendall tau-a rank correlation). The RSA tool box was used for correlation calculation (Nili et al., 2014).

position (Isik et al., 2013). This invariant representation of objects is evolved across the ventral visual hierarchy (e.g., Isik et al., 2013; Yamins et al., 2014). An important, yet unanswered, question is whether different types of variations need different processing times and which one is more difficult to solve? From a modeling viewpoint, 3D variations (i.e., rotation in-depth and in-plane) are thought to be more difficult than others (Pinto et al., 2011). However, there are very few studies addressing this problem using real-world naturalistic objects with systematically controlled variations (e.g., see Pinto et al., 2008; Yamins et al., 2014). To reach this goal, we need to explore the behavioral and neural responses to different types of variations applied to real-world objects.

Another question is whether the time course of responses depend on the strength of the variations, the lower the variation, the faster the responses? Here we behaviorally showed that as the complexity level of image variation increases, the performance decreases and the RT increases. This suggests that the responses depend on the strength of variations. One potential future research would be measuring the neural responses to the strength of variations using different recording tools (e.g., EEG/MEG, fMRI and electrophysiology—e.g., Yamins et al., 2014) in different species. It would also be interesting to look at the extent to which feedforward pathway can solve invariant object recognition and whether the visual system requires prolonged exposure of object images and a supervised learning to learn invariance.

MODELS ARE MISSING A FIGURE-GROUND SEGREGATION STEP

We observed a significant increase in human RTs when objects were presented on natural backgrounds compared to plain backgrounds (**Figure 5**, pink curves compared to green curves). This suggests that some further ongoing processes occur when objects have cluttered natural backgrounds. To detect a target in a cluttered background, visual system needs to extract the boarder of the target object (object contours). This process is performed by the mechanism of figure-ground segregation in the visual cortex (Lamme, 1995). Grouping a set of collinear contour segments into a spatially extended object requires sufficient time (Roelfsema et al., 1999), even in plain background. This task is more difficult and time consuming when objects are presented in cluttered natural backgrounds. Therefore, the increase in RTs in the case of natural backgrounds could be due to the time needed for figure-ground segregation (Lamme et al., 1999; Lamme and Roelfsema, 2000).

Studies also suggest that recurrent processing is involved in figure-ground segregation (Roelfsema et al., 2002; Raudies and Neumann, 2010). This may explain why we observe a dramatic decrease in the categorization performance of the feedforward models in the natural background condition. The models are missing a figure-ground segregation step that seems to arise from interlayer and between layers feedback signals.

THE ROLE OF FEEDBACK AND FUTURE MODELING INSIGHTS

As studies show, if models can represent object categories similar to IT, they can achieve higher performances in object categorization (Khaligh-Razavi and Kriegeskorte, 2013). Moreover, the timing of several studies indicates that feedback projections may strengthen the semantic categorical clustering in IT neural representations–where objects from the same category, regardless of their variations, are clustered together (Kiani et al., 2007; Kriegeskorte et al., 2008b; Carlson et al., 2013). Therefore, considering the role of feedback in models may lead to better categorization performances when image variation is high.

Recurrent processing can play a pivotal role in object recognition and can help the visual system to make responses that

are more robust to noise and variations (Lamme and Roelfsema, 2000; Wyatte et al., 2012; O'Reilly et al., 2013). Having said that, the results of our behavioral experiments demonstrated that even with very fast presentation of images with different levels of variations, human observers perform considerably well. One explanation is that the high categorization performances are not simply the results of initial responses in higher visual areas due to the feedforward sweep. Indeed early category-related responses, which emerge at about 150 ms after stimulus onset, may already involve recurrent activity between higher and lower areas (Koivisto et al., 2011). Another explanation could be that the IT representational geometry in this condition is not strongly categorical—this can be tested with fMRI in future studies—and so object categories are not linearly separable, but perhaps in later stages of the hierarchy (i.e., in PFC) the categorical representation gets stronger, which allows subjects to perform well. It would be interesting to investigate whether a linear read-out can decode the presented objects from the IT representation when recurrent processing is disrupted. Understanding the role of feedforward vs. recurrent processing in invariant object recognition opens a new avenue toward solving the computational crux of object recognition.

FUTURE DIRECTIONS FOR UNDERSTANDING HOW/WHEN/WHERE THE INVARIANT REPRESENTATION EMERGES ACROSS THE HIERARCHY OF HUMAN VISUAL SYSTEM

It is of great importance to investigate not only where the categorical information emerges in the ventral visual pathway (Henriksson et al., 2013), but also when the representations of stimuli in the brain reaches to a level that shows categorical information clearly (Cichy et al., 2014). Having accurate temporal and spatial information of object representation in the brain can help us to know where the invariant representations emerge and how long it takes to have sufficient information about them. This can help us to understand how neural representations evolve over time and different stages in the ventral visual system that finally result in this remarkable performance in invariant object recognition without losing specificity to distinguish between similar exemplars. Moreover, it opens new ways for developing models that have similar representations and performance to the primates' brain (Yamins et al., 2014).

We need to exploit new recording technologies, such as high-resolution fMRI, MEG, and cutting-edge cell recording, to simultaneously record large population of neurons throughout the hierarchy, and advanced computational analyses (Kriegeskorte et al., 2008b; Naselaris et al., 2011; Haxby et al., 2014) in order to understand the mechanisms of invariant object recognition. This would help us to understand when and where invariant responses emerge in response to naturalistic object images with controlled image variations such as our database.

ACKNOWLEDGMENTS

We would like to thank Hamed Nili for his help and support in using the RSA toolbox. This work was funded by Cambridge Overseas Trust and Yousef Jameel Scholarship to Seyed-Mahdi Khaligh-Razavi.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fncom.2014.00074/abstract

DEEP SUPERVISED CONVOLUTIONAL NEURAL NETWORK vs. HUMANS

In addition to the models we discussed in the paper, we also tested a recent deep supervised convolutional network (Krizhevsky et al., 2012) that has been shown to be successful in different object classification tasks. The model is trained with extensive supervision (over a million labeled training images).

Given that all the feedforward models discussed so far failed to reach human level performance in higher levels of image variation, we were interested to see how a deeper feedforward model that is supervised with more training images will perform in our invariant object recognition task. Similar to other experiments, we compared the model performance against humans in two binary (animal vs. car and motorcycle vs. car) and one multiclass invariant object categorization tasks, both with plain and natural background. The results show that in high image variations humans perform significantly better than the model (**Figure S3**). Particularly, in all tasks, when the image variation is 4 or higher, humans are always better.

Figure S1 | Sample images in different levels of variation with natural backgrounds. Object images, rendered from 3D planes, vary in four dimensions: size, position (x, y), rotation in-depth, and rotation in plane, superimposed on randomly selected natural background.

Figure S2 | Representational Dissimilarity Matrices (RDM) for multiclass invariant object categorization task with natural background across different levels of variations, obtained based on classifier responses. Each element in a matrix shows the pairwise dissimilarities between the internal representations of a model for pairs of objects (see Materials and Methods). Each column in the figure shows the RDMs for a particular level of variation (from 0 to 6) and each row shows the RDMs of a model in different levels of variation. The first row illustrates the RDMs for human calculated based on responses in psychophysical experiments. The color bar at the top-right corner shows the degree of dissimilarity (measured as: 1-correlation- Spearman's rank correlation). The size of each matrix is 75*75. For visualization, we selected a subset of responses to images in each category (15 images from each category).

Figure S3 | The performance of the Deep Convolutional Neural Network (DCNN) in invariant object categorization tasks. (A) Performances in animal vs. car categorization task across different levels of variation. Black curve shows human performance and green curve shows the performance of DCNN. The top plot illustrates the performances when objects were presented on plain backgrounds and the bottom plot shows the performances when objects were presented on natural backgrounds. *P*-values for comparisons between human and the model across different levels of variation are depicted at the top of each plot (Wilcoxon signed-rank test). **(B)** Performances in motorcycle vs. car invariant categorization task across different levels of variation. The top plot illustrates the performances when objects were presented on plain backgrounds and the bottom plot shows the performances when objects were presented on natural backgrounds. *P*-values for comparisons

between human and the model across different levels of variation are depicted at the top of each plot (Wilcoxon signed-rank test). The results are the average of 15 independent random runs and the error bars show the standard deviation of the mean. **(C)** Performance comparisons between DCNN and human in a multiclass invariant object recognition task. Left plot shows the performance comparison when objects were presented on plain backgrounds while the right plot shows the performances when objects were presented on natural backgrounds. **(D)** Representational Dissimilarity Matrices (RDM) for DCNN in multiclass invariant object recognition with plain (left column) and natural (right column) background across different levels of variation, calculated based on models' feature vector. Each element in a matrix shows pairwise dissimilarities between the internal representations of the model for pairs of objects. The color bar at the top-right shows the degree of dissimilarity (measured as: 1-correlation- Spearman's rank correlation). For visualization, we selected a subset of responses to images in each category (15 images from each category), meaning that the size of each matrix is 75*75. **(E)** Correlation between human and DCNN RDMs (based on DCNN model features from the last layer of the model) in different background conditions and complexity levels. Correlations across all levels are significant (*****$p < 10^{-6}$). Error bars are standard deviations of the mean. *P*-values are obtained by bootstrap resampling the images. The correlation results are the average over 10,000 bootstrap resamples (we used Kendall tau-a rank correlation).

REFERENCES

Alemi-Neissi, A., Rosselli, F. B., and Zoccolan, D. (2013). Multifeatural shape processing in rats engaged in invariant visual object recognition. *J. Neurosci.* 33, 5939–5956. doi: 10.1523/JNEUROSCI.3629-12.2013

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., and Poggio, T. (2013). Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint* arXiv:1311.4158.

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., and Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Comput. Biol.* 9:e1003167. doi: 10.1371/journal.pcbi.1003167

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897X00357

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597. doi: 10.1523/JNEUROSCI.3726-05.2005

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13, 1–19. doi: 10.1167/13.10.1

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., and Turret, J. (2011). High temporal resolution decoding of object position and category. *J. Vis.* 11, 1–17. doi: 10.1167/11.10.9

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Cox, D. D. (2014). Do we understand high-level vision? *Curr. Opin. Neurobiol.* 25, 187–193. doi: 10.1016/j.conb.2014.01.016

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint* arXiv:1310.1531.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Front. Psychol.* 2:243. doi: 10.3389/fpsyg.2011.00243

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* 16, 974–981. doi: 10.1038/nn.3402

Freiwald, W. A., and Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330, 845–851. doi: 10.1126/science.1194908

Ghodrati, M., Khaligh-Razavi, S.-M., Ebrahimpour, R., Rajaei, K., and Pooyan, M. (2012). How can selection of biologically inspired features improve the performance of a robust object recognition model? *PLoS ONE* 7:e32357. doi: 10.1371/journal.pone.0032357

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* 23, 121–134. doi: 10.1007/BF00344744

Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* 37. doi: 10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Heinen, K., Jolij, J., and Lamme, V. A. (2005). Figure-ground segregation requires two distinct periods of activity in V1: a transcranial magnetic stimulation study. *Neuroreport* 16, 1483–1487. doi: 10.1097/01.wnr.0000175611.26485.c8

Henriksson, L., Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2013). Population-code representations of natural images across human visual areas. *J. Vis.* 13, 1035–1035. doi: 10.1167/13.9.1035

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hubel, D., and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866. doi: 10.1126/science.1117593

Isik, L., Meyers, E. M., Leibo, J. Z., and Poggio, T. A. (2013). The dynamics of invariant object recognition in the human visual system. *J. Neurophys.* 111, 91–102. doi: 10.1152/jn.00394.2013

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). "What is the best multi-stage architecture for object recognition?" in *Computer Vision, 12th International Conference on (IEEE)* (Kyoto), 2146–2153.

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2013). "Object-vision models that better explain IT also categorize better, but all models fail at both," in COSYNE (Salt Lake City, UT).

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97, 4296–4309. doi: 10.1152/jn.00024.2007

Kirchner, H., and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Res.* 46, 1762–1776. doi: 10.1016/j.visres.2005.10.002

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., and Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *J. Neurosci.* 31, 2488–2492. doi: 10.1523/JNEUROSCI.3074-10.2011

Kreiman, G., Hung, C. P., Kraskov, A., Quiroga, R. Q., Poggio, T., and DiCarlo, J. J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron* 49, 433–445. doi: 10.1016/j.neuron.2005.12.019

Kreiman, G., Serre, T., and Poggio, T. (2007). On the limits of feed-forward processing in visual object recognition. *J. Vis.* 7, 1041–1041. doi: 10.1167/7.9.1041

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Front. Neurosci.* 3:363. doi: 10.3389/neuro.01.035.2009

Kriegeskorte, N., and Mur, M. (2012). "Representational similarity analysis of object population codes in humans, monkeys, and models," in *Visual Population Codes: Towards a Common Multivariate Framework for Cell Recording and Functional Imaging*, eds N. Kriegeskorte and G. Kreiman (Cambridge, MA: MIT Press), 659.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis–connecting the branches of systems neuroscience. *Front. Sys. Neurosci.* 2,4. doi: 10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *NIPS* (Lake Tahoe, NV), 4.

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* 15, 1605–1615.

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X

Lamme, V. A. F., Rodriguez-Rodriguez, V., and Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cereb. Cortex* 9, 406–413. doi: 10.1093/cercor/9.4.406

LeCun, Y., and Bengio, Y. (1998). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (Cambridge, MA: MIT Press), 255–258.

Liao, Q., Leibo, J. Z., and Poggio, T. (2013). "Learning invariant representations and applications to face verification," in *NIPS* (Lake Tahoe, NV), 3057–3065.

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Mack, M. L., and Palmeri, T. J. (2011). The timing of visual object categorization. *Front. Psychol.* 2:165. doi: 10.3389/fpsyg.2011.00165

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., et al. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* 522, 225–259. doi: 10.1002/cne.23458

Mutch, J., and Lowe, D. G. (2006). "Multiclass object recognition with sparse, localized features," in *CVPR*, 11–18. doi: 10.1109/CVPR.2006.200

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80, 45–57. doi: 10.1007/s11263-007-0118-0

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Comp. Biol.* 10:e1003553. doi: 10.1371/journal.pcbi.1003553

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., and Jilk, D. J. (2013). Recurrent processing during object recognition. *Front. Psychol.* 4:124. doi: 10.3389/fpsyg.2013.00124

Peissig, J. J., and Tarr, M. J. (2007). Visual object recognition: do we know more now than we did 20 years ago? *Annu. Rev. Psychol.* 58, 75–96. doi: 10.1146/annurev.psych.58.102904.190114

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442. doi: 10.1163/156856897X00366

Pinto, N., Barhomi, Y., Cox, D. D., and DiCarlo, J. J. (2011). "Comparing state-of-the-art visual features on invariant object recognition tasks," in *Applications of Computer Vision (WACV)* (Kona, HI), 463–470. doi: 10.1109/WACV.2011.5711540

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput. Biol.* 4:e27. doi: 10.1371/journal.pcbi.0040027

Poggio, T., and Ullman, S. (2013). Vision: are models of object recognition catching up with the brain? *Ann. N.Y. Acad. Sci.* 1305, 72–82. doi: 10.1111/nyas.12148

Potter, M. C., and Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *J. Exp. Psychol.* 81, 10. doi: 10.1037/h0027470

Potter, M. C., Wyble, B., Hagmann, C. E., and McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Atten. Percep. Psychophys.* 76, 270–279. doi: 10.3758/s13414-013-0605-z

Rajaei, K., Khaligh-Razavi, S.-M., Ghodrati, M., Ebrahimpour, R., and Shiri Ahmad Abadi, M. E. (2012). A stable biologically motivated learning mechanism for visual feature extraction to handle facial categorization. *PLoS ONE* 7:e38478. doi: 10.1371/journal.pone.0038478

Raudies, F., and Neumann, H. (2010). A neural model of the temporal dynamics of figure–ground segregation in motion perception. *Neural Netw.* 23, 160–176. doi: 10.1016/j.neunet.2009.10.005

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., et al. (2012). Toward a unified theory of visual area V4. *Neuron* 74, 12–29. doi: 10.1016/j.neuron.2012.03.011

Roelfsema, P. R., Lamme, V. A., Spekreijse, H., and Bosch, H. (2002). Figure—ground segregation in a recurrent network architecture. *J. Cogn. Neurosci.* 14, 525–537. doi: 10.1162/08989290260045756

Roelfsema, P. R., Scholte, H. S., and Spekreijse, H. (1999). Temporal constraints on the grouping of contour segments into spatially extended objects. *Vision Res.* 39, 1509–1529. doi: 10.1016/S0042-6989(98)00222-3

Rolls, E. T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Front. Comp. Neurosci.* 6:35. doi: 10.3389/fncom.2012.00035

Rust, N. C., and DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* 30, 12978–12995. doi: 10.1523/JNEUROSCI.0179-10.2010

Rust, N. C., and DiCarlo, J. J. (2012). Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *J. Neurosci.* 32, 10170–10182. doi: 10.1523/JNEUROSCI.6125-11.2012

Scholte, H. S., Jolij, J., Fahrenfort, J. J., and Lamme, V. A. (2008). Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging. *J. Cogn. Neurosci.* 20, 2097–2109. doi: 10.1162/jocn.2008.20142

Serre, T., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Sharpee, T. O., Kouh, M., and Reynolds, J. H. (2013). Trade-off between curvature tuning and position invariance in visual area V4. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11618–11623. doi: 10.1073/pnas.1217479110

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687. doi: 10.1038/nn870

Vanrullen, R., and Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci.* 13, 454–461. doi: 10.1162/08989290152001880

Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194. doi: 10.1016/S0301-0082(96)00054-8

Wood, J. N. (2013). Newborn chickens generate invariant object representations at the onset of visual object experience. *Proc. Natl. Acad. Sci. U.S.A.* 110, 14000–14005. doi: 10.1073/pnas.1308246110

Wyatte, D., Curran, T., and O'Reilly, R. (2012). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *J. Cogn. Neurosci.* 24, 2248–2261. doi: 10.1162/jocn_a_00282

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* 27, 12292–12307. doi: 10.1523/JNEUROSCI.1897-07.2007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 April 2014; accepted: 28 June 2014; published online: 18 July 2014.
Citation: Ghodrati M, Farzmahdi A, Rajaei K, Ebrahimpour R and Khaligh-Razavi S-M (2014) Feedforward object-vision models only tolerate small image variations compared to human. Front. Comput. Neurosci. ***8****:74. doi: 10.3389/fncom.2014.00074*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Ghodrati, Farzmahdi, Rajaei, Ebrahimpour and Khaligh-Razavi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

MINI REVIEW ARTICLE
published: 22 October 2014
doi: 10.3389/fncom.2014.00135

Why vision is not both hierarchical *and* feedforward

Michael H. Herzog* and Aaron M. Clarke

Laboratory of Psychophysics, Brain, Mind Institute, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Edited by:
Antonio J. Rodríguez-Sánchez, University of Innsbruck, Austria

Reviewed by:
Michael Zillich, Vienna University of Technology, Austria
Norbert Küger, The Maersk Mc-Kinney Moller Institute, Denmark

****Correspondence:***
Michael H. Herzog, Laboratory of Psychophysics, Brain, Mind Institute, École Polytechnique Fédérale de Lausanne, EPFL SV BMI LPSY SV-2807, Station 19, CH-1015 Lausanne, Switzerland
e-mail: michael.herzog@epfl.ch

In classical models of object recognition, first, basic features (e.g., edges and lines) are analyzed by independent filters that mimic the receptive field profiles of V1 neurons. In a feedforward fashion, the outputs of these filters are fed to filters at the next processing stage, pooling information across several filters from the previous level, and so forth at subsequent processing stages. Low-level processing determines high-level processing. Information lost on lower stages is irretrievably lost. Models of this type have proven to be very successful in many fields of vision, but have failed to explain object recognition in general. Here, we present experiments that, first, show that, similar to demonstrations from the Gestaltists, figural aspects determine low-level processing (as much as the other way around). Second, performance on a single element depends on all the other elements in the visual scene. Small changes in the overall configuration can lead to large changes in performance. Third, grouping of elements is key. Only if we know how elements group across the entire visual field, can we determine performance on individual elements, i.e., challenging the classical stereotypical filtering approach, which is at the very heart of most vision models.

Keywords: feedback, object recognition, crowding, Verniers, Gestalt

Object recognition traditionally proceeds from the analysis of simple to complex features. The Gestaltists proposed a number of basic rules, such as spatial proximity and good continuation, that underlie the grouping of elements into objects. Whereas the Gestalt rules work well for very basic stimuli, they grossly fail for slightly more complex stimuli. For this reason, research on Gestalt principles almost disappeared after the 1930's. After world war II, the discovery of the receptive field advanced vision science by revealing fundamental principles of retinal and cortical processing, which has led to a core scenario that is often, explicitly or implicitly, behind most models in visual neuroscience and the psychology of perception, and provides the basis for most models in computer vision.

The model is characterized by its hierarchical and feedforward organization (**Figure 1**). Neurons in lower visual areas, with small receptive fields, are sensitive to basic visual features. For example, neurons in V1 respond predominantly to edges and lines. These neurons project to neurons at the next stage of the hierarchy, which code for more complex features. By V4, the neurons are selective for basic shapes, and by IT they respond in a viewpoint-invariant manner to full objects. Decisions making happens in the frontal cortex. This basic scenario has a well-defined set of characteristics. Processing is hierarchical, feedforward, and local on each level, i.e., only neighboring neurons, coding for neighboring parts in the visual field, project to a common higher-level neuron (**Figure 1**). In addition, processing at one stage is fully determined by processing at the previous stage. Information lost at previous stages is irretrievably lost. Processing follows an atomistic, Lego® building block type of encoding. For example, a hypothetical "square neuron" is created by feedforward projections from "lower" neurons coding for vertical and horizontal lines (**Figure 1**; Riesenhuber and Poggio, 1999; Hung et al., 2005; Serre et al., 2005, 2007a,b). Finally, there is an isomorphism between objects of the outer world (e.g., a blue line), basic neural circuitry (analyzing the blue line), and the corresponding percept ("blue line"). And this is exactly the beauty of these models: naturalizing the subjectivity of perception by identifying the basic neural circuits of perception.

Evidence for fast, hierarchical feedforward processing comes from experiments showing that humans can detect animals in a scene in less than 150 ms. Calculations based on neural conduction velocity show that there are only one or two spikes per cortical area before a decision is made, arguing strongly against feedback processing (Thorpe et al., 2001).

Computer vision models often follow closely the philosophy of neurobiological feedforward hierarchies. In these, as in neurobiological models, first, basic features are extracted, for example, through V1-style Gabor filtering or Haar wavelets. Often, the downstream hierarchical stages (V2, V4) are collapsed into one processing stage, where a classifier is trained to detect specialized objects such as faces or cars. Similar to IT neurons, these detectors are often scale- and viewpoint-invariant (Biederman, 1987; Ullman et al., 2002; Fink and Perona, 2003; Torralba, 2003; Schneiderman and Kanade, 2004; Viola and Jones, 2004; Felzenszwalb and Huttenlocher, 2005; Fei-Fei et al., 2006; Amit and Trouvé, 2007; Fergus et al., 2007; Heisele et al., 2007; Hoiem et al., 2008; Wu et al., 2010).

Here, we will present experiments from crowding research that challenge classical feedforward hierarchy models. In crowding, target discriminability strongly deteriorates when neighboring elements are presented (**Figure 2**). Crowding is often seen as a breakdown of object recognition and most models

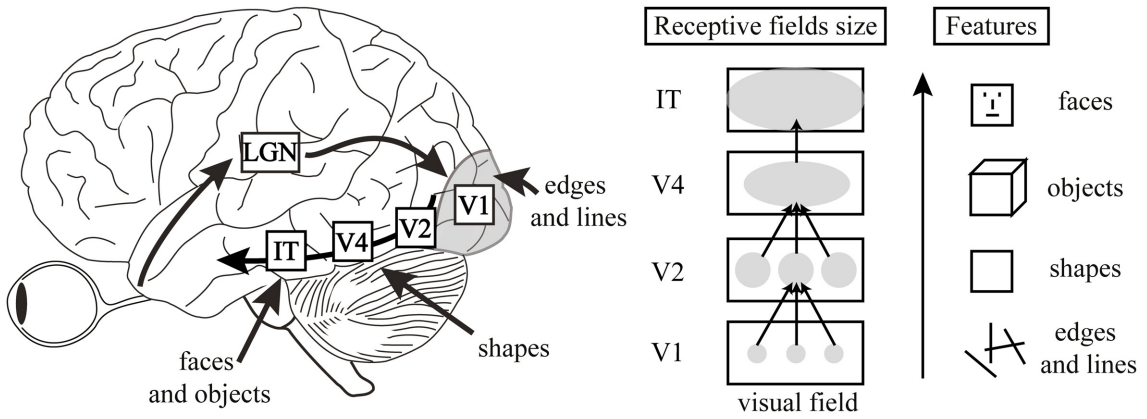


FIGURE 1 | Left: A typical hierarchical, feedforward model, where information processing starts at the retina, proceeds to the LGN, then to V1, V2, V4, and IT. Decisions about stimuli are made in the frontal cortex. **Center:** Lower visual areas have smaller receptive fields, while neurons in higher areas have gradually increasing receptive field sizes, integrating information over larger and larger regions of the visual field. **Right:** Lower visual areas, such as V1, code for basic features such as edges and lines. Higher-level neurons pool information over multiple low-level neurons with smaller receptive fields and code for more complex features. There is thus a hierarchy of features. Figure adapted from Manassi et al. (2013).

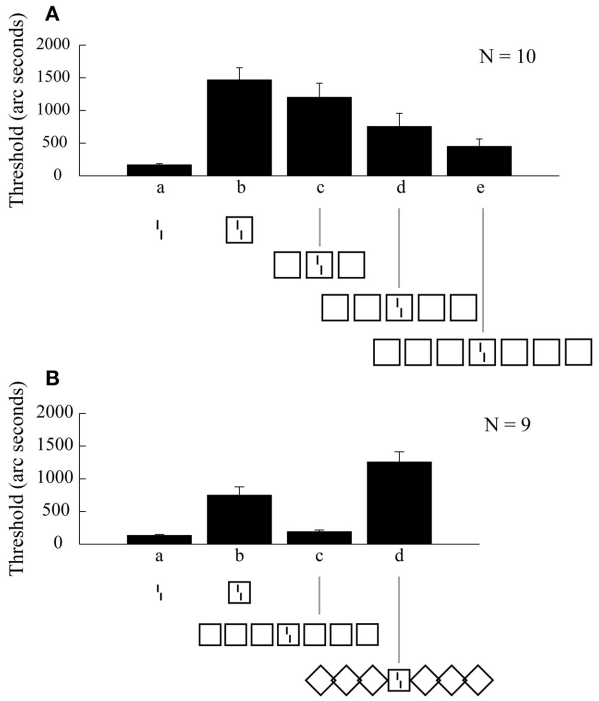


FIGURE 2 | Vernier offset discrimination as a function of stimulus configuration. (A). The reference stimulus is the un-flanked vernier shown in **a**. Enclosing the vernier in a square deteriorates performance. Adding additional squares leads to increasingly better performance. **(B)**. We replicated the results with the squares (b,c). In addition, rotating the flanking squares to form diamonds (d) undoes the effect of grouping and reinstates the crowding effect. From Manassi et al. (2013).

of crowding are very much in the spirit of object recognition models. In pooling models, information from lower-level neurons is pooled by higher-level neurons, to see wholes at the cost of more poorly perceiving the parts. Indeed, observers can clearly *detect* a crowded target, it is only its features and spatial relationships that are jumbled with flanker features (e.g., Pelli et al., 2004). Target-feature perception is lost because target and distracter features are pooled. A prediction made by pooling models is that, because spatial integration is local at each stage, only nearby elements deteriorate target discriminability (Bouma's law). In addition, if more flankers are added within Bouma's window, performance should deteriorate (or at least not improve) because the signal-to-noise ratio decreases. A third prediction is that adding more flankers should deteriorate performance.

In previous experiments, we presented a vernier stimulus, which consists of two vertical lines, offset slightly to the left or right (Manassi et al., 2012, 2013). Observers indicated the offset direction. Verniers were presented in the periphery, 9 degrees (of visual angle) to the right of fixation. Performance strongly deteriorated when the vernier was surrounded by a square (**Figures 2Aa, b**). This is a classic crowding effect and is well-explained by traditional crowding models. Next, Manassi et al. (2012, 2013) presented 2 × 3 neighboring squares (**Figure 2Ae**). According to pooling models, and most object recognition models, more flankers should deteriorate performance. However, the opposite was the case. Crowding almost disappeared. Interestingly, this *un*crowding effect increased with the number of squares that were presented (**Figure 2A**). Importantly, the fixation dot was only 0.5 degrees apart from the left-most square, i.e., the stimulus configuration extended over large parts of the right visual field. Hence, vernier offset discrimination is influenced by elements far outside the integration region predicted by Bouma's law. Second, and more importantly, vernier offset discrimination is influenced by the overall stimulus configuration. This becomes evident when turning the flanking squares by 90° creating diamonds, resulting in the return of the crowding effect (**Figure 2B**). Hence, figural aspects determine basic feature processing (Wolford and Chambers, 1983; Livne and Sagi, 2007; Malania et al., 2007; Sayim et al., 2010).

Our results clearly show that simple pooling models cannot explain crowding and the same seems to be true for most basic models of object recognition. Figural processing determines low-level processing as much as low-level processing determines figural processing. It seems that first the squares are computed from their constituting lines. Next square representations interact with each other and the outputs of this processing determine the vernier offset discriminability. This is reminiscent of the famous quote by Wertheimer that "the whole determines the appearance of the parts" (Wertheimer, 1938). In our example, the whole determines even low-level processing. It also agrees with more modern sentiments suggesting that feedback is crucial for normal vision at all levels of the processing hierarchy (Krüger et al., 2013). We propose that it is only when we know how elements group together that we will be able to accurately predict performance on even the simplest tasks, i.e., without understand grouping across the entire visual field, it is impossible to understand human object recognition.

Note here, that we are not claiming that the visual system is not hierarchical. Nor are we claiming that there is no feedforward sweep through the cortex. We are arguing against models that are both feedforward *and* contain a strict feature hierarchy. For example, classic models posit that low-level features (such as verniers) are encoded at an early cortical level and that shapes (such as squares) are encoded at a later cortical level. Square-square interactions are crucial, as we have shown. However, since there are no feedback connections, the classic models cannot explain how square-square interactions change low-level processing of the vernier. One solution is to give up feedforward processing and have *recurrent* interactions between lower and higher levels of processing.

Evidence for recurrent processing comes from timing experiments on the dynamics of grouping in crowding (Manassi and Herzog, 2013; Manassi, 2014). A vernier target was flanked by either two vertical lines, or by two vertical lines that formed the edges of two cuboids. In both cases, the vertical lines were identical and only the surrounding context differed—the lines grouped with the vernier, but when they were part of the cuboids, the lines segmented from the vernier. Vernier offset discrimination thresholds were measured as a function of stimulus presentation time for seven fixed durations ranging from 20 ms to 640 ms. Under brief presentation times ($\leq$120 ms) performance in the two stimulus conditions did not significantly differ. Beyond 160 ms, however, performance with the cuboids was significantly better than with the lines. These results indicate that perceptual grouping evolves with time, even for such basic stimuli as verniers. Current models of vernier offset discrimination show that this task can be achieved in a feedforward way by reading out the responses of orientation-tuned V1 neurons (Wilson, 1986)—a process that takes on the order of 50 ms (Cottaris and De Valois, 1998; Gershon et al., 1998). Sending spikes to additional synapses requires at least 10 ms per spike. Thus, the long time required for vernier discrimination in the cuboid flanker condition to be differentiated from line flanker condition ($\geq$160 ms, i.e., more than double the arrival time of the stimulus at V1) indicates significant additional cortical processing for perceptual grouping. Since 160 ms $-$50 ms $=$ 110 ms, at least 11 additional synaptic connections could be activated. Recent electrophysiological evidence suggests that the additional time can be accounted for by feedback connections from the lateral occipital cortex to earlier cortical areas, the result of which is the promotion of perceptual grouping (Shpaner et al., 2013).

Our results are not restricted to crowding but occur in many other visual paradigms including overlay masking (Saarela and Herzog, 2008, 2009), backward masking (Herzog, 2007; Hermens and Herzog, 2007; Dombrowe et al., 2009), letter recognition (Saarela et al., 2010), in haptics (Overvliet and Sayim, 2013), and in audition (Oberfeld et al., 2012).

Why does the processing of an element's basic features depend on remote elements? Vision is ill-posed. For example, the light (luminance) that arrives at the retina is a product of the light shining on an object (illuminance) and the material properties of the object (reflectance). Hence, on the photoreceptor level, it is impossible to determine whether or not a banana is yellow and ready to eat. The brain tries to solve this problem by discounting the illuminance, taking contextual information into account. This becomes obvious in the case of computing material properties. Glossy objects, for example, reflect bright spots (specularities) in regions of high curvature. Removal or addition of an object's specularities completely changes the object's perceived material, in spite of the fact that the rest of the object remains the same. To compute the material properties, integrating information across the visual field is crucial: where is the illuminance coming from? What is the shape of the object?

Key then, is that without knowing the whole one cannot know the parts. To the best of our knowledge, very few models adopt this approach of including recurrent processing and effectively integrating information over large parts of the visual field. Not surprisingly, these models are highly effective at modeling human data, not only from crowding, but also from many other areas of cognitive science, hinting at their general ability to explain cortical processing. For example, they effectively explain data pertaining to attention (Tsotsos, 1995; Tsotsos et al., 1995; Cutzu and Tsotsos, 2003; Bruce and Tsotsos, 2005, 2009; Rodriguez-Sanchez et al., 2007), and visual object learning (Bengio et al., 2013; Goodfellow et al., 2013; Salakhutdinov et al., 2013). They also do well at scene-segmentation, where successful models typically use a global approach, such as coarse-to-fine image pyramids (Estrada and Elder, 2006) or normalized cuts over extended graphs (Malik et al., 1999, 2001; Shi and Malik, 2000; Ren and Malik, 2002; Martin et al., 2004), which are leveraged to produce human-like scene segmentations. Here, again, computations are not purely local and feedforward, but rather global and iterative. Grossberg has also produced similar models in terms of their ability to do grouping that extends over a scene (Grossberg and Mignolla, 1985; Dresp and Grossberg, 1997), as has Francis (Francis et al., 1994; Francis and Grossberg, 1995). Future work will show whether these models can explain our particular crowding results.

In summary, there is a wealth of evidence suggesting that cortical processing is not purely hierarchical *and* feed-forward. In order to know how the visual system processes fine-grained

information at a particular location it is necessary to integrate information about the surrounding context over the entire visual field. Grouping and segmentation are crucial to understanding vision, and must be understood on a global scale.

FUNDING

Grant number 320030_135741.

ACKNOWLEDGMENTS

Aaron Clarke was funded by the Swiss National Science Foundation (SNF) project "Basics of visual processing: what crowds in crowding?" (Project number: 320030_135741).

REFERENCES

Amit, Y., and Trouvé, A. (2007). Pop: Patchwork of parts models for object recognition. *Int. J. Comput. Vis.* 75, 267–282. doi: 10.1007/s11263-006-0033-9

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *Patt. Analys. Mach. Intell. IEEE Trans.* 35, 1–30. doi: 10.1109/TPAMI.2013.50

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115. doi: 10.1037/0033-295X.94.2.115

Bruce, N., and Tsotsos, J. (2005). Saliency based on information maximization. *Adv. Neural Inf. Process. Syst.* 18, 1–8. Available online at: http://papers.nips.cc/paper/2830-saliency-based-on-information-maximization

Bruce, N. D., and Tsotsos, J. K. (2009). Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* 9:5. doi: 10.1167/9.3.5

Cottaris, N. P., and De Valois, R. L. (1998). Temporal dynamics of chromatic tuning in macaque primary visual cortex. *Nature* 395, 896–900. doi: 10.1038/27666

Cutzu, F., and Tsotsos, J. K. (2003). The selective tuning model of attention: psychophysical evidence for a suppressive annulus around an attended item. *Vis. Res.* 43, 205–219. doi: 10.1016/S0042-6989(02)00491-1

Dombrowe, I., Hermens, F., Francis, G., and Herzog, M. H. (2009). The roles of mask luminance and perceptual grouping in visual backward masking. *J. Vis.* 9:22. doi: 10.1167/9.11.22

Dresp, B., and Grossberg, S. (1997). Contour integration across polarities and spatial gaps: from local contrast filtering to global grouping. *Vis. Res.* 37, 913–924. doi: 10.1016/S0042-6989(96)00227-1

Estrada, F. J., and Elder, J. H. (2006). "Multi-scale contour extraction based on natural image statistics," in *Proc. IEEE Workshop on Perceptual Organization in Computer Vision* (New York, NY:IEEE), 183.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *Patt. Analys. Mach. Intell. IEEE Trans.* 28, 594–611. doi: 10.1109/TPAMI.2006.79

Felzenszwalb, P. F., and Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *Int. J. Comput. Vis.* 61, 55–79. doi: 10.1023/B:VISI.0000042934.15159.49

Fergus, R., Perona, P., and Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. *Int. J. Comput. Vis.* 71, 273–303. doi: 10.1007/s11263-006-8707-x

Fink, M., and Perona, P. (2003). "Mutual boosting for contextual inference," in *NIPS*, Vol. 2 (Vancouver, BC), 7.

Francis, G., and Grossberg, S. (1995). *Cortical Dynamics of Boundary Segmentation and Reset: Persistence, Afterimages, and Residual Traces.* Technical report, Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems.

Francis, G., Grossberg, S., and Mingolla, E. (1994). Cortical dynamics of feature binding and reset: control of visual persistence. *Vis. Res.* 34, 1089–1104. doi: 10.1016/0042-6989(94)90012-4

Gershon, E. D., Wiener, M. C., Latham, P. E., and Richmond, B. J. (1998). Coding strategies in monkey v1 and inferior temporal cortices. *J. Neurophysiol.* 79, 1135–1144.

Goodfellow, I. J., Courville, A., and Bengio, Y. (2013). Scaling up spike-and-slab models for unsupervised feature learning. *Pattern Anal. Mach. Intell. IEEE Trans.* 35, 1902–1914. doi: 10.1109/TPAMI.2012.273

Grossberg, S., and Mignolla, E. (1985). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychol. Rev.* 92, 173–211. doi: 10.1037/0033-295X.92.2.173

Heisele, B., Serre, T., and Poggio, T. (2007). A component-based framework for face detection and identification. *Int. J. Comput. Vis.* 74, 167–181. doi: 10.1007/s11263-006-0006-z

Hermens, F., and Herzog, M. H. (2007). The effects of the global structure of the mask in visual backward masking. *Vis. Res.* 47, 1790–1797. doi: 10.1016/j.visres.2007.02.020

Herzog, M. H. (2007). Spatial processing and visual backward masking. *Adv. Cogn. Psychol.* 3, 85–92. doi: 10.2478/v10053-008-0016-1

Hoiem, D., Efros, A. A., and Hebert, M. (2008). Putting objects in perspective. *Int. J. Comput. Vis.* 80, 3–15. doi: 10.1007/s11263-008-0137-5

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866. doi: 10.1126/science.1117593

Krüger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., et al. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1847–1871. doi: 10.1109/TPAMI.2012.272

Livne, T., and Sagi, D. (2007). Configuration influence on crowding. *J. Vis.* 7:4. doi: 10.1167/7.2.4

Malania, M., Herzog, M. H., and Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *J. Vis.* 7, 1–7. doi: 10.1167/7.2.1

Malik, J., Belongie, S., Leung, T., and Shi, J. (2001). Contour and texture analysis for image segmentation. *Int. J. Comput. Vis.* 43, 7–27. doi: 10.1023/A:1011174803800

Malik, J., Belongie, S., Shi, J., and Leung, T. (1999). "Textons, contours and regions: Cue integration in image segmentation," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, Vol. 2 (Kerkyra: IEEE), 918–925. doi: 10.1109/ICCV.1999.790346

Manassi, M. (2014). *Crowding, Grouping and Object Recognition*. Ph. D. thesis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Manassi, M., and Herzog, M. (2013). Crowding and grouping: how much time is needed to process good gestalt? *Perception* 42, 229. doi: 10.1068/v130301

Manassi, M., Sayim, B., and Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *J. Vis.* 12:13. doi: 10.1167/12.10.13

Manassi, M., Sayim, B., and Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *J. Vis.* 13:10. doi: 10.1167/13.13.10

Martin, D. R., Fowlkes, C. C., and Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Anal. Mach. Intell. IEEE Trans.* 26, 530–549. doi: 10.1109/TPAMI.2004.1273918

Oberfeld, D., Stahn, P., and Kuta, M. (2012). Binaural release from masking in forward-masked intensity discrimination: evidence for effects of selective attention. *Hear. Res.* 294, 1–9. doi: 10.1016/j.heares.2012.09.004

Overvliet, K., and Sayim, B. (2013). Contextual modulation in haptic vernier offset discrimination. *Percept. 42 ECVP Abstr. Suppl.* 175. Available online at: http://www.perceptionweb.com/abstract.cgi?id=v130322

Pelli, D. G., Palomares, M., and Majaj, N. J. (2004). Crowding is unlike ordinary masking: distinguishing feature integration from detection. *J. Vis.* 4:12. doi: 10.1167/4.12.12

Ren, X., and Malik, J. (2002). A probabilistic multi-scale model for contour completion based on image statistics. *Comput. Vis. ECCV 2002*, 7, 312–327. doi: 10.1007/3-540-47969-4_21

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rodriguez-Sanchez, A. J., Simine, E., and Tsotsos, J. K. (2007). Attention and visual search. *Int. J. Neural Syst.* 17, 275–288. doi: 10.1142/S0129065707001135

Saarela, T. P., and Herzog, M. H. (2008). Time-course and surround modulation of contrast masking in human vision. *J. Vis.* 8, 1–10. doi: 10.1167/8.3.23

Saarela, T. P., and Herzog, M. H. (2009). Size tuning and contextual modulation of backward contrast masking. *J. Vis.* 9, 1–12. doi: 10.1167/9.11.21

Saarela, T. P., Westheimer, G., and Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *J. Vis.* 10, 1–7. doi: 10.1167/10.10.17

Salakhutdinov, R., Tenenbaum, J. B., and Torralba, A. (2013). Learning with hierarchical-deep models. *Pattern Anal. Mach. Intell. IEEE Trans.* 35, 1958–1971. doi: 10.1109/TPAMI.2012.269

Sayim, B., Westheimer, G., and Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychol. Sci.* 21, 641–644. doi: 10.1177/0956797610368811

Schneiderman, H., and Kanade, T. (2004). Object detection using the statistics of parts. *Int. J. Comput. Vis.* 56, 151–177. doi: 10.1023/B:VISI.0000011202.85607.00

Serre, T., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Serre, T., Wolf, L., and Poggio, T. (2005). "Object recognition with features inspired by visual cortex," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005*, Vol. 2 (San Diego, CA: IEEE), 994–1000.

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Anal. Mach. Intell. IEEE Trans.* 22, 888–905. doi: 10.1109/34.868688

Shpaner, M., Molholm, S., Forde, E., and Foxe, J. J. (2013). Disambiguating the roles of area v1 and the lateral occipital complex (loc) in contour integration. *Neuroimage* 69, 146–156. doi: 10.1016/j.neuroimage.2012.11.023

Thorpe, S., Delorme, A., and Van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Netw.* 14, 715–725. doi: 10.1016/S0893-6080(01)00083-1

Torralba, A. (2003). Contextual priming for object detection. *Int. J. Comput. Vis.* 53, 169–191. doi: 10.1023/A:1023052124951

Tsotsos, J. K. (1995). Toward a computational model of visual attention. *Early Vis. Beyond* 1, 207–218.

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artif. Intell.* 78, 507–545. doi: 10.1016/0004-3702(95)00025-9

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687. doi: 10.1038/nn870

Viola, P., and Jones, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vis.* 57, 137–154. doi: 10.1023/B:VISI.0000013087.49260.fb

Wertheimer, M. (1938). *Gestalt Theory*. London: Hayes Barton Press.

Wilson, H. R. (1986). Responses of spatial mechanisms can explain hyperacuity. *Vis. Res.* 26, 453–469. doi: 10.1016/0042-6989(86)90188-4

Wolford, G., and Chambers, L. (1983). Lateral masking as a function of spacing. *Percept. Psychophys.* 33, 129–138. doi: 10.3758/BF03202830

Wu, Y. N., Si, Z., Gong, H., and Zhu, S.-C. (2010). Learning active basis model for object detection and recognition. *Int. J. Comput. Vis.* 90, 198–235. doi: 10.1007/s11263-009-0287-0

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 April 2014; accepted: 03 October 2014; published online: 22 October 2014.
Citation: Herzog MH and Clarke AM (2014) Why vision is not both hierarchical and feedforward. Front. Comput. Neurosci. ***8****:135. doi: 10.3389/fncom.2014.00135*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Herzog and Clarke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

PERSPECTIVE ARTICLE
published: 04 November 2014
doi: 10.3389/fncom.2014.00138

The proactive brain and the fate of dead hypotheses

Amir Tal* and Moshe Bar

The Leslie and Susan Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat-Gan, Israel

Edited by:
Ales Leonardis, University of Birmingham, UK

Reviewed by:
John Lisman, Brandeis University, USA
Sen Song, Tsinghua University, China

****Correspondence:***
Amir Tal, The Leslie and Susan Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Building #901, Ramat-Gan 52900, Israel
e-mail: amir.tal@biu.ac.il

A substantial portion of information flow in the brain is directed top-down, from high processing areas downwards. Signals of this sort are regarded as conveying prior expectations, biasing the processing and eventual perception of incoming stimuli. In this perspective we describe a framework of top-down processing in the visual system in which predictions on the identity of objects in sight aid in their recognition. Focus is placed, in particular, on a relatively uncharted ramification of this framework, that of the fate of initial predictions that are eventually rejected during the process of selection. We propose that such predictions are rapidly inhibited in the brain after a competing option has been selected. Empirical support, along with behavioral, neuronal and computational aspects of this proposal are discussed, and future directions for related research are offered.

Keywords: visual processing, object recognition, predictions, competition suppression, negative priming, top-down, ambiguity resolution

INTRODUCTION

The hierarchical nature of information processing in the brain, particularly that of the visual system, has long been acknowledged. Originally, work focused on the accumulating complexity and sophistication added by each level of the hierarchy to the one preceding it, as information propagates upstream (Hubel and Wiesel, 1962). In recent years, however, a growing body of research has established the opposite direction of processing as well, from higher cortical areas downwards.

Top-down influences on perception are ubiquitous. From context (Biederman et al., 1982) and mood (Basso et al., 1996), to esthetic preference (Chen and Scholl, 2014) and inherent perceptual biases (Ramachandran, 1988), numerous factors intermix and contribute in shaping the subjective percept of a single objective stimulus. Many of these influences may be regarded as "predictions", in the sense that they express prior expectations concerning the incoming stimulus. In Bar (2003), such a framework for top-down visual processing was offered, in which rapid implicit predictions are formed and utilized to facilitate object recognition. In this paper we shall first present this framework, and then discuss one aspect of it in particular—the fate of alternative, competing predictions that are not chosen.

A FRAMEWORK FOR TOP-DOWN FACILITATION OF VISUAL PROCESSING

Visual information is distributed across a range of spatial frequencies. Low spatial frequency (LSF) information carries gross outlines and object contours, while information of high spatial frequency (HSF) encodes edges and finer details of an image. LSFs of visual stimuli have been found to elicit early synchronized activity between primary visual areas and the orbitofrontal cortex (OFC), followed by a synchronized coupling of the OFC and object recognition regions of the inferior temporal (IT) cortex (Bar et al., 2006). This pattern of activation therefore seems to bypass the ventral bottom-up visual processing chain in reaching the prefrontal cortex (PFC), affording it early access to coarse general aspects of visual stimuli. It has been offered that this bypass pathway is enabled by magnocellular projections (Bar et al., 2006), which are quicker conductors and more attuned to LSF information compared with their parvocellular counterparts (Tootell et al., 1988; Merigan and Maunsell, 1993). Indeed, stimuli biasing magnocellular processing preferably activate the OFC and evoke a pattern of functional coupling between visual, OFC and IT regions (Kveraga et al., 2007). Robust afferent projections connect early visual areas and the OFC, both directly and indirectly (Barbas, 1995; Fuster, 2008), and several possible routes have been proposed to account for the rapid flow of activation in this pathway (Kveraga et al., 2007).

According to the proposed framework, the OFC uses rudimentary versions of an input to rapidly activate familiar object categories resembling it. These activations act as hypotheses of the input's identity, subsequently combined with HSF bottom-up processing in temporal regions to facilitate recognition (Bar et al., 2006; **Figure 1**). LSF-based "initial guesses" assist, therefore, in narrowing down the search space of identity matching. An image of a tennis racket, by this account, will evoke a cursory image in the OFC, which will cause a signal corresponding to a racket, a guitar, a spoon and other object prototypes sharing a relatively similar outline, to be subsequently passed downwards. Finding the best match between these options and additional bottom-up incoming information will constitute recognition. Combining bottom-up with top-down signals in this manner has indeed been demonstrated to yield optimal efficiency in a computational model of visual recognition (Graboi and Lisman, 2003). Bottom-up information confines the hypothesis space to a selected subset of options, which are then passed back downwards to subsequently confine the breadth of lower-level processing required.

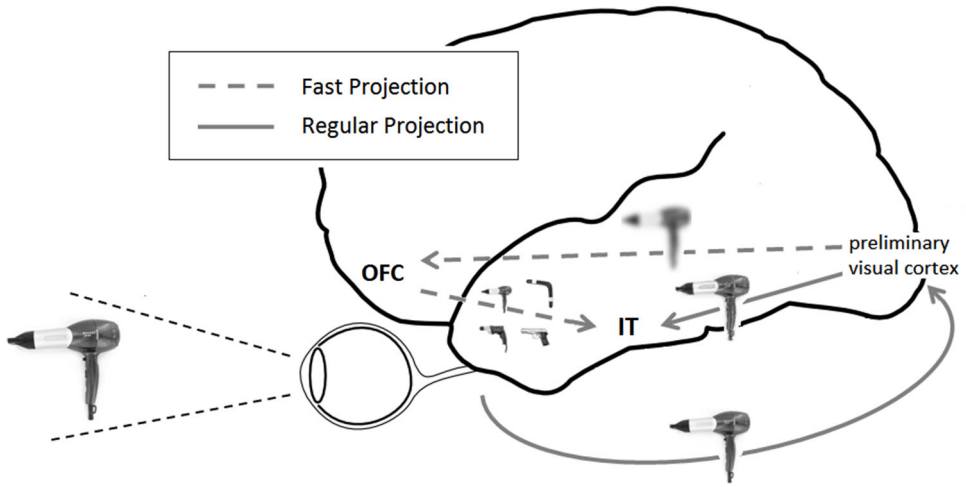


FIGURE 1 | Framework for top-down visual processing. LSF information is rapidly projected from preliminary occipital visual areas to the OFC. Based on this rudimentary display, predictions are formed and projected downwards to IT regions, where they coincide with the slower ventral visual stream of processing and facilitate recognition. In this illustration, an image of hairdryer supposedly prompts predictions of a drill, a gun, a hairdryer and a boomerang.

Due to its cursory nature, LSF information is highly suitable for forming visual predictions of the sort described above. Hypotheses derived from it may be sufficiently guiding and at the same time not too specific or constrained. Such LSF-based hypotheses may be seen as object categories in that a blurred representation of an object typically encapsulates most of its category members. Indeed, LSF information facilitates object recognition that is flexible and resilient to changes in exemplars and viewpoints (Cheung and Bar, 2014). In this light, all LSF-based images are at least somewhat ambiguous. However, although some outlines are more informative and distinguishable than others (consider, for example, the outline of a bicycle), most LSF-based images would be expected to resemble and therefore prime not only different versions of the same object, but more than one object as well. A recognition process using LSF-based predictions foretells substantial constant ambiguity dealt with in the visual system. In this perspective we would like to extend our top-down visual processing framework by addressing a relatively overlooked ramification of it: the fate of the initial guesses that first compete but are eventually not chosen.

PASSIVE DECAY OR ACTIVE SUPPRESSION

LSFs of objects in our visual environment typically evoke multiple possible interpretations of their identity. One of these guesses will ultimately be selected as the correct one as more details are conveyed by the HSFs and combined in analysis to constitute recognition. But what happens to the other alternative representations that have also been activated? Do these obsolete hypotheses gradually decay and die-out, neglected in the background while processing is concentrated elsewhere? Or is there rather active effort exerted for suppressing them rapidly? This question may provide promising new testing grounds for studying the strategy employed in the brain for conscious perception. Although no compelling evidence exists so far to support either possible account, in this perspective we would like to advocate as our hypothesis the latter, perhaps less intuitive, option of active suppression based on functional considerations.

At first glance, a mechanism in the brain for actively extinguishing activity that has been found irrelevant seems wasteful. Being irrelevant, such activity should lack encouragement and autonomously decay, so there is no apparent need in investing additional energy in putting it out quickly. In the next section we shall describe why we believe this mechanism is plausible and in fact necessary in the brain, by considering the possible benefits it might hold from a system point of view. Next, we shall support our claim by describing various findings of such a strategy employed in the brain in other domains. In the last section of this perspective we shall put forward our prediction regarding the mechanism of visual initial-guess competition suppression, and its testable manifestations.

BENEFITS OF SUPPRESSING UN-CHOSEN PREDICTIONS

Visual predictions as suggested by our framework are helpful in parsing a visual scene. However, once one of the predictions had been chosen, the remaining alternatives become a potential interference to network dynamics. Unless quickly extinguished, such activations may, presumably, distort processing of new information in light of some ambiguity that had already been settled. Their efficient and abrupt dismissal seems therefore important for ongoing performance. This is potentially true regarding any predictive system in the brain. In the case of visual recognition, however, chosen predictions become conscious perceptions, and so the potential threat of distracting activations may be most pronounced.

An elementary characteristic of conscious experience is its unequivocal explicitness. Despite the bombardment of information we are constantly confronted with, which is both noisy and partial, our personal sensation is typically unitary, coherent and

unambiguous. This is rather striking considering the breadth of subliminal activity we know is evoked and processed simultaneously in the brain at any given time. Indeed, consciousness is portrayed in many cases as an all-or-nothing mechanism, compared with subliminal workings that are statistical and inconclusive (Charles et al., 2013). To achieve this task, the nervous system must therefore be quick and decisive regarding interpretation of the world, selecting only a single percept to dominate consciousness at any given time. To this end, inhibition of the un-chosen options seems cardinal.

Current theories portray consciousness as the product of a large-scale connected component in neural network dynamics (Dehaene et al., 2003; Dehaene and Changeux, 2011). A mechanism for focusing activity on one selected option by suppressing irrelevant surroundings is important for the stability of such a component. Strong surrounding activations could otherwise potentially sway dynamics, even to the point of joining the locus of activation, which would result in a different conscious percept according to this theory. Therefore, efficient inhibition should take effect, and materialize rather immediately after selection when competition is highest and stability most fragile. This would be crucial for keeping our perception coherent. An extreme illustration of this idea may be found in the case of binocular rivalry, where strong and evenly matched competing activations are artificially created. The result is a non-stable perception alternating between the two of them (Blake and Logothetis, 2002).

The mechanism we propose in this perspective, therefore, is one aimed at "conscious decisiveness" and not ambiguity resolution *per se*. In our opinion, inhibition does not accentuate a single option over all others and thus helps in choosing between them, but rather it swiftly removes activations once such a choice has been made. The role ascribed to it is in "protecting" an implicit decision from interference once one is taken, by teasing out the to-be-explicit from the pool of implicit activations as quickly as possible to eliminate competition. Our argument is therefore for a post-selection and not a pre-selection mechanism, in agreement with most findings of competition suppression in the brain (May et al., 1995). In the next section we will overview several such findings, describing existing empirical evidence that provide initial support to our proposal.

EVIDENCE OF ACTIVE COMPETITION SUPPRESSION

In the realm of linguistics, ambiguity resolution has drawn considerable interest over the years, and has been studied as a cognitive phenomenon to the greatest extent. It therefore provides good grounds from which to build our argument dealing with LSF-based ambiguity resolution.

SUPPRESSION OF COMPETING INTERPRETATIONS

Numerous studies have shown that when an ambiguous word is encountered, several possible meanings of that word are immediately primed (Simpson, 1994), but access to all but the correct meaning of the word is significantly degraded within under a second of stimulus onset (Seidenberg et al., 1982; Gernsbacher and Faust, 1991b). Such decline in activation seems neither due to natural signal decay nor to competitive mutual inhibition between concepts (Gernsbacher and Faust, 1991b). Access to incorrect meanings may even become slower than access to neutral meanings that were not activated to begin with, providing further challenge to a mere decay account. This was shown in a lexical ambiguity experiment in which stimuli did not repeat, suggesting it is also not due to a prevalent memory-based explanation of such phenomena (Nievas and Marí-Beffa, 2002). It is interesting to note that the extent of suppression had been found in that study to be modulated by the strategy employed by subjects. Mirroring similar findings from selective attention (Neill and Westberry, 1987), suppression was exerted mostly when emphasis was placed on accuracy over speed, possibly indicating that it is mostly recruited when distractors are most harmful (Nievas and Marí-Beffa, 2002). Active inhibition, in any case, seems to take part in suppression of ambiguous word meanings, and its effect develops within a certain limited delay of selection.

Suppression in a linguistic task had been found when response was probed close to, but not immediately after, ambiguous primes. This was tested and shown as close as 467 ms from prime onset (Gernsbacher and Faust, 1991b), and analogous effects were also found under other modalities and tasks, within 1 s or less (Gernsbacher and Faust, 1991a). In one visual experiment, for example, subjects were presented with arrays of objects on screen for 250 ms, and then later shown a target and asked whether it was present in that array or not (Gernsbacher and Faust, 1991a). As expected, it was found that when targets were not part of the preceding array it took longer to reject them if they were contextually related to the array than if they were not (following an array of farm-related objects it took longer to reject the target "tractor" than it did to reject the target "kettle"). An interesting difference, however, emerged within subjects between "less-skilled comprehenders" and "more-skilled comprehenders" when targets were shown not immediately after the array display, but 1 s later. After a 1 s delay, the prolonged response times incurred by contextual relatedness remained in less-skilled comprehenders, while they had completely disappeared in skilled comprehenders. Successful comprehension of a visual scene seems to rely to some extent on the efficient suppression of potential distractors. This paradigm has proven so fruitful that suppression of inappropriate options was argued to be an overarching pivotal skill in comprehension in general (Gernsbacher and Faust, 1991a).

SUBLIMINALLY INDUCED COMPETITION

In the studies discussed above, competing alternatives stemmed from conscious perception of an ambiguous stimulus. Their suppression may therefore be seen as a form of cognitive inhibition, known to be a major factor in a wide array of decision-making tasks, supposedly controlled by executive functions in the PFC (Miyake et al., 2000; Aron et al., 2004). LSF of visual objects, however, are presumed to elicit a fleeting subliminal perception. Here we present findings from motor control research that support our proposal of predictive activation and subsequent inhibition that are triggered subliminally.

In Eimer and Schlaghecken (1998), subjects were to press either a left or a right key, as fast and accurately as possible, according to given cues. Each cue was preceded by an additional masked prime cue that was not consciously perceived. It was found that primes harmed performance when the subsequent

task was congruent with them, and on the other hand improved performance when the responses associated with the prime and the task were incongruent. Researchers called this phenomenon the "negative compatibility effect" (NCE), and offered that it is a result of inhibition acting on responses which were only partially activated (Eimer and Schlaghecken, 1998, 2003). A key insight from this study is that performance in this type of task depends not only on prime-target compatibility, but also on their relative timing. Event-related potential recordings of motor cortex during this task revealed that primes elicit activation corresponding with the movement they denote within roughly 200 ms of their presentation, but this activity reverses polarity 100 ms later and dwells below baseline for around 100 ms (between 300–400 ms of prime onset). The researchers found that when a motor choice is made in this latter time frame, performance is impaired if a compatible cue was given before the target, and conversely enhanced by a "misleading cue". If timed right, the primed alternative will in fact be suppressed beneath threshold.

This effect had been replicated and proven rather robust in the motor system. Importantly, when varying time intervals between primes and targets, NCE was only observed following longer delays (96–192 ms), while short delays of up to 32 ms generated positive priming, in compliance with a rise-and-fall activation pattern (Schlaghecken and Eimer, 2000). Eimer and Schlaghecken (2003) propose stimulus-driven inhibition as a faculty complimentary to conscious cognitive control. They call it "exogenous" inhibition, in contrast to the conscious cognitive inhibition of responses, which is "endogenous" (Eimer and Schlaghecken, 2003). Competition suppression of the NCE phenomenon validates the possibility of competing elements subliminally activated and then inhibited in brain functioning, and demonstrates such inhibition acting fast (within 100 ms).

EXPERIMENTAL PREDICTIONS OF INITIAL-GUESS SUPPRESSION

Building on the two paradigms described above, the supporting evidence for multiple-alternative activation and suppression in ambiguity resolution, and the possible evocation of them subliminally, we now turn to describe the testable manifestations we expect suppression of visual initial-guesses will have.

TIME FRAME

Active suppression of a concept in memory will cause an increased difficulty in accessing that concept for a certain amount of time thereafter. This is termed negative priming, as it is the exact opposite of classic positive priming in which recently activated concepts enjoy advantageous processing. Initial-guess suppression, as we propose, should therefore join the varied multitude of cognitive tasks that behaviorally manifest as negative priming for a certain amount of time. Because the speed of visual object recognition depends on numerous aspects of the stimulus (image complexity, display duration, contextual information, familiarity, and so on) we align the temporal description of our model to the time point of recognition.

We expect concepts activated by visual LSF information to be subsequently suppressed when additional information arrives and confirms one of the competing hypotheses. Visual LSF information creates enhanced activity in OFC areas around 50 ms before major activation of object-recognition areas of the IT begins (peaking at 130 and 180 ms, respectively, in the paradigm used in Bar et al. (2006)). This is the time window LSF-based generation of initial guesses happens, according to our proposition, so at this time interval multiple concepts should be activated and positively primed (**Figure 2**, between "OFC activation" and "recognition").

Next, following strong coupling between the OFC and IT regions, recognition is achieved. It is this moment, or slightly earlier, that deems all but one activated guesses irrelevant, and so presumably ignites inhibition. Based on findings from the linguistic domain and others, we expect activation levels to drop below baseline level within less than 500 ms of this moment. The duration in which excitation should persist below baseline is hard to estimate. Various different negative priming paradigms have found an effect lasting between 0 and 8 s (May et al., 1995). Persistence would be particularly interesting to examine in the domain of vision, because visual context is typically rather stable over time, and so a mechanism for longer lasting suppression of irrelevant guesses might be desirable. A significant negative level of activation, in any case, is reasonable to expect for at least 100 ms (see **Figure 2B** for a summary of the predicted activation timeline of an un-chosen initial-guess).

INHIBITORY MECHANISM

Inhibition of un-chosen initial guesses, as hypothesized in this perspective, would operate only on the highest mode of visual processing and not trickle down to earlier regions of the visual pathway. Since selection is made between visually similar concepts (sharing major LSF features) activity patterns corresponding to them in low-level visual regions would be more similar than not. Inhibiting a low-level region for one concept and not the other would therefore be difficult, but mainly negligible. Behaviorally, as stated earlier, inhibition would manifest as negative priming between LSF-similar representations. However, alternative interpretations of such a finding exist and would have to be accounted for.

Whether negative priming indicates underlying inhibition has been the subject of a long and rich debate (May et al., 1995), but modern accounts tend to agree that both forward-acting inhibition and backward-acting memory mechanisms may give rise to negative priming under different experimental settings (Tipper, 2001; Mayr and Buchner, 2007). According to memory-based accounts, negative priming occurs when a task evokes previous processing episodes from memory and these episodes conflict with current settings. The conflict may lay in the response associated with the target of the task (previously ignored but currently requiring a response) (Neill et al., 1992) or in the different features the target object had in both episodes (Park and Kanwisher, 1994). A prerequisite for these accounts, in either case, is that the target serves as a good retrieval cue to a previous prime. This would seem less probable in an experiment examining suppression of LSF-based predictions. Unlike most negative priming paradigms, primes and targets of our framework would be dissimilar (a guitar and a tennis racket, for example). The supposed activation and

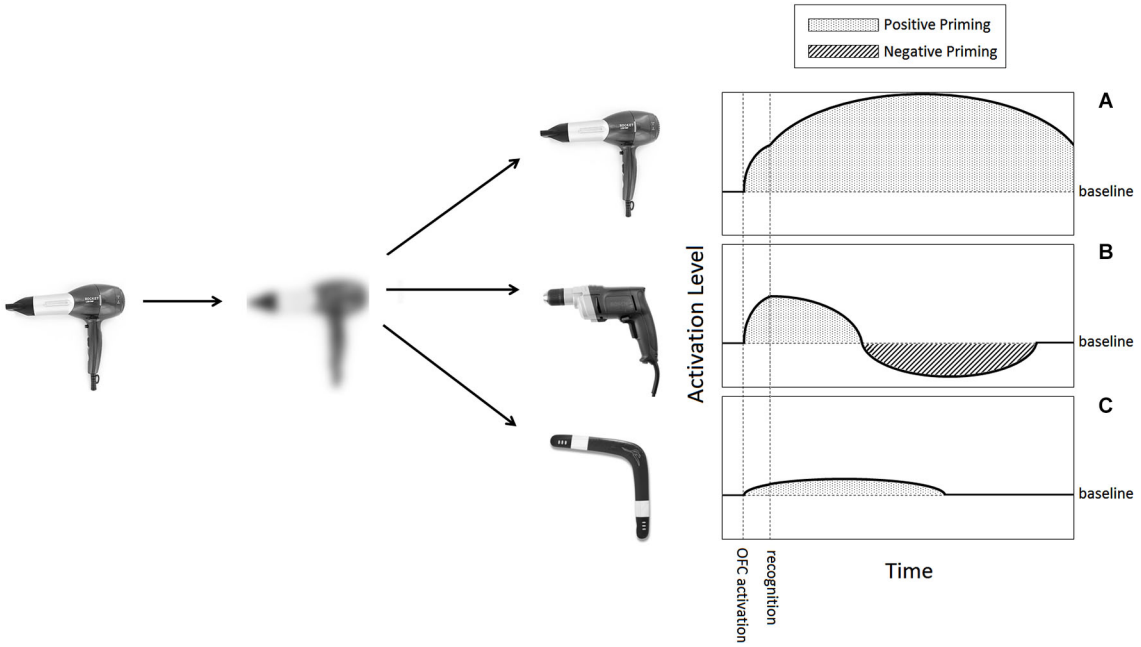


FIGURE 2 | Proposed activation patterns of LSF-induced hypotheses. When presented an object (a hairdryer in this example), an LSF version of it is extracted and projected to the OFC. This cursory image activates resembling objects in memory, acting as hypotheses for recognition. In this case three such hypotheses are depicted. **(A)** The correct hypothesis is activated when an image reaches the OFC, like any other hypothesis. When additional information combines with it for recognition, it receives additional encouragment and will therefore display a classic positive priming affect for substantial time. **(B)** Competing initial hypotheses (a drill in this example), will initially be activated, but once recognition of a competing option will be made, our model proposes they will go under inhibition. Activation of an obsolete hypothesis of this sort should begin dropping and cross baseline activity. Undershooting the baseline will manifest as negative priming for a certain short period of time. **(C)** An example of an especially weak hypotheses, bearning minimal resemblance to the original image. According to several findings of threshold-dependent behavior of inhibition, a weak activation of this sort may not be inhibited at all, despite it being an un-chosen hypothesis like **(B)**.

rejection of the target in the prime episode, moreover, is implicit and not encoded in memory. Therefore, under experimental settings in which targets are fully visible and processed, it is unlikely that a reliance on memory retrieval would be promoted (May et al., 1995), and negative priming in these cases would be better explained by neuronal inhibition. Settings examining our framework would therefore favor an inhibition explanation, in our opinion, but careful design would nevertheless have to regard the alternative accounts.

Lastly, an additional research path, considering associative strength, may shed light on the characteristics of the inhibition we propose. In several studies, neural inhibition behaves in a seemingly threshold-dependent manner. In such cases, surprisingly no inhibition is applied when activation strength is particularly low. It has been found that activations that are especially weak are allowed to linger, spared the neural inhibition their counterparts receive, as if going "beneath the radar" of neural inhibition (Eimer and Schlaghecken, 2003; Tsushima et al., 2006). This aspect allows postulating that certain guesses may not be inhibited if their activation was particularly small to begin with. In our case, activations would be weak if the outline of the display and the outline of the guess are only remotely similar, analogous to a guess that is less probable (**Figure 2C**). This could be an interesting research path to follow, one that could yield firm evidence that the processes supporting this type of inhibition are similar to the ones supporting resembling phenomena mentioned here from selective attention and motor control.

SUMMARY

In this perspective we have overviewed our framework for top-down processing in the visual system, and focused on a particularly intriguing and understudied implication of it: the fate of un-chosen predictions. Albeit counter-intuitive, we believe that irrelevant activations of this sort undergo active inhibition in the brain. Such conduct seems important for maintaining stability and coherence in ongoing conscious perception. Reviewing evidence from different research areas of neuroscience, it seems that activation of numerous options, followed by coincided facilitation of one and inhibition of the rest, is characteristic of normal brain functioning. We hope future research will build on this proposal and elaborate it, as we believe further scientific study of this phenomenon will improve our understanding of efficient strategies for visual and non-visual information processing, and of human perception in general.

ACKNOWLEDGMENTS

Our work is supported by The Israeli Center of Research Excellence in Cognitive Sciences (ICORE grant No. 51/11).

REFERENCES

Aron, A. R., Robbins, T. W., and Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends Cogn. Sci.* 8, 170–177. doi: 10.1016/j.tics.2004.02.010

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci. U S A* 103, 449–454. doi: 10.1073/pnas.0507062103

Barbas, H. (1995). Anatomic basis of cognitive-emotional interactions in the primate prefrontal cortex. *Neurosci. Biobehav. Rev.* 19, 499–510. doi: 10.1016/0149-7634(94)00053-4

Basso, M. R., Schefft, B. K., Ris, M. D., and Dember, W. N. (1996). Mood and global-local visuai processing. *J. Int. Neuropsychol. Soc.* 2, 249–255. doi: 10.1017/s1355617700001193

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* 14, 143–177. doi: 10.1016/0010-0285(82)90007-x

Blake, R., and Logothetis, N. K. (2002). Visual competition. *Nat. Rev. Neurosci.* 3, 13–21. doi: 10.1038/nrn701

Charles, L., Van Opstal, F., Marti, S., and Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage* 73, 80–94. doi: 10.1016/j.neuroimage.2013.01.054

Chen, Y.-C., and Scholl, B. J. (2014). Seeing and liking: biased perception of ambiguous figures consistent with the "inward bias" in aesthetic preferences. *Psychon. Bull. Rev.* doi: 10.3758/s13423-014-0617-2. [Epub ahead of print].

Cheung, O. S., and Bar, M. (2014). The resilience of object predictions: early recognition across viewpoints and exemplars. *Psychon. Bull. Rev.* 21, 682–688. doi: 10.3758/s13423-013-0546-5

Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Dehaene, S., Sergent, C., and Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl. Acad. Sci. U S A* 100, 8520–8525. doi: 10.1073/pnas.1332574100

Eimer, M., and Schlaghecken, F. (1998). Effects of masked stimuli on motor activation: behavioral and electrophysiological evidence. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1737–1747. doi: 10.1037//0096-1523.24.6.1737

Eimer, M., and Schlaghecken, F. (2003). Response facilitation and inhibition in subliminal priming. *Biol. Psychol.* 64, 7–26. doi: 10.1016/s0301-0511(03)00100-5

Fuster, J. (2008). *The Prefrontal Cortex.* 4th Edn. London: Academic Press.

Gernsbacher, M. A., and Faust, M. E. (1991a). The mechanism of suppression: a component of general comprehension skill. *J. Exp. Psychol. Learn. Mem. Cogn.* 17, 245–262. doi: 10.1037//0278-7393.17.2.245

Gernsbacher, M. A., and Faust, M. (1991b). "The role of suppression in sentence comprehension," in *Understanding Word and Sentence*, ed G. B. Simpson (Amsterdam: North-Holland), 97–128.

Graboi, D., and Lisman, J. (2003). Recognition by top-down and bottom-up processing in cortex: the control of selective attention. *J. Neurophysiol.* 90, 798–810. doi: 10.1152/jn.00777.2002

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Kveraga, K., Boshyan, J., and Bar, M. (2007). Magnocellular projections as the trigger of top-down facilitation in recognition. *J. Neurosci.* 27, 13232–13240. doi: 10.1523/jneurosci.3481-07.2007

May, C. P., Kane, M. J., and Hasher, L. (1995). Determinants of negative priming. *Psychol. Bull.* 118, 35–54. doi: 10.1037/0033-2909.118.1.35

Mayr, S., and Buchner, A. (2007). Negative priming as a memory phenomenon. *J. Psychol.* 215, 35–51. doi: 10.1027/0044-3409.215.1.35

Merigan, W. H., and Maunsell, J. H. (1993). How parallel are the primate visual pathways? *Annu. Rev. Neurosci.* 16, 369–402. doi: 10.1146/annurev.neuro.16.1.369

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Neill, W. T., Valdes, L. A., Terry, K. M., and Gorfein, D. S. (1992). Persistence of negative priming: II. Evidence for episodic trace retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 993–1000. doi: 10.1037//0278-7393.18.5.993

Neill, W. T., and Westberry, R. L. (1987). Selective attention and the suppression of cognitive noise. *J. Exp. Psychol. Learn. Mem. Cogn.* 13, 327–334. doi: 10.1037//0278-7393.13.2.327

Nievas, F., and Marí-Beffa, P. (2002). Negative priming from the non-selected meaning of the homograph. *Br. J. Psychol.* 93, 47–66. doi: 10.1348/000712602162445

Park, J., and Kanwisher, N. (1994). Negative priming for spatial locations: identity mismatching, not distractor inhibition. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 613–623. doi: 10.1037//0096-1523.20.3.613

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature* 331, 163–166. doi: 10.1038/331163a0

Schlaghecken, F., and Eimer, M. (2000). A central-peripheral asymmetry in masked priming. *Percept. Psychophys.* 62, 1367–1382. doi: 10.3758/bf03212139

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., and Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: some limitations of knowledge-based processing. *Cogn. Psychol.* 14, 489–537. doi: 10.1016/0010-0285(82)90017-2

Simpson, G. B. (1994). "Context and the processing of ambiguous words," in *Handbook of Psycholinguistics*, ed M. A. Gernsbacher (San Diego, CA: Academic Press), 359–374.

Tipper, S. P. (2001). Does negative priming reflect inhibitory mechanisms? A review and integration of conflicting views. *Q. J. Exp. Psychol. A* 54, 321–343. doi: 10.1080/02724980042000183

Tootell, R. B., Silverman, M. S., Hamilton, S. L., Switkes, E., and De Valois, R. L. (1988). Functional anatomy of macaque striate cortex. V. Spatial frequency. *J. Neurosci.* 8, 1610–1624.

Tsushima, Y., Sasaki, Y., and Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science* 314, 1786–1788. doi: 10.1126/science.1133197

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 May 2014; accepted: 13 October 2014; published online: 04 November 2014.

*Citation: Tal A and Bar M (2014) The proactive brain and the fate of dead hypotheses. Front. Comput. Neurosci. **8**:138. doi: 10.3389/fncom.2014.00138*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Tal and Bar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution and reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 14 August 2014
doi: 10.3389/fncom.2014.00096

Combining segmentation and attention: a new foveal attention model

Rebeca Marfil*, Antonio J. Palomino and Antonio Bandera

ISIS Group, Department of Electronic Technology, University of Málaga, Málaga, Spain

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Sébastien M. Crouzet, Charité University Medecine, Germany
Ozgur Erkent, Innsbruck University, Austria

****Correspondence:***
Rebeca Marfil, ISIS Group, Department of Electronic Technology, University of Málaga, Campus de Teatinos, Málaga 29071, Spain
e-mail: rebeca@uma.es

Artificial vision systems cannot process all the information that they receive from the world in real time because it is highly expensive and inefficient in terms of computational cost. Inspired by biological perception systems, artificial attention models pursuit to select only the relevant part of the scene. On human vision, it is also well established that these units of attention are not merely spatial but closely related to perceptual objects (proto-objects). This implies a strong bidirectional relationship between segmentation and attention processes. While the segmentation process is the responsible to extract the proto-objects from the scene, attention can guide segmentation, arising the concept of foveal attention. When the focus of attention is deployed from one visual unit to another, the rest of the scene is perceived but at a lower resolution that the focused object. The result is a multi-resolution visual perception in which the fovea, a dimple on the central retina, provides the highest resolution vision. In this paper, a bottom-up foveal attention model is presented. In this model the input image is a foveal image represented using a Cartesian Foveal Geometry (CFG), which encodes the field of view of the sensor as a fovea (placed in the focus of attention) surrounded by a set of concentric rings with decreasing resolution. Then multi-resolution perceptual segmentation is performed by building a foveal polygon using the Bounded Irregular Pyramid (BIP). Bottom-up attention is enclosed in the same structure, allowing to set the fovea over the most salient image proto-object. Saliency is computed as a linear combination of multiple low level features such as color and intensity contrast, symmetry, orientation and roundness. Obtained results from natural images show that the performance of the combination of hierarchical foveal segmentation and saliency estimation is good in terms of accuracy and speed.

Keywords: artificial attention, foveal images, foveal segmentation, saliency computation, irregular pyramids

1. INTRODUCTION

Human vision system presents an interesting set of features of adaptability and robustness that allows it to analyse and process the visual information of a complex scene in a very efficient manner. Research in Psychology and Physiology demonstrates that the efficiency of natural vision has foundations in visual attention, which is a process that filters out irrelevant information and limits processing to salient items (Duncan, 1984). It has been demonstrated by psychophysics studies that, when a human observes a scene, she does not do so as a whole, but rather will make a series of visual fixations at salient locations in the scene using eye saccade movements (Martinez-Conde et al., 2004). These voluntary movements have the main purpose of capturing salient locations using the central region of the retina (fovea), which is the place where the human retina has a high concentration of cones and the image is captured with fine resolution. Psychophysics studies suggest other important role of fixations in how humans perceive a scene (Martinez-Conde et al., 2004). Experiments show that subjects are not able to detect scene changes when they occur at a location away from the fixation, unless they modify the gist of the scene. Because the scene is captured with less resolution in the periphery than in the fovea. In contrast, the changes are detected quickly when they occur in the fixation area or close to it. Then, it is clear that there is a relationship between visual fixation and attention in the human vision system. Attention allows to select salient locations that using a visual fixation are centered in the fovea to be acquired with fine resolution, while the rest of the scene is captured with less resolution. This multi-resolution encoding allows the human visual system to perceive a large field of view, bounding the data flow coming from the retina.

In the Computer Vision community, the non-uniform encoding of images has been emulated through methods such as the Reciprocal Wedge Transform (RWT), or the log-polar or Cartesian Foveal Geometries (CFG) (Traver and Bernardino, 2010). Also the selection of salient regions from an image has been widely studied, appearing different artificial attention models (Frintrop et al., 2010). However, the combination of attention and foveal image representation has been very little studied. This combination implies a close bidirectional relationship between foveal image segmentation and attention. This relationship comes from the fact that the location of human fixation is closely related to perceptual objects or proto-objects instead of disembodied spatial locations of the image (Rensink, 2000). Proto-objects can be defined as units of visual information that can be bounded into

a coherent and stable object and they can be extracted using a perceptual segmentation algorithm. So, it seems logical to place the fovea in the location of the most salient proto-object in each moment. The saliency of each proto-object is obtained using an artificial attention model. Therefore the relationship between foveal segmentation and attention in one direction is clear: foveal segmentation provides the proto-objects to attention. But also the reverse relationship is very important. Segmentation essentially refers to a process that divides up a scene into non-overlapping, compact regions. Each region encloses a set of pixels that are bound together on the basis of some similarity or dissimilarity measure. A large variety of approaches for image segmentation has been proposed by the Computer Vision community in the last decades. And simultaneously, this community has been asked for a definition of what a correct segmentation is. As several authors have argued, the conclusion about this problem definition is that it is not well posed (Lin et al., 2007; Singaraju et al., 2008; Mishra et al., 2012). For example, if we see the original image and the segmentations provided by two human subjects in **Figure 1**, a major question arises: which is the correct segmentation? The answer to this question depends on what object we want to segment in the image: the two people (**Figure 1** middle) or certain image details such as faces or hands (**Figure 1** right). As Mishra et al. (2012) pointed out, the answer to this question depends on another question: what is the object of interest on the scene? Attention can be used to provide segmentation with the object of interest, fitting the correct input parameters and making segmentation well-defined (Jung and Kim, 2012; Mishra et al., 2012). These methods make use of the influence of attention in segmentation, but they do not take into account the reverse relation: how segmentation can influence attention.

In this paper, we propose a foveal attention mechanism which illustrates the bidirectional relation among attention and foveal segmentation. It uses a hierarchical image encoding where foveal segmentation and bottom-up attention processes can be simultaneously performed. As other approaches, this structure resembles the one of the human retina: it will only capture a small region of the scene in high resolution (fovea), while the rest of the scene will be captured in lower resolution on the periphery. Specifically, we use an adaptive CFG where the fovea can be located in any place of the scene and its size can be dynamically modified. The structure of the CFG is very suitable for hierarchical processing, allowing to encode the multi-resolution image within a *foveal polygon*. The foveal polygon represents the image at different resolution levels and is built using the irregular decimation process of the Bounded Irregular Pyramid (BIP) (Marfil et al., 2007) applied to perceptual segmentation. The saliency of each proto-object is computed following the *Feature Integration Theory* (Treisman and Gelade, 1980) as a linear combination of a set of low level features which clearly influences attention. While the computation of the low level features is independent of the task, being a pure bottom-up process, the linear combination of features is computed as a weighted summation where the weights can be set depending on the task in a top-down way. This attention mechanism is able to manage dynamic scenarios by adding an Inhibition of Return (IOR) mechanism which keeps permanently updated the position of each already attended proto-object and avoids revisiting an already attended one.

1.1. RELATED WORK

According to the taxonomy of computational models of visual attention proposed by Tsotsos and Rothenstein (2011), the method proposed in this paper can be considered as a saliency-based one. From the psychological point of view, the development of saliency-based computational models of visual attention is mainly based on the so-called *early-selection* theories. These theories postulate that the selection of a relevant region precedes pattern recognition. Therefore, attention is drawn by simple features (such as color, location, shape or size) and attended entities do not have full perceptive meaning, i.e., they could not correspond to real objects. Two complementary biological theories or descriptive models are the most influential ones regarding saliency-based computational models of visual attention: Treisman's *Feature Integration Theory* (FIT) (Treisman and Gelade, 1980) and Wolfe's *Guided Search* (Wolfe et al., 1989; Wolfe, 1994). FIT suggests that the human vision system detects separable features in parallel in an early step of the attention process. According to this model, methods compute image features in a number of parallel channels in a *pre-attentive* task-independent stage. Then, the extracted features are integrated through a *bottom-up* process into a single saliency map which codes the relevance of each image entity. The first saliency-based computational models mainly followed these guidelines. For example, the models proposed by Itti et al. (1998) or Koch and Ullman (1985) compute the saliency of each pixel based on a set of basic features. They were pure bottom-up, static models. Several years later, Wolfe proposed that a *top-down*

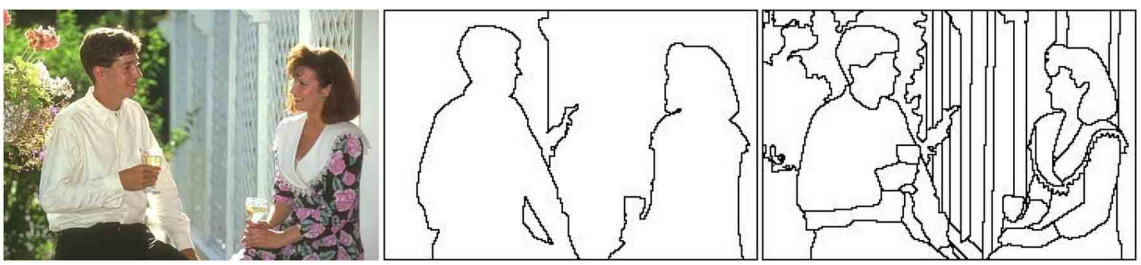


FIGURE 1 | (Left) Test image #157055 from the Berkeley Segmentation Data Set (BSDS300) (Martin et al., 2001), **(Middle)** segmentation by user #1109 (8 segments), and **(Right)** segmentation by user #1123 (61 segments).

component in attention can increase the speed of the process giving more relevance to those parts of the image corresponding to the current task. These two approaches are not mutually exclusive and, nowadays, several efforts in computational attention are being conducted to develop models which combine a bottom-up processing stage with a top-down selection process. Thus, Navalpakkam and Itti (2005) modified Itti's original model in order to add a multi-scale object representation in a long-term memory. The multi-scale object's features stored in this memory determine the relevance of the scene features depending on the current executed task, implementing, therefore, a top-down behavior. As an alternative to *space-based* models, where attention deploys on an unstructured region of the scene rather than on an object, *object-based* models of visual attention provide a more efficient visual search. These models are based on the assumption that the boundaries of segmented objects, and not just spatial position, determine what is selected and how attention is drawn (Scholl, 2001). Therefore, these models reflect the fact that perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. Orabona et al. (2007) propose a model of visual attention based on the concept of *proto-objects* (Rensink, 2000) as units of visual information that can be bound into a coherent and stable object. They compute these proto-objects by employing the watershed transform to segment the input image using edge and color features in a pre-attentive stage. The saliency of each proto-object is computed taking into account top-down information about the object to perform a task-driven search. Yu et al. (2010) propose a model of attention that segments the scene into proto-objects in a bottom-up strategy based on Gestalt theories. After that, in a top-down way, the saliency of the proto-objects is computed taking into account the current task to accomplish by using models of objects which are relevant to this task. These models are stored in a long-term memory. These proto-object based models compute in a firs step the set of proto-objects from the scene and then they compute their saliency. There exist other type of methods that first compute the saliency map from the scene and then, the most salient proto-object is computed from the saliency map (Walther and Koch, 2006).

Attention theories introduce another important concept: the *Inhibition of Return* (IOR) (Posner et al., 1985). Human visual psychophysics studies have demonstrated that a local inhibition is activated in the saliency map to avoid attention being directed immediately to a previously attended region. In the context of computational models of visual attention, this IOR has been usually implemented using a 2D inhibition map that contains suppression factors for one or more focuses of attention that were recently attended (Itti et al., 1998; Frintrop, 2006). However, this 2D inhibition map is not able to handle the situations where inhibited objects are in motion or when the vision system itself is in motion. In this situation, establishing a correspondence between regions of the previous frame with those of the successive frame becomes a significant issue. In order to allow that the inhibition can track an object while it changes its location, the model proposed by Backer et al. (2001) relates the inhibitions to features of activity clusters. However, the scope of dynamic inhibition becomes very limited as it is related to activity clusters rather than objects themselves (Aziz and Mertsching, 2007). Thus, it is a better option to attach the inhibition to moving objects (Tipper, 1991). Aziz and Mertsching (2007) utilizes a queue of inhibited region features to maintain object inhibition in dynamic scenes.

Finally, Psychophysics studies also refer to how many elements can be attended at the same time. Bundesen establishes in his *Theory of Visual Attention* (Bundesen et al., 2011) that there exists a short-term memory where recently attended elements are stored. This memory has a fixed capacity usually reduced up to 3 or 5 elements.

All the attention models presented in this section have focused in different aspects such as e.g., the identification of features which influence attention, the combination of these features to generate the saliency map or how an specific task drives attention. But they neglect the foveal nature of the human attention system. The methods following a multi-resolution strategy usually employ two images of different resolution (Meger et al., 2008): A low-resolution image for computing the saliency map of the scene and a high resolution one for studying in detail the most salient region. Foveation has been typically proposed as an efficient way for image encoding (Geisler and Perry, 1998; Guo and Zhang, 2010). Built over the foveal encoding by Geisler and Perry (1998), the Gaze Attentive Fixation Finding Framework (GAFFE) (Rajashekar et al., 2008) employs four low-level local image saliency features (luminance, contrast, and bandpass outputs of both luminance and contrast) to build saliency maps and predict gaze fixations. It works on a sequential process in which the stimulus is foveated at the current fixation point and saliency features are obtained from circular patches from this foveated image to predict the next fixation point. This strategy has been recently evaluated by Gide and Karam (2012), replacing these saliency features with features from other models such as AIM (Attentive Information Maximization) (Bruce and Tsotsos, 2009) or SUN (Saliency Using Natural Image Statistics) (Zhang et al., 2008). Evaluated under a quality assessment task for different types of distortions (Gaussian blur, white noise and JPEG compression), Gide and Karam (2012) showed that the performance of all saliency models significantly improved with foveation over all distortion types. It should be noted that Rajashekar et al. (2008) and Gide and Karam (2012) do not obtain the fixation points from a saliency model, but from features extracted of the foveated images. Following a different strategy, Advani et al. (2013) propose to encode the image as a three level Gaussian pyramid. The higher level represents the whole field-of-view at a lower resolution, meanwhile the lower one only encodes the 50% of the field-of-view at the resolution of the original image. The AIM model is run at these three levels, which returns corresponding information maps. These maps represent the salient regions at different resolutions and are fused within an unique saliency map using weighted summation.

1.2. OVERVIEW OF THE PROPOSED ATTENTION MODEL

In this paper, a bottom-up foveal attention model is presented. The input of this model is a foveal image represented in an adaptive CFG where the focus of attention, or Region of Interest (ROI), is located at the fovea which is surrounded by a set of concentric rings with decreasing resolution. In this model the

attention is deployed to proto-objects instead of disembodied spatial locations. These proto-objects are defined as the blobs of uniform color and disparity of the image which are bounded by the edges obtained using a Canny detector. They are extracted using a perceptual segmentation algorithm which is conducted using an extension of the BIP (Marfil et al., 2007). The saliency of each proto-object is computed in a bottom-up framework in order to obtain the ROI for the next frame. This saliency value is the combination of a set of low level features that according to psychological studies clearly influences saliency computation (Treisman and Souther, 1985; Wolfe et al., 1992). Specifically, it is computed in terms of the following features: color contrast, intensity contrast, proximity, symmetry, roundness, orientation and similarity to skin color. To have an homogenized calculus, all features values are normalized in the range $[0 \ldots 255]$.

Hence, contrary to all previous approaches to foveal attention, our approach merges within the same hierarchical framework the segmentation and saliency estimation processes. The levels of the hierarchy are not obtained by blurring and downsampling the content on the level below and adding additional information to increase the field-of-view. In our approach, each level of the hierarchy is able to provide a segmentation of the encoded field-of-view. Then, the highest level of the hierarchy, that encodes the full field-of-view, provides a segmentation S^t where the fovea details are present but those at the peripheral regions are not. This segmentation S^t depends on the fovea location provided by the attention process at $t-1$ and drives the next location of the fovea. Once the saliency of each proto-object is computed, the ROI at $t+1$ is extracted as the location of the most salient proto-object in the current frame. In order to compute this ROI and to avoid revisiting or ignoring proto-objects, it is necessary to implement an Inhibition of Return mechanism (IOR). This IOR is very important in the case of dynamic environments where there are moving objects. It is typically implemented using a 2D inhibition map which contains suppression factors for one or more recently attended focuses of attention. This approach is valid to manage static scenarios, but it is not able to handle dynamic environments where inhibited proto-objects or the vision system itself are in motion. In the proposed system, a tracker module keeps permanently updated the position of recently attended proto-objects or focuses of attention. The features and location of these already attended proto-objects are stored in a Working Memory. Thereby, it is avoided to attend an already selected proto-object even if the proto-object changes its location in the image. Specifically, the tracker is based on the Comaniciu mean-shift approach (Comaniciu et al., 2003) , a method which allows to track non-uniform color regions in an image.

Figure 2 shows the main stages involved in the proposed attentional model and **Figure 3** shows an example. First, a foveal image is captured with the fovea located in the Region of Interest (ROI) computed in the previous frame. In frame t of **Figure 3** the fovea is located in the woman's face, in $t+1$ the fovea is located in the man's face. It must be noted that in the first frame the fovea is located at the image center. After that, the foveal image is segmented by building the Foveal Polygon using the BIP. In this stage the set of proto-objects is extracted from the foveal image and the fovea could be processed by further attentional stages (that are out of the scope of this paper). Then, saliency of each obtained proto-object is computed. These saliency values are used to compute the ROI of the next frame taking into account the output of the tracking module. This tracker computes the locations of the previously attended proto-objects in the current frame. These locations and the location of the current ROI are inhibited in order to extract the new ROI (black squares in **Figure 3**).

1.3. CONTRIBUTIONS

The main contributions of this work are:

- The use of foveal images as inputs of the attentional mechanism.

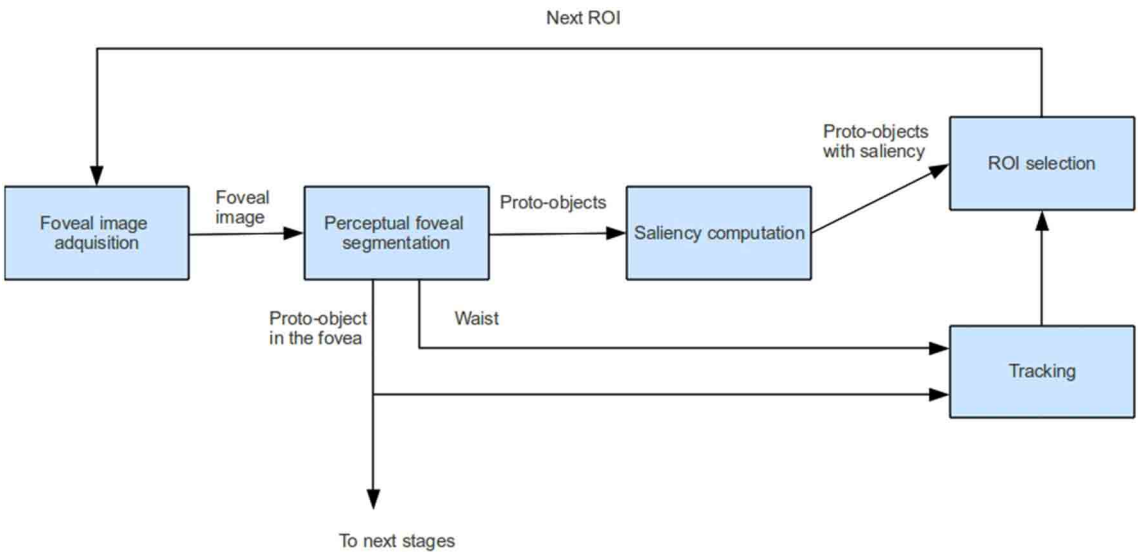


FIGURE 2 | Overview of the proposed foveal attention model.

- The hierarchical representation of the foveal image that allows to simultaneously built the foveal polygon and perceptually segment the input image extracting the proto-objects.
- The combination of foveal segmentation and attention: the attention process allows to select the next position of the fovea and segmentation allows to extract the units of attention.

1.4. ORGANIZATION OF THE PAPER

After providing a brief overview of the proposed approach in this Section 1, the rest of the paper is organized as follows: Sections 2, 3 provide a more detailed description of the two main processes (perceptual foveal segmentation and bottom-up attention) tied within our framework. Section 2 introduces the Cartesian Foveal Geometries and the concept of the Foveal Polygon. Then, it describes the data structure and decimation strategy that define the foveal Bounded Irregular Pyramid (foveal BIP). Section 3 describes how the saliency is computed and the ROI is chosen, including a description of our implementation of the IOR. Section 4 evaluates the performance of the foveal attention system. Three kinds of tests have been conducted: a comparison of the uniform and foveated models of attention, an evaluation of the ability of our approach for actively driving an image exploration process, and a quantitative evaluation of the attention and fixation prediction models.

2. PERCEPTUAL FOVEAL SEGMENTATION

In this paper, we propose an artificial attentional system which uses a hierarchical image encoding where segmentation and bottom-up attention processes are simultaneously performed. This image encoding resembles the one of the human retina by using a foveal representation: only a small region of the scene is captured with high resolution (fovea), while the rest of the scene is captured in lower resolution on the periphery. Specifically, an adaptive Cartesian Foveal Geometry is used to capture the input image which is hierarchically encoded by means of a Perceptual Segmentation approach. It allows to extract the proto-objects from the visual scene and it is conducted using the Bounded Irregular Pyramid (BIP) (Marfil et al., 2007).

2.1. CARTESIAN FOVEAL GEOMETRIES (CFG) AND FOVEAL POLYGONS

Cartesian Foveal geometries (CFG) encode the field of view of the sensor as a fovea surrounded by a set of concentric rings with decreasing resolution (Arrebola et al., 1997). In the majority of the Cartesian proposals, this fovea is centered on the geometry and the rings present the same parameters. Thus, the geometry is characterized by the number of rings surrounding the fovea (m) and the number of subrings of resolution cells (*rexels*) found in the directions of the Cartesian axes within any of the rings. **Figure 4** shows an example of a fovea-centered CFG.

Among other advantages, there are CFGs that are able to provide a shiftable fovea of adaptive size (Arrebola et al., 1997) (adaptive CFGs). Vision systems which use the fovea-centered CFG require to place the region of interest in the center of the image. That is usually achieved by moving the cameras. A shiftable fovea can be very useful to avoid these camera movements. Furthermore, the adaptation of the fovea to the size of the

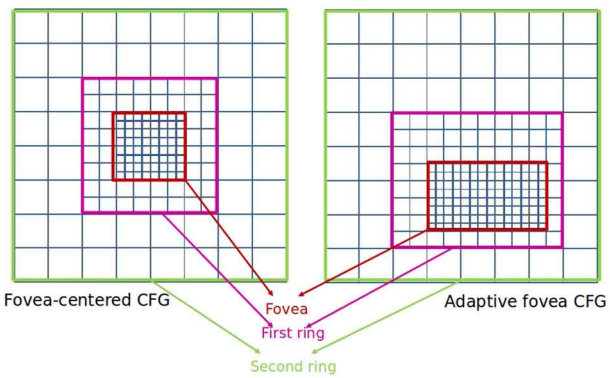


FIGURE 4 | Cartesian Foveal Geometries (CFG).

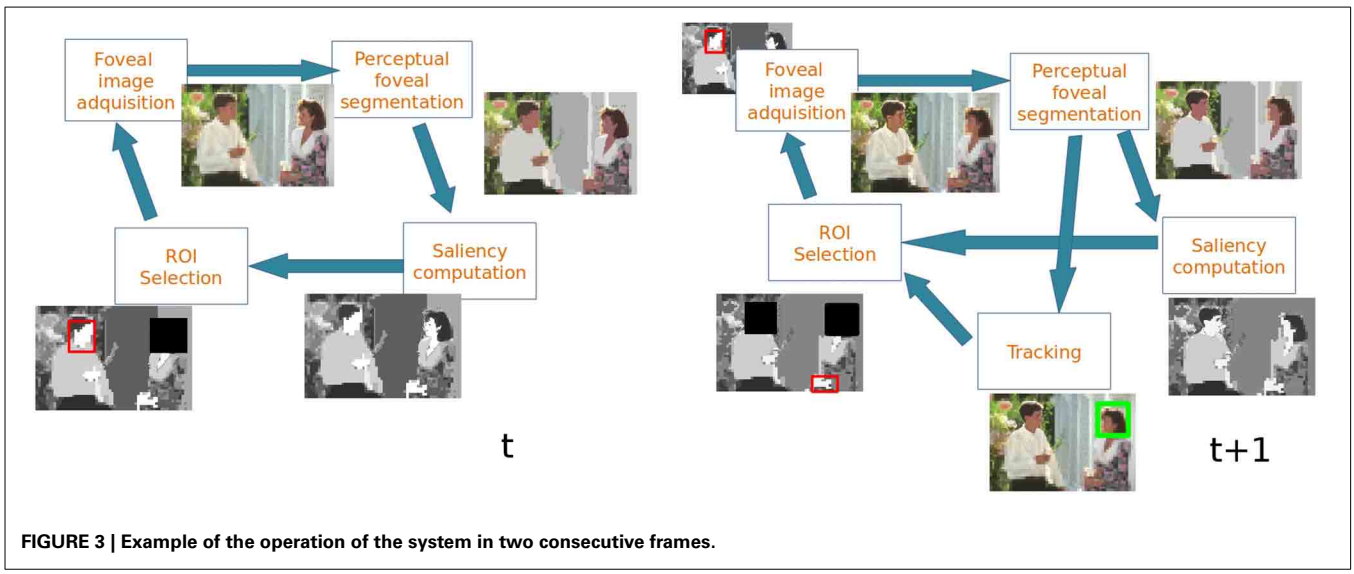


FIGURE 3 | Example of the operation of the system in two consecutive frames.

region of interest can help to optimize the consumption of computational resources. **Figure 4** shows the rectangular structure of an adaptive fovea. The geometry is now characterized by the subdivision factors at each side of the fovea. It should be noted that the foveal geometry is not adequate for processing planar images. On the contrary, the aim is to use it for hierarchical processing. Thus, a hierarchical representation of the foveal image (the foveal polygon) is built like **Figure 5** shows. This foveal polygon has a first set of levels of abstraction built from the fovea to the waist (the first level where the complete field of view is encoded). In the figure, levels 1 and 2 on this hierarchy are built by decimating the information from the level below and adding the data from the corresponding ring of the multi-resolution image. Over the waist, there are a second set of levels. All these levels encode the whole field of view and are built by decimating the level below.

Typically, the decimation process inside the CFGs have been conducted using regular approximations (Arrebola et al., 1997). Then, all levels of the foveal polygon can be encoded as images. The problems of regular decimation processes were early reported (Antonisse, 1982; Bister et al., 1990), but here, these processes were justified due to the simplicity for processing (Traver and Bernardino, 2010).

In this work, we propose to build the foveal polygon using the irregular decimation process provided by the Bounded Irregular Pyramid (BIP) (Marfil et al., 2007).

2.2. PERCEPTUAL FOVEAL SEGMENTATION USING BIP

The BIP is an irregular pyramid which is defined by a data structure and an irregular decimation process. This irregular decimation is applied to build the foveal polygon by segmenting the foveal input image using a perceptual segmentation approach which allows to extract the proto-objects from the visual scene.

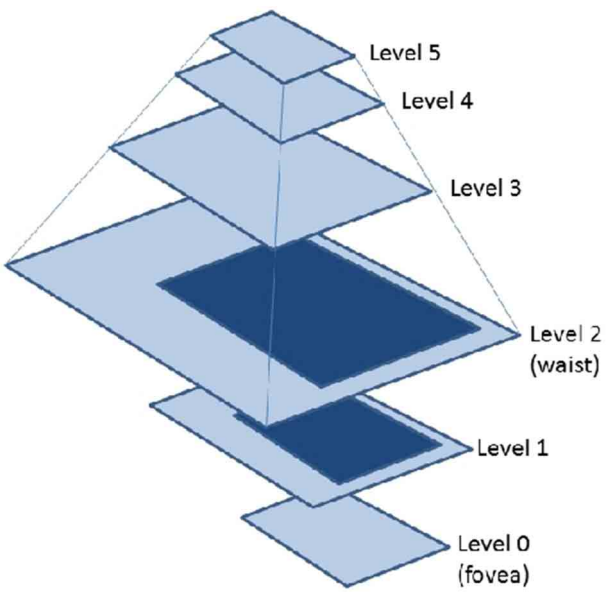


FIGURE 5 | Foveal Polygon associated to an adaptive CFG with two rings.

2.2.1. Data structure of the BIP

The data structure of the BIP is a mixture of regular and irregular data structures: a 2 × 2/4 "incomplete" regular structure and a simple graph. The regular structure of the BIP is said to be incomplete because, although the whole storage structure is built, only the homogeneous regular nodes (see subsection 2.2.2) are set in it. Therefore, the neighborhood relationships of these nodes can be easily computed. The mixture of both regular and irregular structures generates an irregular configuration which is described as a graph hierarchy. In this hierarchy, there are two types of nodes: nodes belonging to the 2 × 2/4 structure, named *regular nodes* and *irregular nodes* or nodes belonging to the irregular structure. Therefore, a level l of the hierarchy can be expressed as a graph $G_l = (N_l, E_l)$, where N_l stands for the set of regular and irregular nodes and E_l for the set of arcs between nodes (intra-level arcs). Each node $n_i \in N_l$ is linked with a set of nodes $\{n_k\}$ of N_{l-1} using inter-level arcs, being $\{n_k\}$ the reduction window of n_i. A node $n_i \in N_l$ is neighbor of other node $n_j \in N_l$ if their reduction windows w_{n_i} and w_{n_j} are connected. Two reduction windows are connected if there are at least two nodes at level l-1, $n_p \in w_{n_i}$ and $n_q \in w_{n_j}$, which are neighbors.

2.2.2. Decimation process of the foveal BIP

Two nodes x and y which are neighbors at level l are connected by an intra-level arc $(x, y) \in E_l$. Let ε_l^{xy} be equal to 1 if $(x, y) \in E_l$ and equal to 0 otherwise. Then, the neighborhood of the node x ($\xi_{\mathbf{x}}$) can be defined as $\xi_{\mathbf{x}} = \left\{\mathbf{y} \in N_l : \varepsilon_l^{\mathbf{xy}}\right\}$. It can be noted that a given node $\mathbf{x}$ is not a member of its neighborhood, which can be composed by regular and irregular nodes. Each node $\mathbf{x}$ has associated a $v_{\mathbf{x}}$ value. Besides, each regular node has associated a boolean value $h_{\mathbf{x}}$: the homogeneity (Marfil et al., 2007). At the base level of the hierarchy G_0, the fovea, all nodes are regular, and they have $h_{\mathbf{x}}$ equal to 1 (they are homogeneous). Only regular nodes which have $h_{\mathbf{x}}$ equal to 1 are considered to be part of the regular structure. Regular nodes with an homogeneity value equal to 0 are not considered for further processing. The proposed decimation process transforms the graph G_l in G_{l+1} using the pairwise comparison of neighbor nodes. Then, a pairwise comparison function, $g(v_{\mathbf{x}_1}, v_{\mathbf{x}_2})$ is defined. This function is true if the $v_{\mathbf{x}_1}$ and $v_{\mathbf{x}_2}$ values associated to the $\mathbf{x}_1$ and $\mathbf{x}_2$ nodes are similar according to some criteria and false otherwise. When G_{l+1} is obtained from G_l, being $l <$ waist, this graph is completed with the regular nodes associated to the ring $l + 1$. This process will require to compute the neighborhood relationships among the regular nodes coming from the ring and the rest of nodes at G_{l+1}. Over the waist level, G_{l+1} is built by decimating the level below G_l.

The building process of the foveal BIP consists of the following steps:

1. Regular decimation process. The $h_{\mathbf{x}}$ value of a regular node $\mathbf{x}$ at level $l + 1$ is set to 1 if the four regular nodes immediately underneath $\{\mathbf{y}_i\}$ are similar according to some criteria and their $h_{\{\mathbf{y}_i\}}$ values are equal to 1. That is, $h_{\mathbf{x}}$ is set to 1 if

$$\left\{ \bigcap_{\forall \mathbf{y}_j, \mathbf{y}_k \in \{\mathbf{y}_i\}} g\left(v_{\mathbf{y}_j}, v_{\mathbf{y}_k}\right) \right\} \cap \left\{ \bigcap_{\mathbf{y}_j \in \{\mathbf{y}_i\}} h_{\mathbf{y}_j} \right\} \quad (1)$$

Besides, at this step, inter-level arcs among homogeneous regular nodes at levels l and $l+1$ are established. If $\mathbf{x}$ is an homogeneous regular node at level $l+1$ ($h_{\mathbf{x}} == 1$), then the set of four nodes immediately underneath $\{\mathbf{y_i}\}$ are linked to $\mathbf{x}$ and the v_x value is computed.

2. Irregular decimation process. Each irregular or regular node $\mathbf{x} \in N_l$ without parent at level $l+1$ chooses the closest neighbor $\mathbf{y}$ according to the $v_{\mathbf{x}}$ value. Besides, this node $\mathbf{y}$ must be similar to $\mathbf{x}$. That is, the node $\mathbf{y}$ must satisfy

$$\left\{||v_{\mathbf{x}} - v_{\mathbf{y}}|| = \min\left(||v_{\mathbf{x}} - v_{\mathbf{z}}|| : \mathbf{z} \in \xi_{\mathbf{x}}\right)\right\} \cap \left\{g\left(v_{\mathbf{x}}, v_{\mathbf{y}}\right)\right\} \quad (2)$$

If this condition is not satisfied by any node, then a new node $\mathbf{x}'$ is generated at level $l+1$. This node will be the parent node of $\mathbf{x}$ and it will constitute a root node. Its $v_{x'}$ value is computed. On the other hand, if $\mathbf{y}$ exists and it has a parent $\mathbf{z}$ at level $l+1$, then $\mathbf{x}$ is also linked to $\mathbf{z}$. If $\mathbf{y}$ exists but it does not have a parent at level $l+1$, a new irregular node $\mathbf{z}'$ is generated at level $l+1$ and $v_{z'}$ is computed. In this case, the nodes $\mathbf{x}$ and $\mathbf{y}$ are linked to $\mathbf{z}'$.

This process is sequentially performed and, when it finishes, each node of G_l is linked to its parent node in G_{l+1}. That is, a partition of N_l is defined. It must be noted that this process constitutes an implementation of the union-find strategy.

3. Definition of intra-level arcs. The set of edges E_{l+1} is obtained by defining the neighborhood relationships between the nodes N_{l+1}. As aforementioned, two nodes at level $l+1$ are neighbors if their reduction windows are connected at level l.
4. For $l <$ waist

- The set of nodes N_{l+1} is completed with the rexels of the ring $l+1$. These rexels are added as homogeneous regular nodes, N_{l+1}^{ring}.
- The intra-level arcs between nodes of N_{l+1}^{ring} and the rest of nodes of N_{l+1} are computed as in step 3. Nodes of N_{l+1}^{ring} do not have a real reduction window at level l, they present a *virtual reduction window*. The virtual reduction window of a node $\mathbf{x} \in N_{l+1}^{\text{ring}}$ is computed by quadrupling this node at level l . Therefore, the reduction window of $\mathbf{x}$ is formed by the four nodes immediately underneath at level l.

In **Figure 6** the whole process to build the structure of the BIP associated to a foveal image with one ring is shown. Homogeneous regular nodes are represented by squares or cubes and irregular ones by spheres. In the first row, the process to build the first level is shown. From left to right: original image, nodes of the first level generated after the regular and irregular decimation processes (only some inter-level arcs are shown), structure of the first level after the definition of the intra-level arcs and final structure of the first level after adding the nodes of the ring (the virtual reduction window of one node of the ring is shown). In the second row of the figure, the rest of levels are shown.

2.2.3. Perceptual segmentation

As the process to group image pixels into higher-level structures can be computationally complex, perceptual segmentation approaches typically combine a pre-segmentation step with a subsequent perceptual grouping step. The pre-segmentation step performs the low-level definition of segmentation. It groups pixels into homogeneous clusters. Thereby, pixels in input image are grouped into blobs of uniform color, replacing the pixel-based image representation. Besides, these regions preserve the image geometric structure because each significant feature contains at least one region. The perceptual grouping step conducts a domain-independent grouping which is mainly based on properties such as proximity, closure or continuity. Both steps are conducted using the aforementioned decimation process but employing different similarity criteria between nodes.

In order to compute the pre-segmentation stage, a basic color segmentation is applied. In this case, a distance based on the HSV color space is used. Two nodes n_i and n_j are similar (they share a similar color) if their HSV values are less or equal than a similarity threshold τ_{color}:

$$g(v_{\mathbf{n}_i}, v_{\mathbf{n}_j}) = (d(n_i, n_j)) \leq \tau_{\text{color}}) \quad (3)$$

being $v_{\mathbf{n}_i}$ and $v_{\mathbf{n}_j}$ the HSV color of nodes n_i and n_j in cylindrical coordinates, and $d(n_i, n_j)$ is the color distance between them.

$$d(n_i, n_j) = \sqrt{d_v(n_i,\ n_j) + d_c(n_i,\ n_j)} \quad (4)$$

where

$$d_v(n_i, n_j) = |V_i - V_j| \quad (5)$$

$$d_c(n_i, n_j) = \sqrt{S_i + S_j + 2 \cdot S_i \cdot S_j \cdot \cos\theta} \quad (6)$$

with $\theta = |H_i - H_j|$.

In the perceptual grouping step, the roots of the pre-segmented blobs are considered the first level of a new segmentation process. In this case, two constraints are taken into account for an efficient grouping process: first, although all groupings are tested, only the best groupings are locally retained; and second, all the groupings must be spread on the image so no part of the image takes advantage. As segmentation criterion, a more complex distance is employed instead of a simple color threshold. This distance has three main components: the color contrast between blobs, the edges of the original image, obtained using a Canny detector, and the depth information of the image blobs in form of disparity. To avoid working at pixel resolution, which decreases the computational speed, a global contrast measurement is used instead of a local one. Then, the distance $\phi(n_i, n_j)$ between two nodes n_i and n_j is defined as:

$$\phi(n_i, n_j) = \sqrt{\omega_1 \left[\frac{d(n_i, n_j) \cdot b_i}{\alpha \cdot C_{ij} + \beta(b_{ij} - c_{ij})}\right]^2 + \omega_2 \left[\delta(n_i) - \delta(n_j)\right]^2} \quad (7)$$

where $d(n_i, n_j)$ is the HSV color distance between n_i and n_j, $\delta(x)$ is the mean disparity associated to the base image region represented by node x, b_i is the perimeter of n_i, b_{ij} is the number of pixels in the common boundary between n_i and n_j and c_{ij} is the set of pixels in this common boundary which corresponds to pixels of the boundary obtained using the Canny detector. α

and β are two constant values used to control the influence of the Canny edges in the grouping process. ω_1 and ω_2 are two constants which weight the terms associated with color and disparity. These parameters should be manually tuned depending on the application and the environment. Two nodes are similar if the distance $\phi(n_i,\ n_j)$ between them is equal or less than a threshold τ_{percep}:

$$g(v_{\mathbf{n}_i}, v_{\mathbf{n}_j}) = \left(\phi(n_i, n_j)\right) \leq \tau_{\text{percep}}) \tag{8}$$

The grouping process is iterated until the number of nodes remains constant among two consecutive levels, because it is not possible to group together more nodes because they are not similar. After the perceptual grouping, the nodes of the BIP with no parents are the roots of the proto-objects. **Figure 7** shows an example of the result of a perceptual segmentation.

3. SALIENCY COMPUTATION AND ROI SELECTION

Once the scene is divided into proto-objects, the next step is the selection of the most relevant one. According to Treisman and Gelade (1980), this process is based on the computation of a set of low-level features. But, what features must be taken into consideration? What features really guide attention?

According to psychological studies, some features, such as color (Treisman and Souther, 1985), motion (McLeod et al., 1988) or orientation (Wolfe et al., 1992), clearly influence in saliency computation. These three features, plus size, are cataloged by Wolfe and Horowitz (2004) as the only undoubted attributes that can guide attention. Wolfe also offers in his work a complete list of features that might guide the deployment of attention, grouped by their likelihood to be an effective source of attentional guiding. He differentiates among the aforementioned undoubted attributes, probable attributes, possible attributes, doubtful cases and probable non-attributes.

Another important issue when selecting features to develop an artificial attention system is concerned with computational cost. Computing a large number of features provides a richer description about elements in the scene. However, the associated computing time could be unacceptable. Hence, it is necessary a trade-off between computational efficiency and the number and type of the selected features.

Following the previous guidelines, seven different features have been selected to compute saliency in the proposed system. From the undoubted attributes, orientation and color have been chosen. Because there is no background subtraction in the perceptual segmentation, larger proto-objects usually correspond to non-relevance parts of the image (e.g., walls, floor or empty tables). Therefore, size feature is not employed to avoid an erroneous highlighting of irrelevant elements. Motion is discarded

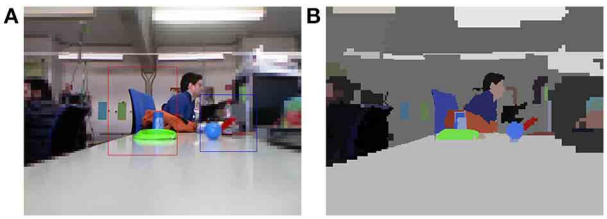


FIGURE 7 | (A) Foveal image; **(B)** Perceptual segmentation associated to **(A)** ($\tau_{\text{color}} = 50$, $\tau_{\text{percep}} = 100$).

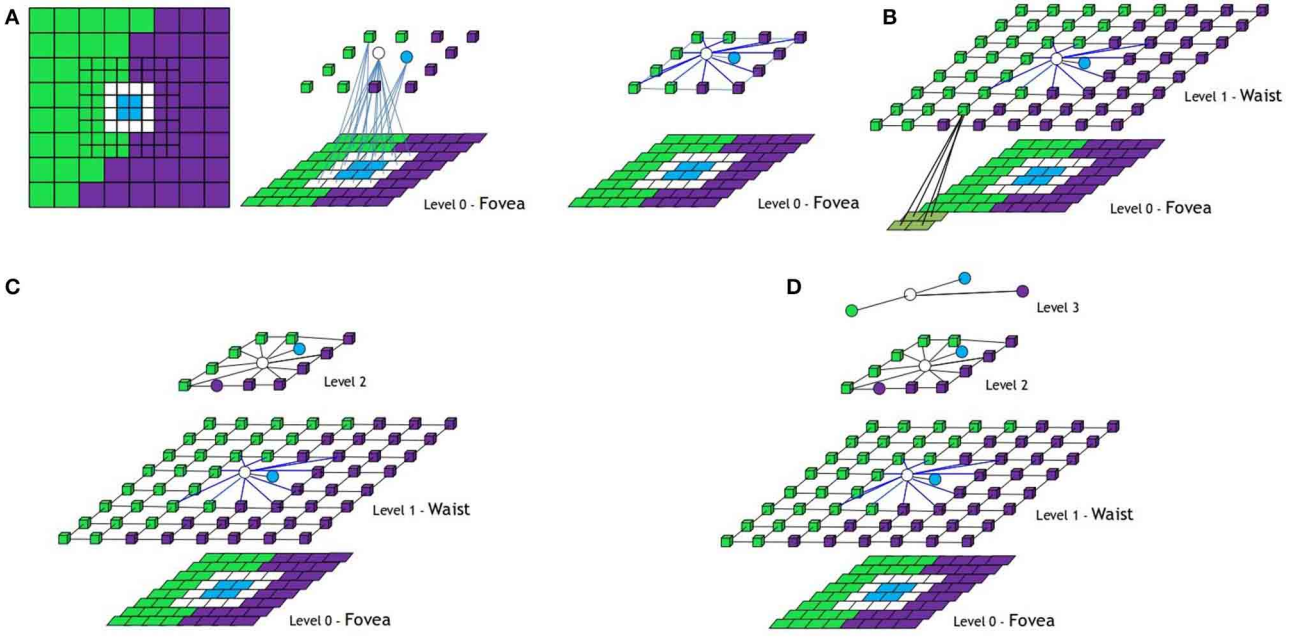


FIGURE 6 | Foveal image with one ring and how the structure of the Bounded Irregular Pyramid associated to it is built. (A) Building the central part of Level 1 from the fovea (Level 0), **(B)** definition of the intra-level edges at this part and adding new nodes from Ring 1, **(C)** Building Level 2, and **(D)** Building Level 3. Regular nodes are drawn as 3d cubes and irregular ones as spheres (see text).

due to computational cost restriction. Although intensity contrast is not considered an undoubted feature, it has also been included as a special case of color contrast (intensity deals with gray, black, and white elements). From the remainder of available possible attributes, those describing shape and location have been considered as more suitable for a complete description of the objects in scene. Location is calculated in terms of proximity to the visual sensor. Regarding the shape, two features are taken into account: symmetry, which allows to discriminate between symmetric and non-symmetric elements, and roundness, a measure about the closure and the contour of an object. Finally, in order to reach a social interaction with humans, it seems to be reasonable to include features able to pop out people from a scene. Although some works directly consider faces as a feature (Judd et al., 2009), experimental studies differ (Nothdurft, 1993; Suzuki and Cavanagh, 1995). Faces themselves do not guide attention but they can be separated into basic features that really achieve the guidance (Wolfe and Horowitz, 2004). In general, global properties are correlated with low-level features that explain search efficiency (Greene and Wolfe, 2011). Consequently, the proposed model uses similarity with skin color as an undoubted feature to guide attention to human faces in combination with other features as roundness.

To summarize, saliency is computed in terms of the following features: color contrast, intensity contrast, proximity, symmetry, roundness, orientation and similarity to skin color. All features values are normalized in the range $[0 \ldots 255]$ in order to have an homogenized calculus. As most of the artificial attention systems following Treisman's *Feature Integration Theory* (Treisman and Gelade, 1980), the total saliency of an element in an image is the result of a linear combination of its low-level features. **Figure 8** shows an example of foveal image and its associated feature maps. These feature maps represent the value of the corresponding feature for each proto-object. The final saliency map is also shown.

In the proposed attention system, the final saliency value, sal_i, for each proto-object, $\mathcal{P}_i$, is obtained as a weighted sum of all the previously described features:

$$sal_i = \vec{\lambda} \cdot \vec{f} \tag{9}$$

where $\vec{\lambda}$ is a set of weights, verifying $\sum_i \lambda_i = 1$, and $\vec{f}$ is the feature vector formed by the different features computed as explained in the following subsections. As it was previously commented in the Introduction section the weights can be set depending on the task in a top-down way. For example, in **Figure 9** two saliency maps obtained with a different set of weights are shown. While, in the left saliency map all the weights are set to the same value, in the right map the weight associated to the proximity feature is higher than the rest, and therefore, the proto-objects closer to the camera have a bigger saliency value than those who are far away. This variation in the saliency values causes a modification in the location of the next fovea (blue boxes in a). Therefore, the sequence of fixations of a scene can be modified by varying the values of the weights.

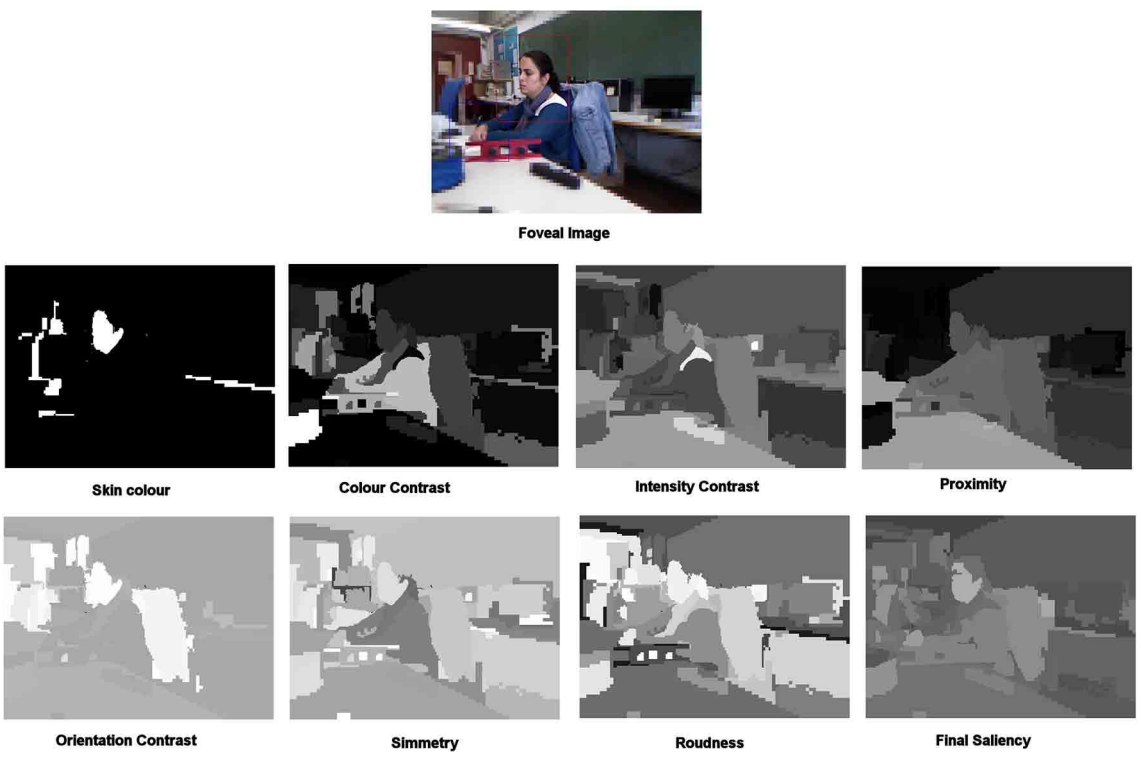


FIGURE 8 | Foveal image and its associated feature maps.

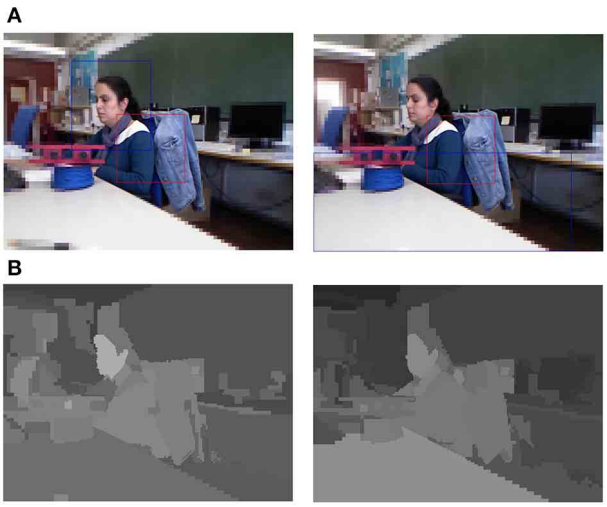


FIGURE 9 | (A) First frames of two very similar sequences where the red box corresponds with the current fovea and the blue box corresponds with the next ROI; **(B)** Saliency maps obtained with all the weights set to 1/7 (left image) and with the weight corresponding to proximity equal to 0.5 and the rest to 0.5/6 (right image).

3.1. COLOR CONTRAST AND INTENSITY CONTRAST

These features measure how different a proto-object is with respect to its surrounding in terms of color and luminosity. The color contrast, (*ColCON*), of a specific proto-object, $\mathcal{P}_i$, can be computed as the mean color gradient along its boundary to the neighbors in the segmentation hierarchy:

$$ColCON_i = \frac{S_i}{b_i} \sum_{j \in N_i} b_{ij} \cdot d\left(< C_i >, < C_j >\right) \tag{10}$$

where b_i is the perimeter of $\mathcal{P}_i$, N_i is the set of proto-objects that are neighbors of $\mathcal{P}_i$, b_{ij} is the length of the perimeter of $\mathcal{P}_i$ in contact with proto-object $\mathcal{P}_j$, $d\left[< C_i >, < C_j >\right]$ is the HSV color distance between the color mean values $< C >$ of proto-objects $\mathcal{P}_i$ and $\mathcal{P}_j$ and S_i is the mean saturation value of proto-object $\mathcal{P}_i$.

Because of the use of S_i in the color contrast equation, white, black and gray proto-objects are suppressed. Thus, a feature about intensity contrast is also introduced. The intensity contrast, (*IntCON*), of a proto-object, $\mathcal{P}_i$, is computed as the mean luminosity gradient along its boundary to the neighbors:

$$IntCON_i = \frac{1}{b_i} \sum_{j \in N_i} b_{ij} \cdot d\left(< I_i >, < I_j >\right) \tag{11}$$

being $< I_i >$ the mean luminosity value of the proto-object $\mathcal{P}_i$.

3.2. PROXIMITY

Another important parameter in order to characterize a proto-object is to determine its distance to the vision system. Nowadays, not only stereo pairs of cameras but also cheaper devices like Microsoft Kinect or ASUS Xtion provide accurate depth information of the captured image.

When using a sensor able to directly provide depth information (e.g., a RGBD camera or similar), the proximity, (*PROX*), of a proto-object, $\mathcal{P}_i$, is directly obtained as the inverse of the mean of the depth values provided by the sensor in the area of the proto-object $depth_i$:

$$PROX_i = \frac{1}{depth_i} \tag{12}$$

In the case of using a stereo pair of cameras as depth sensor, the proximity can be obtained directly from disparity information.

3.3. ROUNDNESS

Roundness measurement reflects how similar to a circle a proto-object is. This feature provides information about convexity, closure and dispersion. Roundness is obtained employing a traditional technique based on image moments. Concretely, three different central moments are used:

$$\mu^i_{1,1} = \sum (x - \overline{x})(y - \overline{y}) \quad \forall (x, y) \in \mathcal{P}_i \tag{13}$$

$$\mu^i_{2,0} = \sum (x - \overline{x})^2 \quad \forall (x, y) \in \mathcal{P}_i \tag{14}$$

$$\mu^i_{0,2} = \sum (y - \overline{y})^2 \quad \forall (x, y) \in \mathcal{P}_i \tag{15}$$

being $(\overline{x}, \overline{y})$ the center of the proto-object $\mathcal{P}_i$.

From the combination of the equations above, it is possible to measure the difference between a region and a perfect circle. This measure is known as eccentricity and can be calculated as follows:

$$ecc_i = \frac{\left(\mu^i_{2,0} - \mu^i_{0,2}\right)^2 + \left(2\mu^i_{1,1}\right)^2}{\left(\mu^i_{2,0} + \mu^i_{0,2}\right)^2} \tag{16}$$

being the result in the range $[0 \ldots 1]$.

Finally, the roundness, ($ROUND_i$), for a proto-object, $\mathcal{P}_i$, is obtained from the definition of eccentricity as:

$$ROUND_i = 1 - ecc_i \tag{17}$$

3.4. ORIENTATION

The orientation of a region in a image can also be obtained from central moments computed in (13–15):

$$\varphi_i = \frac{1}{2} \arctan\left(\frac{2\mu^i_{1,1}}{\mu^i_{2,0} - \mu^i_{0,2}}\right) \tag{18}$$

But the orientation of a proto-object, by itself, does not provide any useful information about its relevance. Only when comparing its orientation with the orientation of the rest of proto-objects in the image, a feasible measure of relevance is obtained. Thus, in fact, it is more interesting to compute saliency in terms of

contrast with the surrounding elements. The orientation contrast, ($OriCON$), of a proto-object, $\mathcal{P}_i$, is obtained as:

$$OriCON_i = \sum_{j \in N_i} |\varphi_i - \varphi_j| \tag{19}$$

where N_i is the set of proto-objects that are neighbors of $\mathcal{P}_i$.

Although pure orientation information is not employed to calculate relevance, it is saved as a descriptor of the proto-object for further use (for example, to compute symmetry).

3.5. SYMMETRY

To compute the symmetry of a proto-object, an approach similar to Aziz and Mertsching (2008) is followed. They propose a method to obtain symmetry using a scanning function $\psi(L, P_s)$ that counts the symmetric points around a point P_s along a line L. This procedure is repeated employing different lines of reference. For each line, the measure of symmetry is computed as:

$$S^{\theta} = \sum_{s=1}^{l} \frac{\psi(L, P_s)}{\alpha(R_i)} \tag{20}$$

where l and θ are the length and the angle of the line of reference and $\alpha(R_i)$ is the area of the region in order to normalize the result between 0 and 1.

Only an approximation of symmetry is needed in terms of attention systems. Thus, only 4 different angles for symmetry axes are considered: 0, 45, 90, and 135° respect to the orientation, φ_i, of the image [obtained in (18)]. In Aziz and Mertsching (2008), the total measure of symmetry is computed as an average of the symmetry values in the different lines of reference. Nevertheless, such strategy can define a region with only one axis of symmetry as asymmetric, because non-symmetric axes cancel out the contribution of the symmetric one.

As relevance is given to symmetry independently of the axis of symmetry, the maximum symmetry, ($SYMM$), for a proto-object, $\mathcal{P}_i$, is computed as:

$$SYMM = max_{\theta}(S^{\theta}) \tag{21}$$

3.6. SKIN COLOR

The computation is based on the skin color chrominance model proposed by Terrillon and Akamatsu (1999). First, the image is transformed into the TSL color space. Then, the Mahalanobis distance between the color of the proto-object and the mean vector of the skin chrominance model is computed. If this distance is less than a threshold Θ_{skin}, the skin color feature is marked with a value of 255. Otherwise, it is set to 0.

$$SKN_i = \begin{cases} 255 \text{ if } d_M\left(< C_i^{TSL} >, < C_{yellow}^{TSL} >\right) \leq \Theta_{skin} \\ 0 \quad \text{otherwise} \end{cases} \tag{22}$$

3.7. INHIBITION OF RETURN AND ROI SELECTION

Once the saliency of each proto-object has been computed, the most salient one is selected as the next ROI where the fovea will be located in the next frame. In this process it is necessary to take into account that revisiting already attended proto-objects and ignoring not attended ones must be avoided. To do that an inhibition of return algorithm should be implemented.

Psychophysics studies about human visual attention have established that a local inhibition is activated in the saliency map when a region is already attended. This mechanism avoids directing focus of attention to a region immediately visited and it is normally called *inhibition of return (IOR)* (Posner et al., 1985). In order to handle dynamic environments, this IOR mechanism needs to establish a correspondence between regions among consecutive frames. In order to associate this inhibition to the computed proto-objects and not only to activity clusters as in Backer et al. (2001) or to object features as in Aziz and Mertsching (2007), an object-based inhibition of return applying image tracking is employed instead in the proposed work. To do that, recently attended proto-objects are stored in a Working Memory (WM). When the vision system moves, the proto-objects stored in the WM are kept tracked. In the next frame, a new set of proto-objects is obtained from the image and the positions of the previously stored ones are updated. Then, from the new set of proto-objects, those occupying the same region than the already attended ones are suppressed. Discarded proto-objects are not taken into account in the selection of the most salient one.

A tracker based on Dorin Comaniciu's *mean-shift* approach (Comaniciu et al., 2003) is employed to achieve the inhibition of return. Mean-shift algorithm is a non-parametric density estimator that optimizes a smooth similarity function to find the direction of movement of a target. A mean-shift based tracker is specially interesting because of its simplicity, efficiency, effectiveness, adaptability and robustness. Moreover, its low computational cost allows to track several objects in a scene maintaining a reasonable frame rate (real-time tracking of multiple objects). In the proposed system, the target model is represented by a 16-bin color histogram masked with an isotropic kernel in the spatial domain. Specifically, the Epanechnikov kernel is employed.

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed foveal attention system, the experiments have been divided into three parts: the comparison between uniform and foveal attention models; the evaluation of the ability of the approach for actively driving an image exploration process; and finally the evaluation of the attention and fixation prediction model. All tests have been conducted on an Intel(R) Core(TM)2 Duo CPU T8100 2.10 GHz.

4.1. UNIFORM vs. FOVEAL ATTENTION

One of the main reasons for using a foveal strategy is the reduction of the computational costs. In our tests, running the system within different platforms, the foveal attention approach demonstrated to be approximately 4 times faster than its uniform counterpart. All tests were conducted using a Microsoft Kinect as input and working with images of 640 × 480 pixels. Within this framework, the algorithm is able to run at 10–12 frames per second (fps). The reduction on computational cost is significant, specially if we consider that the foveal image generation (the Kinect sensor provides an uniform image) is included in

the computational costs associated to the foveal approach. If we remove these costs, the foveal approach is approximately 6 times faster.

Then, the question is: what is the cost to pay for being faster? **Figure 10** assesses the sequence of fixations obtained by an attention model that uses (top) uniform images and (bottom) foveal images. It must be noted that they are not the same video sequence, and although the scenario is the same for both trials (with the same relevant items), some differences can be presented due to light variations or slightly motions. In both cases the same set of weights has been employed for the saliency computation and the results are then very similar. There are significant differences on the peripheral part of the image, but the fovea is in both cases at the same resolution. And the fovea includes the object to attend.

On the contrary, the drawbacks of being slow are clear when dealing with real scenarios. Thus, **Figure 11** shows how the use of foveal images is not sufficient to attend on time to a region marked as relevant (on the second frame of the sequence). When the fovea moves to this position (third frame), it does not find the searched region. The active exploration continues and the fovea will move to a new coherent position (the blue cup) on the next frame.

4.2. ACTIVE EXPLORATION USING THE FOVEAL ATTENTION APPROACH

As it has been illustrated in the previous section, due to its foveal nature, the proposed approach does not provide a single saliency map for a given scenario but a sequence of saliency maps. Thus, there is an iterative flow whose steps imply (a) to move the fovea to a new location, (b) to obtain a new saliency map, and (c) to determine the new location of the fovea according to this map. The foveal approach should then be understood within the framework of video processing, i.e., scenarios where visual information constantly changes due to ego-centric movements or dynamics of the world (Borji and Itti, 2013b). When we use this approach for exploring a static scene, the result will be the same: it is necessary more than one iteration to explore it (unless this has only one relevant object). **Figure 12** shows scanpath results for three images from the Saliency ToolBox (http://www.saliencytoolbox.net/). The left column shows the results obtained using the approach by Walther and Koch (2006). The right one the set of proto-objects obtained using our approach. Gaze ordering is drawn over the images. Each iteration provides a foveal region to be analyzed in detail. This exploration is an active process which is completed in a finite number of iterations (when all the relevant parts of the image have been located at the fovea). This behavior is due to the existence of an IOR mechanism

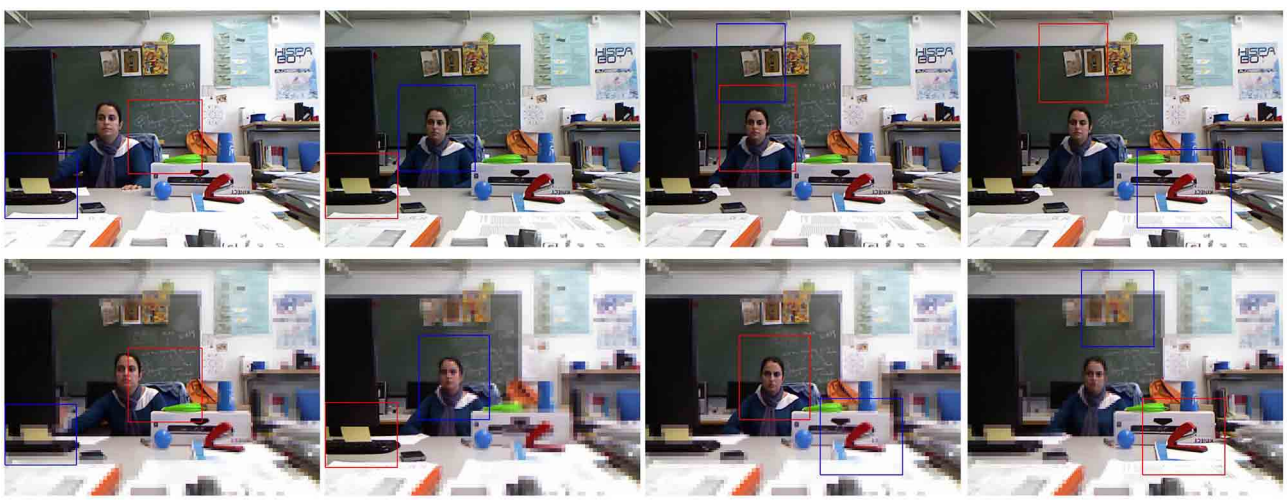


FIGURE 10 | Active exploration of a video sequence. (Top) uniform images, and **(Bottom)** foveal images. In both cases the used color parameters have been $\tau_{color} = 50$ and $\tau_{percep} = 100$.

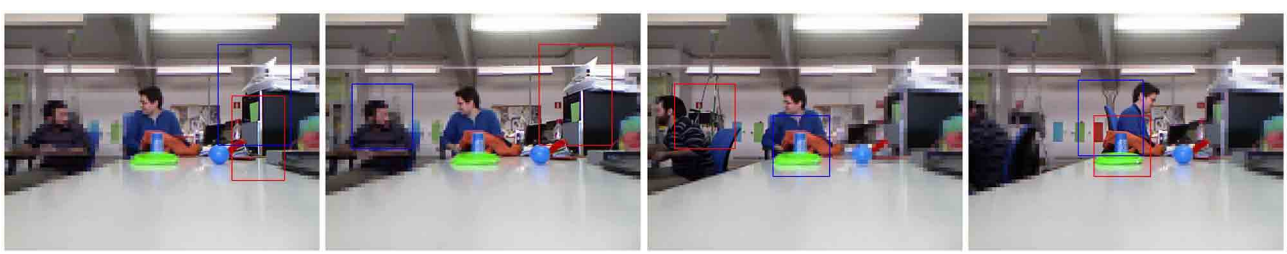


FIGURE 11 | Active exploration of a video sequence (see text for details) ($\tau_{color} = 50$, $\tau_{percep} = 100$).

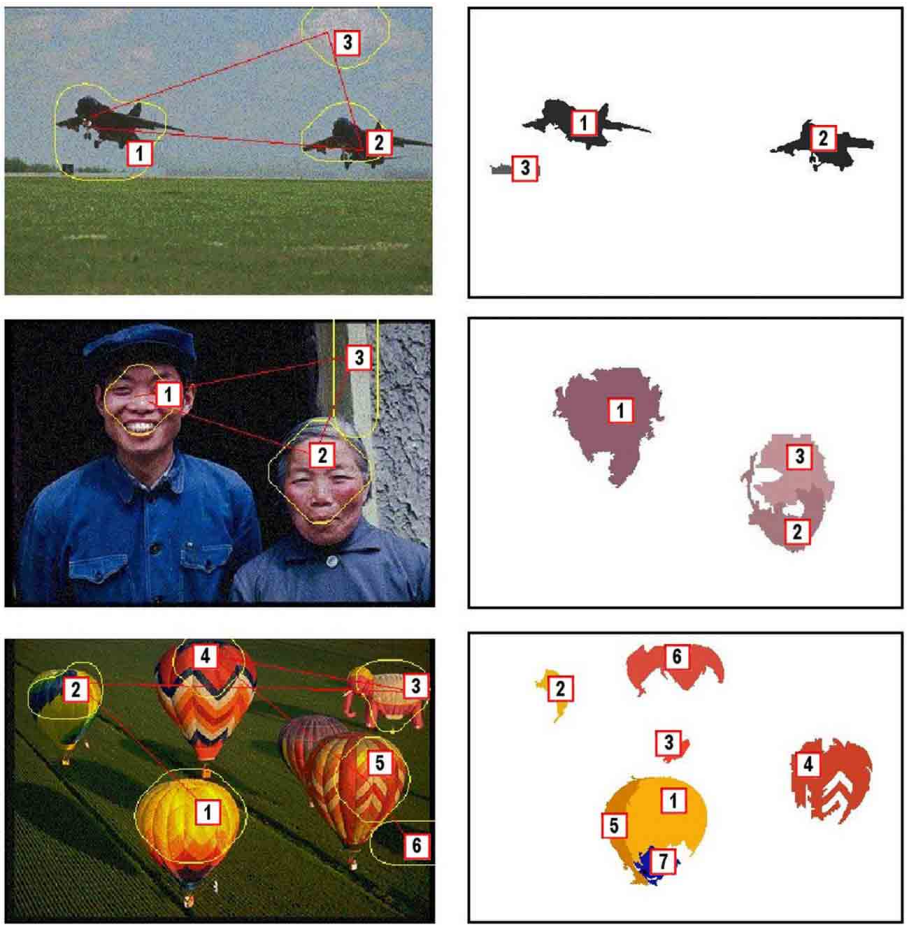


FIGURE 12 | Scanpath results for three images from the Saliency ToolBox. (Left) Results obtained using the approach by Walther and Koch (2006), and **(Right)** sets of proto-objects obtained using our approach. Gaze ordering is drawn over the images.

but also to the existing differences among foveal segmentations results depending on the location of the fovea. The foveal region is segmented in detail while the level of detail decreases with the distance to the fovea. That is, the segmentation of the same region can be very different between iterations. This is illustrated in **Figures 13**, **14**.

Figure 13 shows that the approach outcomes a fixation region in each iteration. This fixation region is the most salient proto-object inside the fovea. These proto-objects are usually among the set of segments in which the people divides up the image (face, one hand, one leg…). The top-middle image of the figure represent the first seven fixation regions. At the bottom (left and middle), the figure shows the first two segmentations. Although there is certain constancy on the boundaries, they are not identical. Segmentations will be more different when fixation regions are more distant on the image. For instance, this occurs in **Figure 14**. From top-left to bottom-middle, this figure shows a sequence of fixations. The current fovea is marked within a red rectangle and the next within a blue one. The first fovea is over the face of the man, then it moves to a salient flower on the top-left corner, then to the hand of the man… Sometimes, this scan-path does not follow the path we could desire: from the hand it now moves to the elbow of the man and, from here, to the dress of the woman. But we are dealing with an active process, and it will return to "relevant" (from our point-of-view) regions quickly. Finally, this image also shows how the IOR works. After some frames, the fovea returns to previously visited regions (the face of the man, his hand...). Results are similar to the ones provided by the approach by Walther and Koch (2006) (see the bottom-right images at **Figures 13**, **14**).

The effectiveness of our approach has been verified with experiments performed on human eye gaze data. As ground truth scan-paths, we use the JUDD publicly available eye tracking dataset (Judd et al., 2009). This dataset records human gaze in a free viewing setting (1003 images with scan-paths of 15 subjects). Our estimated scan-paths are obtained as an ordered sequence of region's centroids. The comparison between an estimated scan-path and one of these ground truth scan-paths is performed using the similarity index described by Liu et al. (2013). In this measuring metric, there is a parameter (*gap*), which is the penalty value employed when it is necessary to add a gap (deletion or insertion operation) in any of the scan-paths during local alignment. It is

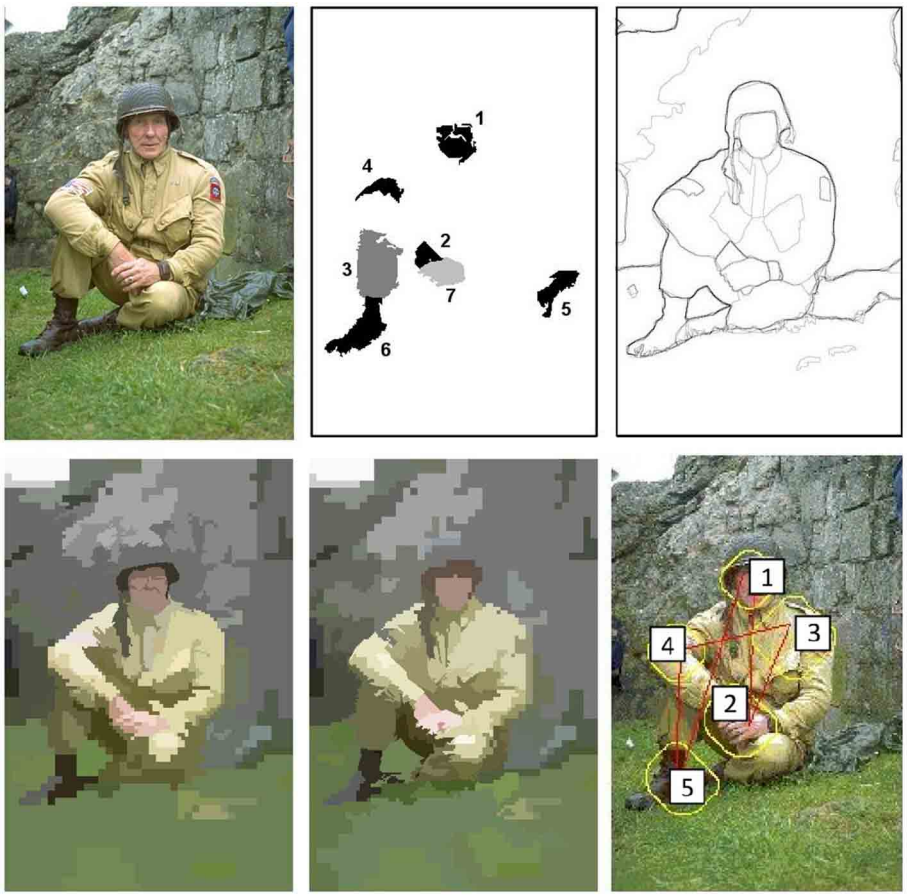


FIGURE 13 | Active exploration of the image #376043 of the Berkeley Segmentation Dataset. (Top left) original image, **(Top middle)** set of seven first fixation regions, **(Top right)** human segmentations, **(Bottom left-middle)** first two segmentations from the proposed approach and **(Bottom right)** scanpath result using the approach by Walther and Koch (2006).

set to -1/2 in our tests. Finally, for each image at the JUDD dataset, we have 15 ground truth scan-paths (one from each users). Then, we compare each scan-path with all these ground truth ones, providing the average similarity value. Our result is close to 1.05. It can be noted that our approach provides a better result in this framework than the approaches by Itti et al. (1998) and Walther and Koch (2006) (both under 0.9). On the other hand, this result is under the Liu et al. (2013)'s scores (close to 1.15). However, it should be appreciated that the Liu et al. (2013)'s approach does not only use low-level feature saliency, but also spatial position and semantic content. Our approach does not take into account these factors.

4.3. EXPERIMENTS WITH ATTENTION AND FIXATION PREDICTION

The approach has been evaluated using the Toronto database (Bruce and Tsotsos, 2009). This dataset was recently defined as the most widely used image data set in the review paper by Borji and Itti (2013b). The dataset contains 120 images (681×511 px) with eye-tracking data from 20 people. The subjects saw the images for four seconds, and they had no assigned task (i.e., free-viewing). **Figure 15** shows four images of the data set. Fixations are drawn over the images. A fixation density map is generated for each image based on these fixation points (Bruce and Tsotsos, 2009). They are also shown at **Figure 15** under each original image.

Contrary to the most attention approaches, our saliency maps should be also estimated from a set of fixations. However, contrary to the density maps obtained from experimental human eye tracking data, our fixations cannot be associated to points, but to regions. The fixation density maps shown at the bottom row of the **Figure 15** were built by the sum of the most saliency regions on n fixations. The number n was equal to the mean of the number of human fixations recorded for this image in the original data set.

Then, we use the well-known receiver operating characteristic (ROC) area under curve (AUC) measure to assess the performance of the approach. Each saliency map can be thresholded and then considered to be a binary classifier that separates positive samples (fixation points of all subjects on that image) from negative samples (fixation points of all subjects on all other images in the database). This process avoids the center-bias effect (Borji and Itti, 2013b). Then, we can sweep over all thresholds to estimate the ROC curve for each saliency map and calculated the area beneath the ROC curve. This area provides a good measure to



FIGURE 14 | Active exploration of the image #157055 of the Berkeley Segmentation Dataset. From **(Top left)** to **(Bottom middle)**, the figure shows a sequence of fixations (each image shows the current fovea, marked with a red rectangle, and the next one, within a blue rectangle). **(Bottom right)** Scanpath result using the approach by Walther and Koch (2006).

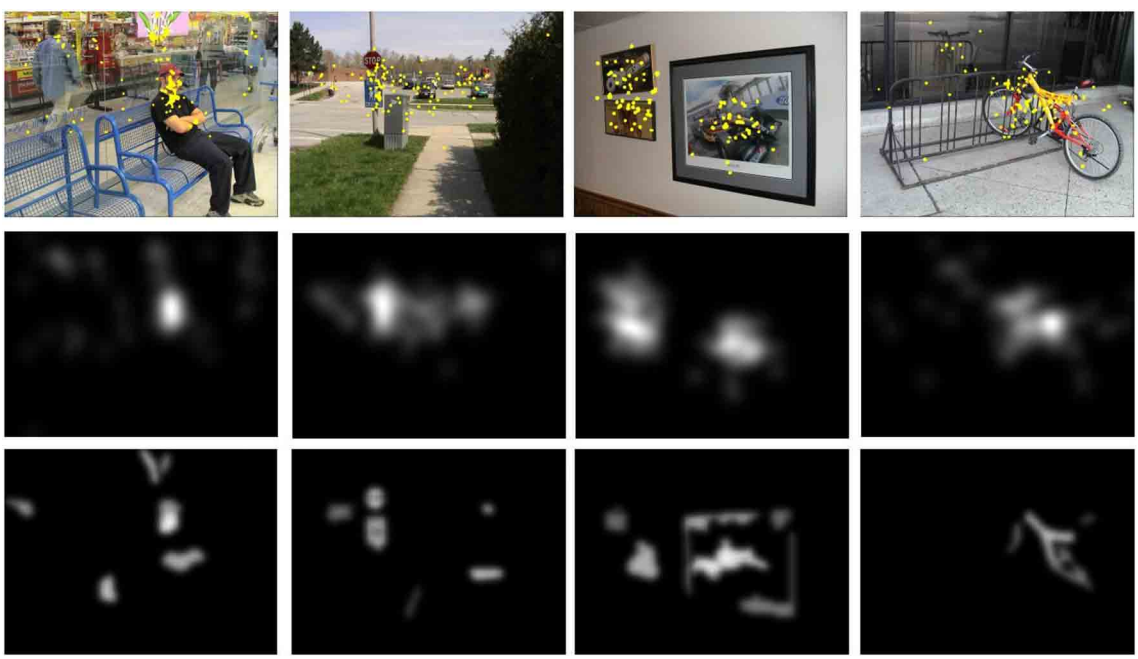


FIGURE 15 | Toronto Database. (Top) original images and fixation points, **(Middle)** fixation density maps obtained from the human fixations, and **(Bottom)** fixation density maps obtained by the proposed foveal attention approach.

assess how accurately the saliency map predicts the eye fixations on the image. An AUC value greater than 0.5 indicates positive correlation. As a performance baseline we can estimate an ideal AUC measuring how well the fixations of one subject can be predicted by the fixations of the rest of subjects. The ideal AUC for the data set is 0.878 (Borji and Itti, 2013b). In our experiments, the obtained score was 0.669. This value is similar to the ones provided by other methods. In the ranking documented by Borji and Itti (2013a), it will be the fifth best value of 28 evaluated models.

5. CONCLUSIONS AND FUTURE WORK

We proposed in this paper a foveal model of attention which combines static cues with depth and tracking to deal with dynamic scenarios. The framework was developed for an active observer, but this paper shows that it can also be applied to image databases. These static images were preferably employed to compare or evaluate the approach. Contrary to other approaches (such as the recently proposed by Mishra et al., 2012), we do not pursuit here a novel formulation of segmentation. Thus, in Section 4.2, we prefer to speak about active exploration and not segmentation. Active segmentation will probably require an additional (and better) algorithm that will try to extract the whole object from the fixation region. We refer the reader to the excellent work by Mishra et al. (2012) to understand the whole problem of active segmentation.

With respect to previous approaches to object-based attention, this work must be classified with those methods that compute the saliency of scene regions and not of isolated pixels. For this end, these approaches segment the input image before to evaluate and obtain the saliency map. As a main difference with previous works such as the ones by Orabona et al. (2007) and Yu et al. (2010), our approach performs this segmentation as a multi-resolution process, where only the fovea is processed with details. Thus, this segmentation depends on the position of the last fovea or ROI. Furthermore, our framework provides a complete approximation for closing the loop that involves segmentation and saliency estimation, including an inhibition of return mechanism. We consider that analyzing this loop closing is basic to understand an object-based attention mechanism working on a real, dynamic scenario.

This approach should be extended in several ways. Launched as a system to endow into a mobile robot, the foveal approach needs to be faster and to take into consideration top-down factors. We are working on both research direction. The speed will be improved by implementing the approach in a Zedboard platform. This is allowing to move part of the code to a FPGA, meanwhile the main function continues running on a processor. Top-down component of attention will initially come from the adjustment of the weights used to bias the saliency maps. Further work should be addressed to add object models on this process.

ACKNOWLEDGEMENT

This paper has been partially supported by the Spanish Ministerio de Economía y Competitividad TIN2012-TIN2012-38079-C03 and FEDER funds.

REFERENCES

Advani, S., Sustersic, J., Irick, K., and Narayanan, V. (2013). "A multi-resolution saliency framework to drive foveation," in *International Conference on Acoustics, Speech and Signal Processing*, 2596–2600. doi: 10.1109/ICASSP.2013.6638125

Antonisse, H. (1982). Image segmentation in pyramids. *Comput. Graph. Image Proc.* 19, 367–383. doi: 10.1016/0146-664X(82)90022-3

Arrebola, F., Camacho, P., and Sandoval, F. (1997). "Generalization of shifted fovea multiresolution geometries applied to object detection," in *ICIAP1997*, Vol. 2, (Florence), 477–484.

Aziz, M. Z., and Mertsching, B. (2007). "Color saliency and inhibition using static and dynamic scenes in region based visual attention," in *WAPCV, Lecture Notes in Computer Science*, Vol. 4840 (Hyderabad: Springer), 234–250

Aziz, Z., and Mertsching, B. (2008). Fast and robust generation of feature maps for region-based visual attention. *IEEE Trans. Image Proc.* 17, 633–644. doi: 10.1109/TIP.2008.919365

Backer, G., Mertsching, B., and Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system. *IEEE Trans. Patt. Anal. Mach. Intell.* 23, 1415–1429. doi: 10.1109/34.977565

Bister, M., Cornelis, J., and Rosenfeld, A. (1990). A critical view of pyramid segmentation algorithms. *Patt. Recogn. Lett.* 11, 605–617. doi: 10.1016/0167-8655(90)90013-R

Borji, A., and Itti, L. (2013a). Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *Image Proc. IEEE Trans.* 22, 55–69. doi: 10.1109/TIP.2012.2210727

Borji, A., and Itti, L. (2013b). State-of-the-art in visual attention modeling. *Patt. Anal. Mach. Intell. IEEE Trans.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Bruce, N., and Tsotsos, J. (2009). Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* 9, 1–24. doi: 10.1167/9.3.5

Bundesen, C., Habekost, T., and Kyllingsbaek, S. (2011). A neural theory of visual attention and short-term memory (ntva). *Neuropsychologia* 49, 1446–1457. doi: 10.1016/j.neuropsychologia.2010.12.006

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Trans. Patt. Anal. Mach. Intell.* 25, 564–575. doi: 10.1109/TPAMI.2003.1195991

Duncan, J. (1984). Selective attention and the organization of visual information. *J. Exp. Psychol.* 113, 501–517. doi: 10.1037/0096-3445.113.4.501

Frintrop, S. (2006). *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, Vol. 3899, Berlin; Heidelberg: Springer.

Frintrop, S., Rome, E., and Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: a survey. *ACM Trans. Appl. Percept.* 7, 1–39. doi: 10.1145/1658349.1658355

Geisler, W., and Perry, J. (1998). "A real-time foveated multiresolution system for low-bandwidth video communication," in *Proc. of the SPIE: The International Society for Optical Engineering*, Vol. 3299 (San Jose, CA).

Gide, M., and Karam, L. (2012). "Improved foveation- and saliency-based visual attention prediction under a quality assessment task," in *Workshop on Quality of Multimedia Experience* (Melbourne, VIC), 200–205.

Greene, M., and Wolfe, J. (2011). Global image properties do not guide visual search. *J. Vis.* 11, 1–9. doi: 10.1167/11.6.18

Guo, C., and Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its application in image and video compression. *IEEE Trans. Image Proc.* 19, 185–198. doi: 10.1109/TIP.2009.2030969

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 20, 1254–1259. doi: 10.1109/34.730558

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). "Learning to predict where humans look," in *IEEE 12th international Conference on Computer Vision* (Kyoto: Kyoto Prefecture), 2106–2113.

Jung, C., and Kim, C. (2012). A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. *IEEE Trans. Image Proc.* 21, 1272–1283. doi: 10.1109/TIP.2011.2164420

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlaying neural circuitry. *Hum. Neurobiol.* 4, 219–227.

Lin, Z., Davis, L. S., Doermann, D. S., and DeMenthon, D. (2007). "An interactive approach to pose-assisted and appearance-based segmentation of humans," in *IEEE 11th International Conference on Computer Vision, ICCV 2007* (Rio de Janeiro), 1–8.

Liu, H., Xu, D., Huang, Q., Li, W., Xu, M., and Lin, S. (2013). "Semantically-based human scanpath estimation with hmms," in *IEEE International Conference Computer Vision* (Sydney, NSW).

Marfil, R., Molina-Tanco, L., Bandera, A., and Sandoval, F. (2007). "The construction of bounded irregular pyramids with a union-find decimation process," in *GbRPR 2007. LNCS*, Vol. 4538, eds F. Escolano M. and Vento (Heidelberg: Springer), 307–318.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, Vol. 2 (Vancouver, BC), 416–423.

Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* 5, 229–240. doi: 10.1038/nrn1348

McLeod, P., Driver, J., and Crisp, J. (1988). Visual search for a conjunction of movement and form is parallel. *Nature* 332, 154–155. doi: 10.1038/332154a0

Meger, D., Forssn, P., Lai, K., Helmer, S., McCann, S., Southey, T., et al. (2008). Curious george: an attentive semantic robot. *Robot. Auton. Syst.* 56, 503–511. doi: 10.1016/j.robot.2008.03.008

Mishra, A. K., Aloimonos, Y., Cheong, L.-F., and Kassim, A. A. (2012). Active visual segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* 34, 639–653. doi: 10.1109/TPAMI.2011.171

Navalpakkam, V., and Itti, L. (2005). Moelling the influence of task on attention. *Vision Res.* 45, 205–231. doi: 10.1016/j.visres.2004.07.042

Nothdurft, H.-C. (1993). Faces and facial expressions do not pop out. *Perception* 22, 1287–1298. doi: 10.1068/p221287

Orabona, F., Metta, G., and Sandini, G. (2007). "A proto-object based visual attention model," in *WAPCV 2007. LNCS (LNAI)*, Vol. 4840, eds L. Paletta and E. Rome (Heidelberg: Springer), 198–215

Posner, M., Rafal, R., Choate, L., and Vaughan, J. (1985). Inhibition of return: neural basis and function. *Cogn. Neuropsychol.* 2, 211–228. doi: 10.1080/02643298508252866

Rajashekar, U., van der Linde, I., Bovik, C., and Cormack, L. (2008). Gaffe: a gaze-attentive fixation finding engine. *IEEE Trans. Image Proc.* 17, 564–573. doi: 10.1109/TIP.2008.917218

Rensink, R. (2000). Seeing, sensing, and scrutinizing. *Vision Res.* 40, 1469–1487. doi: 10.1016/S0042-6989(00)00003-1

Scholl, B. (2001). Objects and attention: the state of art. *Cognition* 80, 1–46. doi: 10.1016/S0010-0277(00)00152-9

Singaraju, D., Grady, L., and Vidal, R. (2008). "Interactive image segmentation via minimization of quadratic energies on directed graphs," in *CVPR* (Anchorage, AK: IEEE Computer Society).

Suzuki, S., and Cavanagh, P. (1995). Facial organization blocks access to low-level features: an object inferiority effect. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 901–913. doi: 10.1037/0096-1523.21.4.901

Terrillon, J., and Akamatsu, S. (1999). "Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images," in *Proceedings of the 12th Conference on Vision Interface* (Trois-Rivières, QC) 180–187.

Tipper, S. (1991). Object-centred inhibition of return of visual attention. *Q. J. Exp. Psychol.* 43, 289–298. doi: 10.1080/14640749108400971

Traver, J., and Bernardino, A. (2010). A review of log-polar imaging for visual perception in robotics. *Robot. Auton. Syst.* 58, 378–398. doi: 10.1016/j.robot.2009.10.002

Treisman, A., and Gelade, G. (1980). A feature integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Treisman, A., and Souther, J. (1985). Search asymmetry: a diagnostic for preattentive processing of separable features. *J. Exp. Psychol. Gen.* 114, 285–310. doi: 10.1037/0096-3445.114.3.285

Tsotsos, J. K., and Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia* 6:6201. doi: 10.4249/scholarpedia.6201

Walther, D., and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Netw.* 19, 1395–1407. doi: 10.1016/j.neunet.2006.10.001

Wolfe, J. (1994). Guided search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238. doi: 10.3758/BF03200774

Wolfe, J., Cave, K., and Franzel, S. (1989). Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 419–433. doi: 10.1037/0096-1523.15.3.419

Wolfe, J., and Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501. doi: 10.1038/nrn1411

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., and O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 34–49.

Yu, Y., Mann, K., and Gosine, R. (2010). An object-based visual attention model for robotic applications. *IEEE Trans. Syst. Man Cybern. B* 40, 1–15. doi: 10.1109/TSMCB.2009.2038895

Zhang, L., Tong, M., Marks, T., Shan, H., and Cottrell, G. (2008). Sun: a bayesian framework for saliency using natural statistics. *J. Vis.* 8, 1–20. doi: 10.1167/8.7.32

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2014; accepted: 24 July 2014; published online: 14 August 2014.
Citation: Marfil R, Palomino AJ and Bandera A (2014) Combining segmentation and attention: a new foveal attention model. Front. Comput. Neurosci. ***8****:96. doi: 10.3389/fncom.2014.00096*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Marfil, Palomino and Bandera. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 09 September 2014
doi: 10.3389/fncom.2014.00109

Object recognition with hierarchical discriminant saliency networks

***Sunhyoung Han*[1]* and *Nuno Vasconcelos*[2]**

[1] *Analytics Department, ID Analytics, San Diego, CA, USA*
[2] *Statistical and Visual Computing Lab, Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA*

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Neil Bruce, University of Manitoba, Canada
Seyed-Mahdi Khaligh-Razavi, University of Cambridge, UK

***Correspondence:**
Sunhyoung Han, Analytics Department, ID Analytics, 15253 Avenue of Science, San Diego, CA 92128, USA
e-mail: shan@idanalytics.com

The benefits of integrating attention and object recognition are investigated. While attention is frequently modeled as a pre-processor for recognition, we investigate the hypothesis that attention is an intrinsic component of recognition and vice-versa. This hypothesis is tested with a recognition model, the hierarchical discriminant saliency network (HDSN), whose layers are top-down saliency detectors, tuned for a visual class according to the principles of discriminant saliency. As a model of neural computation, the HDSN has two possible implementations. In a biologically plausible implementation, all layers comply with the standard neurophysiological model of visual cortex, with sub-layers of simple and complex units that implement a combination of filtering, divisive normalization, pooling, and non-linearities. In a convolutional neural network implementation, all layers are convolutional and implement a combination of filtering, rectification, and pooling. The rectification is performed with a parametric extension of the now popular rectified linear units (ReLUs), whose parameters can be tuned for the detection of target object classes. This enables a number of functional enhancements over neural network models that lack a connection to saliency, including optimal feature denoising mechanisms for recognition, modulation of saliency responses by the discriminant power of the underlying features, and the ability to detect both feature presence and absence. In either implementation, each layer has a precise statistical interpretation, and all parameters are tuned by statistical learning. Each saliency detection layer learns more discriminant saliency templates than its predecessors and higher layers have larger pooling fields. This enables the HDSN to simultaneously achieve high selectivity to target object classes and invariance. The performance of the network in saliency and object recognition tasks is compared to those of models from the biological and computer vision literatures. This demonstrates benefits for all the functional enhancements of the HDSN, the class tuning inherent to discriminant saliency, and saliency layers based on templates of increasing target selectivity and invariance. Altogether, these experiments suggest that there are non-trivial benefits in integrating attention and recognition.

Keywords: object recognition, object detection, top-down saliency, discriminant saliency, hierarchical network

1. INTRODUCTION

Recent research in computational neuroscience has enabled significant advances in the modeling of object recognition in visual cortex. These advances are encoded in recent object recognition models, such as HMAX (Riesenhuber and Poggio, 1999; Serre et al., 2007; Mutch and Lowe, 2008) the convolutional networks of Pinto et al. (2008); Jarrett et al. (2009) and a number of deep learning models (Hinton et al., 2006; Krizhevsky et al., 2012). When compared to classical sigmoid networks (LeCun et al., 1990, 1998), these models reflect an improved understanding of the neurophysiology of visual cortex (Graham, 2011), recently summarized by the standard neurophysiological model of Carandini et al. (2005). This consists of hierarchical layers of simple and complex cells (Hubel and Wiesel, 1962). Simple cells implement a combination of filtering, rectification, divisive contrast normalization, and sigmoidal non-linearity, which makes them *selective* to certain visual features, e.g., orientation. Complex cells pool information from multiple simple cells, producing an *invariant* representation. While the receptive fields of cells at the lower hierarchical levels resemble Gabor filters of limited spatial extent, cells at the higher layers have much more complex receptive fields, and pool information from larger regions of support (Poggio and Edelman, 1990; Perrett and Oram, 1993). This makes them more *selective* and *invariant* than their low-level counterparts. Extensive experiments have shown that accounting for simple and complex cells (Serre et al., 2007), using normalization and rectification (Jarrett et al., 2009), optimizing the sequence of these operations (Pinto et al., 2009), or learning deep

networks with multiple layers (Krizhevsky et al., 2012) can be highly beneficial in terms of recognition performance.

There are, nevertheless many aspects of cortical processing that remain poorly understood. In this work, we consider the role of attention in object recognition, namely how attention and recognition can be integrated in a shared computational architecture. We consider, in particular, the saliency circuits that drive the attention system. These circuits are usually classified as either bottom-up or top-down. Bottom-up mechanisms are stimulus driven, driving attention to image regions of conspicuous stimuli. Many computational models of bottom-up saliency have been proposed in the literature. They equate saliency to center-surround operations (Itti et al., 1998; Gao and Vasconcelos, 2009), frequency analysis (Hou and Zhang, 2007; Guo et al., 2008), or stimuli with specific properties, e.g., low-probability (Rosenholtz, 1999; Bruce and Tsotsos, 2006; Zhang et al., 2008), high entropy (Kadir and Brady, 2001), or high complexity (Sebe and Lew, 2003). An extensive review of bottom-saliency models is available in Borji and Itti (2013) and an experimental comparison of their ability to predict human eye fixations in Borji et al. (2013). While these mechanisms can speed up object recognition (Miau and Schmid, 2001; Walther and Koch, 2006), by avoiding an exhaustive scan of the visual scene, they are not intrinsically connected to any recognition task. Instead, bottom-up saliency is mostly a pre-processor of the visual stimulus, driving attention to regions that are likely to be of general vision interest. On the other hand, top-down saliency mechanism are task-dependent, and emphasize the visual features that are most informative for a given visual task. These mechanisms assign different degrees of saliency to different components of a scene, depending on the recognition task to be performed. For example, it is well known since the early studies of Yarbus (1967) that, when subjects are asked to search for different objects in a scene, their eye fixation patterns can vary significantly. It has also long been known that attention has a feature based component. More precisely, human saliency judgments can be manipulated by enhancement or inhibition of the feature channels of early vision, e.g., color or orientation (Maunsell and Treue, 2006). This type of feature selection should, in principle, be useful for recognition.

Overall, there are several reasons to study the integration of recognition and top-down saliency. First, the ability to simultaneously achieve selectivity and invariance is a critical requirement of robust image representations for recognition. By increasing the selectivity of neural circuits to certain classes of stimuli, the addition of top-down saliency, which increases selectivity to the object classes of interest, could potentially improve recognition performance. Second, there is some evidence that adding an attention mechanism to computational models of object recognition can improve their performance. For example, spatially selective units are known to substantially improve HMAX performance (Mutch and Lowe, 2008). In fact, as we will show later in this work, some of the recent object recognition advances in computer vision, such as the now widely used SIFT descriptor, can be interpreted as saliency mechanisms. Although these are purely stimulus driven, i.e., bottom-up, the gains with which they are credited again suggest that saliency has a role to play in recognition. Third, the connection to saliency provides the intermediate layers of a recognition network with a functional justification. Rather than a side effect of a holistic network optimization with respect to a global recognition criterion, they become individual saliency detectors, each attempting to improve on the saliency detection performance of their predecessors. This has a simpler evolutionary justification, under which (1) visual systems would evolve one layer at a time and (2) the search for improved performance in attention tasks leads naturally to object recognition networks.

All these observations suggest the hypothesis that, rather than a simple bottom-up pre-processor that determines conspicuous locations to be sequentially analyzed by the visual system, saliency could be embedded in object recognition circuits. Our previous work has also shown that, under the discriminant saliency principle, the computations of saliency can be mapped to the standard neurophysiological model (Gao et al., 2008; Gao and Vasconcelos, 2009). While we have exploited this mapping extensively for modeling bottom-up saliency, the underlying computations can be naturally extended to top-down saliency. In fact, under this extension, the saliency operation boils down to the discrimination between an object class and the class of natural images. This is intrinsically connected to object recognition. It, thus, appears that biology could have chosen to embed saliency in the recognition circuitry, if this had an evolutionary benefit, i.e., if embedding saliency in object recognition networks improves recognition performance. One of the goals of this work is to investigate this question. For this, we propose a family of *hierarchical discriminant saliency networks (HDSNs),* which jointly implement attention and recognition. More precisely, HDSNs are networks whose layers implement top-down saliency detection, based on features of increasing selectivity and invariance. These layers are stacked, so as to enhance the saliency detection of their predecessors. Since higher layers become more selective for the target objects, object recognition should be enhanced as a by-product of the saliency computation. All saliency detectors are derived from the discriminant saliency principle of Gao and Vasconcelos (2009) and explicitly minimize recognition error, using the top-down saliency measure of Gao et al. (2009). This is implemented with the biologically plausible computations of Gao and Vasconcelos (2009). In this way, HDSNs are consistent with the standard neurophysiological model (Carandini et al., 2005), but have a precise computational justification, and a statistical interpretation for all network computations. All parameters can thus be tuned by statistical learning, enabling the explicit optimization of the network for recognition.

A number of properties of HDSNs are investigated in this work. We start by showing that HSDNs can be implemented in multiple ways. In addition to the biologically plausible implementation, they can be interpreted as an extension of convolutional neural network models commonly used for recognition. This extension consists of a new type of rectifier function, which is a generalization of the recently popular rectified linear unit (ReLU) (Nair and Hinton, 2010; Krizhevsky et al., 2012). The generalization is parametric and can be tuned according to the statistics of the object classes of interest. This tuning enables the network to implement behaviors, such as switching from selectivity to feature presence to selectivity to feature absence, that are not possible with the units in common use. The computation

of saliency also enables the network to learn more discriminant receptive fields. In result, receptive fields at the higher network layers become tuned for configurations of salient low-level features, improving both saliency and object recognition performance. Overall, HDSNs are shown to exhibit the ability to model both salient features and their configurations, to replicate the human ability to identify objects due to both feature presence and absence, to modulate saliency responses according to the discriminant power of the underlying features, and to implement optimal feature denoising for recognition. The introduction of HDSNs is complemented by the analysis of several recognition methods from computer vision (Vasconcelos and Lippman, 2000; Lazebnik et al., 2006; Zhang et al., 2007; Boiman et al., 2008; Yang et al., 2009; Zhou et al., 2009), which are mapped to a canonical architecture with many of the attributes of the biological models. This enables a clear comparison of methods from the two literatures. A rigorously controlled investigation, involving models from both computational neuroscience and computer vision, shows that there are recognition benefits to both the class-tuning inherent to discriminant saliency and the hierarchical organization of the HDSN into saliency layers of increased target selectivity and invariance. Experiments on standard visual recognition datasets, as well as a challenging dataset for saliency, involving the detection of panda bears in a cluttered habitat, show that these advantages can translate into significant gains for object detection, localization, recognition, and scene classification.

2. METHODS

We start with a brief review of discriminant saliency.

2.1. DISCRIMINANT SALIENCY

Discriminant saliency is derived from two main principles: that (1) neurons are optimal decision-making devices and (2) optimality is tuned to the statistics of natural visual stimuli. The visual stimulus is first projected into the receptive field of a neuron, through a *linear transformation* $\mathcal{T}$, which produces a *feature response* X. The neuron then attempts to classify the stimulus as either belonging to a *target* or *background* (also denoted *null*) class. The definitions of target and background class define the saliency operation. For bottom-up saliency, they are the feature responses in a pair of center (target) and surround (background) windows co-located with the receptive field (Gao et al., 2008; Gao and Vasconcelos, 2009). In this work we consider the problem of top-down saliency, where the target class is defined by the feature responses to a stimulus class of interest and the background class by the feature responses to the class of natural images (Gao et al., 2009). In the object recognition context, the stimulus class of interest is a class of objects. Neurons implement the optimal decision rule for stimulus classification in the minimum probability of error (MPE) sense (Duda et al., 2001; Vasconcelos, 2004a). Saliency is then formulated as the discriminability of the visual stimulus with respect to this classification. Stimuli that can be easily assigned to the target class are denoted salient, otherwise they are not salient. The discriminability score used to measure stimulus saliency is computed in two steps, implemented by two classes of neurons that comply with the classical grouping into simple and complex cells. Simple cells first compute the optimal decision rule for stimulus classification into target and background, at each location of the visual field. Complex cells then combine simple cell outputs to produce a discriminability score.

2.1.1. Statistical model

Consider a simple cell, whose receptive field is centered at location l of the visual field. The visual stimulus at l is drawn from class $Y(l)$, where $Y(l) = 1$ for target and $Y(l) = 0$ for background. The goal of the cell is to determine $Y(l)$. For this, it applies a linear transformation $\mathcal{T}$ to the stimulus in a neighborhood of l (the receptive field of the cell) , producing a feature response $X(l)$ at that location. The details of the transformation are not critical, the only constraint is that it is a bandpass transformation. Using the well know-fact that bandpass feature responses to natural images follow the generalized Gaussian distribution (GGD) (Buccigrossi and Simoncelli, 1999; Huang and Mumford, 1999; Do and Vetterli, 2002), the feature distributions for target and background are

$$P_{X|Y}(x(l)|i) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-\left(\frac{|x(l)|}{\alpha_i}\right)^{\beta}} \qquad i \in \{0, 1\}. \tag{1}$$

The parameters α_i are the *scales* (variances) of the two distributions, while β is a parameter that determines their shape. For natural imagery, β is remarkably consistent, taking values around 0.5 (Srivastava et al., 2003). This value is assumed in the remainder of this work. The scales α_i are learned from two training samples R_1, R_o of examples from target and null class, respectively, by maximum a posteriori (MAP) estimation, using a conjugate Gamma prior of hyper-parameters η, ν. As described in Gao and Vasconcelos (2009) the MAP estimates of α_1 and α_0 are

$$\alpha_i^{\beta} = \frac{1}{\kappa} \sum_{x_j \in R_i} |x_j|^{\beta} + \nu, \quad i \in \{0, 1\}, \quad \kappa = \frac{n + \eta}{\beta}. \tag{2}$$

The values of the prior parameters are not critical. They are used mostly to guarantee that the estimates of α_i are non-zero. In this work, we use $\eta = 1$ and $\nu = 10^{-3}$.

2.1.2. Saliency measure

A simple cell uses the above model of natural image statistics to compute the posterior probability of the target class, given the observed feature response $X(l)$

$$P_{Y|X}(1|x(l)) = \sigma\left(g[x(l)]\right), \tag{3}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function and $g(x)$ the log-likelihood ratio (LLR)

$$g(x) = \log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} = \left(\frac{|x|}{\alpha_0}\right)^{\beta} - \left(\frac{|x|}{\alpha_1}\right)^{\beta} + T, \tag{4}$$

with $T = \log\left(\frac{\alpha_0}{\alpha_1}\right)$. Simple cells are organized into convolutional layers, which repeat the simple cell computation at each location of the visual field. Each layer produces a retinotopic map

of posterior probabilities $P_{Y|X}(1|x(l))$ given the feature responses derived from a common transformation $\mathcal{T}$. The computation is repeated for various transformations $\mathcal{T}_i$, producing several *channels* of simple cell response. As illustrated in the left of **Figure 1**, these channels are computed at multiple resolutions, by applying each transformation to re-scaled replicas of the visual stimulus. In our implementation, we use 10 scales, with subsampling factors of $2^{i/4}$, $i \in \{0, \ldots, 9\}$.

The saliency of the stimulus at location l is evaluated by a complex cell that combines the responses of afferent simple cell responses in a neighborhood $N(l)$ (its pooling neighborhood) into the discriminability score

$$S(l) = E_{X(l)}\left(\lfloor g(X) \rfloor_+\right), \tag{5}$$

where $E_{X(l)}$ denotes the expectation with respect to the distribution of X in $N(l)$ and $\lfloor x \rfloor_+ = \max(x, 0)$ is the half-wave rectification function. This rectification assures that the score is non-negative, by zeroing the LLR $g(x)$ at all locations where the outcome of the Bayes decision rule for MPE classification

$$\hat{Y}(l) = \begin{cases} 1, & \text{if } g[x(l)] \geq 0 \\ 0, & \text{if } g[x(l)] < 0 \end{cases} \tag{6}$$

assigns the response to the background class (i.e., chooses $\hat{Y}(l) = 0$). Large values of the score $S(l)$ indicate that the feature response $X(l)$ can be clearly assigned to the target class, i.e., the LLR $g(x)$ is both positive and large. For such stimuli, the posterior probability of (3) is close to one. In this case, the visual stimulus is salient. Small scores indicate that this is not the case. The computation of the saliency score of (5) is implemented by replacing the expectation with a sample average over $N(l)$

$$S(l) = \frac{1}{|N(l)|} \sum_{j \in N(l)} \lfloor g[x(j)] \rfloor_+. \tag{7}$$

This is computed as a combination of the responses of simple cells in $N(l)$, since (7) can be written as Han and Vasconcelos (2010)

$$S(l) = \frac{1}{|N(l)|} \sum_{j \in N(l)} \xi\{P_{Y|X}[1|x(j)]\} \tag{8}$$

with

$$\xi(x) = \begin{cases} \frac{1}{2} \log \frac{x}{1-x}, & x \geq .5 \\ 0, & x < .5. \end{cases}$$

Hence, a complex cell applies the non-linear transformation $\xi(x)$ to the responses of the afferent simple cells and pools the transformed responses into the saliency measure $S(l)$. The neighborhood $N(l)$ is thus denoted as the *pooling neighborhood* of the complex cell. Like simple cells, the complex cell computation is

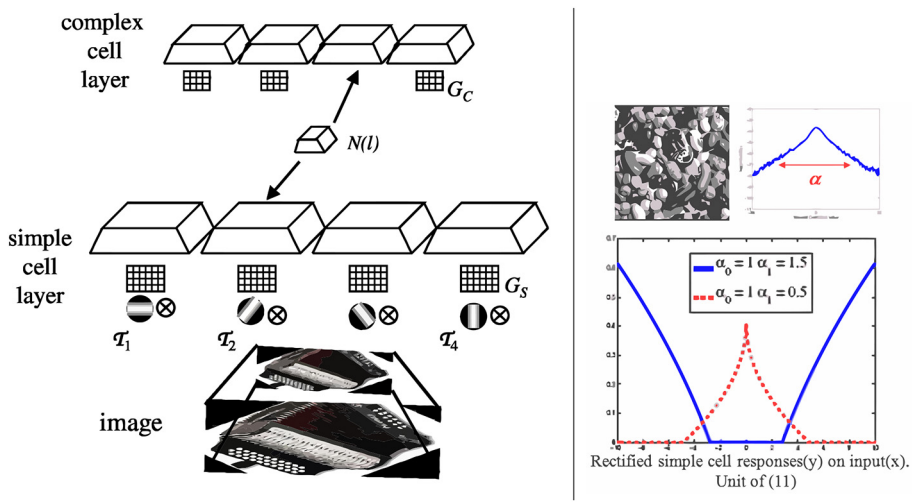


FIGURE 1 | Left: saliency is computed by a pair of layers of simple and complex cells. In the simple cell layer, the visual stimulus is first subject to a number of linear transformations, which are repeated at various image scales, illustrated by chopped pyramids. In the example of the figure, the set of transformations consist of four oriented filters $\mathcal{T}_i$. Each simple cell computes the optimal decision rule for the classification of the filter response at one scale and location of a simple cell grid G_S. A channel consists of all retinotopic maps of simple cell response derived from a common transformation (4 channels in the figure). A complex cell computes the saliency score of (7), using a pooling neighborhood $N(l)$ that spans locations and scales. The retinotopic maps of complex cell response are in one to one correspondence to those of simple cell response, but the grid G_C of complex cell locations is a subsampled replica of its simple cell counterpart. The simple and complex cell computations can be implemented in two ways. In a biologically plausible implementation, simple cells compute the posterior probabilities of (3), while complex cells implement the pooling operator of (8). In an artificial neural network implementation, simple cells implement the parametric ReLU units of (11), while complex cells perform simple averaging. **Right**: the top inset shows the histogram of responses of a bandpass filter to the natural image on the left. The scale parameter α characterizes the spread of the distribution and is large (small) for filters that match (do not match) structures in the image, i.e., features that are "present" ("absent"). The plot in the bottom shows the function of (11) for different values of α_i. The behavior of the parametric ReLU can change from the detection of feature presence to the detection of feature absence, depending on the scales of the target and background GGD distributions. The curve in red (blue) corresponds to a feature present (absent) in the target but absent (present) in the null class.

replicated at a grid of locations G_C (usually a subset of the simple cell grid G_S) to produce a retinotopic channel of saliency response. Each channel is associated with a common feature transformation $\mathcal{T}$, i.e., complex cells only combine the responses of simple cells of common transformation $\mathcal{T}$. As illustrated in the left of **Figure 1**, the number of channels of complex cell response is identical to that of simple cell response.

2.2. SALIENCY DETECTOR IMPLEMENTATIONS

The saliency measure of (5) can be implemented in three different ways, which are of interest for different applications of the saliency model.

2.2.1. Biologically plausible implementation

The saliency computations can be mapped into a network that replicates the standard neurophysiological model of visual cortex (Carandini et al., 2005). In biology, rather than the static analysis of a single image, recognition is usually combined with object tracking or some other dynamic visual process. In this case, saliency is not strictly a feedforward computation. In particular, the training sets R_i of (2), used to learn the GGD parameters of a cell, are composed by responses of other cells, i.e., the target and background classes are defined by the lateral connections of a simple cell. An implementation of object tracking, by continuously adaptive recognition of the objects to track, using this type of mechanism is presented in Mahadevan and Vasconcelos (2013). In this implementation, the lateral connections are organized in a center surround manner, defining (1) the target class as the visual stimulus in a window containing the object to track and (2) the background class as the stimulus in a surrounding window. Under this type of implementation, a simple cell computes the LLR $g[x(l)]$ by combining (4) and (2) into the divisive normalization operation

$$g[x(l)] = \frac{|x(l)|^\beta}{\frac{1}{\kappa}\sum_{j\in R_0}|x(j)|^\beta + \nu} - \frac{|x(l)|^\beta}{\frac{1}{\kappa}\sum_{j\in R_1}|x(j)|^\beta + \nu} + T, \quad (9)$$

characteristic of simple cell computations (Heeger, 1992; Carandini et al., 1997, 2005). The LLR is then transformed into the posterior probability of (3) by application of a sigmoid transformation to the divisively normalized responses. An illustration of the simple cell computations is given in **Figure 2A**. Complex cells then implement the computations of (8), as illustrated in **Figure 2B**. When equipped with these units, the network of **Figure 1** has a one to one mapping with the standard neurophysiologic model of the visual cortex (Carandini et al., 2005).

2.2.2. Neural network implementation

Neural networks are commonly used to solve the computer vision problem of object recognition. In this setting, network parameters are learned during a training stage, after which the network operates in a feedforward manner. For these type of applications, the GGD parameters of (4) can be learned from a training set, using (2), and kept constant during the recognition process. This allows the simplification of the saliency operations. Namely, by combining (7) and (4) it follows that

$$S(l) = \frac{1}{N(l)}\sum_{j\in N(l)} \lfloor \gamma |x(j)|^\beta + T \rfloor_+ \quad (10)$$

where $\gamma = \left(\frac{1}{\alpha_0^\beta} - \frac{1}{\alpha_1^\beta}\right)$, and $T = \log\frac{\alpha_1}{\alpha_0}$. This can again be mapped to the two layer network of **Figure 1**, but simple cells now simply rectify feature responses, according to

$$\psi(x) = \lfloor \gamma |x|^\beta - T \rfloor_+, \quad (11)$$

while complex cells perform a simple average pooling operation. The resulting network is similar to the stages of rectifier linear units (ReLU) that have recently become popular in the deep learning literature (Nair and Hinton, 2010; Krizhevsky et al., 2012). When compared to the ReLU computation, $f(x) = \lfloor x \rfloor_+$, the parametric rectifier of (11) replaces static rectification by an adaptive rectification, tuned to the scales α_i of the feature distributions under target and background hypotheses.

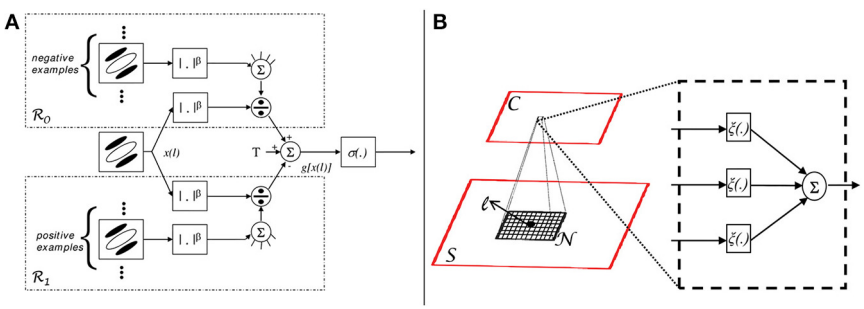


FIGURE 2 | Discriminant saliency computations. (A) Simple cell (S unit). A unit of receptive field centered at location l computes a feature response $x(l)$. This is then rectified, differentially divisively normalized by feature responses from areas R_0 and R_1, and fed to a sigmoid. The responses from the two areas act as training sets for the binary classification of $x(l)$. More precisely, responses in R_0 (R_1) act as training examples for the negative (positive) class. The output $g[x(l)]$ of the differential divisive normalization operator is the log-likelihood ration for the classification of $x(l)$ with respect to the two classes (under the assumption of GGD statistics), as in (9). The sigmoid finally transforms this ratio into the posterior probability of the positive class, as in (3). **(B)** Complex cell (C unit). The bottom plane symbolizes the output of a layer of S-units, the top one the output of a layer of C-units. S-unit responses within a neighborhood $N(l)$ are passed through non-linearity $\xi(x)$ and pooled additively, to produce the response of a C unit. This implements the saliency measure of (8).

This adaptation is illustrated in the right side of **Figure 1**. When $\alpha_1 = \alpha_0$, target and null distributions are identical and $\psi(x) = 0$ for all x. Hence, non-informative features for target detection are totally inhibited. When $\alpha_1 > \alpha_0$, the target distribution has heavier tails than the null distribution, i.e., the feature is *present* in the target. In this case (blue curve), the rectifier enhances large responses and inhibits small ones, acting as a detector of feature presence. Conversely, the null hypothesis has heavier tails when $\alpha_1 < \alpha_0$, i.e., when the feature is *absent* from the target. In this case (red dashed curve), the rectifier enhances small responses and inhibits large ones, acting as a detector of feature absence. In summary, the rectification introduced by the simple cells of (11) varies with a measure of discrimination of the feature X, based on the parameters γ and T. In result, the cell responses adapt to the feature distributions under the two hypotheses, allowing simple cells to have very diverse responses for different features. This is beyond the reach of the conventional ReLU rectifier. The adaptive behavior of $\psi(x)$ is also reminiscent of optimal rules for image denoising (Chang et al., 2000). Like these rules, it thresholds the feature response, exhibiting a dead-zone (region of zero output) which depends on the feature type. Note that this results from (8), which is the Bayes decision rule for classification of the response $x(l)$ into target and background. Hence, $\psi(x)$ can be seen as an optimal feature denoising operator for the detection of targets embedded in clutter. The dead-zone depends on the relative scales of target and background distribution, according to

$$\begin{array}{l} |x|^\beta \leq T/\gamma \text{ when } \alpha_1 > \alpha_0 \\ |x|^\beta \geq T/\gamma \text{ when } \alpha_1 < \alpha_0. \end{array} \tag{12}$$

2.2.3. Algorithmic implementation

It is also possible to compute the discriminant saliency measure with an algorithm that has little resemblance to any biological computation but provides insight into the saliency score. This follows from rewriting (5) as

$$\begin{aligned} S(l) &= \sum_{i=0}^{1} E_{X(l)|Y(l)} \left(\lfloor g(X) \rfloor_+ | i \right) P_{Y(l)}(i) \\ &= E_{X(l)|Y(l)} \left(\lfloor g(X) \rfloor_+ | 1 \right) P_{Y(l)}(1) \propto E_{X(l)|Y(l)} \left(g(X) | 1 \right) \\ &= \int_{N(l)} P_{X|Y}(x|1) \log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} dx, \end{aligned}$$

where we have used the fact that $\lfloor g(x(l)) \rfloor_+ = 0$ whenever $Y(l) = 0$ and $g(x(l)) \geq 0$ otherwise. Hence, the saliency score can be interpreted as the computation, over the neighborhood $N(l)$, of the Kullback-Leibler (KL) divergence between the probability distributions of the feature responses under the target and background distributions. Since the KL divergence is a well-known measure of distance between probability distributions, this confirms the discriminant nature of the saliency measure. Using (4), the KL divergence can be written as

$$S(l) \propto E_{X(l)|Y(l)}[|x|^\beta|1] \left(\frac{1}{\alpha_0^\beta} - \frac{1}{\alpha_1^\beta} \right) + T \tag{13}$$

$$\propto \frac{\gamma}{\beta} \alpha_1^\beta(l) + T, \tag{14}$$

where $\alpha_1^\beta(l)$ is the scale parameter of a GGD distribution with the responses observed in $N(l)$. This enables a very simple computation of the saliency measure, using the following procedure.

(1) From the feature responses $x_i(j)$ in the neighborhood $N(l)$ estimate $\alpha_1^\beta(l)$, using (2).
(2) Use (14) with $\alpha_1^\beta(l)$ and the model parameters α_i^β learned from the training samples R_i to compute the saliency score $S(l)$.

2.2.4. Discussion on different implementations

The three implementations above are equivalent, in the sense that they produce similar results on a given saliency task. They are suitable for different applications of the saliency measure of (5). In general, any model of biological computation has several implementations. For example, the convolution $y(l)$ of a visual stimulus $x(l)$ with a linear filter $h(l)$ can be computed in at least two ways: (1) the classical convolution formula

$$y(l) = \sum_k x(k) h(k - l) \tag{15}$$

or (2) the response to the stimulus $x(l)$ of a convolutional neural network layer (Fukushima, 1980; LeCun et al., 1998) of linear units with identical weights, derived from the filter $h(l)$. In this case, each network unit computes the output $y(l)$ for a particular value of l. We refer to the first as the mathematical implementation and to the second as the biological implementation. While any biologically plausible network has an equivalent mathematical implementation, it is generally not true that all mathematical formulas can be implemented with biological circuits. Even when this is possible, the implementation may occur at different levels of abstraction. In general, an algorithm is considered biologically plausible if it can be mapped to a realistic model of neural computations (mapping from neuron stimuli to responses). This does not mean that it actually simulates neurons at the molecular level. It should, however, be able to predict the behavior of the neuron in neuroscience experiments.

In the discussion above, the algorithmic implementation of Section 2.2.3 is a mathematical implementation of the proposed saliency measure. It does not explicitly define units or neurons and is most suitable for the implementation of the measure as a computer vision algorithm, in a standard sequential processor. On the other hand, because it does not make explicit the input-output relationship of any particular neuron, it is not of great interest as a model of neuroscience. The biologically plausible implementation of Section 2.2.1 has the reverse role. Because it is fully compliant with the standard neurophysiological model of the visual cortex (Carandini et al., 2005), it predicts a large set of non-linear neuron behaviors which this model has been documented to capture (Carandini and Heeger, 2011). It could,

thus, be used to study the role of these behaviors in object recognition. On the other hand, because it explicitly implements the computations of each neuron, its implementation on a sequential processor is much slower than the mathematical implementation of Section 2.2.3. Hence, it makes little sense to adopt it if the goal is simply to produce an efficient computer vision system. Finally, the neural network implementation of Section 2.2.2 is somewhere in between. It is a more abstract implementation than that of Section 2.2.1, in the sense that it does not explicitly include operations like divisive normalization. This makes it faster to compute and establishes a connection to recent models in the deep learning literature (Krizhevsky et al., 2012), which have been shown to achieve impressive object recognition results. These models can also be efficiently implemented in a GPU computer architecture, but are much slower on a traditional processor. Since this implementation achieves the best trade-off between fidelity to the neural computations and speed, we adopt it in the remainder of the paper. In particular, a CPU-based implementation of the neural network of Section 2.2.2 was used in all experiments of Section 4.

2.3. HIERARCHICAL DISCRIMINANT SALIENCY NETWORKS

A *hierarchical discriminant saliency network* (HDSN) is a neural network whose layers are implemented by the saliency detector of **Figure 1**.

2.3.1. HDSN architecture

The architecture of the HDSN is illustrated in **Figure 3**, for a two layer network. In general, a HDSN has M layers. As in **Figure 1**, layer m has two sub-layers: $S^{(m)}$ of S units (simple cells) and $C^{(m)}$ of C units (complex cells). S-units are located in a coordinate grid $G_S^{(m)}$, C-units in a coordinate grid $G_C^{(m)}$. Each sub-layer is organized into C *channels*. Channel c is based on the convolution of the layer input with a *template*, $\mathcal{T}_c^{(m)}$, shared by all its units. The processing of each channel is repeated at $R^{(m)}$ image resolutions. The network of **Figure 3**, has $C^{(1)} = 4$ channels in layer 1 and $C^{(2)} = N$ in layer 2.

Let $y^{(0)}$ be the network input, and $y_c^{(m-1)}$ the output of cth channel of layer $m-1$. At layer m, $y^{(m-1)}$ is first contrast normalized

$$\bar{y}_c(l) = \frac{y_c^{(m-1)}(l)}{\sum_{j \in Z(l)} \sum_i y_i^{(m-1)}(j)} \tag{16}$$

where $Z(l)$ is a window, centered at l, with the size of template $\mathcal{T}_c^{(m)}$. The normalized input is then processed by the sub-layer of S-units, which first convolves it with the filters $\mathcal{T}_c^{(m)}$. This produces feature responses $x_c^{(m)}(l)$, which are then sampled at S-unit locations $G_s^{(m)}$, and rectified by the parametric ReLU of (11),

$$\psi_c^{(m)}(x) = \left\lfloor \gamma_c^{(m)} |x|^{\beta} - T_c^{(m)}, \right\rfloor_+, \tag{17}$$

with parameters

$$\gamma_c^{(m)} = \left(\frac{1}{\left(\alpha_{c,0}^{(m)}\right)^{\beta}} - \frac{1}{\left(\alpha_{c,1}^{(m)}\right)^{\beta}} \right) \quad T_c^{(m)} = \log \frac{\alpha_{c,1}^{(m)}}{\alpha_{c,0}^{(m)}}. \tag{18}$$

The rectified filter responses are then fed to the sub-layer of C-units. Each C-unit computes the saliency score of (7) by simple averaging over its pooling window, i.e.,

$$y_c^{(m)}(l') = S_c^{(m)}(l') = \frac{1}{|N^{(m)}(l')|} \sum_{l \in N^{(m)}(l')} \psi_c^{(m)}\left(x_c^{(m)}(l)\right) \tag{19}$$

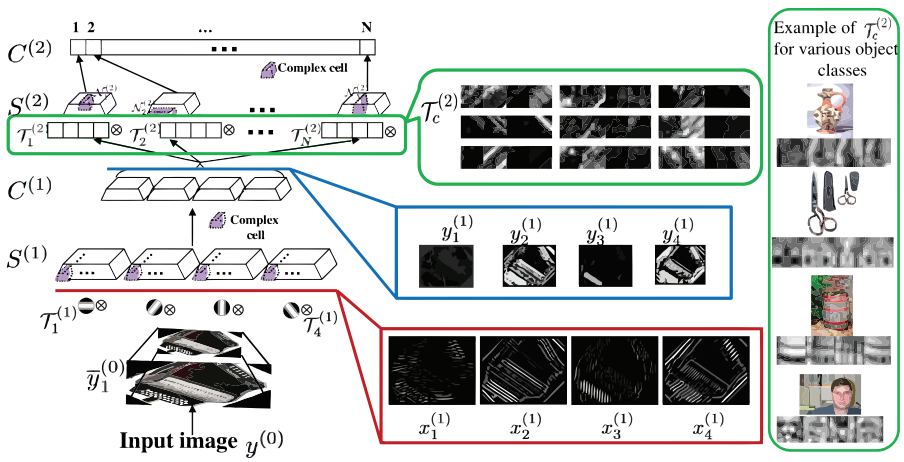


FIGURE 3 | Left: HDSN with two layers. Each layer consists of a DSN, as in **Figure 1**. Layer i contains a sub-layer of simple ($S^{(i)}$) and a sub-layer of complex ($C^{(i)}$) units. The network has 4 channels in layer 1 and N in layer 2. Channel c is obtained by convolving the input of a layer with a template $\mathcal{T}_c$, at several resolutions. Templates $\mathcal{T}_c^{(1)}$ of layer 1 are Gabor filters, templates $\mathcal{T}_c^{(2)}$ of layer 2 are learned during training. **Center**: Gabor channels $x_c^{(1)}$ derived from the input image, corresponding saliency channels $y_c^{(1)}$ at the output of the first network layer, and example saliency templates $\mathcal{T}_c^{(2)}$ learned by the second layer. **Right**: most discriminant template learned for each of four classes of Caltech101 (an example image is also shown for each class). Note that each template is composed of four image patches, derived from the four channels of the image representation in the first network layer.

The cth channel of this representation is the saliency map with respect to template $\mathcal{T}_c^{(m)}$ and the cth channel of the output of layer m. The locations l' are defined by the C-unit grid $G_C^{(m)}$. The pooling neighborhood $N(l')$ is usually smaller than the output of the afferent S sub-layer. Hence, both S and C-units have limited spatial support. However, $N^{(m)}(l')$ can be location adaptive, i.e., depend on l'.

2.3.2. Learning

The training of a HDSN consists of learning the templates $\mathcal{T}_c^{(m)}$ and the GGD scales $\alpha_{c,0}^{(m)}, \alpha_{c,1}^{(m)}$ per layer m. Many approaches are possible to learn the templates $\mathcal{T}_c^{(2)}$, including the back-propagation algorithm (LeCun et al., 1998), restricted Boltzmann machines (Hinton et al., 2006), clustering (Coates et al., 2011), multi-level sparse decompositions (Kavukcuoglu et al., 2010), etc. In this work, we adopt the simple procedure proposed for training the HMAX network in Serre et al. (2007); Mutch and Lowe (2008), where the templates $\mathcal{T}_c^{(m)}$ of layer m are randomly sampled patches from the responses $y_c^{(m-1)}$ of layer $m-1$, normalized to zero mean and unit norm. Given $\mathcal{T}_c^{(m)}$, the network is exposed to images from class $i \in \{0, 1\}$, and training samples $R_{c,i}^{(m)}$ collected. These consist of the responses $x_c^{(m)}(l)$ across locations l and training images from class i. The scale parameters are then computed with (2).

2.3.3. Object recognition

The HDSN is a hierarchical feature extractor, which maps the input image into a vector of responses of layer $C^{(M)}$. For object recognition, this vector is fed to a linear classifier. In our implementation this is a support vector machine (SVM). The network topology is characterized by the parameters $\Theta^{(m)} = \{R^{(m)}, G_S^{(m)}, G_C{}^{(m)}, \mathcal{T}^{(m)}, N^{(m)}\}$, $m \in \{1, \ldots, M\}$. As is usual in the hierarchical network literature, a good trade-off between object selectivity and invariance can be achieved by using (1) sparser grids $G_S^{(m)}$, $G_C{}^{(m)}$, (2) filters $\mathcal{T}^{(m)}$ of larger spatial support, and (3) larger pooling neighborhoods $N^{(m)}$, as m increases. This results in higher layer templates that are more selective for the target objects than those of the lower layers, without compromise of invariance. Since the selectivity-invariance trade-off of deep networks has been demonstrated by many prior works (Riesenhuber and Poggio, 1999; Serre et al., 2007; Krizhevsky et al., 2012), we do not discuss it here. In fact, the goal of this work was not to test the benefits of deep learning *per se*, which have now been amply demonstrated in the literature, but to investigate the benefits of augmenting the network with the saliency computations. Since, as we will see in the next section, many of the computer vision methods for object recognition can be mapped into two-layer networks, our study was limited to the network of **Figure 3**. This also had the advantage of enabling training from much smaller training sets.

In our implementation, $S^{(1)}$ units use the 11×11 Gabor filters proposed in Mutch and Lowe (2008),

$$\mathcal{T}_c^{(1)}(x, y) = \exp\left(-\frac{X^2 + \gamma^2 Y^2}{2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda} x\right) \qquad (20)$$

where $X = x\cos\theta_c - y\sin\theta_c$, $Y = x\sin\theta_c + y\cos\theta_c$, $\theta_c \in \{0, \pi/4, \pi/2, 3\pi/4\}$, and γ, σ, and λ are set to 0.3, 4.5, and 5.6, respectively. This makes the first layer a detector of characteristic edges of the target. The training samples $R_{c,i}^{(1)}$ for learning the scale parameters $\alpha_{c,i}^{(1)}$ are the set of Gabor responses $x_c^{(1)}$ to images of class i over the entire channel c. On the other hand, the templates of $S^{(2)}$, $\mathcal{T}_c^{(2)} = \{\mathcal{T}_{c,1}^{(2)}, \ldots, \mathcal{T}_{c,C^{(1)}}^{(2)}\}$, span the $C^{(1)}$ channels of the first layer, and are learned by random sampling, as discussed above. Since these templates are saliency patterns produced by layer 1 in response to the target, they are usually more complex features. The different complexity of the templates of the two layers warrants different pooling neighborhoods for C-units. Since simple features are homogeneous, layer 1 relies on a fixed neighborhood $N^{(1)}$. On the other hand, to accommodate the diversity of its complex features, layer 2 uses template specific pooling neighborhoods $N_c^{(2)}$. Templates $\mathcal{T}^{(2)}$ have dimension $n \times n \times 4$, for $n \in \{4, 8, 12, 16\}$, and are normalized to zero mean and unit norm (over the 4 channels). Pooling neighborhoods have area $S \in \{10, 20, 30\%\}$ of the size of layer 2 channels, and span $d \in \{3, 5, 7\}$ scales. Like the templates, they are sampled randomly. These neighborhoods are also used to collect the training samples $R_{c,i}^{(2)}$ for learning the scale parameters associated with each of the templates. The network configuration is summarized in **Table 1**.

Figure 3 illustrates the computations of the HDSN. It shows an image and the corresponding responses $x_c^{(1)}$ of the layer 1 Gabor

Table 1 | Configuration of the network used in all our experiments.

	$R^{(m)}$	$G_S^{(m)}$	$G_C{}^{(m)}$	$T^{(m)}$	$N^{(m)}$
$m = 1$	10 resolutions $r \in \{2^{i/4} \mid i = 0, \ldots, 9\}$	$1 \times 1 \times 1$	$3 \times 3 \times 1$ subsampling	Gabor filters of (20)	$5 \times 5 \times 2$ window
$m = 2$	same	$1 \times 1 \times 1$	Location and scale from which template is originally sampled	Randomly selected $n \times n \times 4$ templates, $n \in \{4, 8, 12, 16\}$, with zero mean and unit norm	S% of image area and depth d in scale, where $S \in \{10, 20, 30\}$, $d \in \{3, 5, 7\}$

Unless otherwise noted, $n \times m \times l$ means a spatial step of $n \times m$ and a step of l across scales.

filters, and $y_c^{(1)}$ of the layer 1 C-units. Note that, due to the class adaptive rectification of (17), the saliency responses $y_c^{(1)}$ amplify the filter responses $x_c^{(1)}$ of certain channels and inhibit the remaining. This allows the layer to produce a response that is more finely tuned to the discriminant features of the target class (in this example, the Caltech class "accordion"). Or, in other words, the layer highlights the features that are most distinctive of the target class. This, in turn, allows layer 2 to learn templates that are more discriminant of the target class than would be possible in the absence of the saliency computation. Note how the example templates $\mathcal{T}^{(2)}$ are selective for some of the feature channels. The inset on right of the figure presents the most discriminant template learned for four classes of Caltech101 (an example image of each of the classes is also shown). Note how the network has learned templates that are highly selective for the target objects. These templates are complex features (Vidal-Naquet and Ullman, 2003; Gao and Vasconcelos, 2005), which capture the spatial configuration of low-level features in target objects, resembling the receptive fields of cells in area IT (Riesenhuber and Poggio, 1999; Brincat and Connor, 2004; Yamane et al., 2008). Overall, while layer 1 processes edges, layer 2 captures shape information. When combined with the ability of the parametric ReLU rectifiers of (17) to behave as detectors of both feature presence and absence, this hierarchical learning of increasingly more selective templates enables the HDSN to compute saliency in challenging scenes. This is illustrated in **Figure 4**, using the pandaCam dataset, where background textures can be much more complex than the target object (panda bear). To be successful, the network must learn that the distinctive panda property is the absence of many of the features present in the background. The figure compares saliency maps produced by a HDSN with a single-layer (center column) and two layers (right column). Note how the latter produces saliency maps with less false positives and a much more precise localization of the target bears. The combination of (1) hierarchical learning of discriminant templates, and (2) detection of feature absence by parametric ReLUs, is critical for the network's effectiveness as a saliency detector.

3. RELATIONSHIPS TO RECOGNITION MODELS

In this section we compare HDSNs to previous object recognition models. We start by considering saliency models, then neural networks proposed for object recognition, and finally models from the computer vision literature.

3.1. SALIENCY MODELS

Many stimulus driven, bottom-up, saliency models have been proposed in the literature. They implement center-surround operations (Itti et al., 1998; Gao and Vasconcelos, 2009), frequency analysis (Hou and Zhang, 2007; Guo et al., 2008), or detect stimuli with specific properties, e.g., low-probability (Rosenholtz, 1999; Bruce and Tsotsos, 2006; Zhang et al., 2008), high entropy (Kadir and Brady, 2001), or high complexity (Sebe and Lew, 2003). These models cannot account for the well-known fact that, beyond the stimulus, saliency is

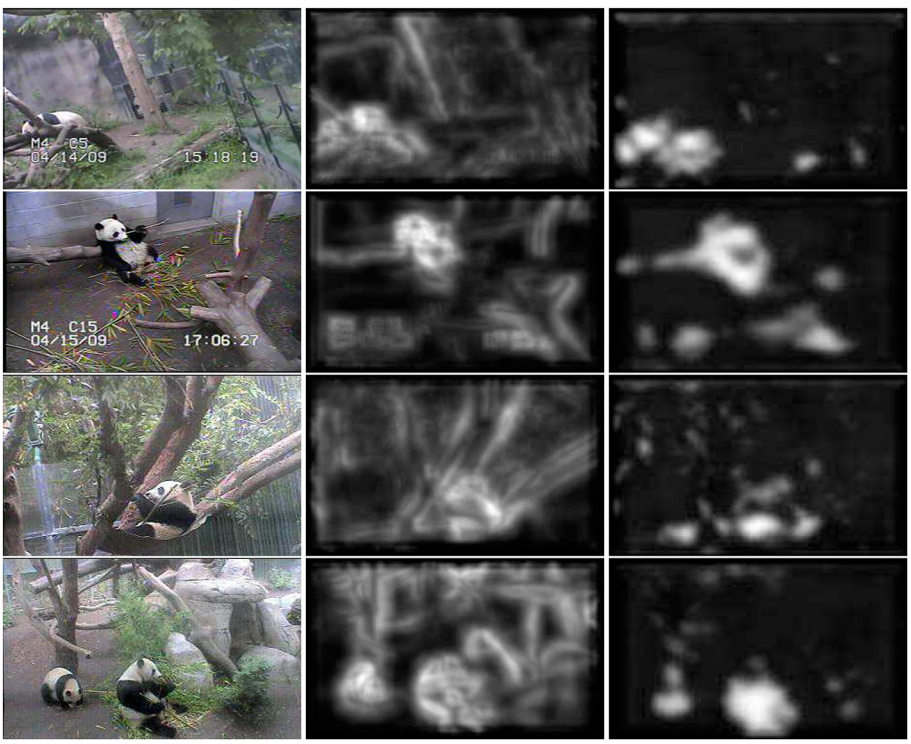


FIGURE 4 | Localization of panda bears in a complex environment. Left: bear images. Note the highly variable pose of the bears and the strongly textured backgrounds. **Center**: saliency maps produced by a single layer HSDN. **Right**: saliency maps produced by a two-layer HDSN. The ability of the second network layer to learn discriminant saliency patterns reduces the number of false positives and enables significantly superior target localization.

influenced by the task to be performed. For example, knowledge of target features increases the efficiency of visual search for a target among distractors (Tsotsos, 1991; Wolfe, 1998). This *top-down* component of saliency is classically modeled by modulating features responses (Treisman, 1985; Wolfe, 1994; Desimone and Duncan, 1995; Navalpakkam and Itti, 2007), i.e., global feature selection. This, however limits the ability to localize targets, since the selected filters respond to stimuli across the visual field. More recent top-down saliency models estimate distributions of feature response to target and background, and use them to derive optimal decision rules. These rules modulate feature responses spatially, according to the stimuli at different locations. A top-down saliency detector of this type is that of Elazary and Itti (2010). It differs from discriminant saliency through two simplifications: (1) assumption of Gaussian instead of generalized Gaussian responses ($\beta = 2$), and (2) use of the target log likelihood

$$S'(l) = \log P_{X_c^{(1)}|Y}(x_c^{(1)}(l)|1) \tag{21}$$

instead of (5), as saliency criterion (Elazary and Itti, 2010). In terms of the biological implementation discussed above, this corresponds to eliminating (1) C units, (2) the sigmoid $\sigma(x)$, and (3) the top divisive normalization branch (see **Figure 2A**) of S units. We refer to such S units as target likelihood (TL) units, and the resulting network as likelihood saliency network (LSN).

3.2. NEURAL NETWORKS FOR RECOGNITION

HDSNs have commonalities with many neural network models proposed for object recognition.

3.2.1. HMAX

Like the HDSN, the HMAX network follows the general architecture of **Figure 3** (Serre et al., 2007). $S^{(1)}$ units are Gabor filters, whose responses are pooled by $C^{(1)}$ units, using a maximum operator

$$y_c^{(1)}(l) = \max_{j \in N^{(1)}(l)} x_c^{(1)}(j), \tag{22}$$

where we again denote filter responses by $x_c^{(m)}(l)$ and pooling window by $N^{(m)}$. The $S^{(2)}$ sub-layer is a radial basis function (RBF) network with outputs

$$s_c^{(2)}(l) = \exp\left(-\beta \sum_i ||y_i^{(1)}(l) - \mathcal{T}_{c,i}^{(2)}||^2\right) \tag{23}$$

where β determines the sharpness of the RBF-unit tuning and $\mathcal{T}_c^{(2)}$ is a template. Similarly to the proposed implementation of the HDSN, these templates are randomly selected during training, and have as many components $\mathcal{T}_{c,i}^{(2)}$ as the number of layer 1 channels. $C^{(2)}$ units are again max-pooling operators

$$y_c^{(2)}(l) = \max_{j \in M^{(2)}} s_c^{(2)}(j), \tag{24}$$

where $M^{(2)}$ is the whole visual field. A number of improvements to the HMAX architecture have been proposed in Mutch and Lowe (2008): a lateral inhibition that emulates divisive normalization, the restriction of $M^{(2)}$ to template-specific neighborhoods [to increase localization of $C^{(2)}$ units], a single set of templates shared by all object classes, and a support vector-machine (SVM)-based feature selection mechanism to select the most discriminant subset.

3.2.2. Convolutional neural networks

Both the HDSN and the HMAX networks are members of the broader family of convolutional neural networks. These are again networks with the hierarchical structure of **Figure 3**, which date back to Fukushima's neocognitron (Fukushima, 1980). While early models lacked an explicit optimality criterion for training, convolutional networks trained by backpropagation became popular in the 1980s (LeCun et al., 1998). Classical models had no C units and their S units were composed uniquely of filtering and the sigmoid of (3). Recent extensions introduced S and C-like units per network layer (Pinto et al., 2008; Jarrett et al., 2009). While many variations are possible, modern S-units tend to include filtering, rectification, and contrast normalization. C-units then pool their responses. These extensions have significantly improved performance, sometimes producing staggering improvements. For example, Jarrett et al. (2009) reports that simply rectifying the output of each convolutional network unit drastically improves recognition accuracy. In fact, a network with random filters, but whose S-units include rectification and normalization, performs close to a network with extensively optimized filters. More recently, it has been shown that replacing the sigmoid of (3) by the ReLU nonlinearity $f(x) = \lfloor x \rfloor_+$ can significantly speed-up network training (Krizhevsky et al., 2012).

In this work, we consider in greater detail the network of Jarrett et al. (2009), which implements the most sophisticated S-units. The input of layer m is first convolved with a set of filters $\mathcal{T}_c^{(m)}$, producing feature responses $x_c^{(m)}$. These are then passed through a squashing non-linearity, absolute value rectification, subtractive, and divisive normalization, according to

$$a_c^{(m)}(l) = |g_c \tanh x_c^{(m)}(l)| \tag{25}$$

$$v_c^{(m)}(l) = a_c^{(m)}(l) - \sum_{c=1}^{C} \sum_{j \in M(l)} w(j) a_c^{(m)}(j) \sum_{j \in M(l)} w(j) = 1/C \tag{26}$$

$$u_c^{(m)}(l) = \frac{v_c^{(m)}(l)}{\max\left(\epsilon, \sum_{c=1}^{C} \sum_{j \in M(l)} w(j) \left(v_c^{(m)}\right)^2 (j)\right)}, \tag{27}$$

where $M(l)$ is a 9×9 window. The normalized responses are finally fed to a layer of C-units, which implement spatial pooling

$$y_c^{(m)}(l) = \sum_{j \in N(l)} u_c^{(m)}(j) \tag{28}$$

and subsampling. It is shown that unsupervised learning of the filters $\mathcal{T}_c^{(m)}$ is marginally better than adopting a random filter set, and relatively small gains result from global filter learning.

More recently, Krizhevsky et al. (2012) have shown that state of the art results on large scale recognition problems can be obtained with a deep network, whose layers are slightly simpler than those of Jarrett et al. (2009). This is a network of five convolutional and three fully connected layers. Its convolutional stages consist of a sub-layer of S-units, which implement a sequence of filtering, divisive normalization with (27) and ReLU rectification, and a sub-layer of C-units, which implement the max pooling operation of (22). The filters $\mathcal{T}_c^{(m)}$ are learned by back-propagation.

3.3. COMPUTER VISION MODELS

Many object recognition methods have been proposed in the computer vision literature. Over the last decade, there has been a convergence to a canonical architecture, consisting of three stages: descriptor extraction, descriptor encoding, and classification. While the classification stage is usually a linear SVM, many of the recent object recognition methods differ on the details of the first two stages (Chatfield et al., 2011). We next show that this architecture can be mapped to the network of **Figure 3**.

3.3.1. Canonical recognition architecture

Figure 5 shows the two-stage canonical architecture for object recognition in computer vision. The first stage transforms an image into a collection of descriptors, usually denoted a bag-of-features. The descriptors $y^{(1)}(l)$ are calculated at image locations l, e.g., per pixel, in a regular pixel grid (dense sampling), or at keypoint locations (Lowe, 1999). We assume dense sampling, which produces best results (Zhang et al., 2007) and is more widely used. Descriptors are high-dimensional vectors, obtained by application of spatially localized operators at each image location. If each descriptor dimension $y_c^{(1)}$ is used to define a channel of this representation, descriptor channels can be interpreted as the channels of $C^{(1)}$ output in **Figure 3**. The second stage computes an encoding of the descriptors extracted by the first. This is based on a set of descriptor templates, $\mathcal{T}_c^{(2)}$, learned from a training dataset. Descriptor templates can be the components of a model of the descriptor probability distribution, e.g., a Gaussian mixture model (GMM), kernel density, vector quantizer, or RBF network (Duda et al., 2001) or the basis functions of a sparse representation of descriptor space. When the former are used, we denote the encoding as probabilistic, while the term sparse encoding is used for the latter. Examples of probabilistic encodings include the minimum probability of error (MPE) architecture of Vasconcelos and Lippman (1997, 2000); Vasconcelos (2004a); Carneiro et al. (2007), the spatial pyramid matching kernel (SPMK) of Lazebnik et al. (2006), the naive-Bayes nearest neighbor (NBNN) classifier of Boiman et al. (2008), the hierarchical Gaussianization (HGMM) of Zhou et al. (2009), and many variants on these methods. Sparse encodings include, among others, the sparse SPMK method of Yang et al. (2009) and the locality-constrained linear (LLC) encoding of Wang et al. (2010).

The most popular encoding is probabilistic, namely a GMM with templates learned by either k-means or the expectation-maximization algorithm. In this case, the descriptor encoding reduces to computing a measure of descriptor-template similarity $s(y^{(1)}(l), \mathcal{T}_c^{(2)})$ and assigning the descriptor the closest template. It is also possible to rely on a soft assignment, where a descriptor is assigned to multiple templates with different weights. This is, for example, the case of sparse encodings. In all cases, the map of descriptor assignments to the cth template, $\mathcal{T}_c^{(2)}$, is the cth channel of the stage 2 representation. Assignment channels are then pooled spatially, to produce the final image representation.

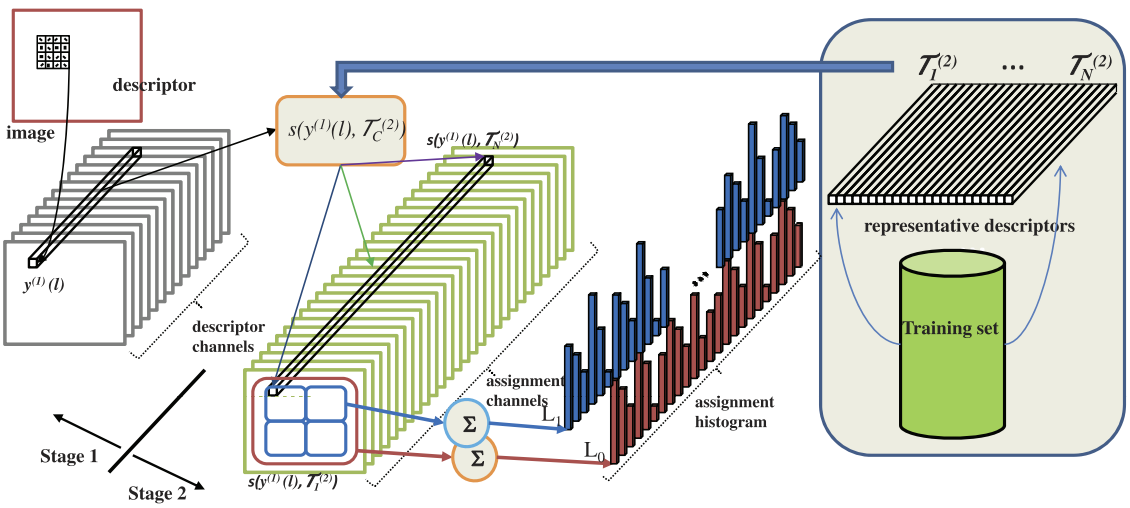


FIGURE 5 | Canonical architecture implemented by various popular object recognition methods. Images are represented by sets of descriptors. A set of representative descriptors $\{\mathcal{T}_c^{(2)}\}$ is learned from an image training set. The descriptors $y^{(1)}(l)$ extracted from the image to classify are then encoded, with respect to this set of representatives. The encoding consists of assigning each descriptor to a subset of the representatives, using a similarity function $s\left(y^{(1)}(l), \mathcal{T}_c^{(2)}\right)$. This could be a probabilistic function, e.g., probability under a Gaussian mixture model, or a sparse encoding. The assignments are finally pooled spatially to produce assignment histograms, which are fed to a classifier, e.g., a support vector machine.

For hard assignments, this is equivalent to representing the input image as a histogram of stage 2 assignments. The pooling operation can be performed over the entire image, sub-areas, or both. We next discuss how different computer vision methods map into this architecture.

3.3.2. Stage 1: descriptors

Popular descriptors, e.g., SIFT (Lowe, 1999) or HoG (Dalal and Triggs, 2005), are measures of orientation dominance. While we discuss SIFT in detail, a similar analysis applies to others. The SIFT descriptor $y \in \mathbb{R}^{128}$ is a set of 8-bin histograms of orientation response computed from intensity gradients. Location l contributes to histogram bin k with $a_k(l) = r(l)g(l)b_k[\theta(l)]$, where $r(l), \theta(l)$ are the gradient magnitude and orientation at l, $g(l)$ a Gaussian that penalizes locations farther from the descriptor center, and $b_k(\theta)$ a trilinear interpolator, based on the distance between θ and the orientation of bin k. The kth histogram entry is

$$h_k = \sum_{l \in B} a_k(l), \tag{29}$$

where B is a 4×4 pixel cell. The descriptor concatenates histograms of 4×4 cells into a 128-dimensional vector, which is normalized, fed to a saturating nonlinearity $\tau(x) = \max(x, 0.2)$ and normalized again to unit length. Using superscripts $q \in \{1, \ldots, 16\}$ for cells, and subscripts $k \in \{1, \ldots, 8\}$ for orientation bins, this is the sequence of computations

$$\overline{h_k^q} = \tau\left[\frac{h_k^q}{\sum_{q,k} h_k^q}\right] = \tau\left[\sum_{l \in B^q} \frac{a_k(l)}{\sum_{q,k}\sum_{l \in B^q} a_k(l)}\right] \tag{30}$$

$$s_k^q = \frac{\overline{h_k^q}}{\sum_{q,k} \overline{h_k^q}^2} \qquad y = (s^1, \ldots, s^{16})^T. \tag{31}$$

Note that (31) is a combination of divisive normalization (of $a_k(l)$ by responses in all cells B^q), average pooling, and squashing non-linearity, similar to the sequence of (27) and (28). The main difference is the application of the non-linearity after pooling vs. after filtering, as in (25). (31) can be seen as pre-processing for stage 2, contrast normalizing stage 1 responses. This is identical to (16), the normalization of HDSN layer inputs. In summary, the SIFT computations can be mapped to a network layer similar to those discussed above.

In fact, the descriptor can be interpreted as a saliency measure, if $a_k(l)$ is replaced by the response magnitude $|x_k^{(1)}(l)|$ of a Gabor filter with the kth orientation, a conceptually equivalent measure of oriented image energy. Defining

$$\begin{aligned}\alpha &= \sum_{q,j}\sum_{l \in B^q} |x_j^{(1)}(l)| = \sum_{q,l \in B^q} |x_k^{(1)}(l)| + \sum_{j \neq k}\sum_{q,l \in B^q} |x_j^{(1)}(l)| \\ &= \sum_{q,l \in B^q} |x_k^{(1)}(l)| + \nu\end{aligned}$$

(31) reduces to $\overline{h_k^q} = \tau[\epsilon_k^q]$ where

$$\epsilon_k^q = \sum_{l \in B^q} \frac{|x_k^{(1)}(l)|}{\alpha} \tag{32}$$

$$\propto -\sum_{l \in B^q} \log P_{X_k^{(1)}}(x_k^{(1)}(l); \alpha, 1) \tag{33}$$

$$\approx -\int_{B^q} P_{X_k}(x; \alpha^q, 1) \log P_{X_k}(x; \alpha, 1) dx \tag{34}$$

with $P_X(x; \alpha, 1)$ as given in (1), and $\alpha^q = \sum_{l \in B^q} |x_k^{(1)}(l)|$. Hence, up to constants, ϵ_k^q is the cross-entropy between the responses of filter $X_k^{(1)}$ within cell B^q and across the support of the descriptor. Assuming that the distributions are identical, this is the response entropy, a common saliency measure (Rosenholtz, 1999; Kadir and Brady, 2001; Bruce and Tsotsos, 2006; Zhang et al., 2008) that equates salient to rare (low-probability) events. Hence, SIFT can be interpreted as a saliency measure, which identifies as salient stimuli of rare orientation within a local image neighborhood.

3.3.3. Stage 2: descriptor assignments

Under this interpretation, the templates $\mathcal{T}_c^{(2)}$ are saliency templates[1]. For probabilistic models, the descriptor-to-template assignment of stage 2 is always a variation on layer 2 of the HMAX network. The likelihoods $s_c^{(2)}(l)$ of the descriptor $y^{(1)}(l)$ under the components of a Gaussian mixture whose means are the templates $\mathcal{T}_c^{(2)}, c \in \{1, \ldots, N\}$ are first computed with (23). These likelihoods are then mapped into posterior probabilities of component given descriptor, by a divisive normalization across channels

$$p_c^{(2)}(l) = \frac{s_c^{(2)}(l)}{\sum_{c=1}^N s_c^{(2)}(l)}. \tag{35}$$

The RBF precision parameter β of (23) controls the softness of the assignments. When $\beta \to 0$ the mixture model becomes a vector quantizer (Vasconcelos, 2004b) and $p_c^{(2)}(l) = 1$ for the template closest to $y^{(1)}(l)$, and zero for all others, i.e., assignments are hard. When $\beta > 0$ descriptors are assigned to multiple components, according to the posteriors $p_c^{(2)}(l)$, i.e., assignments are soft. Some methods, e.g., MPE, HGMM, or NN, learn descriptor templates per object class and compute the posterior class probability

$$P_{Y|X}(c|y^{(1)}(l)) = \sum_{j \in I_c} p_j^{(2)}(l) \tag{36}$$

where I_c is the set of indices of templates from class c. In summary, for probabilistic models, the second stage of the canonical architecture consists of the RBF network of HMAX plus the divisive normalization of (35), and can be complemented by (36). Overall,

[1]This is a terminology for descriptor templates alternative to *visual words* (Sivic and Zisserman, 2003; Csurka et al., 2004), *textons* (Leung and Malik, 2001), *visemes* (Ezzat and Poggio, 2000), or others used in the literature.

there are three types of layer 2 units: HMAX uses the likelihood units (LU) of (23), while the remaining approaches rely on the posterior units (PU) of (35), or the class-posterior units (CPU) of (36).

For sparse models, the assignments $p_c^{(2)}(l)$ are obtained by minimizing a sparseness inducing assignment cost. For example, the assignments of SPMK are the solution of

$$p^{(2)}(l) = \arg\min_p ||y^{(1)}(l) - \mathbf{T}^{(2)}p||^2 + \lambda||p||_1 \quad (37)$$

where $\mathbf{T}^{(2)}$ is a dictionary with templates $\mathcal{T}_c^{(2)}$ as columns, $||p||_1$ the ℓ_1 norm of p, and λ a regularization parameter. This produces a soft assignment, of sparsity (number of non-zero entries) controlled by λ. While sparse assignments can improve recognition performance, they have increased computational cost, since the optimization of (37) has to be repeated for each descriptor of the image to classify. This is frequently done with greedy optimization by matching pursuits (Mallat and Zhang, 1993), which involve multiple iterations over all templates in $\mathbf{T}^{(2)}$. We denote the units of sparse representation as *projection pursuit* (PP) units.

For both probabilistic and sparse models, the final step of stage 2 is an assignment histogram, computed by either average

$$y_c^{(2)}(l) = \frac{1}{|N^{(2)}(l)|} \sum_{m \in N^{(2)}(l)} p_c^{(2)}(m), \quad (38)$$

or maximum

$$y_c^{(2)}(l) = \max_{m \in N^{(2)}(l)} p_c^{(2)}(m), \quad (39)$$

pooling. The neighborhood $N^{(2)}(l)$ can be the entire image, in which case there are as many pooling units as descriptor templates, i.e., N, but is usually repeated for a number of subregions, using the pyramid structure introduced by SPMK and shown in **Figure 5**. This is usually a three-layer pyramid, containing the full image at level 0, and its partition into 2×2, and 4×4 equal sized cells at levels 1 and 2, respectively. In this case, there are a total of $21N$ pooling units.

3.4. DISCUSSION

Table 2 summarizes the operations of various popular recognition methods. The table is organized by the type of saliency (none, bottom-up, or top-down) implemented by each of the methods. It should be noted that the template learning procedures are not necessarily tied to the network architecture. For example, HMAX could use k-means, and SPMK could use codebooks of randomly collected examples. In fact, many alternative methods have been proposed for codebook learning (Sivic and Zisserman, 2003; Csurka et al., 2004; Fei-Fei and Perona, 2005; Winn et al., 2005; Moosmann et al., 2007) or sparse representation (Mairal et al., 2008; Wang et al., 2010). It is, nevertheless, clear that the different methods perform similar sequences of operations. In all cases, these operations can be mapped into the network architecture of **Figure 1** and implement at least some aspects of the standard neurophysiologic model (Carandini et al., 2005). However, the basic operations can differ in substantive details, such as the types of non-linearities, the order in which they are applied, etc. Since any combinations are in principle possible, the space of possible object recognition networks is combinatorial. This is amplified by the combinatorial possibilities for the number of parameters of any particular network configuration, e.g., receptive field sizes, subsampling factors, size of pooling regions, normalizing connections, etc. In result, it is nearly impossible to search for the best configuration for any particular recognition problem.

From a theoretical point of view, the main benefit of the HDSN is the statistical interpretation (e.g., computation of target probabilities) and functional justification (e.g., saliency detection) that it provides for all network computations. This results in clear guidelines for the sequence of network operations to be implemented, namely the S and C-units of **Figure 2**, clear semantics for normalizing connections (training feature responses under the target and background classes), and an abstract characterization of the unit computations, as in the algorithmic implementation of Section 2.2.3. It is thus possible to design network architectures

Table 2 | Mapping of various popular recognition algorithms to the canonical architecture of Figure 5.

Method	Stage 1		Stage 2			
	Units	Saliency	Templates	Units	Assignment	Pooling
HMAX	Filter responses	–	Random	LU	Soft	Max
MPE	Filter responses	–	GMM	CPU	Soft	Sum
NBNN	SIFT	Bottom-up	Training set	CPU	Hard	Sum
SPMK	SIFT	Bottom-up	Codebook	PU	Hard	Sum
HGMM	SIFT	Bottom-up	GMM	CPU	soft	sum
Sparse SPMK	SIFT	Bottom-up	Sparse dictionary	PP	Soft	Max
LSN	SL	Top-down		–		
HDSN	DS	Top-down	Random	DS	–	Sum

HMAX: (Serre et al., 2007), MPE: (Carneiro et al., 2007), NBNN: (Boiman et al., 2008), SPMK: (Lazebnik et al., 2006), HGMM: (Zhou et al., 2009), sparse SPMK: (Yang et al., 2009), LSN: (Elazary and Itti, 2010).

for specific tasks, without the need for exhaustive search. In fact, the statistical nature of the underlying computations could be used to expand network functionality, e.g., by resorting to model adaptation techniques (Saenko et al., 2010; Dixit et al., 2011; Kulis et al., 2011) in order to reduce training set sizes, or belief propagation to enable more sophisticated forms of statistical inference, such as Markov or conditional random fields (Geman and Geman, 1984; He et al., 2004). For object recognition, some form of model adaptation is already enabled by the divisive normalization connections of **Figure 2A**) or, equivalently, the scale parameters α_i of the target and background distributions. As mentioned in Section 2.2.2, these enable the interpretation of S-units as the *parametric* rectification units $\psi(x)$ of (11), which support a much richer set of network behaviors (e.g., sensitivity to feature absence) than commonly used non-linearities (such as the sigmoid or ReLU operations). By changing its scale parameters, the network can *adapt* to new recognition tasks without having to relearn new filters. This adaptation is also quite simple: it reduces to collecting samples of filter response to the target classes of interest and using (2) to estimate the scales α_i. None of the other networks (or even computer vision algorithms) discussed above has this property.

Of all the recognition architectures discussed above, the HDSN is also unique in its explicit modeling of discriminant saliency, based on statistical modeling of the target and background distributions. In most other models, the saliency computation does not even involve the notions of target and background class, and the GGD scale is simply estimated from a neighborhood of the image to classify, as in (27) or (32). This strictly bottom-up definition of saliency cannot be tuned for recognition. On the other hand, the saliency maps of the HDSN identify feature responses discriminant for target detection, with all the advantages previously discussed: optimal feature denoising, modulation of saliency responses by the discriminant power of the underlying features, and ability to detect both feature presence and absence. These differences in turn have a non-trivial impact in the saliency templates $\mathcal{T}_c^{(2)}$ of stage 2. SIFT templates are usually much less discriminant than those of **Figure 3**. By implementing saliency in layer 2, the HDSN complements this advantage with the identification of saliency configurations discriminant for target recognition. We next show that these properties make the HDSN more *efficient* in terms of image representation than all other models, achieving higher accuracies with fewer layer 2 units and a fairly simple training procedure.

4. RESULTS

An extensive set of experiments was conducted to evaluate HDSN performance on saliency, object recognition, and localization tasks. All experiments were performed on datasets available in the literature, including Caltech101 (C101) (Fei-Fei et al., 2005), 15 scenes (N15) (Lazebnik et al., 2006), ALOI (Geusebroek et al., 2005), and the pandaCam dataset of Han and Vasconcelos (2011). Details of these datasets are given in the Supplementary Material.

4.1. OBJECT RECOGNITION EXPERIMENTS

We start with object recognition. While, as shown in **Table 2** the different approaches can be mapped to a common network form, the standard configurations of the different methods disagree even in the most elementary parameters, e.g., number of layer 2 units. For example, SPMK usually relies on a dictionary of size 1024 and a pyramid of 21 pooling regions. While this should be compared to an HMAX model of $21K$ units, only $4K$ are usually adopted in the HMAX literature. Methods that learn a codebook per class increase the number of units by a few orders of magnitude. In the worst case of Boiman et al. (2008) (as many units as training examples), the layer 2 RBF has 10 million units. This lack of uniformity makes it difficult to compare the different approaches. To overcome this problem, we implemented all units discussed in the previous sections and used them to build networks that are otherwise identical, i.e., have the same configuration, use the same learning procedure, etc. We then compared network performance on C101 and N15. A first experiment measured the impact of each unit of **Table 2** on recognition accuracy. This experiment used a relatively small network, with fixed (Gabor) templates in the bottom layer and randomly sampled (from the first layer responses) templates in the second layer. In a second experiment, we built a larger HDSN and compared its performance to the results reported for the various recognition algorithms in the literature. This was mostly a sanity check, to ensure that the HDSN could achieve the results reported for these methods, using the parameters with which they were proposed. It is assumed that these parameters were optimized to guarantee the best results per method, of the network components, but allowing an unbiased estimate of the best possible performance per architecture.

4.1.1. Impact of network units on recognition performance

To test the impact of network units on recognition accuracy, we started from a base network with the configuration of **Table 1** and the following operations:

(1) Local normalization of image intensities, according to (16);
(2) $S^{(1)}$ *units:* Gabor filters, no saliency;
(3) $C^{(1)}$ *units:* average pooling;
(4) $S^{(2)}$ *units:* 40 LUs with randomly selected templates $\mathcal{T}_c^{(2)}$ per class, for a total of $C^{(2)} = 600$ channels for N15 and $C^{(2)} = 4040$ channels for C101.
(5) $C^{(2)}$ *units:* average pooling.

In a first experiment, we compared the impact of layer 1 units on network performance. This was done, by replacing the $S^{(1)}$ and $C^{(1)}$ units with those on the left of **Table 3**. The same Gabor channels were used across settings, the convolutional network layer (CN) was implemented with (25)–(28), SIFT with (31)–(31), and discriminant saliency (DS) with (17) and (2). The pooling operator was that which performed best for each network. Note that the type II network is identical to HMAX (Serre et al., 2007), and the first layer of the networks of type III, IV, and V is, respectively, layer 1 of the convolutional network of Jarrett et al. (2009), the first stage (SIFT) of the computer vision methods of Lazebnik et al. (2006); Boiman et al. (2008); Yang et al. (2009); Zhou et al. (2009), and layer 1 of the HDSN.

The table supports several conclusions. First, pooling significantly enhances recognition performance, as all methods with C

Table 3 | Recognition accuracy of a 2-layer network with different units.

Type	Simple unit	Pooling	L2	N15	C101	Type	L1	L2	N15	C101
I	Filter	–	LU	58.7 ± 0.3	40.5 ± 0.8	I	DS	CPU	67.4 ± 1	61 ± 0.8
II	Filter	Max	LU	65.6 ± 1.3	52.8 ± 1	II	DS	PU	68.1 ± 1	62.1 ± 1.1
III	CN	Max	LU	67.1 ± 1.5	58.8 ± 1.3	III	DS	LU	68.3 ± 0.6	64.2 ± 1.3
IV	SIFT	Average	LU	67.5 ± 0.6	62.8 ± 0.9	IV	DS	DS	80 ± 0.6	69.2 ± 1.3
V	DS	Average	LU	68.3 ± 0.6	64.2 ± 1.3	V	DS*	DS	82.2 ± 0.7	69.9 ± 1.7

Left: impact of layer 1 units on recognition performance. Starting from a network (type I) with Gabor filters and no pooling in layer 1 and LU units with average pooling in layer 2, several enhancements were added to layer 1. These consisted of convolutional network (CN), SIFT, or discriminant saliency (DS) units. Max and average pooling operators were also tested, results are reported for the most effective. Right: impact of layer 2 units on recognition performance. In all cases, layer 1 consists of DS units. Layer 2 is implemented with CPU, PU, LU, or DS units. DS reports to an enhanced layer 1, including feature selection.*

units substantially outperformed the type I network. This finding confirms the importance of the spatial invariance attributed to this operation, and of C units in general. However, we *did not* find an advantage for either average or max pooling. Second, the addition of divisive normalization across features (bottom-up orientation saliency) implemented by both the CN and SIFT layers, further improved recognition accuracy. The gains of this operation were particularly significant on C101. This can be explained by the fact that shape is a more important cue for recognition in C101 (an object database) than in N15 (a database of scenes). Since this type of divisive normalization enhances edges with a dominant orientation, it produces crisper layer 2 templates, which are more informative about object shape. This enables large gains in C101 (from 52.8 to 62.8% for SIFT) but is also beneficial on N15 (from 65.6 to 67.5%). Third, for both datasets, the performance of the SIFT layer was superior to that of the convolutional network layer. This suggests that the sequence of S-unit operations of (31)–(31) is more effective than that of (25)–(28), but it is difficult to ascertain why. Finally, DS units had the best overall performance. It is worth noting that, while the SIFT and CN layers perform normalization *both* within (spatially) and across channels, DS units only require within channel normalization. This enables independent channel processing, considerably simplifying the implementation of this network. In fact, the HDSN layer has very little computational overhead with respect to the HMAX layer of the type II network. As discussed in Section 2.2.2, the only difference is the addition of the parametric ReLU units of (11). On C101, this boosts recognition accuracy from 52.8 to 64.2%. Overall, the HDSN layer has the lowest complexity among the top performing networks (types III to V).

To test the impact of the configuration of layer 2, we used a network with a layer 1 of DS units. Besides likelihood units (LU), layer 2 was implemented with posterior units (PU), class-posterior units (CPU), and DS units. Since the number of layer 2 channels is drastically reduced when CPU units are used (from the number of templates to the number of classes, e.g., 600 to 15 in N15 and 4040 to 101 in C101), and this reduces the effectiveness of the SVM that follows the network, we tried alternative CPU configurations. Best results were obtained, in preliminary experiments, by weighing PU units according to the posterior class probability, i.e., multiplying (35) by (36). The resulting accuracies are summarized in the right of **Table 3**, network types I–IV. Interestingly, neither PU nor CPU improved the performance of LU. Unlike layer 1, cross-channel normalization did not show any benefits in the second layer. Again, DS units achieved the best performance, substantially improving the recognition of LUs (68.3 to 80% on N15 and 64.2 to 69.2% on C101). In summary, both the adoption of DS units and the hierarchical computation of saliency produced substantial recognition gains. Note that the HDSN (type IV of right column of **Table 3**) is a fairly simple extension of HMAX (type II of the left column), both conceptually (addition of saliency) and algorithmically [addition of the parametric ReLU units of (11)]. A comparison to the HMAX performance, or even to an HMAX network with a DSN in the first layer (type V, left column) shows very significant improvements: from 66–68 to 80% on N15 and 53–64 to 69% on C101.

4.1.2. Large network

The previous experiments were based on a relatively small network. We next compared the performance of a larger HDSN to the results reported in the literature for the methods of Section 3.3. We note that, when compared to these approaches, the features implemented by the HDSN (Gabor filters and randomly selected saliency templates) are fairly simple. The published results for the other algorithms are frequently based on much more complex features and feature selection. Examples include independent component analysis (ICA) (Kanan and Cottrell, 2010), sparse decompositions (Yang et al., 2009), or very large sets of random features (Jarrett et al., 2009). Pinto et al. (2008) has shown that a single layer network with many channels can outperform hierarchical networks with few channels per layer. We considered a limited set of enhancements of this type. The filter pool of the first layer was first augmented with 63 discrete cosine transform (DCT) filters of size 8×8 (the DCT set minus the average -DC-filter). This is a proxy for the expansion of Jarrett et al. (2009), who showed that a set of random projections can outperform a Gabor decomposition. Feature selection was then implemented by pooling the saliency measure of (7) across the visual field, per feature X. The 4 channels of largest saliency were selected, maintaining the dimensionality of layer 1 identical to HMAX (Mutch and Lowe, 2008). The resulting recognition accuracy is shown as type V in the right of **Table 3**, where DS* means "DS with feature selection." The more elaborate feature set had gains of 2.2% in N15 and 0.7% in C101. No further extensions were considered.

Table 4 compares these results to the literature, where different methods have very different numbers of layer 2 units. These are

Table 4 | Comparison of a 2-layer HDSN to various methods from the literature, on the 15 scenes and Caltech101 Datasets.

Method	# Layer 2 units		Recognition rate		# Layer 2 units		Recognition rate	
	N15	C101	N15	C101	N15	C101	N15	C101
SPMK($L = 0$)	400	200	74.8 ± 0.3	41.2	–	–	–	–
SPMK ($L = 2$)	–	4200	–	64.6	8400	–	81.4 ± 0.5	–
kNN-SVM	–	3030	–	66.2 ± 0.5	–	–	–	–
V1 model	–	4000	–	42 ± 0.5	–	86,000	–	65
HMAX	–	4075	–	56	–	–	–	–
NBNN	–	–	–	–	–	10 M	–	70.4
Sparse SPMK	450	5120	75.3 ± 0.5	64.8 ± 0.7	21,504	21,504	80.28 ± 0.9	$\mathbf{73.2 \pm 0.5}$
convNN	–	4096	–	65.5	–	–	–	–
HGMM	–	–	–	–	46,080	310,272	85.2	73.1
HDSN	450	4040	$\mathbf{82 \pm 0.5}$	$\mathbf{70 \pm 0.5}$	22,500	20,200	$\mathbf{85.4 \pm 0.3}$	73.1 ± 0.6

Results are presented for different numbers of layer 2 units, and grouped into small (left) and large (right) networks. The comparison includes SPMK (Lazebnik et al., 2006), kNN-SVM (Zhang et al., 2006), V1 model (Pinto et al., 2008), HMAX (Mutch and Lowe, 2008), NBNN (Boiman et al., 2008), sparse SPMK (Yang et al., 2009), convNN (Jarrett et al., 2009), and HGMM (Zhou et al., 2009).

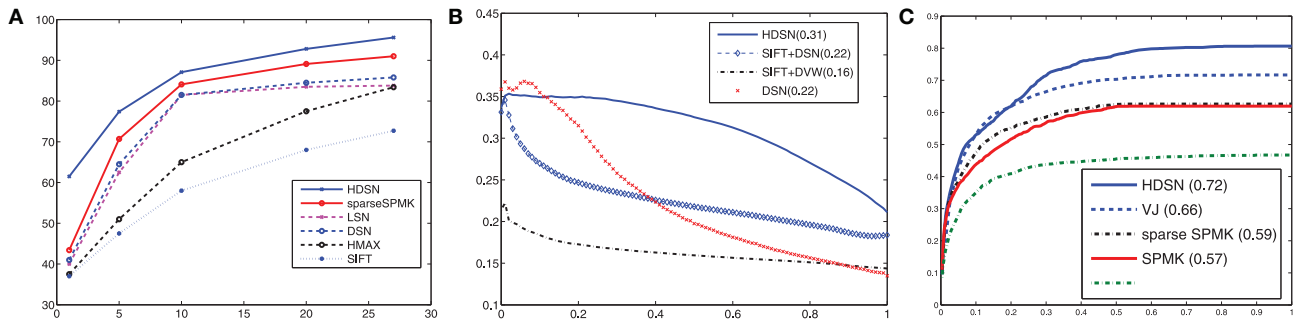


FIGURE 6 | (A) classification accuracy vs. training set size on ALOI. **(B)** Precision-recall curves for object localization on pandaCam. **(C)** Detection rate vs. number of false positives per image for panda detection.

also shown in the table, which we organized by network dimensionality. The left half reports to "small" networks (≈400 units for N15, 4000 for C101), the right to "large" (≈20 K for both). The HDSN performs well in both regimes. The most interesting observation is, however, its performance among small networks, where it is far superior to the next best methods (82 vs. 75% on N15, 70 vs. 66% on C101). In fact, in N15, the small version of HDSN outperforms the large versions of SPMK, NBNN, and sparse SPMK. In C101, it is only outperformed by the large versions of HGMM, and sparse SPMK. It should be pointed that these are not the best results reported on these datasets. Better performance can usually be obtained using SVM classifiers with non-linear kernels, which we have not considered in our implementation. For example, on C101, the accuracy of a 4000 unit SPMK classifier can be boosted to 74.4% by addition of a chi-square kernel (Chatfield et al., 2011). This is slightly superior to the results reported in **Table 4** for the combination of a 20,200 HDSN with a linear SVM. In summary, the HDSN of 20,000 units has learned a high-dimensional embedding similar to that of the kernel-SVM, which has orders of magnitude higher implementation complexity. This is particularly impressive given the simplicity of the random sampling procedure used to learn the HDSN templates. Again, the comparison to an equivalent network with no saliency computation (HMAX) shows very large gains (from 56 to 70% recognition rate on C101 with 4000 units).

4.2. COMPARISON TO SALIENCY MEASURES

These results show that the HDSN outperforms architectures that use no saliency, or bottom-up saliency measures such as SIFT. The next comparison was to the top-down measure (LSN) of Elazary and Itti (2010). Since no software is available for this network, we compared the two approaches on ALOI, where LSN was originally evaluated. In addition to the HDSN (1000 units) and the methods evaluated in (Elazary and Itti, 2010)—LSN, HMAX (1000 units), and SIFT-based image matching (Lowe, 1999)—we also considered a single layer HDSN (denoted DSN) and sparse SPMK (1024 units). **Figure 6A**) compares the recognition rates of all methods, showing that the HDSN has the best performance. For example, with 27 training images per class, it has a recognition rate of 95.6%, while sparse SPMK achieves 91%, DSN 85.8%, LSN 83.8%, HMAX 83.4%, and SIFT 72.7%. These results confirm that both the addition of discriminant saliency (HDSN vs. HMAX) and its hierarchical computation (single-layer

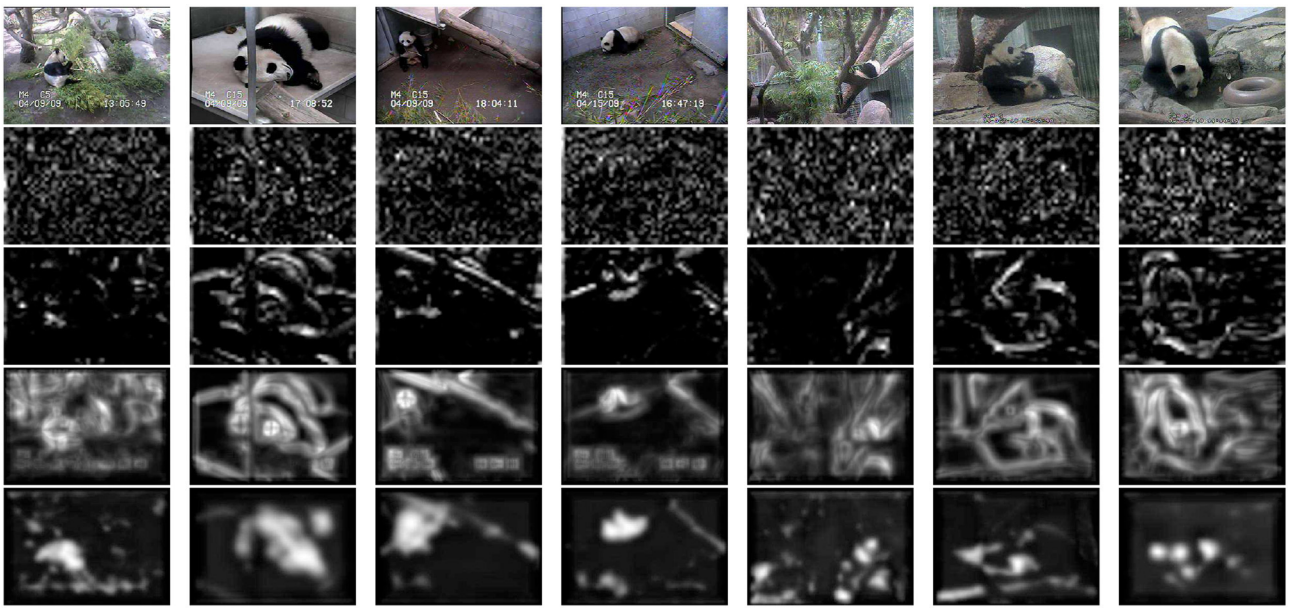


FIGURE 7 | Object localization on the pandaCam dataset. Top row: example images. **Second row**: saliency maps produced by the combination of SIFT descriptors (layer 1) and discriminant visual words (layer 2). **Third row**: same for a combination of SIFT descriptors (layer 1) and discriminant saliency (layer 2). **Fourth row**: saliency maps produced by a single layer HDSN. **Fifth row**: same for a two-layer HDSN. In all cases, the saliency map is obtained by summing simple unit outputs across all channels.

DSN vs. two-layer HDSN) lead to substantial gains in recognition performance.

4.3. OBJECT LOCALIZATION AND DETECTION

We next considered the problem of object localization, on the pandaCam dataset, where we compared the performance of the HDSN to those of a saliency method based on SIFT in layer 1 and discriminant visual words in layer 2 (Dorko and Schmid, 2005) (SIFT+DVW), a HDSN with layer 1 replaced by SIFT units (SIFT+DS), and a single layer HDSN. SIFT+DVW is an intermediate between an RBF and a layer of DS units: it is based on visual words but emphasizes those that are discriminant for each class. **Figure 7** shows saliency maps produced by the four methods, by simply summing the $S^{(2)}$-unit responses across all feature channels. SIFT+DVW produces very noisy maps, with many false positives on the background, and few strong responses at target locations. The replacement of the DVW by the DS layer (SIFT+DS) suppresses most of this noise, but mostly produces edge maps, illustrating the limitations of SIFT: detection of simple features, failure to respond to the object interior, and poor selectivity for the target. While improving on DVW, the use of DS units in layer 2 cannot compensate for all these limitations. In fact, the single-layer HDSN produces better saliency maps than SIFT+DSN. Its maps are more selective for the target, have greater response toward the object interior, and respond more strongly to complex features such as the panda face. Finally, HDSN achieves the best performance, with saliency maps that are active in the target interior and have few false positives. These observations are confirmed by the precision recall curves of **Figure 6B**). The average precision is 0.31 for HDSN, 0.22 for single layer HDSN, 0.22 for SIFT+DS, and 0.16 for SIFT+DVW.

A final set of experiments was performed on object detection. An object detector was implemented by applying a box filter and non-maximum suppression to the saliency map of an HDSN with 200 layer 2 units. This was compared to a 6 component part model (partModel) of Felzenszwalb et al. (2009), sparse SPMK, SPMK, and the Viola-Jones (VJ) detector (Viola and Jones, 2004). Sparse SPMK, SPMK, and VJ used a sliding window, with windows of seven scales, and step size of 10 pixels. Non-maximum suppression was implemented as in Felzenszwalb et al. (2009), and applied to all approaches. SPMK and sparse SPMK used a spatial pyramid of 2 levels, and a codebook of 1000 visual words. Curves of detection rate vs false positives per image (fppi) are shown in **Figure 6C**). The partModel was unable to model pandas with the finite set of poses available, achieving the worst performance. Both sparse SPMK and SPMK produced a significant improvement, with sparse SPMK achieving slightly better performance. Another performance boost was achieved with the VJ detector. Finally, HDSN had the overall best performance. The detection rates at 0.3 fppi were 71.5% for HDSN, 66% for VJ, 58.6% for sparse SPMK, 56.8% for SPMK, and 43.8% for the partModel.

5. CONCLUSIONS

In this work, we have investigated the evolutionary benefits of integrating attention and object recognition, by introducing a joint model, the HDSN, for saliency and recognition. HDSNs are networks whose layers implement top-down saliency detectors based on features of increasing selectivity and invariance. This

is accomplished by (1) learning saliency templates of increasing complexity and (2) adopting pooling operators of increasing support, in higher network layers. It was shown that HDSNs are consistent with the standard neurophysiologic model of the visual cortex but have a precise computational justification, and a statistical interpretation for all network computations. This enables the statistical learning of all network parameters and the explicit optimization of the network for recognition. The learning of HDSN parameters requires very simple mechanisms and has minimal computational cost over previous models, such as HMAX or convolutional neural networks, that lack an explicit connection to saliency. When compared to these models, HDSNs have a more precise mapping to the cortical neurophysiology, and explicitly account for both target and background hypotheses in the computation of all network layers. This results in saliency templates that are highly selective for the object classes of interest. The HDSN also introduces a new type of non-linearity, the parametric ReLU, whose parameters can be tuned for the detection of object classes of interest. This enables a number of functional enhancements, including optimal feature denoising mechanisms for recognition, modulation of saliency responses by the discriminant power of the underlying features, and ability to detect both feature presence and absence. A detailed experimental evaluation has provided evidence for the advantages of all these functional enhancements, as well as for the class-specific tuning inherent to discriminant saliency, and the gains of saliency layers using templates of increasing complexity, target selectivity, and invariance. It was also shown that normalization across orientation channels does not necessarily benefit recognition. This is an interesting finding, which enables much simpler networks and justifies the known cortical organization into orientation selective hypercolumns. Perhaps more importantly, the experiments presented suggest that there are non-trivial benefits in integrating attention and recognition. While attention is frequently modeled as a pre-processor (selector of regions), e.g., the classical dichotomy between pre-attentive and attentive vision, HDSNs assume that recognition *is* a component of attention and vice-versa. This was shown to substantially improve performance in core attention tasks, such as object localization, and core recognition tasks, such as object detection. In fact, it was shown that a single network can perform effectively in the problems of object localization, recognition, and detection, by a simple rearrangement of how the saliency maps produced by the different templates are processed: in parallel for recognition, and additively for localization and detection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fncom.2014.00109/abstract

REFERENCES

Boiman, O., Shechtman, E., and Irani, M. (2008). "In defense of nearest-neighbor based image classification," in *IEEE Conference in Computer Vision and Pattern Recognition* (Anchorage, AK).

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans. Patt. Anal. Mach. Intell.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Borji, A., Sihite, D., and Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans. Image Process.* 22, 55–69. doi: 10.1109/TIP.2012.2210727

Brincat, S., and Connor, C. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* 7, 880–886. doi: 10.1038/nn1278

Bruce, N., and Tsotsos, J. (2006). "Saliency based on information maximization," in *Neural Information Processing Systems* (Vancouver, BC: MIT Press).

Buccigrossi, R., and Simoncelli, E. (1999). Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Process.* 8, 1688–1701. doi: 10.1109/83.806616

Carandini, M., Demb, J., Mante, V., Tolhurst, D., Dan, Y., Olshausen, B., et al. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597. doi: 10.1523/JNEUROSCI.3726-05.2005

Carandini, M., and Heeger, D. (2011). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* 17, 8621–8644.

Carneiro, G., Chan, A., Moreno, P., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Patt. Anal. Mach. Intell.* 29, 394–410. doi: 10.1109/TPAMI.2007.61

Chang, S., Bin, Y., and Vetterli, M. (2000). Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Process.* 9, 1532–1546. doi: 10.1109/83.862633

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011). "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference* (Dundee).

Coates, A., Ng, A. Y., and Lee, H. (2011). "An analysis of single-layer networks in unsupervised feature learning," in *International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, FL), 215–223.

Csurka, G., Dance, C., Fan, L., Williamowski, J., and Bray, C. (2004). "Visual categorization with bags of keypoints," in *In ECCV Workshop on Statistical Learning in Computer Vision* (Prague).

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *IEEE Conference in Computer Vision and Pattern Recognition* (San Diego, CA).

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Ann. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Dixit, M., Rasiwasia, N., and Vasconcelos, N. (2011). "Adapted gaussian models for image classification," in *IEEE Conference in Computer Vision and Pattern Recognition* (Colorado Springs, CO).

Do, M., and Vetterli, M. (2002). Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Trans. Image Process.* 11, 146–158. doi: 10.1109/83.982822

Dorko, G., and Schmid, C. (2005). *Object Class Recognition using Discriminative Local Features*. Technical Report RR-5497, INRIA - Rhone-Alpes, February 2005.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*. New York, NY: John Wiley & Sons.

Elazary, L., and Itti, L. (2010). A bayesian model for efficient visual search and recognition. *Vision Res.* 50, 1338–1352. doi: 10.1016/j.visres.2010.01.002

Ezzat, T., and Poggio, T. (2000). Visual speech synthesis by morphing visemes. *Int. J. Comput. Vis.* 38, 45–57. doi: 10.1023/A:1008166717597

Fei-Fei, L., Fergus, R., and Perona, P. (2005). "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories," in *CVPR Workshop on Generative Model-Based Vision* (San Diego, CA).

Fei-Fei, L., and Perona, P. (2005). "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference in Computer Vision and Pattern Recognition* (San Diego, CA).

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part based models. *IEEE Trans. Patt. Anal. Mach. Intell.* 32, 1627–1645. doi: 10.1109/TPAMI.2009.167

Fukushima, K. (1980). Neocognition: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Gao, D., Han, S., and Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. Patt. Anal. Mach. Intell.* 31, 989–1005. doi: 10.1109/TPAMI.2009.27

Gao, D., Mahadevan, V., and Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *J. Vis.* 8, 1–18. doi: 10.1167/8.7.13

Gao, D., and Vasconcelos, N. (2005). "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *IEEE Conference in Computer Vision and Pattern Recognition* (San Diego, CA).

Gao, D., and Vasconcelos, N. (2009). Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Comput.* 21, 239–271. doi: 10.1162/neco.2009.11-06-391

Geman, S., and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Patt. Anal. Mach. Intell. IEEE Trans.* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596

Geusebroek, J., Burghouts, G., and Smeulders, A. (2005). The amsterdam library of object images. *Int. J. Comput. Vis.* 61, 103–112. doi: 10.1023/B:VISI.0000042993.50813.60

Graham, N. (2011). Beyond multiple pattern analyzers modeled as linear filters (as classical v1 simple cells): useful additions of the last 25 years. *Vision Res.* 51, 1397–1430. doi: 10.1016/j.visres.2011.02.007

Guo, C., Ma, Q., and Zhang, L. (2008). "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *IEEE Conference in Computer Vision and Pattern Recognition* (Anchorage, AK).

Han, S., and Vasconcelos, N. (2010). Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Res.* 50, 2295–2307. doi: 10.1016/j.visres.2010.05.034

Han, S., and Vasconcelos, N. (2011). "Biologically Plausible Detection of Amorphous Objects in the Wild," in *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Colorado Springs, CO).

He, X., Zemel, R., and Carreira-Perpinan, M. (2004). "Multiscale conditional random fields for image labeling," in *IEEE Conference in Computer Vision and Pattern Recognition* (Washington, DC).

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* 9, 181–197. doi: 10.1017/S0952523800009640

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hou, X., and Zhang, L. (2007). "Saliency detection: a spectral residual approach," in *IEEE Conference in Computer Vision and Pattern Recognition* (Minneapolis, MN).

Huang, J., and Mumford, D. (1999). "Statistics of natural images and models," in *IEEE Conference in Computer Vision and Pattern Recognition* (Fort Collins, CO).

Hubel, D. H., and Wiesel, T. N. (1962). Receptive field, binocular interaction, and functional architecture of the cat's visual cortex. *J. Physiol.* 160, 106.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 20, 1254–1259. doi: 10.1109/34.730558

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). "What is the best multi-stage architecture for object recognition?," in *IEEE Conference in Computer Vision and Pattern Recognition* (Miami Beach, FL).

Kadir, T., and Brady, M. (2001). Scale, saliency and image description. *Int. J. Comput. Vis.* 45, 83–105. doi: 10.1023/A:1012460413855

Kanan, C., and Cottrell, G. (2010). "Robust classification of objects, faces, and flowers using natural image statistics," in *IEEE Conference in Computer Vision and Pattern Recognition* (San Francisco, CA).

Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., and LeCun, Y. (2010). "Learning convolutional feature hierarchies for visual recognition," in *Neural Information Processing Systems* (Vancouver, BC).

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems* (Lake Tahoe, NV).

Kulis, B., Saenko, K., and Darrell, T. (2011). "What you saw is not what you get: domain adaptation using asymmetric kernel transforms," in *IEEE Conference in Computer Vision and Pattern Recognition* (Colorado Springs, CO), 1785–1792.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference in Computer Vision and Pattern Recognition* (New York, NY).

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., et al. (1990). "Handwritten digit recognition with a back-propagation network," in *Neural Information Processing Systems* (Denver, CO: MIT Press).

LeCun, Y., Bottou, L., Bengio, Y., and Haffiner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Leung, T., and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vis.* 43, 29–44. doi: 10.1023/A:1011126920638

Lowe, D. G. (1999). "Object recognition from local scale-invariant features," in *International Conference on Computer Vision* (Corfu).

Mahadevan, V., and Vasconcelos, N. (2013). Biologically-inspired object tracking using center-surround mechanisms. *IEEE Trans. Patt. Anal. Mach. Intell.* 35, 541–554. doi: 10.1109/TPAMI.2012.98

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008). "Discriminative learned dictionaries for local image analysis," in *IEEE Conference in Computer Vision and Pattern Recognition* (Anchorage, AK).

Mallat, S., and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* 41, 3397–3415. doi: 10.1109/78.258082

Maunsell, J., and Treue, S. (2006). Feature-based attention in visual cortex. *Trends Nerosci.* 29, 317–322. doi: 10.1016/j.tins.2006.04.001

Miau, K., and Schmid, C. (2001). "Neuromorphic algorithms for computer vision and attention," in *Proceedings of SPIE 46 Annual International Symposium on Optical Science and Technology* (San Diego, CA).

Moosmann, F., Triggs, B., and Jurie, F. (2007). "Randomized clustering forests for building fast and discriminative visual vocabularies," in *Neural Information Processing Systems* (Vancouver, BC: MIT Press).

Mutch, J., and Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80, 45–57. doi: 10.1007/s11263-007-0118-0

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning* (Haifa), 807–814.

Navalpakkam, V., and Itti, L. (2007). Search goal tunes visual features optimally. *Neuron* 53, 605–617. doi: 10.1016/j.neuron.2007.01.018

Perrett, D., and Oram, M. (1993). Neurophysiology of shape processing. *Imaging Vis. Comput.* 11, 317–333. doi: 10.1016/0262-8856(93)90011-5

Pinto, N., Cox, D., and DiCarlo, J. (2008). Why is real-world visual object recognition hard? *PLoS Comput. Biol.* 4:e27. doi: 10.1371/journal.pcbi.0040027

Pinto, N., Doukhan, D., DiCarlo, J., and Cox, D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput. Biol.* 5:e1000579. doi: 10.1371/journal.pcbi.1000579

Poggio, T., and Edelman, S. (1990). A network that learns to recognize 3d objects. *Nature* 343, 263–266. doi: 10.1038/343263a0

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Res.* 39, 3157–3163. doi: 10.1016/S0042-6989(99)00077-2

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. (2010). "Adapting visual category models to new domains," in *European Conference in Computer Vision* (Creta), 213–226.

Sebe, N., and Lew, M. S. (2003). Comparing salient point detectors. *Patt. Recogn. Lett.* 24, 89–96. doi: 10.1016/S0167-8655(02)00192-7

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Patt. Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Sivic, J., and Zisserman, A. (2003). "Video google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision* (Nice).

Srivastava, A., Lee, A., Simoncelli, E., and Zhu, S. (2003). On advances in statistical modeling of natural images. *Math. Imaging Vis.* 18, 17–33. doi: 10.1023/A:1021889010444

Treisman, A. (1985). Preattentive processing in vision. *Comput. vis. Graph. Image Process.* 31, 156–177. doi: 10.1016/S0734-189X(85)80004-9

Tsotsos, J. (1991). Computational resources do constrain behavior. *Behav. Brain Sci.* 14, 506–507. doi: 10.1017/S0140525X00071053

Vasconcelos, N. (2004a). Minimum probability of error image retrieval. *IEEE Trans. Signal Process.* 52, 2322–2336. doi: 10.1109/TSP.2004.831125

Vasconcelos, N. (2004b). On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Trans. Inform. Theory* 50, 1482–1496. doi: 10.1109/TIT.2004.830760

Vasconcelos, N., and Lippman, A. (1997). "Library-based coding: a representation for efficient video compression and retrieval," in *Data Compression Conference* (Snowbird, UT).

Vasconcelos, N., and Lippman, A. (2000). "A probabilistic architecture for content-based image retrieval," in *IEEE Conference in Computer Vision and Pattern Recognition* (Hilton Head Island, SC).

Vidal-Naquet, M., and Ullman, S. (2003). "Object recognition with informative features and linear classification," in *IEEE Conference in Computer Vision and Pattern Recognition* (Madison, WI).

Viola, P., and Jones, M. (2004). Robust real-time face detection. *Int. J. Comput. Vis.* 57, 137–154. doi: 10.1023/B:VISI.0000013087.49260.fb

Walther, D., and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Netw.* 19, 1395–1407. doi: 10.1016/j.neunet.2006.10.001

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). "Locality-constrained linear coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition* (San Francisco, CA), 3360–3367.

Winn, J., Criminisi, A., and Minka, T. (2005). "Object categorization by learned universal visual dictionary," in *International Conference on Computer Vision* (Beijing).

Wolfe, J. M. (1994). Guided search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238. doi: 10.3758/BF03200774

Wolfe, J. M. (1998). Visual memory: what do you know about what you saw? *Curr. Biol.* 8, R303–R304. doi: 10.1016/S0960-9822(98)70192-7

Yamane, Y., Carlson, E., Bowman, K., Wang, Z., and Connor, C. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360. doi: 10.1038/nn.2202

Yang, J., Yu, K., and Gong, Y. (2009). "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference in Computer Vision and Pattern Recognition* (Miami Beach, FL).

Yarbus, A. (1967). *Eye Movements and Vision*. New York, NY: Plenum. doi: 10.1007/978-1-4899-5379-7

Zhang, H., Berg, A., Maire, M., and Malik, J. (2006). "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," in *IEEE Conference on Computer Vision and Pattern Recognition* (New York, NY).

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *Int. J. Comput. Vis.* 73, 213–238. doi: 10.1007/s11263-006-9794-4

Zhang, L., Tong, M., Marks, H., Tim, K., Shan, H., and Cottrell, G. (2008). Sun: a bayesian framework for saliency using natural statistics. *J. Vis.* 8:32. doi: 10.1167/8.7.32

Zhou, X., Cui, N., Li, Z., Liang, F., and Huang, T. (2009). "Hierarchical gaussianization for image classification," in *IEEE Conference in Computer Vision and Pattern Recognition* (Miami Beach, FL).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 April 2014; accepted: 22 August 2014; published online: 09 September 2014.

Citation: Han S and Vasconcelos N (2014) Object recognition with hierarchical discriminant saliency networks. Front. Comput. Neurosci. ***8****:109. doi: 10.3389/fncom.2014.00109*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Han and Vasconcelos. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 12 December 2014
doi: 10.3389/fncom.2014.00158

A conceptual framework of computations in mid-level vision

Jonas Kubilius[1,2]*, Johan Wagemans[2] and Hans P. Op de Beeck[1]

[1] Laboratory of Biological Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium
[2] Laboratory of Experimental Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Jonathan W. Peirce, Notthingham University, UK
Heiko Neumann, Ulm University, Germany

***Correspondence:**
Jonas Kubilius, Laboratories of Biological and Experimental Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102 bus 3714, Leuven 3000, Belgium
e-mail: jonas.kubilius@ppw.kuleuven.be

If a picture is worth a thousand words, as an English idiom goes, what should those words—or, rather, descriptors—capture? What format of image representation would be sufficiently rich if we were to reconstruct the essence of images from their descriptors? In this paper, we set out to develop a conceptual framework that would be: (i) biologically plausible in order to provide a better mechanistic understanding of our visual system; (ii) sufficiently robust to apply in practice on realistic images; and (iii) able to tap into underlying structure of our visual world. We bring forward three key ideas. First, we argue that surface-based representations are constructed based on feature inference from the input in the intermediate processing layers of the visual system. Such representations are computed in a largely pre-semantic (prior to categorization) and pre-attentive manner using multiple cues (orientation, color, polarity, variation in orientation, and so on), and explicitly retain configural relations between features. The constructed surfaces may be partially overlapping to compensate for occlusions and are ordered in depth (figure-ground organization). Second, we propose that such intermediate representations could be formed by a hierarchical computation of similarity between features in local image patches and pooling of highly-similar units, and reestimated via recurrent loops according to the task demands. Finally, we suggest to use datasets composed of realistically rendered artificial objects and surfaces in order to better understand a model's behavior and its limitations.

Keywords: mid-level vision, similarity, pooling, perceptual organization, summary statistics

VISION AS AN IMAGE UNDERSTANDING SYSTEM

The visual system of primates processes visual inputs incredibly rapidly. Within 100 ms observers are capable of reliably reporting and remembering contents of natural scenes (e.g., Potter, 1976; Thorpe et al., 1996; Li et al., 2002; Quiroga et al., 2008). Such fast processing puts tight constraints on models of vision as most computations should be done roughly within the first feedforward wave of information. Efforts to understand how this is possible have led to the so-called standard view of the primate visual system where objects are rapidly extracted from images by a hierarchy of linear and non-linear processing stages, where simple and specific features are combined in a non-linear fashion, resulting in increasingly more complex and more transformation-tolerant features (Fukushima, 1980; Marr, 1982; Ullman and Basri, 1991; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; DiCarlo et al., 2012; see Kreiman, 2013, for a review).

In particular, in primate visual cortex the earliest stages of visual processing are thought to act as simple local feature detectors. For example, retinal ganglion and lateral geniculate nucleus cells preferentially respond to blobs with center-surround organization (Kuffler, 1953; Hubel and Wiesel, 1961), while neurons in primary visual area V1 respond to oriented edges and bars (Hubel and Wiesel, 1962; see Carandini et al., 2005, for a review). These detectors act locally (within their receptive field) and thus are very sensitive to changes in position or size. In contrast, neurons in the final stages of visual processing in the inferior temporal cortex respond to complex stimuli, including whole objects (Tanaka, 1996; Kourtzi and Kanwisher, 2001; Op de Beeck et al., 2001; Huth et al., 2012), faces (Desimone et al., 1984; Kanwisher et al., 1997; Tsao et al., 2006), scenes (Epstein and Kanwisher, 1998; Kornblith et al., 2013), bodies (Downing et al., 2001; Peelen and Downing, 2005) and other categories. At this stage, neurons have large receptive fields and thus are tolerant to changes in position, size, orientation, lighting, and clutter (DiCarlo and Cox, 2007). While the exact details of the properties of neurons at the low and high visual areas remain an area of active research, in our view the most puzzling question is the following: What computations are performed at the intermediate steps of information processing in order to bridge simple local early representations to highly multidimensional representations of objects and scenes?

In primates, inspired by Hubel and Wiesel's (1965) proposal of the hierarchical processing in the visual cortex, a number of studies focused on demonstrating sensitivity to the increasing complexity of features along the visual hierarchy. For example, in V2 angle or curvature detectors have been reported (Dobbins et al., 1987; Ito and Komatsu, 2004). In V4, neurons are sensitive to even more complex curved fragments and three-dimensional parts of surfaces (Pasupathy and Connor, 1999, 2001, 2002; Yamane et al.,

2008). Thus, the idea is that intermediate layers are responsible for gradually combining simpler features into more complex ones (Riesenhuber and Poggio, 1999; Rodríguez-Sánchez and Tsotsos, 2012).

However, building a system that could robustly utilize such a connection scheme on natural images is difficult. On the one hand, combining simpler features into more complex ones is complicated due to the presence of clutter. Robust mechanisms are necessary to combine the "correct" features and leave out the noise. Similarly, in order to detect complex features, enormous dictionaries must be built since the number of possible feature combinations is huge, so this process is highly resource-intensive (but see Fidler et al., 2009, for an inspiring approach to the issue). On the other hand, focusing solely on edges and their combinations into shapes misses a number of other useful cues in the images—such as differences in color, texture, motion and so on—and thus may lack the necessary power both to process object shapes and to be useful for other tasks that the visual system is performing (e.g., interaction with objects in a scene, navigation, or recovering spatial layout; Regan, 2000).

Thus, in computer vision, partially due to the described limitations of the standard view of primate visual system and partially due to the development of robust algorithms for dealing with large numbers of features, the actually implemented models of vision have bypassed thinking about intermediate representations altogether in their implementations. Instead, such models rely solely on the established features of V1 (namely, oriented edge detection) and directly apply sophisticated machine learning techniques (such as support vector machines) to detect what object categories are likely to occur in the given image. Somewhat surprisingly, this idea works very well for a number of complex tasks. For example, in the famous algorithm by Viola and Jones (2001), faces are detected using several simplistic feature detectors, reminiscent of the odd and even filters of V1. In Oliva and Torralba's GIST framework (2001, 2006; Torralba and Oliva, 2003), scene categorization is achieved by computing global histogram statistics of oriented filter outputs. Flat architectures of SIFT (Lowe, 2004) or HoG (Dalal and Triggs, 2005) that largely rely on oriented feature detection have seen a wide adoption for a variety of visual tasks in computer vision, and, in combination with multi-scale processing (Bosch et al., 2007), for a long time these models that have no hierarchies have been the state-of-the-art approach.

However, eventually hierarchical models that contain intermediate representations ultimately proved superior in many complex visual tasks. While such deep networks have been proposed several decades ago, (Fukushima, 1980; LeCun et al., 1989; Schmidhuber, 1992), only recently upon development of more robust procedures for learning from large pools of data (Hinton and Salakhutdinov, 2006; Boureau et al., 2010) such networks managed to achieve state-of-the-art object identification performance on demanding datasets that contain millions of exemplars, such as the Large Scale Visual Recognition Challenge (Deng et al., 2009; Krizhevsky et al., 2012; Sermanet et al., 2013; Szegedy et al., 2014), or that demand fine-grain discrimination as in the case of face recognition (Lu and Tang, 2014; Taigman et al., 2014). Moreover, these networks have been reported to perform extremely well on a number of visual tasks (Razavian et al., 2014). While many challenges remain (Russakovsky et al., 2013), the fact that base-level object categorization and localization have been very successful and in some cases even approaching or superseding human-level performance (Serre et al., 2007; Lu and Tang, 2014; Taigman et al., 2014) is greatly encouraging. Importantly, representations learned by such deep networks have been shown to match well the representations in the primate V4 and IT (Yamins et al., 2014), demonstrating the relevance of these models to understanding biological vision.

Naturally, the success of these object recognition models begs the question whether we now understand how the visual system processes images. It is tempting to conclude that weakly organized collections of features are sufficient for object and scene categorization, and, by extension, scene understanding. However, it is important to realize that, engineering advances aside, each layer in these architectures is based on the same principles characterized in the early visual processing of the primate brain. Is there really nothing more going on in the intermediate stages of processing?

In the following section, we consider what the computational goal of mid-level vision might be (cf. Marr, 1982). Based on these insights, in Section "Intermediate Computations" we propose basic computational mechanisms that we hypothesize to be sufficient to account for processes occurring at intermediate stages. Finally, we discuss what model evaluation procedures could help in guiding the implementation of such a system.

WHAT DO MID-LEVEL VISUAL AREAS DO?

FEATURE INTERPOLATION

Typically, a model of vision is operationalized as a feature extraction system. Features that are present in the input image need to be detected, so that a veridical (or at least useful) representation of the world (or objects in it) can be reconstructed. However, visual inputs are necessarily impoverished (e.g., due to collapsing of the third dimension as the image is projected on the retina), incomplete (e.g., due to some objects partially occluding others), ambiguous (e.g., due to shadows), and noisy. As a consequence, the problem of vision is not only feature detection but also feature inference (Purves et al., 2014).

A number of studies have shown that mid-level vision is heavily involved in feature inference. Consider, for example, the seminal series of studies by von der Heydt et al. (1984), von der Heydt and Peterhans (1989), who compared neural responses to the typical luminance-defined stimuli and the neural responses to the same stimuli defined by cues other than luminance. In one of their conditions, a stimulus was composed of two regions containing line segments but with one region shifted with respect to the other, forming an offset-defined discontinuity in the texture, which we refer to as a second-order edge (**Figure 1A**). Importantly, a simple edge-detecting V1 model would not be able to find such edges, so if some neurons in the visual cortex were responding to such stimuli, it would mean that a higher-order computation is at work that somehow is capable of integrating information across the two regions in the image.

Consistent with the known properties of early visual areas, the researchers observed a robust response to the luminance-defined edges. However, in addition they also demonstrated that

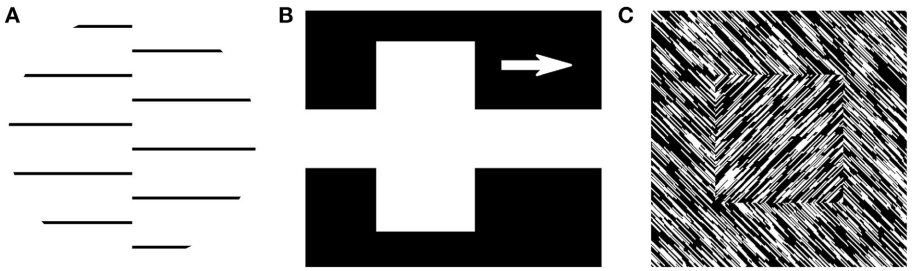


FIGURE 1 | Feature interpolation. (A) A second-order boundary stimulus as used by von der Heydt et al. (1984). **(B)** A stimulus with an illusory contour is perceived in the white gap between the two parts of the white rectangle, as used by von der Heydt et al. (1984). The arrow indicates that the white rectangle was moving. **(C)** A stimulus where a shape is defined entirely by second-order cues (that is, a difference in orientation), used in many figure-ground segmentation studies (e.g., Lamme, 1995).

some neurons in V2 responded to the second-order edges, and, in fact, often with the same orientation preference as to luminance-defined edges. Moreover, Lamme et al. (1999) reported that V1 neurons were also responding to this boundary roughly 60 ms after stimulus onset and suggested iso-orientation suppression as a mechanism behind such fast second-order edge detections. These findings have since been replicated in V2 and V4 (Ramsden et al., 2001; Song and Baker, 2007; El-Shamayleh and Movshon, 2011; Pan et al., 2012) and also reported for discontinuities in orientation (Larsson et al., 2006; Allen et al., 2009; Schmid et al., 2014), motion (Marcar et al., 2000), and contrast (Mareschal and Baker, 1998; Song and Baker, 2007; Li et al., 2014). Taken together, these findings demonstrate that even in the absence of luminance-defined borders in the inputs, mid-level areas infer potential borders from differences in other cues. Importantly, this operation is different from the typical feature detection and combination scheme because in this case a feature is computed that is not present in the input (that is, a second-order border).

An even more extreme example of such feature inference has been demonstrated by another condition in von der Heydt and colleagues' experiments where they used a stimulus inspired by the Kanizsa triangle (Kanizsa, 1955). The stimulus was defined as a white bar moving over two black bars, separated by a white gap (**Figure 1B**)—thus, although physically there were no edges connecting the two halves of the white bar, subjectively observers would nonetheless report seeing the complete white bar, effectively interpolating its borders or surface across the white gap. We refer to such borders as illusory contours. Surprisingly, for this condition, von der Heydt et al. (1984) also reported neurons in V2 responding to these illusory contours, and, in fact, nearly as vigorously as to the luminance-defined ones.

If these examples appear only as curious cases of feature inference in artificial setups, imagine a typical cluttered image where multiple objects are partially occluded. Just like in the two previous cases, the visual system appears to interpolate occluded parts of objects at the early stages of visual information processing (a process known as amodal completion; van Lier et al., 1994; Ban et al., 2013). For example, **Figure 2A** is interpreted as a gray blobby shape partially occluded by the black blobby shape, both on a dotted background, as in **Figure 2C**. In fact, we cannot help but perceive the gray shape inferred behind the black occluder and our phenomenology is most certainly not captured by segmentation into separate non-overlapping regions as in **Figure 2B**.

Similarly, the background appears to continue behind the two shapes even though there is no physical connection between the left and the right portion of it, demonstrating that filling-in is not confined to objects but applies in a more generic manner to any occluded region in the input. Moreover, at least phenomenologically, this filling-in appears to involve not only surface interpolation but also the spread of feature statistics. In our example, observers would report that the occluded part of the background is likely to continue the pattern of polka dots (van Lier, 1999).

Moreover, just like in the other two cases (second-order borders and illusory contours), the amodal interpolation has been reported to be established relatively fast, already in 75–200 ms (Sekuler and Palmer, 1992; Ringach and Shapley, 1996; Murray et al., 2001; Rauschenberger et al., 2006), and has also been observed in the early modulation of the occluded parts of shapes in monkey V4 (Bushnell et al., 2011; Kosai et al., 2014).

Taken together, we see that the visual system actively performs feature inference and it is an early process that may be initiated already with the first wave of information. It is important to note that in all of these cases, the inference does not necessarily produce a complete feature or a shape. Rather, it may reflect a rough estimate of statistical properties of the shape (cf. "fuzzy completions," van Lier, 1999) or the probability of possible completions where the missing part of the shape may occur (D'Antona et al., 2013).

RELATIONAL INFORMATION AND SURFACE CONSTRUCTION

But what is the purpose of feature extraction or interpolation? In many object recognition models, for example, the extracted features are used directly to perform categorization. Notice that such an output lacks the explicit assignment of the features to one object or another, that is, object shapes are not explicitly represented. Such model behavior is strikingly at odds with our phenomenology dominated by explicit object shapes or surfaces. This idea has been nicely illustrated by Lamme (1995) who investigated neural responses to a shape entirely defined by a second-order boundary. His stimulus consisted of a field of oriented noisy elements embedded in a background of an opposite orientation (**Figure 1C**). In order to perceive this shape, the visual system



FIGURE 2 | Seeing is not the same as perceiving. Observers report perceiving the configuration in **(A)** to be composed of full shapes as depicted in **(C)** rather than as in **(B)** which reflects the physical inputs where shapes are fragmented and two portions of background are separate. In **(C)**, the gray shape has been interpolated behind the black shape (depicted in green), indicating that mapping of a two-dimensional surface in a three-dimensional space is already necessary to represent depth relations. Furthermore, the background is also a single surface rather than two separate regions and also with its statistical properties (polka dot pattern) filled in. Some observers will also see the black shape interpolated behind the gray one (depicted in red), but this percept is much less consistent among observers than the completion of the gray shape behind the black one, indicating that surface inference might not be precise and rather indicate probabilities of possible contour and surface properties.

must be able to (i) infer second-order borders and (ii) combine them into the shape as a whole. Lamme (1995) showed that neurons in monkey V1 with receptive fields inside that shape reliably respond more than those outside, that is, the visual system explicitly represents where the figure is. Moreover, the observed enhancement was not instantaneous but rather developed in three stages (as described in Lamme et al., 1999). Early on, only responses to local features were observed. Within a 100 ms, responses to the second-order boundary emerged. Finally, neurons in V1 corresponding to the figural region of the display started responding more than the background. This effect was later shown to be the effect of feedback from higher visual areas such as V4, where such figure-ground assignments are thought to emerge (Poort et al., 2012). Taken together, this example demonstrates that the visual system gradually extracts not only the contour of a shape but also its inside, resulting in a full surface reconstruction.

More broadly, it has been argued that surface-based representations form a critical link between early- and high-level computations (Nakayama et al., 1995; see also Pylyshyn, 2001). Moreover, the presence of a surface strongly influences even the earliest computations of the visual information processing such as the iso-orientation suppression (Joo and Murray, 2014). Finally, surface-based representations can also be beneficial for object identification tasks because surfaces are topologically stable structures and thus largely invariant to affine transformations (Chen, 1982, 2005). For example, a hole in a surface remains present despite drastic changes in its position, orientation or rotation in depth, or to the changes in surface structure (Chen, 1982; Todd et al., 2014).

In general, we argue that encoding spatial relations—whether between features, or deciding which features belong to the same object or surface, or ordering the surfaces in space—provides a tremendous wealth of information (Biederman, 1987; Barenholtz and Tarr, 2007; Oliva and Torralba, 2007): Knowing that a car is on the road or above the road makes a big difference, but using only features without relations between them might fail to capture these differences (Choi et al., 2012). One influencial account of the power of spatial relations has been provided by Biederman (1987), who noticed that certain spatial relations between features, known as non-accidental properties, remain largely invariant to affine transformations in space. For example, short parallel lines nearly always remain parallel despite changes in viewpoint. He proposed that these relations might be used to encode different object categories, and later Hummel and Biederman (1992) developed a model illustrating how such a system might work. While the exact purpose of such structural representations in recognition has been heavily debated since (Barenholtz and Tarr, 2007), consistent with this idea a number of studies demonstrated that observers are very sensitive to changes in these invariant features of a shape (Wagemans et al., 1997, 2000; Vogels et al., 2001; Kayaert et al., 2005a,b; Lescroart et al., 2010; Amir et al., 2012).

Similarly, Feldman (1997, 2003) and van Lier et al. (1994) argued that configural regularities of the inputs are used to organize features into objects, and human visual system has been shown to be sensitive to such configural relations (Kubilius et al., 2014). Moreover, Blum (1973) proposed that the configuration of shapes is encoded in the visual system by representing their skeletal, or medial axis, structure, and Hung et al. (2012) showed that neurons in monkey IT indeed respond both to the contour of a shape and its medial axis structure. Taken together, these studies highlight the fact that the visual system utilizes configural relations between features and surfaces in the higher visual areas, and therefore an explicit encoding of these relations should be supported by mid-level computations.

REPRESENTATIONS FOR MULTIPLE TASKS, NOT ONLY OBJECT RECOGNITION

We argued that mid-level vision was involved in feature detection and surface construction, such that in the end the shape of an object could be reliably extracted from the image. However, the long quest for superior object identification algorithms has somehow overshadowed the fact that visual cortex can achieve more than just object identification. Vision is our means to understanding the world, whereas a mere object-based representation provides only a tiny fraction of information needed for successful behavior in the world. This point is particularly pertinent in lower species such as rodents for whom navigation is a more immediate task than object identification (Cox, 2014). In fact, much

of our visual input is not composed of well-defined objects and thus trying to parse them into objects makes little sense. A richer description is thus needed if we were to capture the essence of information about the world (Gibson, 1979).

To stress the point of the inadequacy of object-based representations, let us consider a series of images in **Figure 3**. In some cases, like **Figure 3A**, where the object ("a car") is clearly separate (self-contained) from the rest (the road), object identification and localization provides the most important information about the scene ("there is a car"). But consider a row of buildings, for example (**Figure 3B**). While one still clearly describes each house as a distinct object, they are impossible to detach from other items (other houses and the ground). A more extreme example is depicted in **Figure 3C**, where even though a mountain is sticking out from the ground surface, it is no longer very clear where the mountain ends and the ground begins. Is the visual system really concerned about finding objects in such images then? In fact, as we go further away from close-up views into panoramic scenes, identifying objects does not appear to be the default mode any longer. In **Figure 3D**, we know that the image is composed of individual trees, grass and other stuff but we no longer can count them. Rather, a percept of various textures and layouts appears to dominate. Thus, talking about individual objects is largely irrelevant in these scenarios and instead describing texture properties and characteristics that allow navigation through the terrain, or a global level semantic labeling of "a forest" or "a lawn" often seems to be the more immediate task for vision (Oliva and Torralba, 2001; Torralba and Oliva, 2003).

Therefore, we point out that surfaces that mid-level areas construct are not only meant to represent the outline of objects in images but also (or primarily) to summarize the properties of textures and surfaces in the environment.

REPRESENTATIONS PRIOR TO IDENTIFICATION

Finally, we point out that intermediate representations do not have to rely on being able to identify the contents, consistent with the idea that they are computed early on. We do not need to know what we are looking at to be able to describe its three-dimensional shape, texture, and spatial relations to other items in an image. For example, notice that in **Figure 2** surface interpolation occurs despite us never having seen these particular shapes before and having no categorical label for them, indicating that this phenomenon could be performed by mid-level computations prior to categorization. This observation also holds for a more realistic image depicted in **Figure 4**, where we can easily agree that five objects situated in different depth planes are depicted. We can describe their shape and imagine acting upon them despite partial occlusions present in the image. This is clearly a more advanced representation of the image contents than a mere V1 filter output, yet not so advanced as to require any categorization, recognition or identification (naming) of the objects in it.

The idea of intermediate representations being established without recognition of contents is well-known in psychology (Witkin and Tenenbaum, 1983; Nakayama et al., 1995). To provide an illustrative example, the famous visual agnosia patient DF cannot report the identity or even orientation of most objects, yet her ability to act on these objects remains intact, a finding that has led Goodale and Milner (1992) to propose the vision-for-action and vision-for-perception division in the visual information processing in the brain. It thus appears that our visual system is adept in processing inputs even lacking knowledge about what they are, pointing to the idea that scene segmentation into objects might be more basic or more immediately performed than recognition. We do not claim that recognition is irrelevant for segmentation, as it has been shown that recognition can bias figure-ground assignment (Peterson, 1994), but our point is that it can largely be done successfully without any knowledge about the identity of objects.

CONCLUSION

Taken together, we claim that the goal of mid-level areas is the construction of surface-based representations that segment the input images into objects, background surfaces, and so on, together with their textural properties, because such format of representations is sufficiently rich for the variety of high-level tasks, including three-dimensional reconstruction of the scene, navigation in it, interaction with objects or restricting attention to them. The idea of the primacy of the surface-based representation is also supported by empirical studies showing that some form of figure-ground organization would be established already shortly after feedforward inputs reach higher visual areas and is consistent with the observation that segmentation does not require knowledge of the identity of the objects involved. Importantly, given the computational complexity, this organization is probably not computed globally but rather is restricted to parts of visual inputs that fall at fixation or where an observer is attending.

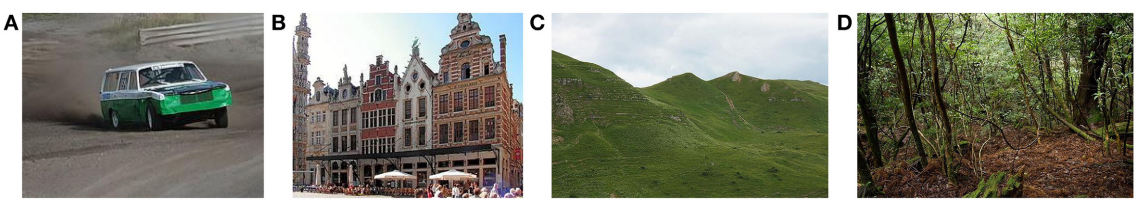


FIGURE 3 | The hierarchy of objecthood. Objects are not the most important piece of information in every image. While **(A)** has a well-defined object, it is already less clear in **(B)** what should count as one: The row of houses? Or each house separately? Or each of the windows? In **(C)**, there are three mountains but where each of them begins and ends is neither clear nor very important, and in **(D)** layout rather than object identity dominates perception, although one can see trees, trunks, etc. (Image credits from left to right: bengt-re, 2009, Snowdog, 2005, Reza, 2009, Σ64, 2012. All images are available under the Creative Commons Attribution License or are in the public domain.).



FIGURE 4 | Recognition is not crucial for scene or object understanding. In this artificially generated scene we see five novel objects, we can describe their three-dimensional shape despite partial occlusions, and navigate around them without having to know the identity of those objects.

It is also important to understand that the segmentation we describe here is not the same as what is commonly meant by this term. Many algorithms of segmentation only divide the image into a mosaic of non-overlapping regions without any information about the depth, that is, which region is in front of another one (see also Section "Current Approaches"). However, whenever something is occluded, that is a cue for depth ordering. Therefore, we consider a process that not only divides the image into separate regions but also infers figure-ground relations between these regions. Since this process often involves the inference of occluded parts, we refer to such interpolated regions as a surfaces.

Finally, such depth ordering is necessarily an oversimplification. For example, observe in **Figure 2C** that we do not perceive the whole of the black shape in front of the gray one. In fact, at least for some observers, part of the black shape (shown in red in **Figure 2C**) appears to be behind the gray shape, suggesting a three-dimensional form of the two shapes (Tse, 1999). This example demonstrates that the resulting representations cannot be captured by splitting an image into several depth planes, and thus require more flexibility. Such representation presumably would be followed by a full rectification of a three-dimensional volume at the later stages of visual information processing.

INTERMEDIATE COMPUTATIONS

We proposed that intermediate processing stages produce surface-based representations from two-dimensional static images. What computations could produce such representations?

CURRENT APPROACHES

In computer vision, many early image segmentation approaches considered segmentation as a global optimization problem of finding the best boundaries, grouped regions, or both. For example, Mumford and Shah (1989) proposed a functional that estimates the difference between the original image and its segmentation with constraints for smoothness and discontinuity at region boundaries (see also Lee et al., 1992). Finding the best segmentation amounts to finding the global minimum of this functional. Similarly, in a boundary-based contour extraction model, Elder and Zucker (1996) considered finding the shortest-path cycles in the graph containing boundary elements.

However, solving for a global optimum deemed to be a complicated task, often leading to unsatisfactory results. In 2000, Shi and Malik proposed a reconceptualization of the image segmentation problem as a graph cut problem. When features in an image are represented in a graph, finding the best segmentation amounts to finding groups of features in this graph that are maximally similar within a group and maximally dissimilar from other groups. Shi and Malik (2000) showed that their normalized cuts algorithm could provide a good optimization of this criterion and, based on this approach, they later developed one of the best-known image segmentation models (Arbeláez et al., 2011; see also Felzenszwalb and Huttenlocher, 2004 and Sharon et al., 2006, for much faster implementations of this idea).

Partitioning a graph in a fixed way, however, cannot capture the inherently hierarchical structure of images (a part can be part of another part; see the windows of houses in **Figure 3B**), nor can it adapt to the task demands. Therefore, in recent years much effort in image segmentation research has been devoted to the development of methods for the probabilistic generation of region proposals (Arbeláez et al., 2014) that could then be refined using a higher-level task such as categorization (Leibe et al., 2008; Girshick et al., 2014; Hariharan et al., 2014) or would be flexibly reconfigured based on Gestalt principles (Ion et al., 2013).

How could such partitioning of an image graph into high-similarity clusters be implemented in a biologically-plausible architecture? Based on behavioral and neural evidence, Nothdurft (1994) hypothesized that image segmentation involves (i)

suppression of responses in homogenous feature fields, and (ii) local pooling of features for boundary detection. Unlike the global optimization approaches considered above, this idea is based on completely local computations that are attractive due to their low complexity and biological plausibility. The implementation of this idea can be found in models by Grossberg (1994) and Thielscher and Neumann (2003), where texture segmentation is performed by enhancing edges that group together by the good continuation cue (using the "bipole cell" idea), and suppressing other locations in the image. Repeated over several iterations, this computation leads to the formation of the outline of the shape. This idea accounts well for Nothdurft's (1994) observations, and also provides an integrated framework of using both texture and boundary information to perform segmentation. Moreover, Thielscher and Neumann (2005) also demonstrated that this approach produces differences in convex and concave boundary appearance, in line with Nothdurft's (1994) observations.

Segmentation into distinct regions is only the first step though. As discussed in the previous section, this is not sufficient because an explicit surface construction and figure-ground relation computation need to occur as well. Some approaches (Roelfsema et al., 2002) attempted to explain figure-ground segmentation simply as an effect of increasing receptive field sizes (thus, decreasing spatial resolution) in higher visual areas. The model operates by initially detecting boundaries in the inputs and then pooling them together in higher visual areas as a result of increasing receptive-field sizes. Eventually, the whole shape is represented by a unit with a sufficiently large receptive field. Then, the figure-ground assignment can be propagated down via feedback to the early visual areas, as observed in the experiments by Lamme (1995).

However, it is unlikely that such scheme would work in more complex displays with more overlapping shapes and more variation in texture. Moreover, smaller shapes always produce higher responses in higher-level areas because their boundaries are closer together. Since these responses represent the figure-ground signal, smaller shapes are always bound to be on top of larger shapes that produce a weaker figure-ground signal. One possibility to resolve some of these issues is to use corners as indicators of the figural side. Since figures tend to be convex, the inside of a corner reliably indicates the boundary of a figure. Based on this observation, Jehee et al. (2007) proposed an extended version of the model by Roelfsema et al. (2002) that could produce more reliable border assignments.

The idea of using convexity can be applied more generally across the entire shape outline and not only at its corners. To illustrate how that could work, consider the two shapes in **Figure 5A**. The two edges shown in red can either be the boundary of the gray surface or the boundary of the white one, as indicated by the green arrows pointing to both directions. Of course, in this case it is clear that these edges must belong to the gray surface because the white one is just the background. But how would a model know? If we assume that objects tend to be convex, edges that are in agreement (the green arrows that are pointing toward each other) might belong to the same surface (**Figure 5B**). This simple computation in the local neighborhood followed by pooling into curved segments (**Figures 5C,D**) results in a largely correct border ownership. If it is further computed globally over a few iterations, local inconsistencies (e.g., a concavity of the lighter gray object) can be resolved (**Figure 5E**; see Figure 5B in Craft et al., 2007, for a working example), resulting in the proper assignment of edges to one of the two objects (**Figure 5F**), which is the desired initial image division into surfaces.

Importantly, because of border-ownership, we also learn which parts of objects are occluded. If a certain surface is partially bounded by a boundary that it does not own, it is a sign of an occlusion. For example, in **Figure 5F**, the yellow object is partially occluding the blue one, and border-ownership assignment indicates that edges along the yellow object belong to it. That leaves the blue object lacking a closed contour, meaning that part of it is occluded. An interpolation of surface results in a more perceptually compelling segmentation into whole shapes (van Lier et al., 1994), and consequently provides an ordering of surfaces in depth (**Figure 5G**).

The existence of such border-ownership cells has been reported in the visual area V2 (Zhou et al., 2000; see Zucker, 2014, for a good overview) and a number of models based on this idea have been proposed since (Zhaoping, 2005; Craft et al., 2007; Layton et al., 2012). Kogo et al. (2010) extended this framework by also using L- and T-junctions to determine not only figure-ground assignment for luminance-defined figures but also to produce the correct output in the case of illusory contours (Kanizsa's figures). Importantly, unlike earlier proposals (e.g., Grossberg, 1994), their approach is capable of yielding the correct representations of comparable yet non-illusory displays without ad hoc deletion of interpolated contours (see **Figure 1B** in Kogo et al., 2010).

Similarly, extending their work on bipole cells, Thielscher and Neumann (2008) showed that T-junctions could be used to infer figure-ground relations for multiple figures (not just figure and ground) in their architecture, and more recently, Tschechne and Neumann (2014) extended their earlier work to a full model of figure-ground segmentation. Initially, bipole cells, curvature and corner detectors are used to produce the consistent outline of a shape. Then, contextual cues are used to compute border-ownership.

Taken together, current biologically-inspired approaches to image segmentation largely concentrate on discovering boundaries in an input image and resolving figure-ground assignment by computing border-ownership of the boundaries in an image. However, unlike purely computer vision algorithms, these approaches are typically not tested with realistic inputs, thus their applicability and robustness on the wide variety of natural images remains unclear. Moreover, some models are better at segmentation but do not perform feature interpolation and figure-ground relation computations, and vice versa, while others focus on using second-order features but are not robust for segmentation using multiple cues, and so on. In other words, each of them only implements several aspects of processes in mid-level vision but the proposed mechanisms are not mutually compatible to build a unified architecture. Could there be several basic mechanisms that could account for the majority of the available data?

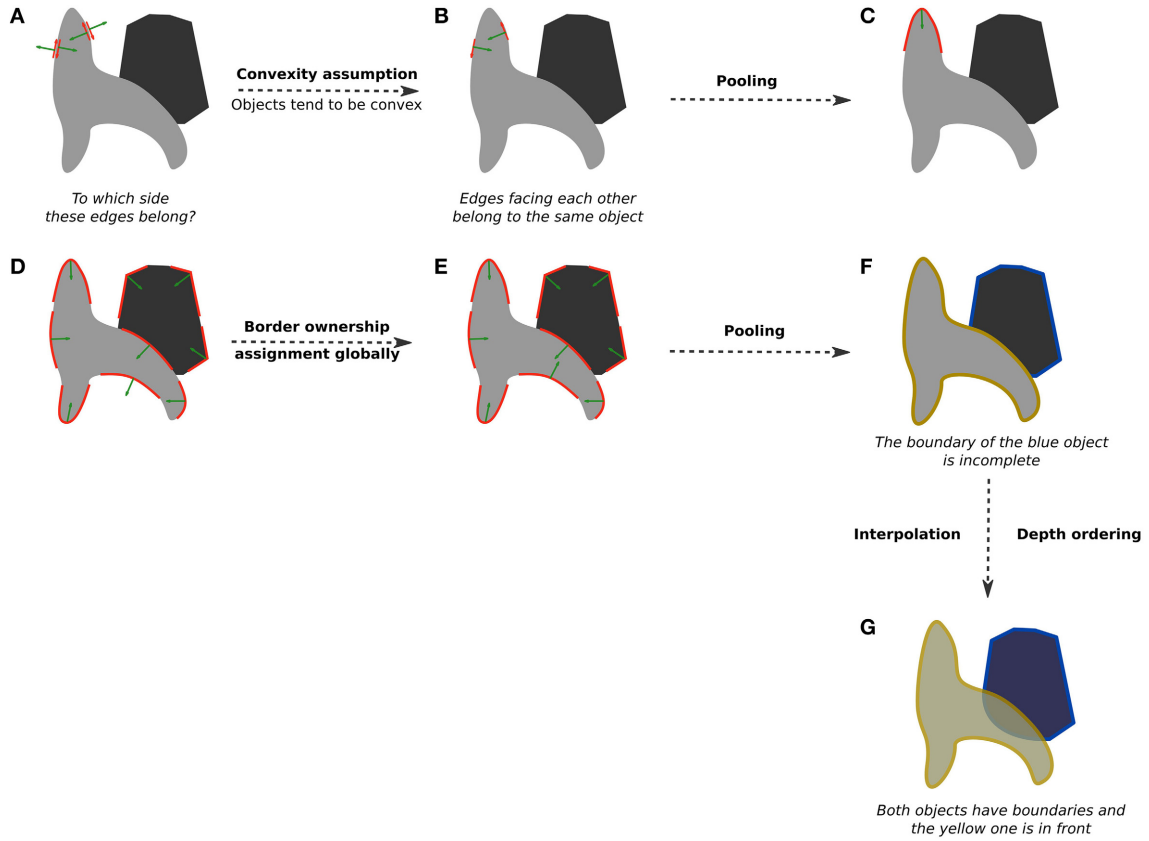


FIGURE 5 | An illustration of border-ownership assignment. (A) Initially, each edge in the image (red arrows) can belong to one of two sides (green arrows): either the gray surface or the white surface. The goal of border ownership computation is to figure out which side is the figural side. In this case, the edges should belong to the gray surface. **(B)** Using the convexity assumption (objects tend to be convex), we can easily determine border ownership in the local neighborhood. Edges that "agree" (green arrows are "looking" at each other) are preferred. **(C)** Pooling these edges together results in a curved segment with the correct border-ownership assignment. **(D)** After this computation is carried out in the local neighborhood, border ownership is largely but not fully correct. We can improve it by using the same convexity assumption over larger areas (e.g., over the entire image). **(E)** Global border-ownership computation results in a correct assignment of all segments. **(F)** With pooling, two separate surfaces emerge. Note that the blue one is missing a boundary at the intersection with the yellow object. This implies that the blue object is partially occluded by the yellow one. **(G)** Using this information, a correct local depth ordering is established and the missing piece of the blue object is interpolated.

OUR APPROACH

In a nutshell, we are interested in understanding *conceptually* what computations could suffice to account for the following biologically-plausible image processing strategy:

1. Region property and boundary extraction.
2. Clustering of boundary and region features into separate surfaces (segmentation).
3. Surface interpolation and depth ordering (figure-ground organization).
4. Representation refinement via recurrent loops.

Moreover, we want these computations to be sufficiently robust such that they would apply across various features in the images and could therefore be used in the typical computer vision setups such as deep networks.

To implement steps 1 and 2, we propose two basic mechanisms for intermediate computations, generalizing the vast majority of approaches discussed in Section "Current Approaches" (**Figure 6**):

- *similarity statistics* that compute correlations between local patches of the input, and
- *pooling* that combines together highly similar (well-correlated) patches.

These two computations are implemented hierarchically, processing over increasingly larger patches of the input image and resulting in a coarse mid-level representation of surfaces and their properties upon the first roughly feedforward processing wave. As a result of feature inference at multiple layers, the constructed surfaces are partially overlapping, providing information for depth ordering at the highest stages of this architecture (step 3). The resulting representations will be very coarse and probably inconsistent, so an iterative refinement of these representations by reapplying similarity and pooling operations over

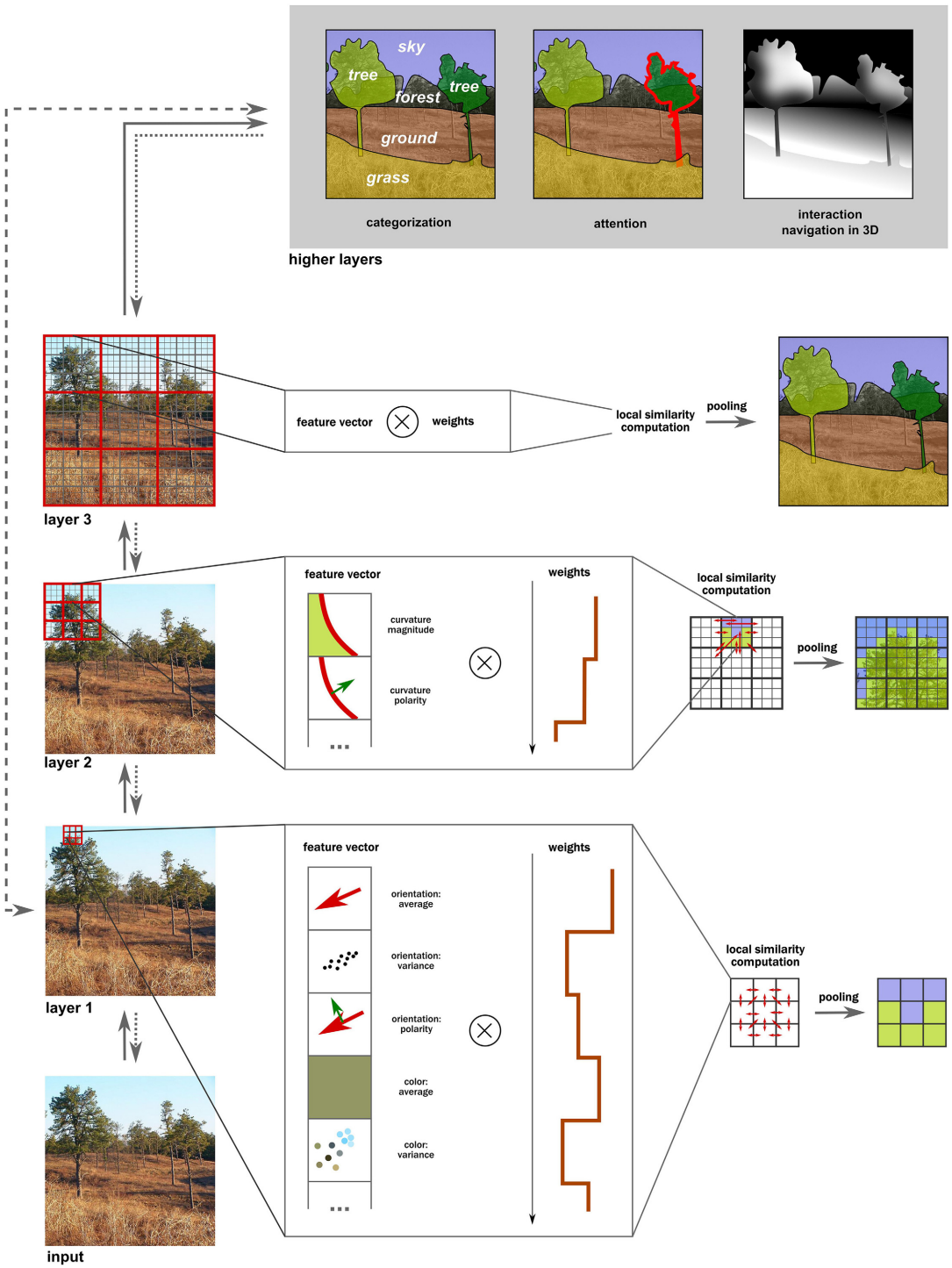


FIGURE 6 | Computation of intermediate representations in the visual hierarchy. In each layer, various features are extracted first at each location, forming a feature vector. Next, correlations are computed in the local neighborhood between each pair of a weighted feature pair, leading to similarity statistics (red arrows). (The optimal weights need to be learned by training the model.) Finally, these patches are pooled together into clusters that contain similar statistics. These new clusters are used in the next layers for the same similarity and pooling over increasingly larger neighborhoods. Note how the resulting intermediate representations are interpolated behind occlusions and are ordered in depth (e.g., the tree is in front of the forest). These representations can now be used for higher-level tasks such as categorization, attention to specific objects or interaction with them, or for navigation. They are also rather coarse initially (e.g., trees on the right are incorrectly lumped together), and can further be refined iteratively via feedback loops (if attention is directed to that region). Moreover, notice that not all steps must necessarily be carried out as certain shortcut routes (e.g., the gist computation) using simpler statistics can occur.

smaller parts of an input image is important as well (step 4; see also Wagemans et al., 2012b). We briefly discuss the role of feedback in Section "The Dynamic Nature of Intermediate Representations."

SIMILARITY ESTIMATION AND POOLING

Let us start by considering the output of a typical low-level computation such as edge detection, as illustrated in **Figure 5A**. The red arrows in this figure show the locations and orientations of salient edges in the image. While this is a useful description of potential boundary positions in the image, this information does not suffice to understand the organization of the image contents. In particular, it does not indicate which edges are likely to define the same surface, as shown in **Figure 5B**. At this stage the system only knows about separate salient edge positions, and further processing is needed to group both boundary and textural elements into coherent surfaces.

Finding which edges might group together can be achieved with a simple *similarity* measure, such as a correlation between two locations in an image. If the similarity is high, the two edges might belong to the same smooth contour (since edges at nearby locations of a smooth curve have similar orientation) or the same surface composed of similarly oriented elements (e.g., the wood texture in **Figure 4**). In contrast, a low similarity indicates a potential discontinuity in an image, or a second-order edge, just like the one between the ground and the object in **Figure 1C**.

Of course, similarity computation need not be restricted to oriented edges only and can be applied across other properties (e.g., spatial frequency, phase bands, color) and even across summary statistics within a local patch (e.g., mean and variance of orientation). Notice that by incorporating multiple cues, this single computation of similarity among the adjacent locations provides a natural approach to dealing with both boundary and textural cues in images. In particular, wherever there is sufficient dissimilarity, textural properties are actively used to generate boundary elements that are further used to construct full surface boundaries.

Freeman et al. (2013) provided evidence that such similarity measures are indeed computed early in the visual system. They constructed synthetic textures with specific higher-order statistical dependencies, such as marginal statistics, local cross-position, orientation, scale and adjacent-phase correlations, and demonstrated that such neurons in primate V2 (but not V1) were particularly sensitive to these built-in statistics, suggesting that V2 computes similarity between features. When used in textures, such summary statistics apparently are sufficient for the synthetic generation of similar-looking textures (Portilla and Simoncelli, 2000). When used on natural images, these statistics appear compatible with percept in peripheral vision (Freeman and Simoncelli, 2011; Freeman et al., 2013) and can also account for certain effects in crowding (Balas et al., 2009) and visual search (Rosenholtz et al., 2012).

Similarity statistics alone are not sufficient, however. While they are clearly useful in providing rich descriptions of the inputs, the number of parameters in the system increases dramatically since these statistics are computed pairwise between many small patches. Maintaining all these parameters does not appear to match our phenomenology where integrated shapes or regions dominate over local fragmented interpretations. Moreover, natural scenes contain substantial redundancy and the visual system appears to take advantage of it via efficient coding strategies (Attneave, 1954; Barlow, 1961; Simoncelli and Olshausen, 2001; Olshausen and Field, 2004; DiCarlo and Cox, 2007). For instance, Vinje and Gallant (2000) demonstrated that V1 neurons use a sparse encoding scheme that matches the sparse structure of natural scenes. Other researchers have demonstrated that sparsity constraint leads to the development of simple and complex cells in computational models (Olshausen and Field, 1996; Hyvärinen and Hoyer, 2000, 2001).

It thus appears that a higher-order statistic, one that would summarize similarity statistics, is necessary. We call this computation *pooling* to reflect the idea that separate units are now pooled together according to the strength of the previously computed pairwise correlations. Computationally, such pooling operation is very simple, for example, a single-link agglomerative clustering of patches that correlate above a certain threshold (Coates et al., 2012) or mean-shift (Paris and Durand, 2007; Rosenholtz et al., 2009). The threshold can be flexible (i.e., a free parameter in the model) reflecting individual differences between participants.

While either similarity or pooling have been utilized in various formats separately by many models, exploring the power of their combination is rare. Geisler and Super (2000) showed that a similar similarity and pooling scheme could account for a number of typical perceptual grouping displays. One successful demonstration of this combination on real images was reported by Yu et al. (2014) who found that a super-pixel segmentation followed by mean-shift clustering accounted surprisingly well for visual clutter perception. In a notable example that such scheme can be both powerful and efficient even for practical applications (due to parallelization), Coates et al. (2012), using K-means and agglomerative clustering, achieved robust unsupervised learning of face features using tens of millions of natural images.

HIERARCHICAL SIMILARITY ESTIMATION AND POOLING

While it would be possible to perform similarity and pooling globally across the whole image, such strategy would be very inefficient and probably not very accurate. Instead, we propose that these computations are performed hierarchically, such that first similarity and pooling are done locally, then over somewhat larger neighborhood using the newly inferred features, and finally globally using few but rather complex features that result from these computations at earlier stages.

The initial computation of a similarity and pooling would yield longer straight or curved segments (**Figure 7A**, right). A low correlation, on the other hand, would indicate the presence of second-order edges that are formed between adjacent surfaces with differently oriented elements. For example, in **Figure 7B**, left, there is no clear edge separating the object from the ground since their overall luminance is quite similar, and thus segmentation could not be done with a simple V1-like edge detection model. The desired segmentation becomes trivial when the difference in orientation content is observed. The dominant orientation of the object is different from that of the ground and

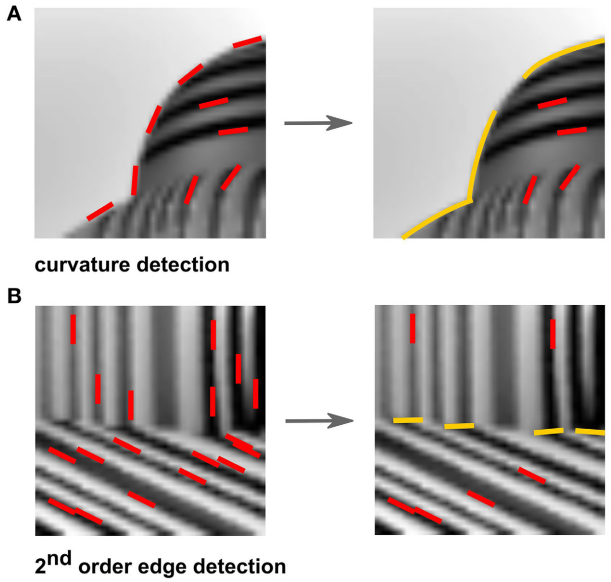


FIGURE 7 | Examples of feature inference. (A) Pooling similarly-oriented features (red) result in the inference of curved fragments (yellow). **(B)** In contrast, a dissimilarity between oriented features (red) result in the inference of second-order edges (yellow) between textures.

can therefore be used to determine a boundary between the two textures, which is indicated by the low similarity measure (**Figure 7B**, right).

Of course, detecting second-order edges in this fashion can also yield spurious results. Boundary element orientation can change significantly at inflection points (i.e., junctions) leading to low similarity measures, and yet these do not imply the presence of a second-order edge. One solution to the problem could be to use only sharp edges for defining boundaries, and otherwise assume that edges define textures (the insides of a surface). Consistent with this idea, Vilankar et al. (2014) reported that edges defining an occlusion tend to have steeper changes in contrast than non-occlusion edges (reflectance difference, surface change, cast shadows) and that a maximum likelihood classifier could predict the type of edge with 83% accuracy in their database. Another possibility is that junctions are not detected during the initial processing and only computed later when the global estimate of the shape is already available from the higher-level areas. Consistent with this idea, McDermott (2004) reported that participants were unable to report T-junctions using local natural image information (small patches of image) only (but see Hansen and Neumann, 2004; Weidenbacher and Neumann, 2009).

However, in general, the visual system is not so much interested in the features as such but in the surfaces they define. Other cues than boundaries can therefore be important in the local computations of which features should be combined into a single surface. As discussed above, convexity is an important cue for border-ownership assignment. Measuring consistency in edge polarity (where the brighter side is) can also provide information if they are likely to belong together (Kogo and Froyen, 2014). In fact, Geisler and Perry (2009) observed that edges with an inconsistent polarity are less likely to belong to the same contour. Recently, it has been reported that even low-level cues, such as the sharpness of an edge or local anisotropies in spectral power can be informative about figure-ground organization (Ramenahalli et al., 2014; Vilankar et al., 2014).

So, at each location where a boundary element has been found or inferred, we can list all these cues as a long vector and then compute the similarity between these vectors in the local neighborhood. Sufficiently similar locations are then pooled together, resulting in new, more complex features at a higher layer of this hierarchy. Now again, the similarity of these new features over larger scales can be computed, and similar features pooled together into even more complex features, such as parts of boundary (Brincat and Connor, 2006) or surface patches (Yamane et al., 2008) with a complex geometry. Finally, these features are pooled again over the entire image, producing the initial segmentation of an image into proto-surfaces.

NEURAL REPRESENTATION OF POOLED UNITS

By definition, a pooling operation combines outputs of several units and treats them as belonging to the same group (same contour, shape, or surface). Several alternatives have been proposed how such groups could be represented in the visual system. Perhaps the most straightforward way to implement this representation is by having dedicated grouping cells. Such idea has been used in a computational model of border-ownership assignment by Craft et al. (2007). They implemented neurons with donut-shaped receptive fields that can pool together units lying on that donut. However, such grouping cells have yet to be found in the visual system. It is possible however that cells with large curved receptive fields that exist in V4 might suffice to perform the border-ownership computation (as the authors themselves suggest on p. 4320 of their paper).

Another simple strategy is an increase of the mean neural response of units belonging to the same group (Roelfsema et al., 2004). However, this strategy also implies that only a single group can be maintained at a time. If another group needs to be processed, such as when shifting attention from one object to another, the integration computation would have to be performed again. While it may appear somewhat limiting, it should also be noted that in many tasks, such as multiple object tracking, observers show a rather poor ability to maintain representations of multiple groups at the same time.

A very different idea has been proposed by von der Malsburg (1981). He hypothesized that representations are held together by synchrony in neuronal firing. Such idea, if true, would in theory allow for multiple stable representations to co-exist in the visual system. While such synchrony has been observed in the visual cortex (Singer and Gray, 1995), its functional role is heavily debated, questioning whether it indeed plays a causal role in representing groups (Roskies, 1999; Roelfsema et al., 2004).

Finally, a similar idea has been put forward by Wehr and Laurent (1996). They provided evidence that locust's olfactory neurons fire in a certain unique temporal patterns to various combinations of scents. For example, while an overall response to an apple and to a mint and an apple scents might appear

comparable, at a finer temporal scale differences emerge in the number and timing of these higher frequency peaks (three peaks for the apple scent but only two for mint and apple). In other words, each stimulus receives a unique code of neural firing which can serve as a tag for belonging to a certain group. Importantly, just like binary code in computers, this code can accommodate a large number of stimuli without running into the combinatorial explosion.

THE DYNAMIC NATURE OF INTERMEDIATE REPRESENTATIONS

The visual processing need not stop with the feedforward formation of the intermediate representations. Probably the best we can expect at this first pass of processing is a very coarse representation capturing the most salient aspects of the input. For example, the initial representations may lack global consistency: it is likely that not all parts of an object will be bound into a single entity, and there can also be errors of the bounding of parts. For instance, the legs, body, and arms of a human body might be separate initially if there is not enough similarity between them. As a result, these parts may also have conflicting figure-ground assignments, such that the body is computed to be behind a chair but the legs are in front. If necessary for the task, a reconfiguration of these components could be formed iteratively until a global minimum is found, resulting in a stable percept of the configuration. For instance, the border-ownership model by Zhaoping (2005) resolves the direction of border-ownership by iteratively computing which side is more likely to be the figural side. The iterations are necessary because, for example, borders in concave parts of a shape might initially have the wrong border-ownership (toward the convex side) but over several iterations the assignment is gradually reversed since other parts of the global shape influence the decision that the concavity should be part of the whole shape. There are also cases where several interpretations are similarly plausible (e.g., the Necker cube or the vase-face figure; see Wagemans et al., 2012a), and thus iterative computations will lead to continuous switches between these interpretations.

In many cases, the refinement of representations will also be necessary. In particular, the initial representation formed in mid-level areas might only capture the gist of the input. Details will be necessarily lost due to agglomerative pooling operations. In order to extract finer details, representations in earlier layers can be reaccessed via feedback loops (indicated by backward arrows in **Figure 6**), as conceptualized by the Reverse Hierarchy Theory (Hochstein and Ahissar, 2002). Such feedback connections are abundant in the primate visual cortex and have been implicated to be important for various purposes (Felleman and Van Essen, 1991; Angelucci et al., 2002; Roelfsema et al., 2010; Arall et al., 2012). For example, intermediate representations could be used as saliency maps to direct attention to a particular part of an image or a particular feature (Walther and Koch, 2006; Russell et al., 2014). Then irrelevant inputs would be inhibited while the relevant ones would receive an enhanced weight (Mihalas et al., 2011; Arall et al., 2012; Wyatte et al., 2012), and the whole similarity and pooling computation would be repeated again. Such approach could be particularly important for resolving complicated parts of images that require high spatial resolution (Bullier, 2001), serial (or incremental) grouping of image features (Roelfsema, 2006), and could play a major role in learning features from input statistics (Roelfsema et al., 2010).

Iterative computations also provide the necessary flexibility for dealing with the inherently hierarchical composition of scenes. Consider, for example, **Figure 3B**, where all buildings could be represented by a single surface, or could be further divided into separate surfaces for each building, or even further for each window or any other detail in the image. Task demands, the mental state of an observer, and other factors can have a strong influence to the percept at any given moment. Utilizing the recurrent connections, the dynamics of the percept could be modeled in our framework by updating the pooling threshold (Sharon et al., 2006; Ion et al., 2013).

Of course, the proposed system need not be strictly hierarchical. For certain computations, it makes sense to have fast bypass routes (indicated by the dashed arrow at the top of **Figure 6**) whenever construction of intermediate representations is too slow or unnecessary, as could be the case for face detection where Viola and Jones' (2001) approach proves sufficient, or for a rapid scene categorization using the gist computation (Torralba and Oliva, 2003). Moreover, including such bypass routes naturally provides the visual system with the flexibility to both build detailed representations gradually and also to produce global impressions of the input statistics rapidly (Bar, 2004). The gist of the scene can provide informative priors (category, context, memory associations, and so on) that could guide processing and segmentation at intermediate layers (Peterson, 1994; Rao and Ballard, 1999; Oliva and Torralba, 2007).

Finally, we want to stress that although recurrent processing can improve surface representations and help in task performance, figure-ground segmentation does not require it. For example, Supèr and Lamme (2007) observed that removing most of feedback connections from higher visual areas to V1 reduced but did not abolish figure-ground perception. In fact, Qiu et al. (2007) reported that border-ownership signals emerge pre-attentively, and a purely feedforward model of figure-ground segmentation has been proposed by Supèr et al. (2010), consistent with a limited role of feedback in figure-ground assignment process (also see Arall et al., 2012, and Kogo and van Ee, 2014, for a discussion).

EVALUATING PERFORMANCE

The proposed architecture is meant to simulate the representations residing in mid-level vision. Given that this is not the final stage of the visual processing, evaluating the model's performance is not trivial. Often, models of vision are evaluated using standard object identification or scene segmentation datasets such as the ImageNet (Deng et al., 2009) or the Berkeley Segmentation Dataset (BSDS500; Arbeláez et al., 2011), where the goal for a model is to produce labels or segmentations as close as possible to the correct answers defined in that dataset. So, one simple solution for testing our architecture could be to extend it to perform one of these tasks. In this section, however, we discuss how blindly applying standard benchmarks can be misleading and highlight the need for good, carefully constructed tests and datasets that would help to detect shortcomings in the model and guide its development (Pinto et al., 2008).

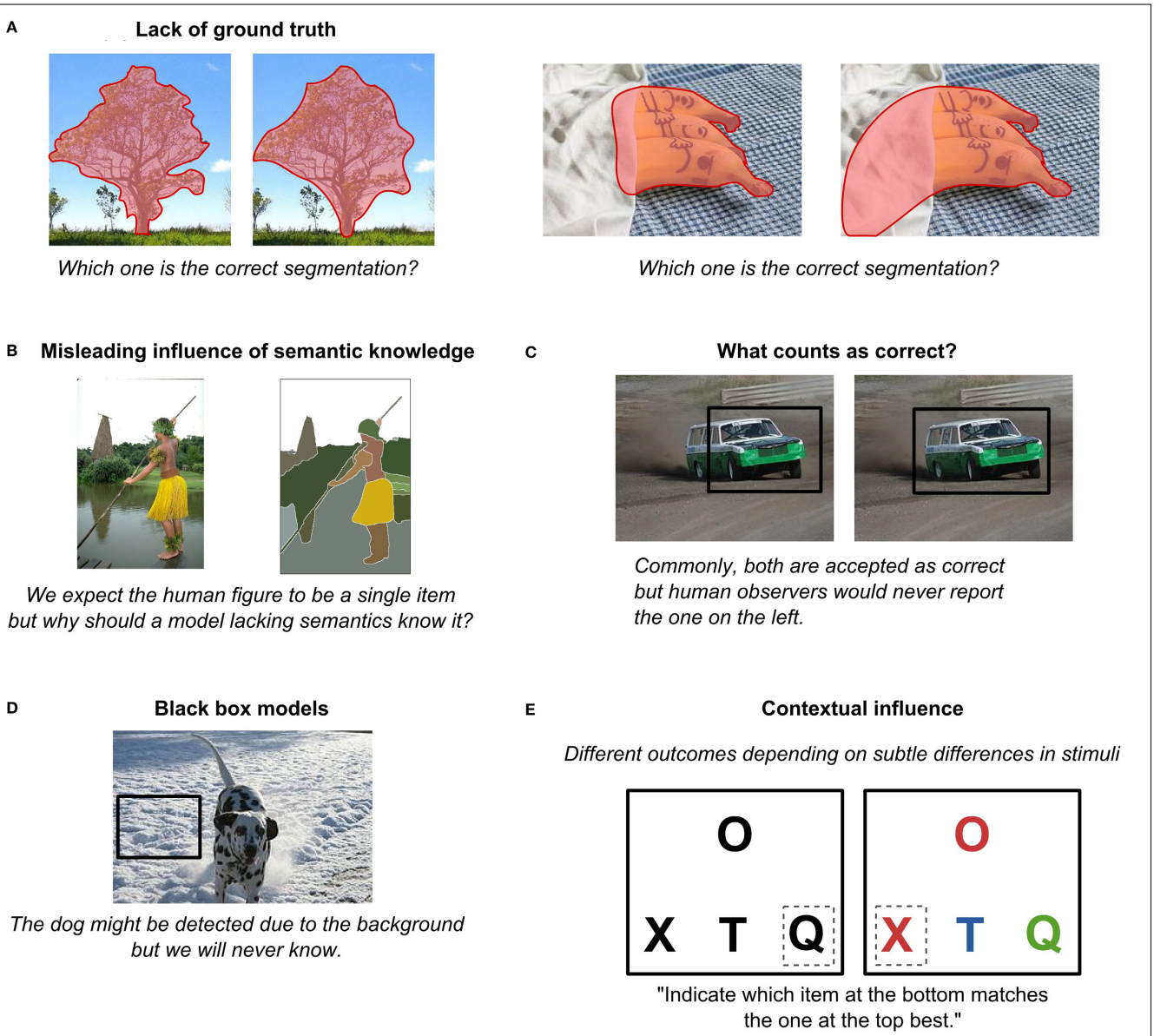


FIGURE 8 | Several potential problems with image datasets. (A) Lack of ground truth (photo credits: Sheila in Moonducks, 2010, and Camera Eye Photography, 2013). **(B)** Misleading influence of semantic knowledge (image from the dataset described in Arbeláez et al., 2011). **(C)** What counts as correct? (photo credit: bengt-re, 2009). **(D)** Black box models (photo credit: Berbezier, 2008; inspired by Landecker et al., 2013). **(E)** Contextual influence. All photos are available under the Creative Commons Attribution License.

First of all, there is always the question of the "ground truth." For example, which of the two segmentations in **Figure 8A**, left, is the ground truth? Both seem reasonable to a human observer and, in fact, they have been annotated by hand, making them, by definition, not objective. For example, smaller objects might be missing, subordinate categories might remain not annotated, and there may even be a disagreement among raters as to what constitutes an object and what is only a part of an object. While it is possible to step away from human raters altogether by obtaining ground-truth data using motion and depth information (Scharstein and Szeliski, 2002), only obtaining more precise measurements is not solving the major issue. In particular, the differences in ratings are largely driven not by imprecise annotation of boundaries but rather reflect individual differences in how people perceive images and what task they think they need to do. In other words, there is no ground truth to natural images because, as we have repeatedly pointed out in this paper, perception (and thus the definition of objects) is observer- and task-dependent. Another pertinent example to illustrate this point are images that contain occlusions (**Figure 8A**, right): What sense does it make to ask about the ground truth if it could be anything behind this occlusion, and we will never be able to tell from the incomplete data in the image? It only makes sense to ask what it looks like to a particular observer, so by forcing models

to match the "ground truth," we may in fact be pushing them to solve the wrong problem.

Similarly, raters are subject to their semantic knowledge. A human figure in a yellow skirt (**Figure 8B**) might be annotated as a human figure rather a body and a skirt separately. But for a model lacking extensive semantic knowledge (or statistical co-occurrences of higher-level entities), there is no reason why that yellow blob that happens to be a skirt could not be an occluder, unrelated to the human (like a flying broomstick). Regardless of whether or not the model combines the two into a single object, it does not mean that the model performed an incorrect initial segmentation. Thus, one needs to be very careful when defining what a correct segmentation is for a given model. A ground truth for one model might not be a ground truth for another.

Perhaps due to the lack of the ground truth, object localization is usually treated as accurate if at least 50% of the box containing the object overlaps with the box proposed by the model (Russakovsky et al., 2013). While finding the bounding box can often provide a good first guess of an object's location, as discussed in Section "Feature Interpolation," it is clear that this measure is far from the explicit human knowledge of the precise boundary and location of an object (**Figure 8C**). As a result, a model that is performing well according to this benchmark might be doing so in a completely different way than we expect or want. For example, an interesting study by Landecker et al. (2013) attempted to track down which parts of an image end up being most important for classification in hierarchical networks. Curiously, they found that sometimes object classification decision was based on completely irrelevant information, such as a background whose statistics happened to match certain object characteristics (**Figure 8D**). Szegedy et al. (2013) provided another striking example where they showed that in a standard deep learning setup for every image it was possible to construct another perceptually indistinguishable image that would nevertheless be categorized incorrectly by the same network. Similarly, analyzing top-performing models in the Image Net Large Scale Visual Recognition Challenge 2012, Russakovsky et al. (2013) observed that while such models tend to provide rather accurate locations of detected objects, their performance deteriorates significantly with more objects or clutter. If object shapes were explicitly represented, clutter would play a much smaller role in localization errors. Finally, Torralba and Efros (2011) showed that models trained on one dataset often perform poorly on another dataset for the same categories of objects. What these models are learning then remains rather questionable. (However, note that there are also examples of models that are capable of generalizing across datasets; see Razavian et al., 2014.)

Finally, a model's output is extremely context dependent. For example, imagine that you are presented with a screen with one stimulus at a top and three below, as in **Figure 8E**, left. You are asked to indicate which item at the bottom matches best the one at the top. Most people would probably choose "Q." But now imagine the stimuli were slightly changed (**Figure 8E**, right). Most people would now go for "X." But how would a model know that? It should somehow take it into account that the colors of "O" and "X" match while "T" and "Q" have some other colors and it should also know that color is more important to the visual system than shape. In other words, it needs a lot of basic knowledge, or basic reasoning skills, that are arguably even harder to build in the system than vision itself.

To avoid some of the listed problems, we suggest using artificially generated scenes, such as the one in **Figure 4**. They can be rendered to contain many difficult features that are abundant in natural images, including shadows, occlusions, clutter, and realistic textures. However, unlike natural images, such scenes do have a well-defined ground truth because they are rendered from three dimensional models. Moreover, since they lack known objects, a good model should be completely capable of dividing an image into surfaces all on its own with little or no mistakes. If the model fails, it is a clear indication that intermediate representations are not being constructed properly yet.

Another possibility to evaluate model's performance is to use the extracted statistics to synthesize new images. This approach was taken by Portilla and Simoncelli (2000) who convincingly showed that their texture synthesis model was accurate by presenting an original texture and synthetically generated ones using the computed statistics. Arguably, such approach would be much trickier to implement for a synthesis of objects (Portilla and Simoncelli's procedure fails to produce coherent objects) but then the model's performance would be more directly observable and would point to issues where the algorithm needs an improvement.

LIMITATIONS AND CONCLUSION

In this paper, we provided a synthesis of the classical works in psychology and recent advances in visual neuroscience and computer vision into a single unified framework of mid-level computations. We hypothesized that two basic mechanisms, namely, similarity estimation and pooling, implemented hierarchically and reiterated via recurrent processing, appear to be sufficient to account for the computational goals of mid-level vision and the available empirical data.

Admittedly, many details in the proposed framework remain speculative at this point. While we provided the sketch of each processing stage (including the initial feature extraction, junction and curvature computation, region growing, border-ownership assignment, and figure-ground organization), it remains to be seen to what extent these computations are robust in natural image processing. Similarly, while the framework can flexibly operate in various feature spaces, we do not propose which features in particular should be included and how different cues could be combined. Learning the weights of these cues is crucial if we want the proposed framework to apply for real images. One possibility is that the proposed computations can be implemented in the standard deep learning networks (by replacing nonlinearity and normalization steps with similarity estimation, and also performing feature inference instead of a simple filtering).

Another possibility, given that, unlike deep networks, the proposed architecture does not require semantic knowledge to be trained, observing certain feature co-occurrences (see Geisler, 2008, for a review) would be a simpler way to learn and adjust these weights. Even more powerful cues would be available from dynamic or stereo-defined inputs, given their tremendous role in bootstrapping the visual system (Ostrovsky et al., 2006, 2009)

Furthermore, we restricted the scope of our discussions to the construction of the initial figure-ground organization briefly after stimulus onset. This choice has been motivated by our interest to advance the idea that image segmentation and figure-ground organization might be rapid, nearly feedforward computations. However, recurrent processing loops are undoubtedly necessary to improve the constructed surfaces and meet task demands. We considered several alternatives for such computations in Section "Evaluating Performance," but the details of such top-down refinement remain to be worked out.

More than anything, this paper is our manifesto on the importance of intermediate computations. We are calling for a reconsideration of the role of mid-level vision and propose that implementing several basic mechanisms might provide an significant step forward in understanding the functioning of primate visual system.

ACKNOWLEDGMENTS

This work was supported in part by a Methusalem Grant (METH/08/02) awarded to Johan Wagemans from the Flemish Government. Jonas Kubilius is a research assistant of the Research Foundation—Flanders (FWO). We thank Naoki Kogo, Bart Machilsen, Lee de-Wit, Pieter Roelfsema, James Elder, Pieter Moors, Maarten Demeyer, and the reviewers of this paper for fruitful discussions and criticism, and Tom Putzeys for generating 3D scenes.

REFERENCES

Σ64. (2012). *The Forest in Yakushima, Kagoshima Pref., Japan*. Available online at: http://commons.wikimedia.org/wiki/File:Forest_in_Yakushima_08.jpg (Accessed November 26, 2014).

Allen, H. A., Humphreys, G. W., Colin, J., and Neumann, H. (2009). Ventral extrastriate cortical areas are required for human visual texture segmentation. *J. Vis.* 9:2. doi: 10.1167/9.9.2

Amir, O., Biederman, I., and Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vision Res.* 62, 35–43. doi: 10.1016/j.visres.2012.03.020

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupé, J.-M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *J. Neurosci.* 22, 8633–8646.

Arall, M., Romeo, A., and Supèr, H. (2012). "Role of feedforward and feedback projections in figure-ground responses," in *Visual Cortex—Current Status and Perspectives*, ed S. Molotchnikoff (InTech). Available online at: http://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/role-of-feedforward-and-feedback-projections-in-figure-ground-responses

Arbeláez, P., Maire, M., Fowlkes, C., and Malik, J. (2011). Contour detection and hierarchical image segmentation. *Pattern Anal. Mach. Intell. IEEE Trans.* 33, 898–916. doi: 10.1109/TPAMI.2010.161

Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., and Malik, J. (2014). "Multiscale combinatorial grouping," in *Computer Vision and Pattern Recognition*. Available online at: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/mcg/

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193. doi: 10.1037/h0054663

Balas, B., Nakano, L., and Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9:13. doi: 10.1167/9.12.13

Ban, H., Yamamoto, H., Hanakawa, T., Urayama, S., Aso, T., Fukuyama, H., et al. (2013). Topographic representation of an occluded object and the effects of spatiotemporal context in human early visual areas. *J. Neurosci.* 33, 16992–17007. doi: 10.1523/JNEUROSCI.1455-12.2013

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629. doi: 10.1038/nrn1476

Barenholtz, E., and Tarr, M. J. (2007). "Reconsidering the role of structure in vision," in *Categories in Use*, Vol. 47, eds A. Markman and B. Ross (San Diego, CA: Academic Press), 157–180.

Barlow, H. B. (1961). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, ed W. Rosenblith (Cambridge, MA: MIT Press), 217–234.

bengt-re. (2009). *Car 132*. Available online at: https://www.flickr.com/photos/bengt-re/3889978926/ (Accessed November 26, 2014).

Berbezier, M. (2008). *IMG_3536*. Available online at: https://www.flickr.com/photos/berbezier/2399217350/ (Accessed November 26, 2014).

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Blum, H. (1973). Biological shape and visual science (part I). *J. Theor. Biol.* 38, 205–287. doi: 10.1016/0022-5193(73)90175-6

Bosch, A., Zisserman, A., and Munoz, X. (2007). "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (New York, NY).

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010). "Learning mid-level features for recognition," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA).

Brincat, S. L., and Connor, C. E. (2006). Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49, 17–24. doi: 10.1016/j.neuron.2005.11.026

Bullier, J. (2001). Integrated model of visual processing. *Brain Res. Rev.* 36, 96–107. doi: 10.1016/S0165-0173(01)00085-6

Bushnell, B. N., Harding, P. J., Kosai, Y., and Pasupathy, A. (2011). Partial occlusion modulates contour-based shape encoding in primate area V4. *J. Neurosci.* 31, 4012–4024. doi: 10.1523/JNEUROSCI.4766-10.2011

Camera Eye Photography. (2013). *Trio, Quartet, or Quintet of Objects... Week #26 [26 of 52]*. Available online at: https://www.flickr.com/photos/camera_is_a_mirror_with_memory/9126653341/ (Accessed November 26, 2014).

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597. doi: 10.1523/JNEUROSCI.3726-05.2005

Chen, L. (1982). Topological structure in visual perception. *Science* 218, 699–700. doi: 10.1126/science.7134969

Chen, L. (2005). The topological approach to perceptual organization. *Vis. Cogn.* 12, 553–637. doi: 10.1080/13506280444000256

Choi, M. J., Torralba, A., and Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognit. Lett.* 33, 853–862. doi: 10.1016/j.patrec.2011.12.004

Coates, A., Karpathy, A., and Ng, A. (2012). "Emergence of object-selective features in unsupervised feature learning," *Presented at the Advances in Neural Information Processing Systems 25*. Available online at: http://nips.cc/Conferences/2012/Program/event.php?ID=3342

Cox, D. D. (2014). Do we understand high-level vision? *Curr. Opin. Neurobiol.* 25, 187–193. doi: 10.1016/j.conb.2014.01.016

Craft, E., Schütze, H., Niebur, E., and von der Heydt, R. (2007). A neural model of figure-ground organization. *J. Neurophysiol.* 97, 4310–4326. doi: 10.1152/jn.00203.2007

D'Antona, A. D., Perry, J. S., and Geisler, W. S. (2013). Humans make efficient use of natural image statistics when performing spatial interpolation. *J. Vis.* 13:11. doi: 10.1167/13.14.11

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, Vol. 1 (San Diego, CA), 886–893. doi: 10.1109/CVPR.2005.177

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (Miami, FL), 248–255. doi: 10.1109/CVPR.2009.5206848

Desimone, R., Albright, T., Gross, C., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4, 2051–2062.

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Dobbins, A., Zucker, S. W., and Cynader, M. S. (1987). Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature* 329, 438–441. doi: 10.1038/329438a0

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science* 293, 2470–2473. doi: 10.1126/science.1063414

Elder, J. H., and Zucker, S. W. (1996). "Computing contour closure," in *Computer Vision—ECCV'96*, eds B. Buxton and R. Cipolla (Berlin Heidelberg: Springer), 399–412. Available online at: http://link.springer.com/chapter/10.1007/BFb0015553

El-Shamayleh, Y., and Movshon, J. A. (2011). Neuronal responses to texture-defined form in macaque visual area V2. *J. Neurosci.* 31, 8543–8555. doi: 10.1523/JNEUROSCI.5974-10.2011

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature* 392, 598–601. doi: 10.1038/33402

Feldman, J. (1997). Regularity-based perceptual grouping. *Comput. Intell.* 13, 582–623. doi: 10.1111/0824-7935.00052

Feldman, J. (2003). What is a visual object? *Trends Cogn. Sci.* 7, 252–256. doi: 10.1016/S1364-6613(03)00111-6

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Felzenszwalb, P. F., and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *Int. J. Comput. Vis.* 59, 167–181. doi: 10.1023/B:VISI.0000022288.19776.77

Fidler, S., Boben, M., and Leonardis, A. (2009). "Learning hierarchical compositional representations of object structure," in *Object Categorization: Computer and Human Vision Perspectives*, *1st Edn.*, eds S. J. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr (New York, NY: Cambridge University Press), 196–215.

Freeman, J., and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1201. doi: 10.1038/nn.2889

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* 16, 974–981. doi: 10.1038/nn.3402

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192. doi: 10.1146/annurev.psych.58.110405.085632

Geisler, W. S., and Perry, J. S. (2009). Contour statistics in natural images: grouping across occlusions. *Vis. Neurosci.* 26, 109–121. doi: 10.1017/S0952523808080875

Geisler, W. S., and Super, B. J. (2000). Perceptual organization of two-dimensional patterns. *Psychol. Rev.* 107, 677–708. doi: 10.1037/0033-295X.107.4.677

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton, MI: Mifflin.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available online at: http://arxiv.org/abs/1311.2524

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Percept. Psychophys.* 55, 48–121. doi: 10.3758/BF03206880

Hansen, T., and Neumann, H. (2004). Neural Mechanisms for the Robust Representation of Junctions. *Neural Comput.* 16, 1013–1037. doi: 10.1162/089976604773135087

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. (2014). "Simultaneous detection and segmentation," in *European Conference on Computer Vision (ECCV)*. Available online at: http://arxiv.org/abs/1407.1808

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochstein, S., and Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 791–804. doi: 10.1016/S0896-6273(02)01091-7

Hubel, D. H., and Wiesel, T. N. (1961). Integrative action in the cat's lateral geniculate body. *J. Physiol.* 155, 385–398.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hubel, D. H., and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* 28, 229–289.

Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517. doi: 10.1037/0033-295X.99.3.480

Hung, C.-C., Carlson, E. T., and Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron* 74, 1099–1113. doi: 10.1016/j.neuron.2012.04.029

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Hyvärinen, A., and Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* 12, 1705–1720. doi: 10.1162/089976600300015312

Hyvärinen, A., and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* 41, 2413–2423. doi: 10.1016/S0042-6989(01)00114-6

Ion, A., Carreira, J., and Sminchisescu, C. (2013). Probabilistic joint image segmentation and labeling by figure-ground composition. *Int. J. Comput. Vis.* doi: 10.1007/s11263-013-0663-7

Ito, M., and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.* 24, 3313–3324. doi: 10.1523/JNEUROSCI.4364-03.2004

Jehee, J. F. M., Lamme, V. A. F., and Roelfsema, P. R. (2007). Boundary assignment in a recurrent network architecture. *Vision Res.* 47, 1153–1165. doi: 10.1016/j.visres.2006.12.018

Joo, S. J., and Murray, S. O. (2014). Contextual effects in human visual cortex depend on surface structure. *J. Neurophysiol.* 111, 1783–1791. doi: 10.1152/jn.00671.2013

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. *Riv. Psicol.* 49, 7–30.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311.

Kayaert, G., Biederman, I., Op de Beeck, H. P., and Vogels, R. (2005a). Tuning for shape dimensions in macaque inferior temporal cortex. *Eur. J. Neurosci.* 22, 212–224. doi: 10.1111/j.1460-9568.2005.04202.x

Kayaert, G., Biederman, I., and Vogels, R. (2005b). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cereb. Cortex* 15, 1308–1321. doi: 10.1093/cercor/bhi014

Kogo, N., and Froyen, V. (2014). "Border-ownership computation reflecting consistency of surface properties," *Presented at the Vision Sciences Society* (St. Pete Beach, FL).

Kogo, N., Strecha, C., Van Gool, L., and Wagemans, J. (2010). Surface construction by a 2-D differentiation–integration process: a neurocomputational model for perceived border ownership, depth, and lightness in Kanizsa figures. *Psychol. Rev.* 117, 406–439. doi: 10.1037/a0019076

Kogo, N., and van Ee, R. (2014). "Neural mechanisms of figure-ground organization: border-ownership, competition and perceptual switching," in *Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford, UK: Oxford University Press).

Kornblith, S., Cheng, X., Ohayon, S., and Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron* 79, 766–781. doi: 10.1016/j.neuron.2013.06.015

Kosai, Y., El-Shamayleh, Y., Fyall, A. M., and Pasupathy, A. (2014). The role of visual area V4 in the discrimination of partially occluded shapes. *J. Neurosci.* 34, 8570–8584. doi: 10.1523/JNEUROSCI.1375-14.2014

Kourtzi, Z., and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science* 293, 1506–1509. doi: 10.1126/science.1061133

Kreiman, G. (2013). "Computational models of visual object recognition," in *Principles of Neural Coding*, Vol. 1-0. (CRC Press), 565–580. Available online at: http://www.crcnetbase.com/doi/abs/10.1201/b14756-33 doi: 10.1201/b14756-33

Krizhevsky, A., Sutskever,I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc), 1097–1105. Available online

at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Kubilius, J., Wagemans, J., and Op de Beeck, H. P. (2014). Encoding of configural regularity in the human visual system. *J. Vis.* 14:11. doi: 10.1167/14.9.11

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.* 16, 37–68.

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* 15, 1605–1615.

Lamme, V. A. F., Rodriguez-Rodriguez, V., and Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cereb. Cortex* 9, 406–413. doi: 10.1093/cercor/9.4.406

Landecker, W., Thomure, M. D., Bettencourt, L. M. A., Mitchell, M., Kenyon, G. T., and Brumby, S. P. (2013). "Interpreting individual classifications of hierarchical networks," in *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (Singapore).

Larsson, J., Landy, M. S., and Heeger, D. J. (2006). Orientation-selective adaptation to first- and second-order patterns in human visual cortex. *J. Neurophysiol.* 95, 862–881. doi: 10.1152/jn.00668.2005

Layton, O. W., Mingolla, E., and Yazdanbakhsh, A. (2012). Dynamic coding of border-ownership in visual cortex. *J. Vis.* 12:8. doi: 10.1167/12.13.8

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lee, T. S., Mumford, D., and Yuille, A. (1992). "Texture segmentation by minimizing vector-valued energy functionals: the coupled-membrane model," in *Computer Vision—ECCV'92*, ed G. Sandini (Berlin Heidelberg: Springer), 165–173. Available online at: http://link.springer.com/chapter/10.1007/3-540-55426-2_19

Leibe, B., Leonardis, A., and Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vis.* 77, 259–289. doi: 10.1007/s11263-007-0095-3

Lescroart, M. D., Biederman, I., Yue, X., and Davidoff, J. (2010). A cross-cultural study of the representation of shape: sensitivity to generalized cone dimensions. *Visual Cogn.* 18, 50–66. doi: 10.1080/13506280802507806

Li, F. F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9596–9601. doi: 10.1073/pnas.092277599

Li, G., Yao, Z., Wang, Z., Yuan, N., Talebi, V., Tan, J., et al. (2014). Form-cue invariant second-order neuronal responses to contrast modulation in primate area V2. *J. Neurosci.* 34, 12081–12092. doi: 10.1523/JNEUROSCI.0211-14.2014

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

Lu, C., and Tang, X. (2014). Surpassing human-level face verification performance on LFW with GaussianFace. *arXiv:1404.3840 [cs, Stat]*. Available online at: http://arxiv.org/abs/1404.3840

Marcar, V. L., Raiguel, S. E., Xiao, D., and Orban, G. A. (2000). Processing of kinetically defined boundaries in areas V1 and V2 of the macaque monkey. *J. Neurophysiol.* 84, 2786–2798.

Mareschal, I., and Baker, C. L. (1998). A cortical locus for the processing of contrast-defined contours. *Nat. Neurosci.* 1, 150–154. doi: 10.1038/401

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

McDermott, J. (2004). Psychophysics with junctions in real images. *Perception* 33, 1101–1127. doi: 10.1068/p5265

Mihalas, S., Dong, Y., von der Heydt, R., and Niebur, E. (2011). Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7583–7588. doi: 10.1073/pnas.1014655108

Mumford, D., and Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.* 42, 577–685. doi: 10.1002/cpa.3160420503

Murray, R. F., Sekuler, A. B., and Bennett, P. J. (2001). Time course of amodal completion revealed by a shape discrimination task. *Psychon. Bull. Rev.* 8, 713–720. doi: 10.3758/BF03196208

Nakayama, K., He, Z. J., and Shimojo, S. (1995). "Visual surface representation: a critical link between lower-level and higher-level vision," in *Visual cognition: An invitation to cognitive science, 2nd Edn*, Vol. 2, eds S. M. Kosslyn and D. N. Osherson (Cambridge, MA: The MIT Press), 1–70. Available online at: http://visionlab.harvard.edu/members/ken/Papers/077NKHeShimojoMIT1995b.pdf

Nothdurft, H. C. (1994). Common properties of visual segmentation. *Ciba Found. Symp.* 184, 245–259. discussion: 260–271.

Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175. doi: 10.1023/A:1011139631724

Oliva, A., and Torralba, A. (2006). "Chapter 2 Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research*, Vol. 155, Part B, eds S. L. Macknik, L. M. Martinez, J. -M. Alonso, P. U. Tse and S. Martinez-Conde (Elsevier), 23–36. Available online at: http://www.sciencedirect.com/science/article/pii/S0079612306550022

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B. A., and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252. doi: 10.1038/nn767

Ostrovsky, Y., Andalman, A., and Sinha, P. (2006). Vision following extended congenital blindness. *Psychol. Sci.* 17, 1009–1014. doi: 10.1111/j.1467-9280.2006.01827.x

Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., and Sinha, P. (2009). Visual parsing after recovery from blindness. *Psychol. Sci.* 20, 1484–1491. doi: 10.1111/j.1467-9280.2009.02471.x

Pan, Y., Chen, M., Yin, J., An, X., Zhang, X., Lu, Y., et al. (2012). Equivalent representation of real and illusory contours in macaque V4. *J. Neurosci.* 32, 6760–6770. doi: 10.1523/JNEUROSCI.6140-11.2012

Paris, S., and Durand, F. (2007). "A topological approach to hierarchical segmentation using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07* (Minneapolis, MN), 1–8.

Pasupathy, A., and Connor, C. E. (1999). Responses to contour features in macaque area V4. *J. Neurophysiol.* 82, 2490–2502.

Pasupathy, A., and Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* 86, 2505–2519.

Pasupathy, A., and Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.* 5, 1332–1338. doi: 10.1038/nn972

Peelen, M. V., and Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *J. Neurophysiol.* 93, 603–608. doi: 10.1152/jn.00513.2004

Peterson, M. A. (1994). Object recognition processes can and do operate before figure-ground organization. *Curr. Dir. Psychol. Sci.* 3, 105–111. doi: 10.1111/1467-8721.ep10770552

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput. Biol.* 4:e27. doi: 10.1371/journal.pcbi.0040027

Poort, J., Raudies, F., Wannig, A., Lamme, V. A. F., Neumann, H., and Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron* 75, 143–156. doi: 10.1016/j.neuron.2012.04.032

Portilla, J., and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40, 49–70. doi: 10.1023/A:1026553619983

Potter, M. C. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. Learn. Mem.* 2, 509–522. doi: 10.1037/0278-7393.2.5.509

Purves, D., Monson, B. B., Sundararajan, J., and Wojtach, W. T. (2014). How biological vision succeeds in the physical world. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4750–4755. doi: 10.1073/pnas.1311309111

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition* 80, 127–158. doi: 10.1016/S0010-0277(00)00156-6

Qiu, F. T., Sugihara, T., and von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10, 1492–1499. doi: 10.1038/nn1989

Quiroga, R. Q., Mukamel, R., Isham, E. A., Malach, R., and Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3599–3604. doi: 10.1073/pnas.0707043105

Ramenahalli, S., Mihalas, S., and Niebur, E. (2014). Local spectral anisotropy is a valid cue for figure-ground organization in natural scenes. *Vision Res.* 103, 116–126. doi: 10.1016/j.visres.2014.08.012

Ramsden, B. M., Hung, C. P., and Roe, A. W. (2001). Real and illusory contour processing in area V1 of the primate: a cortical balancing act. *Cereb. Cortex* 11, 648–665. doi: 10.1093/cercor/11.7.648

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rauschenberger, R., Liu, T., Slotnick, S. D., and Yantis, S. (2006). Temporally unfolding neural representation of pictorial occlusion. *Psychol. Sci.* 17, 358–364. doi: 10.1111/j.1467-9280.2006.01711.x

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. *arXiv:1403.6382 [cs]*. Available online at: http://arxiv.org/abs/1403.6382

Regan, D. (2000). *Human Perception of Objects: Early Visual Processing of Spatial form Defined by Luminance, Color, Texture, Motion, and Binocular Disparity*. Sunderland, MA: Sinauer Associates.

Reza. (2009). *Mountains*. Available online at: https://www.flickr.com/photos/r-z/3690169484/ (Accessed November 26, 2014).

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Ringach, D. L., and Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Res.* 36, 3037–3050. doi: 10.1016/0042-6989(96)00062-4

Rodríguez-Sánchez, A. J., and Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. *PLoS ONE* 7:e42058. doi: 10.1371/journal.pone.0042058

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Roelfsema, P. R., Lamme, V. A. F., and Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nat. Neurosci.* 7, 982–991. doi: 10.1038/nn1304

Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., and Bosch, H. (2002). Figure—ground segregation in a recurrent network architecture. *J. Cogn. Neurosci.* 14, 525–537. doi: 10.1162/08989290260045756

Roelfsema, P. R., van Ooyen, A., and Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends Cogn. Sci.* 14, 64–71. doi: 10.1016/j.tics.2009.11.005

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., and Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *J. Vis.* 12:14. doi: 10.1167/12.4.14

Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., and Wattenberg, M. (2009). "An intuitive model of perceptual grouping for HCI design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1331–1340. Available online at: http://dl.acm.org/citation.cfm?id=1518903

Roskies, A. L. (1999). The binding problem. *Neuron* 24, 7–9. doi: 10.1016/S0896-6273(00)80817-X

Russakovsky, O., Deng, J., Huang, Z., Berg, A. C., and Fei-Fei, L. (2013). "Detecting avocados to zucchinis: what have we done, and where are we going?" in *2013 IEEE International Conference on Computer Vision (ICCV)* (Sydney, NSW), 2064–2071. doi: 10.1109/ICCV.2013.258

Russell, A. F., Mihalaş, S., von der Heydt, R., Niebur, E., and Etienne-Cummings, R. (2014). A model of proto-object based saliency. *Vision Res.* 94, 1–15. doi: 10.1016/j.visres.2013.10.005

Scharstein, D., and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* 47, 7–42. doi: 10.1023/A:1014573219977

Schmid, A. M., Purpura, K. P., and Victor, J. D. (2014). Responses to orientation discontinuities in V1 and V2: physiological dissociations and functional implications. *J. Neurosci.* 34, 3559–3578. doi: 10.1523/JNEUROSCI.2293-13.2014

Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Comput.* 4, 234–242. doi: 10.1162/neco.1992.4.2.234

Sekuler, A. B., and Palmer, S. E. (1992). Perception of partly occluded objects: a microgenetic analysis. *J. Exp. Psychol.* 121, 95–111. doi: 10.1037/0096-3445.121.1.95

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). OverFeat: integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229 [cs]*. Available online at: http://arxiv.org/abs/1312.6229

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Sharon, E., Galun, M., Sharon, D., Basri, R., and Brandt, A. (2006). Hierarchy and adaptivity in segmenting visual scenes. *Nature* 442, 810–813. doi: 10.1038/nature04977

Sheila in Moonducks. (2010). *Tree*. Available online at: https://www.flickr.com/photos/aspis7/5075169756/ (Accessed November 26, 2014).

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 888–905. doi: 10.1109/34.868688

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193

Singer, W., and Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 18, 555–586. doi: 10.1146/annurev.ne.18.030195.003011

Snowdog. (2005). *Leuven - Grote Markt*. Available online at: http://commons.wikimedia.org/wiki/File:Leuven-Grote-Markt.jpg (Accessed November 26, 2014).

Song, Y., and Baker, C. L. (2007). Neuronal response to texture- and contrast-defined boundaries in early visual cortex. *Visual Neurosci.* 24, 65–77. doi: 10.1017/S0952523807070113

Supèr, H., and Lamme, V. A. F. (2007). Altered figure-ground perception in monkeys with an extra-striate lesion. *Neuropsychologia* 45, 3329–3334. doi: 10.1016/j.neuropsychologia.2007.07.001

Supèr, H., Romeo, A., and Keil, M. (2010). Feed-forward segmentation of figure-ground and assignment of border-ownership. *PLoS ONE* 5:e10705. doi: 10.1371/journal.pone.0010705

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). Going deeper with convolutions. *arXiv:1409.4842 [cs]*. Available online at: http://arxiv.org/abs/1409.4842

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*. Available online at: http://arxiv.org/abs/1312.6199

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). "DeepFace: closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available online at: https://www.facebook.com/publications/546316888800776/ doi: 10.1109/CVPR.2014.220

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Thielscher, A., and Neumann, H. (2003). Neural mechanisms of cortico–cortical interaction in texture boundary detection: a modeling approach. *Neuroscience* 122, 921–939. doi: 10.1016/j.neuroscience.2003.08.050

Thielscher, A., and Neumann, H. (2005). Neural mechanisms of human texture processing: texture boundary detection and visual search. *Spat. Vis.* 18, 227–257. doi: 10.1163/1568568053320594

Thielscher, A., and Neumann, H. (2008). Globally consistent depth sorting of overlapping 2D surfaces in a model using local recurrent interactions. *Biol. Cybern.* 98, 305–337. doi: 10.1007/s00422-008-0211-7

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Todd, J. T., Weismantel, E., and Kallie, C. S. (2014). On the relative detectability of configural properties. *J. Vis.* 14:18. doi: 10.1167/14.1.18

Torralba, A., and Efros, A. A. (2011). "Unbiased look at dataset bias," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI), 1521–1528.

Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. *Network (Bristol, England)* 14, 391–412. doi: 10.1088/0954-898X/14/3/302

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science* 311, 670–674. doi: 10.1126/science.1119983

Tschechne, S., and Neumann, H. (2014). Hierarchical representation of shapes in visual cortex–from localized features to figural shape segregation. *Front. Comput. Neurosci.* 8:93. doi: 10.3389/fncom.2014.00093

Tse, P. U. (1999). Volume completion. *Cogn. Psychol.* 39, 37–68. doi: 10.1006/cogp.1999.0715

Ullman, S., and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 992–1006. doi: 10.1109/34.99234

van Lier, R., van der Helm, P., and Leeuwenberg, E. (1994). Integrating global and local aspects of visual occlusion. *Perception* 23, 883–903. doi: 10.1068/p230883

van Lier, R. (1999). Investigating global effects in visual occlusion: from a partly occluded square to the back of a tree-trunk. *Acta Psychol. (Amst.)* 102, 203–220.

Vilankar, K. P., Golden, J. R., Chandler, D. M., and Field, D. J. (2014). Local edge statistics provide information regarding occlusion and nonocclusion edges in natural scenes. *J. Vis.* 14:13. doi: 10.1167/14.9.13

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001* (Vol. 1, I–511–I–518) (Kauai, HI). doi: 10.1109/CVPR.2001.990517

Vogels, R., Biederman, I., Bar, M., and Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *J. Cogn. Neurosci.* 13, 444–453. doi: 10.1162/08989290152001871

von der Heydt, R., and Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J. Neurosci.* 9, 1731–1748.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science* 224, 1260–1262. doi: 10.1126/science.6539501

von der Malsburg, C. (1981). *The Correlation Theory of Brain Function.* Departmental Technical Report, MPI. Available online at: http://cogprints.org/1380/

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012a). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bull.* 138, 1172–1217. doi: 10.1037/a0029333

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., et al. (2012b). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* 138, 1218–1252. doi: 10.1037/a0029334

Wagemans, J., Lamote, C., and Gool, L. V. (1997). Shape equivalence under perspective and projective transformations. *Psychon. Bull. Rev.* 4, 248–253. doi: 10.3758/BF03209401

Wagemans, J., Van Gool, L., Lamote, C., and Foster, D. H. (2000). Minimal information to determine affine shape equivalence. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 443–468. doi: 10.1037/0096-1523.26.2.443

Walther, D., and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Netw.* 19, 1395–1407. doi: 10.1016/j.neunet.2006.10.001

Wehr, M., and Laurent, G. (1996). Odour encoding by temporal sequences of firing in oscillating neural assemblies. *Nature* 384, 162–166. doi: 10.1038/384162a0

Weidenbacher, U., and Neumann, H. (2009). Extraction of surface-related features in a recurrent model of V1-V2 interactions. *PLoS ONE* 4:e5909. doi: 10.1371/journal.pone.0005909

Witkin, A. P., and Tenenbaum, J. M. (1983). On the role of structure in vision. *Hum. Mach. Vis.* 1, 481–543. doi: 10.1016/B978-0-12-084320-6.50022-0

Wyatte, D., Herd, S., Mingus, B., and O'Reilly, R. (2012). The role of competitive inhibition and top-down feedback in binding during object recognition. *Front. Psychol.* 3:182. doi: 10.3389/fpsyg.2012.00182

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., and Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360. doi: 10.1038/nn.2202

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yu, C.-P., Samaras, D., and Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *J. Vis.* 14:4. doi: 10.1167/14.7.4

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron* 47, 143–153. doi: 10.1016/j.neuron.2005.04.005

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611.

Zucker, S. W. (2014). "Border inference and border ownership. The challenge of integrating geometry and topology," in *Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford, UK: Oxford University Press). doi: 10.1093/oxfordhb/9780199686858.013.020

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2014; accepted: 17 November 2014; published online: 12 December 2014.

Citation: Kubilius J, Wagemans J and Op de Beeck HP (2014) A conceptual framework of computations in mid-level vision. Front. Comput. Neurosci. ***8****:158. doi: 10.3389/fncom.2014.00158*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Kubilius, Wagemans and Op de Beeck. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 August 2014
doi: 10.3389/fncom.2014.00093

Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation

Stephan Tschechne* and Heiko Neumann

Faculty of Engineering and Computer Science (with Psychology and Education), Institute of Neural Information Processing, Ulm University, Ulm, Germany

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
John K. Tsotsos, York University, Canada
Jonathan R. Williford, Johns Hopkins University, USA

****Correspondence:***
Stephan Tschechne, Institute of Neural Information Processing, Ulm University, Albert-Einstein-Allee 1, 89069 Ulm, Germany
e-mail: stephan.tschechne@uni-ulm.de

Visual structures in the environment are segmented into image regions and those combined to a representation of surfaces and prototypical objects. Such a perceptual organization is performed by complex neural mechanisms in the visual cortex of primates. Multiple mutually connected areas in the ventral cortical pathway receive visual input and extract local form features that are subsequently grouped into increasingly complex, more meaningful image elements. Such a distributed network of processing must be capable to make accessible highly articulated changes in shape boundary as well as very subtle curvature changes that contribute to the perception of an object. We propose a recurrent computational network architecture that utilizes hierarchical distributed representations of shape features to encode surface and object boundary over different scales of resolution. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in visual cortex, namely areas V1–V4 and IT. We suggest that multiple specialized component representations interact by feedforward hierarchical processing that is combined with feedback signals driven by representations generated at higher stages. Based on this, global configurational as well as local information is made available to distinguish changes in the object's contour. Once the outline of a shape has been established, contextual contour configurations are used to assign border ownership directions and thus achieve segregation of figure and ground. The model, thus, proposes how separate mechanisms contribute to distributed hierarchical cortical shape representation and combine with processes of figure-ground segregation. Our model is probed with a selection of stimuli to illustrate processing results at different processing stages. We especially highlight how modulatory feedback connections contribute to the processing of visual input at various stages in the processing hierarchy.

Keywords: ventral pathway, distributed representation, figure-ground segregation, modulatory feedback, computational model

1. INTRODUCTION

We visually perceive our environment as a stable and comprehensive combination of objects, where we can easily identify objects and persons and we efficently analyse geometrical cues that allow a precise navigation and interaction. This happens so effortlessy and accurately that it is absolutely counterintuitive that this is an extraordinary achievement of our brain. The visual system of mammals achieves this result from input that is captured at the retinal level after light has been projected through the eye and hits light-sensitive neurons. The perception of our environment starts at this local level where our position, the direction of our gaze, the current illumination, an object's surface properties and its location relative to others causes a set of neurons in the retina to respond with increased activation that is a function of received light intensity. How the visual system transforms this concert of local visual inputs into a stable and informative perception of surfaces and objects is subject to intense research. Since the pioneering works on neural principles by Hubel and Wiesel (1959) many insights into cortical processing of visual input has been discovered. Neurophysiologists agree that the processing in the mammalian brain is performed in a hierarchical way and processing is organized into various specialized brain areas (Felleman and Van Essen, 1991). Those brain areas receive connections from preceding processing stages, but also from regions later in the processing stream (Markov et al., 2013). Early areas in visual cortex are retinotopically arranged (Hubel and Wiesel, 1962), which means that juxtaposed retinal locations are mapped to juxtaposed locations in visual cortex, with foveal positions being represented at a higher resolution. Individual assemblies of neurons become activated when their preferred stimulus is presented in their receptive field (Hubel and Wiesel, 1962). With progression in the visual pathway, the size of those RFs increase from sizes smaller than one degree of visual angle to sizes covering a good part of the visual field. In parallel, the tuning toward the preferred stimulus changes from simple features like oriented contrasts (Hubel and Wiesel, 1962) to complex ones like image features, figure-ground-related cues, object categories or faces. Processing along the visual pathway is organized into two streams (Ungerleider and Haxby, 1994), the ventral stream that exhibits a tuning toward movement

and position, whereas the dorsal stream processes shapes and objects.

However, most of the achievements that the visual system exhibits, like the abilities to generalize and its robustness and adaptability, most probably stem from connections that connect higher cortical areas with lower ones (Hupé et al., 1998; Markov et al., 2013). Those feedback connections are believed to play an important role in visual processing, as they enrich local activations with contextual information that is represented at higher visual areas. We propose that on the way from generalizing early local features to higher meaningful representations, the role of object boundaries plays an essential part. Contrasts indicate spatial changes in local illumination which might coincide with object boundaries that allow segregation from background. However, contrasts indicating a real transition from one object to another or from the object to the background must be separated from those indicating an illumination change and those caused by textured regions. This must be accomplished using contextual information. The region delimited by such a boundary is a surface with locally constant parameters, and a set of surfaces forms objects, scenes and eventually our complete visual environment. We believe that the processing capabilities of early and intermediate stages of visual cortex are used to transform local representation into an intermediate, more meaningful representation of contours, shapes and surfaces. Following those ideas, we propose that a stable representation of shape may be established by interacting assemblies that are each devoted to specific features properties. We thus propose a hierarchical model of 2-dimensional shape representation that incorporates processing at low and intermediate areas of visual cortex. Each model area consists of a three-stage processing cascade of initial filtering, application of modulatory feedback effects and center-surround interactions leading to an activity normalization (Carandini and Heeger, 1994; Carandini et al., 1999; Kouh and Poggio, 2008; Carandini and Heeger, 2012). The functional effects of this columnar cascade can roughly be mapped onto compartments of cortical area subdivisions [as suggested in (Self et al., 2012)].

Our model combines the representation of visual shapes with mechanisms for figure-ground segregation on the basis of assigning border ownership and incorporates a distributed representation of local contour curvature over different cortical areas. In our model we emphasize the computational role of feedforward and feedback mechanisms (Grossberg, 1980; Edelman, 1993) to generate a hierarchical distributed representation of shape information. The feedback amplifies the sensory signal such that the subsequent competition between neurons builds a competitive advantage (Tsotsos, 1988; Girard and Bullier, 1989; Desimone, 1998; Roelfsema et al., 2002; Reynolds and Heeger, 2009). Boundaries and their orientation are represented after intial processing in model area V1 and a grouping stage in model area V2. Contextual boundary configurations are also represented at a coarser spatial level at model V2 and V4 to achieve selectivities toward contour curvature. With the influence of feedback, those cells are enhanced at lower stages that contribute to a matching bottom-up signal.

The output of our model is a representation of shapes and shape segments where contextually compatible boundary information benefits from recurrent feedback connections. Such a representation could provide input to subsequent processing stages for e.g., object classification tasks, which would clearly benefit from the enhanced representation.

This model extends previous own works (Neumann and Mingolla, 2001; Hansen and Neumann, 2004; Weidenbacher and Neumann, 2009) but introduces functional properties that have been inspired by the works of other groups. A model of curvature representation can also be found in Cadieu et al. (2007). The authors modeled physiological findings of the same group (Pasupathy and Connor, 1999; Connor et al., 2007) that has focussed on the dynamics of contour processing (Yau et al., 2013). Cell representations from early visual areas are combined to intermediate-level shape descriptors are used in a computational model by Rodríguez-Sánchez and Tsotsos (2012). Riesenhuber and Poggio (1999, 2000); Mutch and Lowe (2008) released very powerful models of object and object class categorization in a hierarchical modeling approach. The physiological (Zhou et al., 2000; O'Herron and von der Heydt, 2011) as well as the computational (Layton et al., 2012) aspects of border ownership are subject to intense research. Models of contour integration and perceptual grouping also exist from Zhaoping (1998) and Jehee et al. (2006); Roelfsema (2006). The role of feedback and physiological investigations are elaborated in Hupé et al. (1998); Markov et al. (2013) and very recently (De Pasquale and Murray Sherman, 2013) found evidence for the modulatory properties of feedback in the visual cortices of mice.

2. MODEL DEFINITION

We propose a biologically inspired model of two-dimensional shape representation that consists of a hierarchical structure of interconnected model areas (see **Figure 1**). These model areas resemble the mechanisms of early and intermediate stages of visual processing in the ventral pathway of visual cortex. Each of the model areas is represented by a staged columnar cascaded model (see **Figure 1**). This cascade consists of (i) initial filtering, (ii) activity modulation, and (iii) center-surround interaction.

2.1. NOMENCLATURE

The following list familiarizes the reader with the nomenclature that is used in our manuscript:

- Names of model areas are written in superscripts to indicate the affiliation to parameters or responses. A response for cells in area V1 thus would read R^{V1}.
- Greek symbols (like α, β, σ)are used for parameters of dynamic functions or shapes of receptive fields.
- N and M are constants indicated the number of orientations and directions used in our model.
- An instance of an orientation is indicated using the step width $\hat{\theta}$ between discretely sampled orientations. The variable i is used as index. A specific orientation in a population is indicatd with $i\hat{\theta}$ with $i \in 0..N-1$.
- $\mathcal{N}$ stands for a normal (Gaussian) distribution that can be isotropic or anisotropic, rotated and spatially shifted, as defined by subscripted parameters.

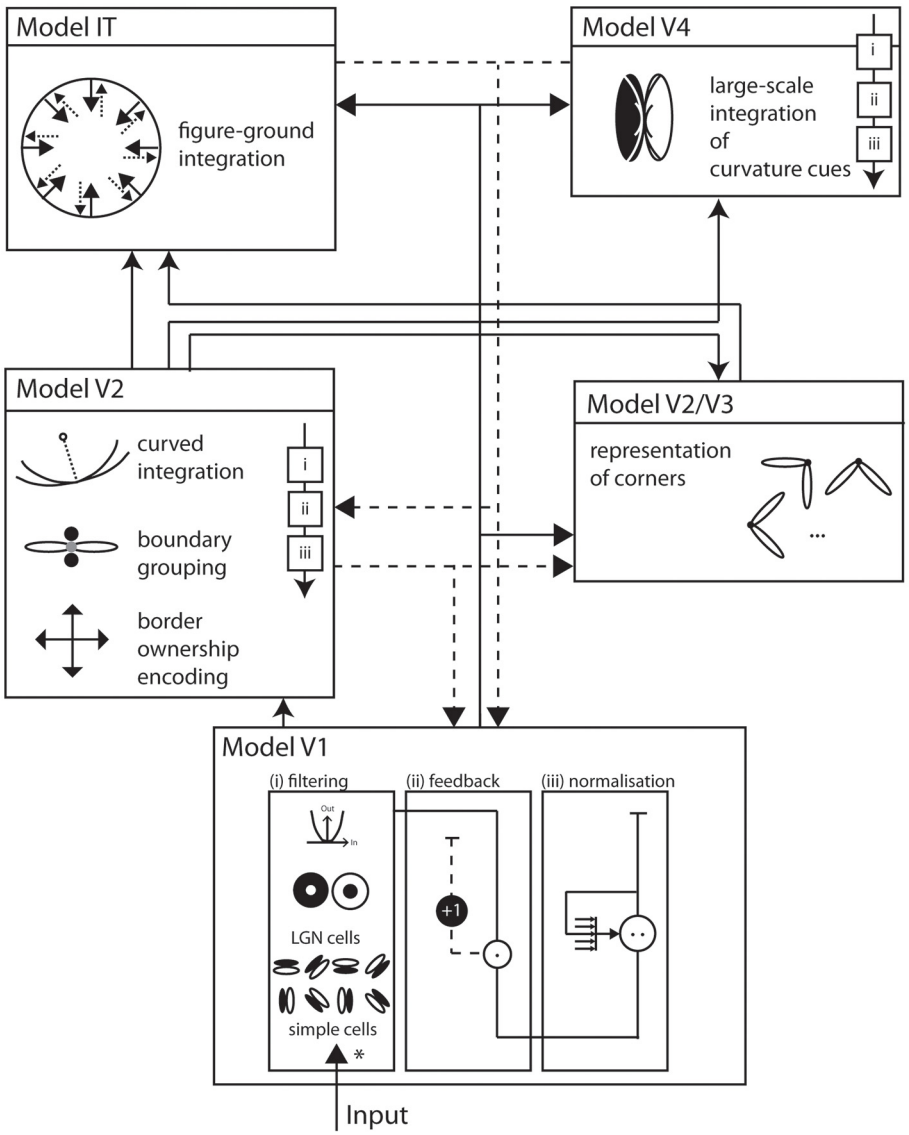


FIGURE 1 | Overall model architecture. Visual input enters the model at the bottom and is subsequently processed by interconnected functional areas with increasingly large receptive field sizes. Solid arrows indicate feedforward, dashed arrows indicate feedback, or modulatory, connections. Each area implements a generic architecture of building blocks that consists of (i) filtering ($*$) of the input, (ii) modulation by feedback, and (iii) response normalization. *Model V1* consists of image filters that resemble properties of early processing in LGN and V1, namely simple and complex cells that are tuned to circular or elongated image contrasts. *Model V2* integrates responses of model V1 with long-range integration cells. A multiplicative combination of subcells responds best to elongated contrasts of one dominant orientation. Also at V2, a population of cells represents border ownership directions. At population of long-range curved integration cells help represent different boundary curvatures. The *Models V2, V3* complex hosts representations of corners by integrating V1 responses from orthogonal configurations over a small spatial surround. *Model V4* consists of cells that asymmetrically integrate responses from V1 and V2 to become curvature selective at an increased spatial scale. In *Model IT*, cells with large receptive fields integrate responses from V1, V2 and V4 at local figure convexities to achieve a contextual segregation into figure and ground. Area V4 allows a description of a shape by means of cues that are represented on distributed areas in the model. Those cues exist at different spatial scales and their mutual interaction generates dynamic processes in the model.

- Spatial positions are denoted in bold latin letters like $\mathbf{x}$ or $\mathbf{p}$.
- To indicate an angle between two vectors we use $\measuredangle(\mathbf{v}_1, \mathbf{v}_2)$.
- The convolution operation is abbreviated using the asterix ($*$).
- A rectification operation is indicated by $\lceil ... \rceil^+$.

2.2. PROCESSING CASCADE

In our model, neural activations or response levels are modeled using a scalar representation of the neural firing rate. For ease of writing, we will in the following refer to the response of *a cell*, keeping in mind that this represents the activation level of a large number of real cells. The first model stage of the cascade is the initial filtering of available input I. To model the response for the preferred stimulus in the visual field, we employ a 2-dimensional convolution operation with the preferred stimulus as the convolution kernel K_{pref}. The response of model cells $R = I * K_{pref}$ is defined as

$$R(x,y) = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} K_{pref}(u,v) I(x-u, y-v) \quad \forall x, y \in D_I \quad (1)$$

A frequently used kernel in our model serves as elementary building block and is a 2-dimensional Gaussian distribution that is elongated along one axis and rotated around its center. We refer to this distribution by $\mathcal{N}$ with parameters for orientation θ, deviation along the axes σ_1, σ_2 and the center of the distribution μ.

$$\mathcal{N}_{\theta,\sigma_1,\sigma-2,\mu}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} exp\left(-\left(\frac{(\hat{x}-\mu_x)^2}{2\sigma_1^2} + \frac{(\hat{y}-\mu_y)^2}{2\sigma_2^2}\right)\right)$$

with (2)

$$\begin{pmatrix}\hat{x}\\ \hat{y}\end{pmatrix} = \begin{pmatrix}x\\ y\end{pmatrix}\begin{pmatrix}cos\,\theta & -sin\,\theta\\ sin\,\theta & cos\,\theta\end{pmatrix} \quad (3)$$

If parameters are not specified they are considered having the following default values: $\theta = 0$, $\mu = (0,0)^T$, $\sigma_1 = 1$, $\sigma_2 = \sigma_1$. In the following, functional filter kernels will often be designed as a combination of multiple such elementary components.

The coefficients of the kernel that models the preferred stimulus might incorporate negative weights to account for the inhibitory connections a cell may receive. This could lead to overall responses that are numerically negative. We thus use a rectification operator after convolution and feedback stages to ensure that numerically the response rate of a population is not negative:

$$\lceil R \rceil^+ = max(0, R). \quad (4)$$

At the second stage of the cascade, response levels are modulated by recurring input from higher visual areas. We propose a feedback mechanism that excerpts a purely modulatory gain control on the input. That means that feedback alone cannot generate activities without activation by the initial filtering step (see **Figure 2**). With R being the unmodulated driving signal and net_{FB} being the strength of the feedback, the modulated response is

$$R_{FB} \propto R \cdot (1 + net_{FB}). \quad (5)$$

Using this approach, given $R = 0$ no signal is generated as output irresponsible of the strength of the feedback net_{FB}. On the other hand, if no feedback signal is available, the right part of the equation leaves the input signal R unchanged (Salin and Bullier, 1995; Hupé et al., 1998; Eckhorn, 1999; Gilbert and Li, 2013).

Before normalization at the final stage of the cascade, we apply a non-linear transfer function to map the computed responses to a cell activation level. In our model, we use a function of type

$$f(R) = R^k \quad (6)$$

with k the non-linearity parameter. At the final stage, we incorporate a mechanism that keeps the response level limited by using a *shunting inhibition* that leads to a non-linear compression of high amplitude activities resembling the *Weber-Fechner-Law* of perceptual thresholds. In its dynamic formulation, the rate of change of the signal $\partial_t R_\theta^{norm}$ depends on the current activation level as well as the amount of input I_{net}:

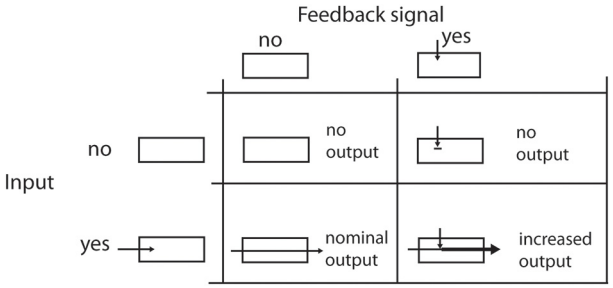


FIGURE 2 | Effects of feedback on the signal flow. The four table cells illustrate the effects when a feedback signal and/or an input signal is available. Please note that a feedback signal alone cannot elicit any cell response in the modeled area. It only enhances the response level when the filtering of the input signal generates some output.

$$\partial_t R_{i\hat{\theta}}^{norm} = -\alpha R_{i\hat{\theta}}^{norm} + \beta R_{i\hat{\theta}} - R_{i\hat{\theta}}^{norm} \cdot I_{net} \quad (7)$$

$$I_{net} = \frac{1}{N} \sum_{i=0}^{N-1} R_{i\hat{\theta}}. \quad (8)$$

With N the size of the used population, respective orientations. When this equation is solved at equilibrium, i.e., when $\partial_t R_{i\hat{\theta}} = 0$, the activation becomes

$$R_{i\hat{\theta}}^{norm} = \beta \frac{R_{i\hat{\theta}}}{\alpha + I_{net}} \quad (9)$$

The constants influence the steepness of the non-linearity (α) and the scale of the normalized signal (β). This model architecture has previously been used in various approaches touching different domains, such as the disambiguation of local motion (Bayerl and Neumann, 2004; Beck and Neumann, 2011), the processing of transparent motion (Raudies and Neumann, 2010) the detection of texture boundaries (Thielscher and Neumann, 2003), the extraction of object boundaries using texture compression (Weidenbacher and Neumann, 2009), and the analysis and representation of biological motion sequences (Layher et al., 2014).

In the following, we describe the forward sweep of our model, from early toward intermediate processing stages. After all areas have been described in detail, we will elaborate on the feedback connections that build the recurrent model structure.

2.3. MODEL AREA V1

The processing starts at early stages of visual cortex where we model the functionality of LGN and V1 cells where LGN cell responses provide feedforward input to V1 cells. Here, the visual input is intially processed to generate a representation of local image contrasts and local contrast orientations (Hubel and Wiesel, 1962).

In general, model cell responses follow first-order dynamics and represent the changes of membrane potentials. Such dynamics are influenced by excitatory and inhibitory inputs and a passive decay of activity. In order to simplify the computations in our large-scale simulations we use steady-state equations in calculations of feedforward filtering stages. Others are numerically integrated using a Euler one-step scheme. The response of LGN cells is calculated using the following linear equation,

$$\partial_t R^{LGN} = -R^{LGN} + (I * \mathcal{N}_\sigma) - (I * \mathcal{N}_k) \tag{10}$$

As pointed above, we assume that such linear feedforward filtering operations quickly relax at their equilibrium state. Therefore, we utilize the steady-state equation

$$R^{LGN} = \lceil I * (\mathcal{N}_\sigma - \mathcal{N}_\kappa) \rceil^+ , \tag{11}$$

with σ and κ denoting the width of center and surround kernel, respectively. To model cells that are tuned to oriented contrasts, we use elongated gaussian kernels that are combined into odd-symmetric simple cell profiles using anisotropic σ_1 and σ_2 and a radius ω_1 for the spatial shift of the integration kernels. The responses of such cells are denoted by the steady-state equation

$$R^{V1}_{i\theta} = \left\lceil R^{LGN} * (\mathcal{N}_{i\hat{\theta},\sigma_1,\sigma_2,\mathbf{x}+\mathbf{p}} - \mathcal{N}_{\theta,\sigma_1,\sigma_2,\mathbf{x}-\mathbf{p}}) \right\rceil^+ \tag{12}$$

with

$$\mathbf{p} = \omega_1 (cos(\theta + \pi), sin(\theta + \pi))^T \tag{13}$$

The filter kernel that is defined that way yields high response activations at positions with local luminance contrasts that match the layout of the filter kernel. To achieve insensitivity against the sign of contrast, pairs of equally oriented filters with opposite sensitivity to contrast polarity are used. Such filters populate a set with evenly distributed orientation tunings that represent possible contrast orientations. The locally dominant orientation can be derived by selecting the orientation channel with maximum response, $i_{max} = argmax_i R^{V1}_{i\hat{\theta}}$.

2.4. MODEL AREA V2/V3 COMPLEX

At the stage of V2 we model cells sensitive to contextual influences of contour segments that are arranged in larger spatial extent compared to V1 receptive fields. The integration of elongated contours in V2 makes use of a mechanism that links cells of like orientations over larger spatial distances. The filters are modeled using elongated Gaussian kernels positioned at $\mathbf{p}$ with offset ω^{V2}_{ex} to the center of the cell. The parameters of the elongated Gaussian kernels are set to build a combined kernel of an elongated integration field, which reflects the highly significant anisotropies of long-range connections in visual cortex (Bosking et al., 1997). The subfields sample the activations generated by V1 complex cells (Grossberg and Mingolla, 1985; Neumann and Sepp, 1999).

The subfields are combined in a multiplicatively. This resembles a logical *and*-operation for the individual subfield activations. Modeled V2 cells only become activated when both subfields receive sufficient input. The response is thus able to bridge local gaps in contours. This is in line with physiological findings, as V2 neurons respond to elongated luminance contrasts as well as to illusory contours (von der Heydt et al., 1984; Heitger et al., 1998) like in the Kanisza square.

This integration mechanism is enhanced by local inhibitory effects. Smaller and isotropic integration fields are positioned along an orthogonal axis from the cell's center with distance ω^{V2}_{inh}, building a cross-like zone of excitatory and inhibitory integration, compare (Piëch et al., 2013). At those positions $\mathbf{p}_\perp$, activity from all orientations is integrated and has an inhibitory effect on the total response. This has a strong suppressive effect on contour fragments that are positioned within a cluttered surround, while isolated boundary segments are not affected. The complete response for an elongated V2 cell is calculated by the steady state equation:

$$\begin{aligned} R^{V2}_{i\hat{\theta}} = \Big\lceil & R^{V1} * \mathcal{N}_{i\hat{\theta},\sigma_1,\sigma_2,\mathbf{x}+\mathbf{p}} \cdot R^{V1} \\ & * \mathcal{N}_{i\hat{\theta},\sigma_1,\sigma_2,\mathbf{x}-\mathbf{p}} - \gamma \cdot R^{V1} \\ & * \mathcal{N}_{\sigma_3,\mathbf{x}+\mathbf{p}_\perp} - \gamma \cdot R^{V1} * \mathcal{N}_{\sigma_3,\mathbf{x}-\mathbf{p}_\perp} \Big\rceil^+_{i\hat{\theta}} \end{aligned} \tag{14}$$

with

$$\mathbf{p} = \omega^{V2}_{ex}(cos(i\hat{\theta}), sin(i\hat{\theta}))^T \tag{15}$$

$$\mathbf{p}_\perp = \omega^{V2}_{inh}(cos(i\hat{\theta} + \pi), sin(i\hat{\theta} + \pi))^T \tag{16}$$

We also model V2 neurons that respond to more complex stimuli like in curved or angular shape outlines. We propose a population of V2 cells tuned to curved contour outlines that allows integration of smooth and even fragmented boundary configurations (Field et al., 1993). We propose a population of V2 cells tuned to a curved contour outline, see **Figure 4**. They resemble the functionality of elongated V2 cells but their integration fields are designed such that they are curved. A curvature direction is defined either to the left or the right of the tangent orientation at the target location. the center of curvature defines an osculating circle with given curvature-radius. The integration weight is modeled by a function w^{dist} that decreases with distance from the cell's center. A second tuning function w^{ori} in the orientation domain specifies the weights for the orientation population. Here, the weight decreases with distance to the main tuning direction which is perpendicular to the dominant orientation. Basically, only those orientations are integrated with maximum that are tangential to the curvature trace at their relative positions. This yields a sharp tuning of the cell for a certain curvature level. The complete response for an curved V2 cell is calculated by the steady state equation:

$$R^{V2Curv}_{i\hat{\theta}} = \sum_{\mathbf{x}} w(\omega^C, \mathbf{x}) \cdot R^{V1}_{x,i\hat{\theta}} \quad \text{with} \tag{17}$$

$$w = w^{dist} \cdot w^{ori} \tag{18}$$

$$w^{dist} = exp(-\frac{(\mathbf{x} - \mathbf{x_0})^2}{\sigma_1^2}) \tag{19}$$

$$w^{ori} = sin(\measuredangle(\overrightarrow{\mathbf{x_0}\mathbf{x}}, \overrightarrow{\mathbf{xc}})) \cdot exp\left(-\frac{(\|\overrightarrow{\mathbf{xc}}\| - \omega^c)^2}{\sigma_2^2}\right) \tag{20}$$

$$\mathbf{c} = \mathbf{x_0} + \omega^c \left(cos(i\hat{\theta}), sin(i\hat{\theta}) \right)^T \quad (21)$$

In this equation w denotes a weighting function for responses in the currently integrated position $\mathbf{x}$. The reference point of the integrating cell is $\mathbf{x_0}$. The weighting functions depends of the curvature radius being integrated w^c. w^{dist} produces a weight depending on the distance from the curvature cell's center $\mathbf{x_0}$. w^{ori} returns a weight given the current angle between integrating position and center of curvature $\mathbf{x_c}$, depending on the reference position x_0. In simple words, orientations orthogonal to the imaginary line between integration position and imaginary curvature center $\mathbf{c}$ receive highest weight. w^{ori} is extended with a function that drops with increased distance of integrating position to imaginary center of curvature.

Cells in visual cortex V2 also show selectivity to the figure-ground arrangement of the scene in the visual field (Williford and von der Heydt, 2013). So-called border ownership cell responses are elicited when figure of arbitrary shape is presented on their preferred side with respect to the center of their receptive field. From the same group, O'Herron and von der Heydt (2013) have also shown that during visual motion caused by eye motion or object motion, these border ownership signals are remapped to different neurons. The visual system uses this information to resolve depth arrangements in the stimulus (Qiu and von der Heydt, 2005). The pointing of border ownership cells indicates the direction of the frontal surface at every image location. This reflects to commonly known Gestalt rule that a boundary is owned by the frontal figure.

We model border ownership cells by a retinotopically arranged population representing four potential directions where the figure can be positioned relative to the cell's center. Border ownership responses are initially isotropic and only occur together with local contrast activations. Cells indicating opponent border ownership direction are mutually rivaling in our model. The complete response for an border ownership V2 cell is calculated by the steady state equation:

$$R_{\lambda}^{V2Bown} = \begin{cases} f(R_{i\hat{\theta}}^{V2}) & \text{when} \quad \lambda \perp i\hat{\theta} \\ 0 & \text{when} \quad \lambda \parallel i\hat{\theta} \end{cases} \quad (22)$$

The mutual competition between activations indicating opposing border ownership directions R_a^{Bown} and R_b^{Bown} is calculated by

$$\partial t R_a^{Bown} = -\alpha \cdot R_a^{Bown} + A(1 - R_a^{Bown}) - \beta \cdot R_b^{Bown} \quad (23)$$

$$\partial t R_b^{Bown} = -\alpha \cdot R_b^{Bown} + A(1 - R_b^{Bown}) - \beta \cdot R_a^{Bown} \quad (24)$$

Based on empirical evidence of neural representations generated by cells selective to multiple orientations (Felleman and Van Essen, 1987; Ito and Komatsu, 2004; Anzai et al., 2007) we incorporate model representations of corners in a dedicated model area V2/V3 complex. We build upon the proposal developed in Weidenbacher and Neumann (2009) that corner and junction configurations can be made explicit by specific readout mechanisms. Here, we employ a simplified version as of Hansen and Neumann (2004) to generate corner representations by grouping V1 responses of orthogonal orientation fields. In a steady-state formalism the response reads

$$R_{i\hat{\theta}}^{V2/V3} = \left\lceil R_{i\hat{\theta}}^{V1} \cdot R_{i\hat{\theta}+\pi}^{V1} \right\rceil^+ \quad (25)$$

2.5. MODEL AREA V4

Inspired by experimental evidence cells in model V4 integrate responses of V1,V2, and V2/V3 to achieve a selectivity that considers large-scale boundary fragments as well as local variations in curvature and a selectivity for corners (Pasupathy and Connor, 1999; Yau et al., 2013). Curvature selective cells are modeled in a two-stage cascade of mechanisms. The first level integrates V2 contour responses and is selective to curvature directions, left or right (relative to the cell's orientation preference). The second level combines opposite curvature directions into one response, like in V1 complex cells. This model mechanism differs from the one proposed by Rodríguez-Sánchez and Tsotsos (2012). which utilizes single stage filter computations. In this approach specific subfield mechanisms sensitive to orientation, tangential contour outline and scale are combined in a non-linear fashion to selectively respond to contour fragments of different curvatures. We develop a mechanism that is distributed over different stages to first group responses to extended contour outlines in V1 and V2 suppressing non-contour clutter. In the case of sharply localized corners and junctions the dedicated representations of localized multi-orientation responses will be activated. Those responses of grouping cells (or the junction representations) are integrated at the subsequent stage. Here, curvature selectivity is made explicit that distinguishes left and right curvatures. Different integration scales generate selectivity to curvature. This distribution allows to associate regions of high contour curvature at an intermediate scale with localized outline details at the finer scale which enhances the selectivity of the model developed by Rodríguez-Sánchez and Tsotsos (2012).

The model cell responses in our model are described by the following equations:

$$\begin{aligned} \partial_t R_{i\hat{\theta}}^{V4,left} &= -\alpha_4 R_{i\hat{\theta}}^{V4,left} + \left(1 - R_{i\hat{\theta}}^{V4,left}\right) \cdot A_{i\hat{\theta}} \\ &\quad - \left(1 + R_{i\hat{\theta}}^{V4,left}\right) \cdot B_{i\hat{\theta}} \end{aligned} \quad (26)$$

$$\begin{aligned} \partial_t R_{i\hat{\theta}}^{V4,right} &= -\alpha_4 R_{i\hat{\theta}}^{V4,right} + \left(1 - R_{i\hat{\theta}}^{V4,right}\right) \cdot A_{i\hat{\theta}} \\ &\quad - \left(1 + R_{i\hat{\theta}}^{V4,right}\right) \cdot B_{i\hat{\theta}} \end{aligned} \quad (27)$$

with

$$A_{i\hat{\theta}} = \left\{ R^{V2} * \mathcal{N}_{\sigma_4, \sigma_{4b}, x+\mathbf{p}} \right\}_{i\hat{\theta}} \quad (28)$$

$$B_{i\hat{\theta}} = \left\{ R^{V2} * \mathcal{N}_{\sigma_4, \sigma_{4b}, x-\mathbf{p}} \right\}_{i\hat{\theta}} \quad (29)$$

$$\mathbf{p} = \omega^{V4}(cos(i\hat{\theta}), sin(i\hat{\theta}))^T \quad (30)$$

These responses are calculated at equilibrium and averaged subsequently, leading to the model V4 filter response

$$R^{V4}_{i\hat{\theta}} = \frac{1}{2}\frac{\left\lceil (A_{i\hat{\theta}} - B_{i\hat{\theta}})\right\rceil^{+} + \left\lceil B_{i\hat{\theta}} - A_{i\hat{\theta}}\right\rceil^{+}}{\alpha_4 + A_{i\hat{\theta}} + B_{i\hat{\theta}}} \tag{31}$$

This integration mechanism yields a response for locally curved boundary segments at a larger spatial scale. For elongated contour segments that show no curvature, the response of individual cells will be equal and the combined response very low.

2.6. MODEL AREA IT

So far, we have described how our model integrated local features from model V1 into elongated, potentially curved boundaries at model V2–V4. Model area IT performs contextual integration that allows a segregation into figure and ground and a representation of prototypical objects at a large spatial scale. As discussed above, a population of V2 cells responds selectively to the direction of figure-ground direction. The local representation of border ownership at model V2 represents a set of available local hypotheses that cannot locally be resolved, as this step requires contextual influence from a larger spatial surround. Cells in IT cortex have been shown to be shape selective with properties generalizing over contrast polarity and mirror reversal (Baylis and Driver, 2001). The authors demonstrate that such cells do not, however, generalize over the assignments of figure-ground direction. The investigation supports the view that the population of probed IT cells is mainly driven by the sidedness of contours and less so by the contour itself. Given the rapidness of ownership selectivity observed in V2, we propose that ownership computation relies on a network of V2–V4–IT cell interaction. Our model uses local shape configuration in the outline of an object to collect confidence about the direction of figure and ground. We adopt an approach of Zhou et al. (2000) and model an integration cell at model IT that integrates border-ownership hypotheses from a larger spatial extent from model V2 input. For each location in the image, border ownership activations in a local neighborhood that point toward the inside of the respective receptive field contribute to the activation of an IT cell. This results in strong responses in model IT where local image regions are surrounded by contour convexities. Local activities of border ownership cells in model V2 then receive a positive enhancement if they contributed to such an integration process. This recurrent architecture resolves the initially ambiguous assignment of border ownership. Taken together, this makes the model belong to the class of feedback architectures according to the categorization in Williford and von der Heydt (2013). The response of cells and their interaction is denoted by the following equations:

$$R^{IT}_{\mathbf{x_0}} = \sum_{i}^{N-1}\sum_{\mathbf{p}} f(\mathbf{x_0}, \mathbf{p}, i\hat{\theta}) \cdot R^{V2}_{i\hat{\theta}, \mathbf{x_0}+\mathbf{p}} \cdot exp\left(-\frac{(\omega_{IT} - \|\overrightarrow{\mathbf{x_0}; \mathbf{x_0}+\mathbf{p}}\|)^2}{\sigma^2_{IT}}\right) \tag{32}$$

with

$$f = cos\left(\measuredangle\left(\overrightarrow{\mathbf{x_0}; \mathbf{x_0}+\mathbf{p}}, i\hat{\theta}\right)\right) \tag{33}$$

Such an IT cell at position $\mathbf{x_0}$ integrates responses of V2 cells in its proximity $\mathbf{p}$. The integration weight f depends on the angle between $\mathbf{x_0}$ and $\mathbf{x_0}+\mathbf{p}$ and the currently integrated orientation $i\hat{\theta}$. This grants orientations parallel to an imaginary line toward x_0 high weights, while orthogonal orientations receive low weights.

This model area receives connections from the early as well as from the intermediate functional stages V1 and V2 where curvature is represented. This means that high-resolution local cues as well as contextual cues like corners from a larger region are available. A shape can thus be described as a set of contributing prototypical elements that contribute to the local configuration at every image location. Those elements are not solely generated through integration of lower areas, but exist as a distributed representation in all modeled areas and profit from mutual interaction through feedback and exhibit dynamic processes when a stimulus is presented.

2.7. FEEDBACK FOR CONTOUR ENHANCEMENT

The mechanisms so far presented contributed to the feedforward sweep of the model. We stated earlier that in visual cortex (and in neural processing in general), the input of cortical areas of higher stages highly contribute to the performance of individual earlier areas. By such recurring connections, contextual information is introduced in lower regions. We are thus now going to focus on the recurrent connections that are incorporated in our model.

Let's briefly recall that we model feedback connections that have a modulatory effect (Girard and Bullier, 1989) as outlined in Section 2, Equation 5. In **Figure 2** we illustrate how a feedback signal alone cannot elicit responses as long as no input activation is present. On the other hand, feedback that matches input configurations will increase those activations. We stick to this convention throughout our following elaborations.

V2 long-range and curved cells represent continuous straight or curved contours. Their multiplicative combination of receptive field subcomponents caused the cells to elicit responses whenever a contour of matching orientation was presented in their receptive fields. Now, those cells in V1 that contributed to the integration process will receive feedback and be thus increased in activity. The following non-linear transformation stage increases the difference in response strength with respect to other oriented contour cells that did not receive feedback. At the subsequent normalization stage, local response levels are now slightly increased by the recurrent input. Now, surrounding activations without feedback have a competitive disadvantage and receive a higher divisive normalization relative to their activation due to the increase response in their neighborhood that contributed to the sum. The dynamics of these interactions are denoted in formal terms. The enhancement of filter responses (Equation 12) via modulating feedback is defined by

$$\partial_t P^{V1}_{i\hat{\theta}} = -\alpha_1 P^{V1}_{i\hat{\theta}} + \beta_1 R^{V1}_{i\hat{\theta}} \cdot \left(1 + \lambda_1 \cdot \left\{R^{V2} * \mathcal{N}^{FB}_{\sigma}\right\}_{i\hat{\theta}}\right) - P^{V1}_{i\hat{\theta}} \cdot Q^{V1}_{i\hat{\theta}} \tag{34}$$

The subsequent competition to accomplish activity normalization is defined as

$$\partial_t Q^{V1}_{i\hat{\theta}} = -Q^{V1}_{i\hat{\theta}} + \left\{P^{V1} * \mathcal{N}^{pool}_{\sigma}\right\}_{i\hat{\theta}} \tag{35}$$

with TODO parameters. **Figure 3** shows an illustrated version of the mechanism with a small numerical model.

2.8. FEEDBACK FOR CURVATURE REPRESENTATION

As stated earlier, the modeled V4 cell do not at all or only marginally respond to straight elongated contours. Responses of V2 cells to curved boundaries are integrated in model V4, where integration cells sensitive to opposite sign of curvature mutually compete for equal orientations. These cells respond at positions with a local curved contour configuration, but are silent at elongated straight contours. Feedback is generated for those V2 cells that contribute to those curved boundary segments the corresponding model V4 cells respond to maximally. Regions with curved boundary segments thus elicit a strong response of V4 cells while regions with mostly straight contours do not elicit such a strong response. This signal can thus be used to differentiate regions of many straight contour segments from regions with many curved contours.

In formal terms, the V2–V4 cell interactions are defined by

$$R^{V2curv} = -\alpha R^{V2curv}_{i\hat{\theta}} + (1 - R^{V2curv}_{i\hat{\theta}}) \cdot A(1 + R_{i\hat{\theta}}) \quad (36)$$

$$A = \{R^{V2curv} * \mathcal{N}_\sigma\} \quad (37)$$

2.9. FEEDBACK FOR FIGURE-GROUND SEGREGATION

The contribution of feedback to figure-ground segregation is twofold in our model. First, local hypotheses of border ownership are generated by intra-area recurrent connections from long-range grouping cells. Contextual feedback from model IT resolves the remaining ambiguities. Initially, all directions of border ownership are equally likely at boundaries. With increasing confidence about local contrast orientations generated by V1 and V2, two options for border ownership directions are discarded and only two orthogonal border ownership directions remain. Activations of long-range V2 cells that indicate elongated surface boundaries and their orientation locally increase activities of those border ownership cells that are directed perpendicularly to the orientation of the boundary. Activity normalization for V2 border ownership cells then leads to a suppression of activities for ownership directions orthogonal to the boundary orientation. Formally, this is accomplished by the dynamics

$$\partial_t R^{BOwn}_{i\hat{\phi}} = -R^{BOwn}_{i\hat{\phi}} + \beta(R^{V2}_{i\hat{\theta}} + h_{tonic}) - R^{BOwn}_{i\hat{\phi}} \cdot \sum_\gamma R^{BOwn}_{i\hat{\phi}} \quad (38)$$

with

$$\theta = \phi + \frac{1}{2}\, mod\, \pi. \quad (39)$$

Second, V2 border ownership cells receive feedback from cells in model IT. Here, border ownership as well as figural cues, e.g., from local junctions, or curvature maxima, were integrated by

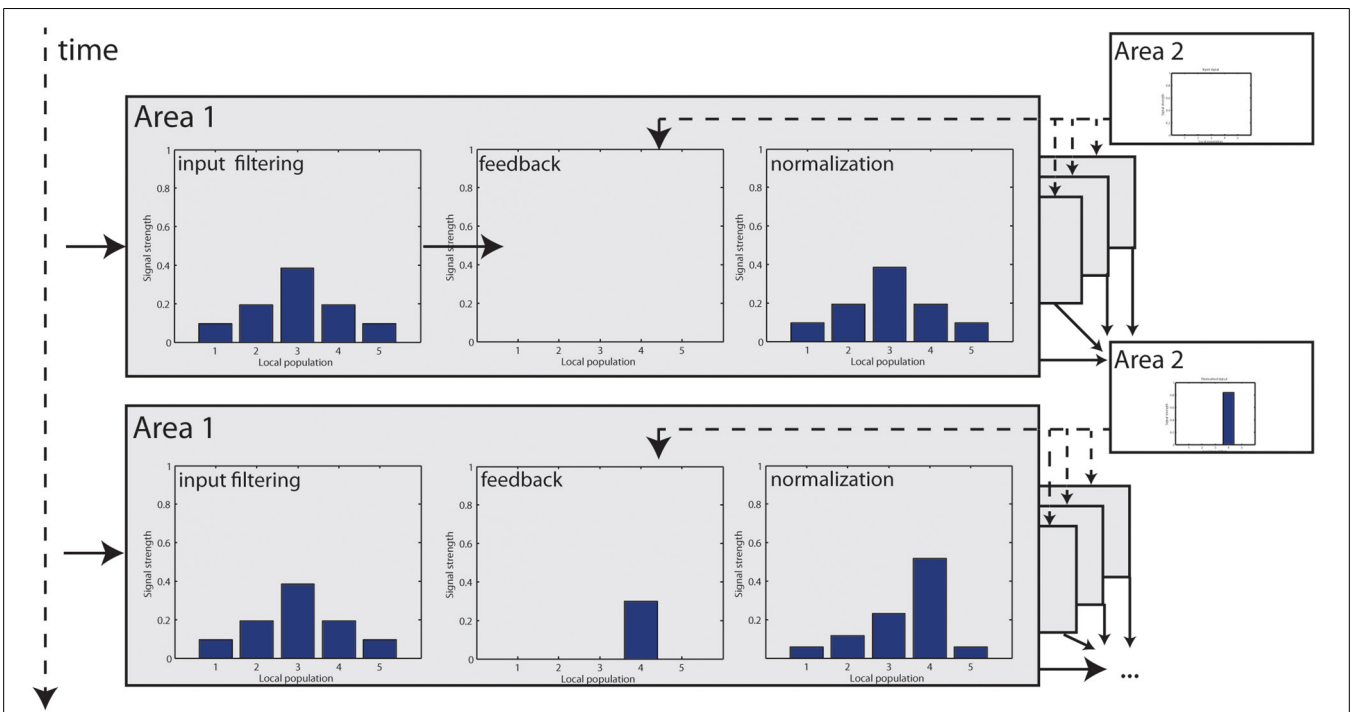


FIGURE 3 | Step-by-step illustration of how feedback modulation is dynamically incorporated in the model. At $t = 1$, visual input is filtered and elicits responses in *Area 1*. Initially, no feedback signal is available which leaves the signal unchanged. Responses are finally normalized. Those responses now become integrated at *Area 2* to elongated contours and a feedback signal is generated, which enters *Area 1* at $t = 2$. Now, some responses are accentuated, resulting in a higher cumulated response level that is used for normalization (which is further intensified by a stage of non-linear transformation). Unmodulated responses are damped in relation to $t = 0$.

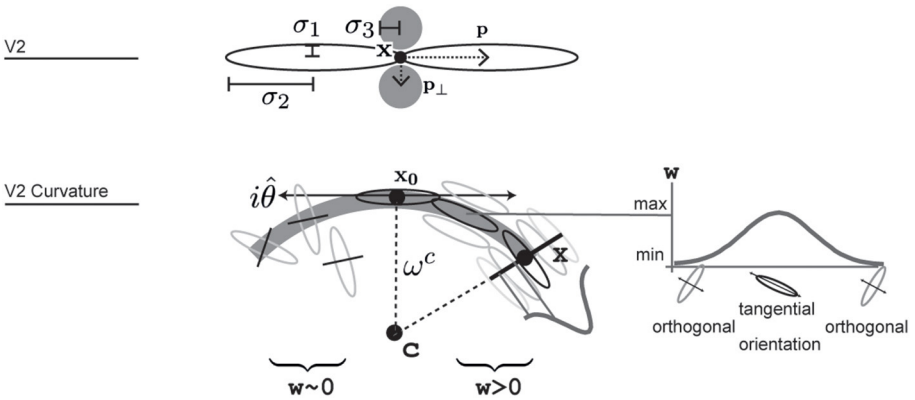


FIGURE 4 | Illustration of how V2 long range and V2 curvature cells are designed and where parameters are used. V2 cells integrate from two larger excitatory and two smaller inhibitory regions. V2 curvature cells integrate responses of oriented cell along a curvilinear path that forms around an imaginary central point **c**. Integration weights additionally depend on the distance from the cell's reference point $\mathbf{x_0}$, angular difference to the tangential trajectory and a function of local radius, indicated here by a gaussian profile. Correctly aligned orientations that result in a large integration weight are shown in the right side of the arc, while some that result in weights close to zero are shown along the left arc in this illustration.

IT cells. For the correct inference of figure and ground, feedback from IT to V2 is essential. Figure-Ground cells at IT level integrate border ownership activations from V2 in a circular fashion to integrate the coherence of directions indicating a convex pattern of figure outline. In the feedback sweep, this contextual information is now fed back to these border ownership cells compatible with the configuration using recurrent connections. In formal terms, this extends the dynamics presented in Equation 38 above by incorporating a modulating feedback signal from model IT cells, namely

$$\partial_t R^{BOwn}_{i\hat{\theta}} = -R^{BOwn}_{i\hat{\theta}} + \beta(R^{V2}_{i\hat{\theta}} + h_{tonic}) \cdot (1 + \lambda_2 \cdot R^{IT}_{i\hat{\theta}}) - R^{BOwn}_{i\hat{\theta}} \cdot \sum_{\gamma} R^{BOwn}_{i\hat{\theta}} \quad (40)$$

This also concludes the feedback sweep of our recurrent model. In the following section, we will show the performance of the model and its individual areas in the Results Section.

3. RESULTS

In this section we illustrate the capabilities of our model in a number of simulations. To demonstrate how the model processes shapes, we use some artificial images to show working principles of various subcomponents of our model. These simple shapes were taken from the *Webdings* font freely available with a Microsoft® Windows™ 8.1 operating system. We also include also a depiction of a *Kanisza* square (Kanizsa, 1955). This is a special stimulus because it elicits the perception of illusory contours at the outline of the occluding square, a sensation our model is also capable to represent.

To demonstrate the abilities of our model to process real world images we acquires the dataset of Fowlkes et al. (2007) and selected a few examples that we included in our Results Sections. These images have a resolution of 321×481 pixels in landscape or portrait orientation. They were converted to grayscale images using the Mathworks® Matlab® `rgb2gray` function which performs a perceptionally weighted combination of the red, green and blue channel. We used 8–12 iteration steps to allow recurrent feedback signals to build up. The angular resolution of cell populations is defined by selecting eight $\frac{\pi}{8}$ steps to encode orientation. Border ownership is represented by a population representing 4 directions. Model V2 curvature cells also used 8 orientations for tangential orientations, but due to two possible curvature directions, our model contains a population of 16 curvature cells. A list of parameters used is given in **Table 1**.

3.1. EARLY PROCESSING STAGES

To begin with, we show how the processing at early stages achieves a representation of the stimulus concerning contrasts and elongated contours. Local contrasts are represented in the early stages by model V1 and V2 cells. However, as can be seen in **Figure 5** the responses rapidly change in the first few iteration steps. The contained contour as well as the added noise signal both elicit responses at the V1 level (*second column*) and cause the shapes outline to be not clearly separated from the background. However, those responses are grouped into elongated contour representations in model V2 (*4th column*). Elongated contour segments are clearly emphasized. From these V2 activations, a recurrent feedback signal is generated that modulates V1 activations. After a few iterations, the representation at V1 dramatically changed, with the outline of the figure now clearly visible.

The effect of the feedback signal is also measurable in a quantitative way, see **Figure 5**, *right*. Along the boundary of an object we plotted the activation levels of the population of V1 neurons that represent the orientation. Initially, the neuron with preferred orientation responds best, but also those with orientation tunings close to the real contour (*first plot*). The situation changes when feedback is added (*second plot*). Now, representations of undesired orientations are attenuated and the activation of the cell representing the contextually valid orientation is highly increased.

Also in **Figure 5**, the representation of illusory contours at V2 stage is depicted. This is illustrated using an input depicting a *Kanisza* square (*last row*). A complete square is highly salient

Table 1 | General model parameters used for simulations.

Description	See equation	Value
Number of orientations used		8
Number of feedback iterations		6
Number of BOwn directions		4
MODEL AREA V1		
Network size		321 × 481
LGN σ	10	1.00
LGN κ		1.50
LGN normalization α	(9 applied)	2.17e-03
LGN normalization β	(9 applied)	2.17e-03
V1 contrast σ_1	12	0.23
V1 contrast σ_2		0.12
for **p**: Excentricity ω_1		3.00
V1 normalization α	(9 applied)	2.16e-06
Non-linearity of V1 responses	(6 applied)	4.00
MODEL AREA V2/V3		
V1:V2 subsampling	14	1: 3
Network size		107 × 161
Filter size of V2 complex		41.00
V2 complex cell $\sigma_1, \sigma_2, \sigma_3$		0.21,0.02,0.10
in **p**: ω_{ex}^{V2}		3.70
in **p**: ω_{inh}^{V2}		2.50
V2 inhibition strength γ		0.10
V2 nonlinearity k	(6 applied)	1.00
Strength of V2–V1 feedback		0.11
Non-linearity of BOwn		1.00
In **c**: curvature radius ω^c	19	15.00
in w^{dist}: σ_1		40.00
in w_{ori}: σ_2		2.00
Strength of R^{V2Curv} feedback		0.15
MODEL AREA V4		
V1:V4 subsampling		1: 4
Network size		81 × 121
V4 filter size		31.00
α_4	26	0.01
σ_4,σ_{4b}	28	0.43,1.35
ω^{V4}	28	−1.00
MODEL AREA IT		
σ_{IT}	32	0.43
σ		29.00
ω_{IT}		17.00
IT Non-linearity		3.00
α	(9 applied)	5.49e-05
Strength of BOwn feedback		50.00

for human observers despite the fact that only a series of circles with cut-out corners are depicted. This is reflected in the grouping responses of V2 neurons, they also show activity in the gap between the real contour fragments. **Figure 5** shows V2 responses for the same parameter set and for a parameter set with changed receptive field sizes, to illustrate the effect even stronger (*framed part*). **Figure 6** shows a result of the corner representation in the model.

3.2. CURVATURE TUNING

Figure 7 illustrates the tuning functions we defined for model V2 curved cells. A curved cell with distinct radius tuning was selected and we presented arcs of different curvature to this cell and simulated the response. We performed this for four cells with curvature tuning to 10, 15, 20, and 25 pixels radius. This curvature definition happens in V2, where the initial resolution of the image had been subsampled. For this reason, the value here correspond to values 40, 60, 80, and 100 in V1 resolution. In each plot, the peak response occured when the stimulus with the matching radius was presented. In this simulation, subsampling artifacts cause the first two plots to elicit some discontinuities.

3.3. SHAPE REPRESENTATION

In the final setup, we show how our model independently represents different elements of a shape, and how this depends on the recurrent feedback connections. **Figure 8** illustrates the results we achieved for an artifical image. Initially, we configured the model to only use feedforward connections from V1 to V2. The model only achieves an representation at model V1 and a representation at V2 where the elongated boundaries are visible, but surrounded by many spurious activations. When recurrent feedback from V2 is added, the representation at V1 improves in the first few iterations before a steady representation is reached. In parallel, elongated boundaries at V2 are integrated and noise is highly reduced.

To represent prototypical objects at an intermediate level of detail, we stated that the model needs to represent different contour properties. In the second row of **Figure 8** we show how the model achieves to emphasize V1 responses when they contribute to a certain contour fragment with desired properties. We deliberately exaggerated the effect and chose a very narrow tuning so that all other responses become almost completely suppressed. On the *left* side, we let the model emphasize contour parts that are oriented almost vertical but in a curved context of a matching radius. As can be seen, the model highlights that parts on the left side of the stimulus that matches and leaves others suppressed, even if their local orientation would match. On the *right* side of **Figure 8**, we perform the same for a different part of the shape outline.

In **Figure 9** we perform the same selection for a realistic photograph depicting an elephant. On the *left* side, we show interaction of model V1–V2 causes an appealing representation of the animal at stages V1–V2. On the *right* side, we configured the model using model area V4 to emphasize parts of the outline of the animal that match a certain context and configuration, here, a part of the outline.

3.4. BORDER OWNERSHIP AND FIGURE-GROUND ASSIGNMENT

In the segregation of a scene into figure and ground the modeled *border ownership* cells participate by indicating the direction where the frontal surface is positioned at a boundary (Zhou et al., 2000). Our model incorporates a mechanism using such border-ownership cells to resolve the direction of a frontal surface from local boundary cues (Zhou et al., 2000). We performed such a assignment for our sample images, see **Figure 10** for an

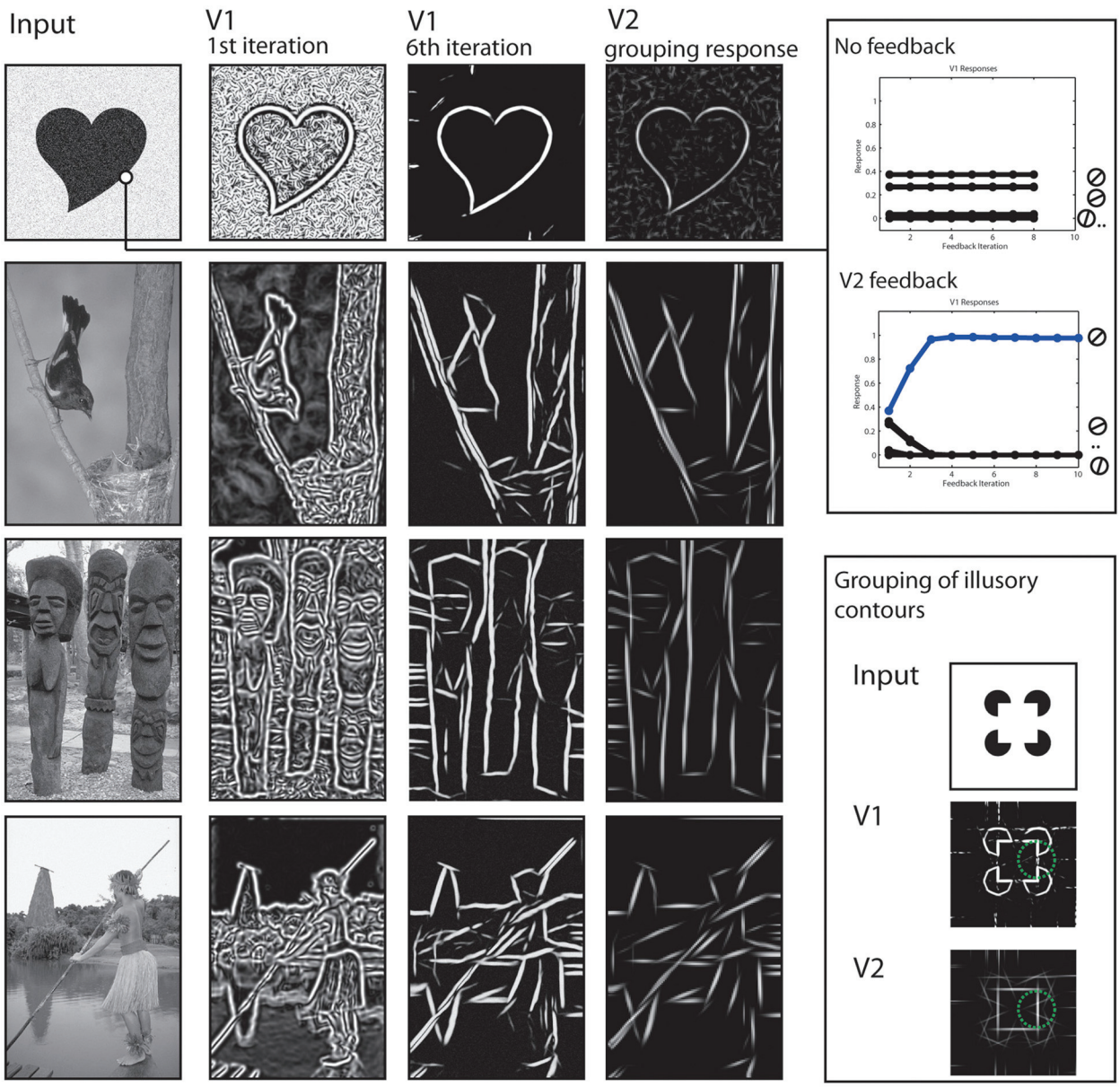


FIGURE 5 | Results of early processing stages V1 and V2. Left column: Initial input images. Second and Third column: Cumulated responses of model V1 neurons at the initial processing iteration and a few iterations steps, respectively. Fourth column: Responses of model V2 neurons. Elongated edges formed by like-oriented contrasts are grouped as reflected by responses at respective locations. This stage also shows activations for illusory contours contours (third row) at the gaps between contrasts. Upper right box: The two plots indicate time courses for V1 activations. Initially, multiple V1 neurons are activated due to a broad tuning width (first plot). Without feedback, this effect prevails through iterations. With feedback, the correct orientation (blue) receives feedback and gradually reduces activations of other orientations (second plot). Lower right box: Example how model V2 neurons show responses at positions formed by illusory contours (in green circle) due to contextual integration.

illustration of the result. The output of model area V1 and of V2 long-range integration cells are acquired to generate initial hypotheses of border ownership direction at image regions where local contrasts are situated. Initially, all four border ownership directions show equal responses at a boundary location. After stimulus onset, three dynamic effects occur and their contribution to the resolution of border ownership is reflected in the time course of cell activation, see **Figure 10** for an illustration.

First, local feedback from V2 cells enhances two hypotheses of border ownership for the directions orthogonal to the local boundary orientations.

A local normalization causes an attenuation of the other two representations **Figure 10**, *second row*; Timestep 0 and 1). Second, shape-level integration at model area IT contributes positive feedback to those border ownership cells that are directed toward the inside of the figural depiction. Again, normalization leaves the

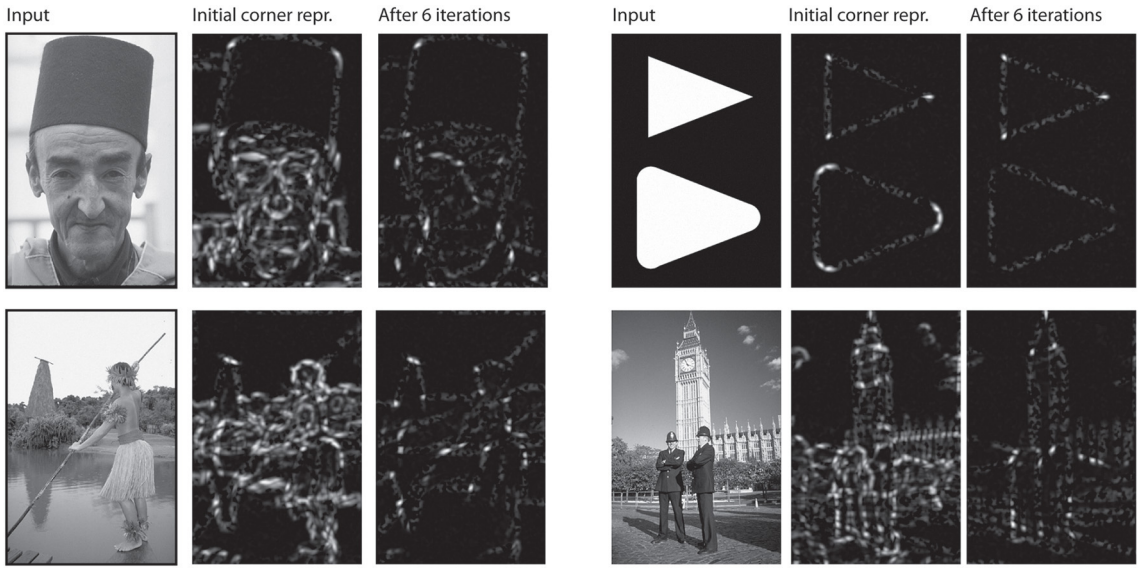


FIGURE 6 | Corner representation in model V2/V3. For each group of three pictures: Initially, responses of model V1 did not yet benefit from contextual feedback of model V2 neurons. Corner representation is thus distorted by noise (second row, middle). After a few iterations, when V1 responses have been modulated V2 feedback, the corner representation is much clearer.

net response of the cells constant (timesteps 2–4). Finally, mutual inhibition among border ownership cells with opposite direction selectivity causes the dominant direction to gain all available net energy (timestep 5–8). At this point, a stable point is reached and the local ambiguity for border ownership direction is resolved using feedback from higher cortical areas. The interpretation of the final representation would be that the frontal surface is to the inside of the curved boundary.

4. DISCUSSION

4.1. SUMMARY OF CONTRIBUTIONS

In this contribution we emphasized the role hierarchical representations have in the organization of shape features and their combinations into a coherent form. Like some previous model developments (Cadieu et al., 2007; Hatori and Sakai, 2012; Rodríguez-Sánchez and Tsotsos, 2012) our model is based on low and intermediate representations of shape features. These proposals are all based on a strictly hierarchical feedforward processing sequence. We propose here that such shape encoding mechanisms may be based on distributed representations that are established by interacting assemblies each devoted to specific feature properties. Such interactions in the model are organized by recurrent interactions of feedforward and feedback signals. The underlying structural principles are based on the cortical architecture of the ventral pathway with mutual interactions between such distributed representations (Markov et al., 2013). The model architecture incorporates principles that have been predicted to minimize the computational efforts of visual systems to successfully deal with the complexity problem of perception (Tsotsos, 1988) [compare also (Tsotsos, 2005)]. Among those, the hierarchical organization of representations in model areas, the specific receptive field properties of model columnar mechanisms, hierarchical pooling of spatially separated input representations, and top-down (modulatory) feedback are proposed here to account for the functional properties of cortical shape processing. We did not discuss complexity advantages in this contribution. However, given the theoretical predictions by such earlier work our proposal of a model architecture provides a evidence how distributed intermediate-level mechanisms may help to shape our understanding of modeling complex visual machinery that captures key cortical principles.

The main contributions of the work presented in the manuscript are twofold. First, we propose a computational network architecture that utilizes a hierarchical distributed representation of shape features. Contour features play a major role to track moving shape in which their strength parametrically change as a function of their saliency (Caplovitz and Tse, 2007). This necessitates global configurational as well as local information to distinguish rather tiny differences in the outline of a 2-dimensional form [such as curved boundaries vs. localized corners (Pasupathy and Connor, 1999; Ito and Komatsu, 2004)]. In order to generate a representation with sufficient spatial resolution combined with spatial context we suggest that multiple specialized component representations interact by feedforward hierarchical processing that is combined with feedback from representations generated at higher stages in the hierarchy. Second, we incorporate grouping mechanisms to integrate like-oriented contour responses that are integrated if they form a smooth outline fragment of a surface boundary (e.g., Grossberg and Mingolla, 1985; Neumann and Sepp, 1999; Ben-Shahar and Zucker, 2004). Such grouping mechanisms operate at the stage of area V2 and are, thus, involved in the hierarchical processing of shape. Given the hierarchical processing and representation of boundary information in the ventral pathway (see the overview

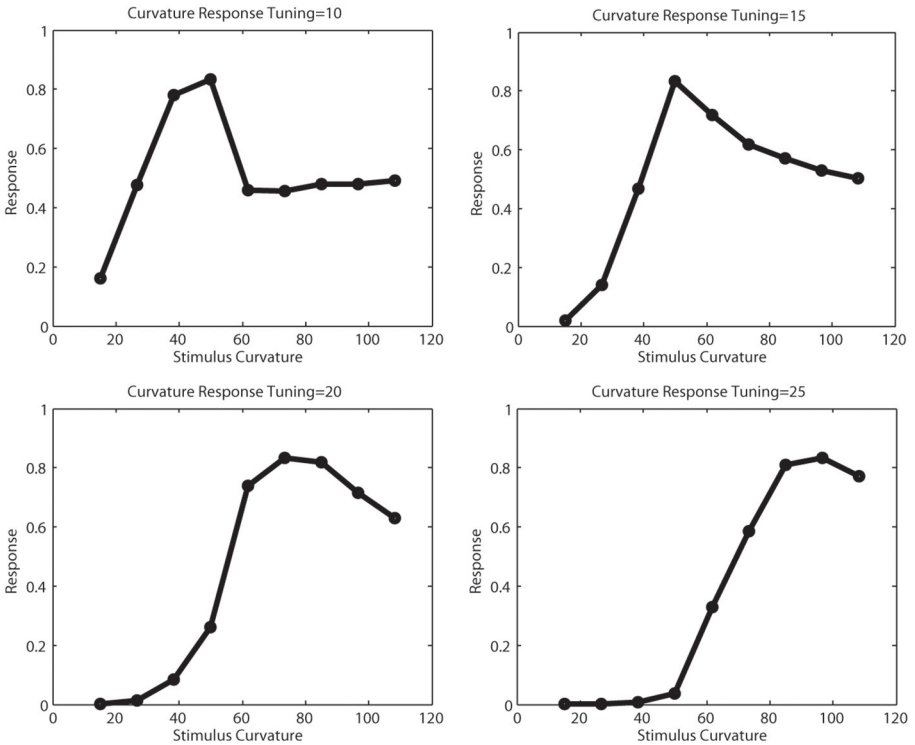


FIGURE 7 | Tunings of different curvature cells. The x-axis show the curvature of the presented stimulus, the y-axis the response strength of a curvature cell tuned for 10, 15, 20, and 25 pixels curvature radius in model V2, which correlates for a curvature of 40, 60, 80, and 100 pixels in model V1. For the smaller curvature radii, subsampling artifacts cause the tuning function to be less smooth.

in Neumann et al., 2007) the shape processing observed in area V4 is mainly driven by the output of grouping responses. It may be supplemented by input from simple/complex cells in V1, a principle of convergent signal streams also used in the models described in Thielscher and Neumann (2003); Rodríguez-Sánchez and Tsotsos (2012). In addition, we suggest that the shape representation built at the stages of V4 and IT influences the assignment of border ownership in surface representation (Zhou et al., 2000) (see overview in Neumann et al., 2007). Model IT cells send modulatory feedback to those V4 cells that provide relevant input (in V4 and V2) such that the net sum of convex corners/curvatures determines the ownership direction. The proposed model thus combines separate findings about the generation of cortical shape representation with figure-ground segregation mechanisms by assigning border ownership.

4.2. RELATION TO PREVIOUS MODELS OF SHAPE REPRESENTATIONS IN CORTEX

Visual shape recognition has already been investigated intensively by considering the 3-dimensional (3D) surface appearance for object recognition (Riesenhuber and Poggio, 1999; Serre et al., 2007; Mutch and Lowe, 2008; Yamane et al., 2008; Serre and Poggio, 2010) as well as 2-dimensional (2D) shape recognition (Schwartz et al., 1983; Mokhtarian and Mackworth, 1986; Mokhtarian, 1995; Rodríguez-Sánchez and Tsotsos, 2012). In the context of view-based models of object recognition stable views (Logothetis et al., 1995) are associated with 2D shapes so that their analysis can be considered as an intermediate stage of object processing (Cadieu et al., 2007). The computational model approaches of 2D shape representation can be subdivided into flat and hierarchical schemes. Examples of flat processing schemes, e.g., utilize Fourier descriptors (Schwartz et al., 1983), multi-scale representations of curvature features in the shape outline (Mokhtarian and Mackworth, 1986; Mokhtarian, 1995), or global schemes for integrating oriented line features (Wilson and Wilkinson, 1998). Hierarchical multi-layer processing schemes are based on different stages to generate an increasingly coarse-grained representation of shape features utilizing repetitive application of local filtering operations (Riesenhuber and Poggio, 1999; Cadieu et al., 2007; Rodríguez-Sánchez and Tsotsos, 2012). In order to resemble the feature selectivity of V4 cells in monkey cortex such cells build coarse-grained orientation-curvature representation of the shape under inspection. The hierarchical organization of a sequence of processing stages follows the idea of the Neocognitron (Fukushima, 1980, 1988) by developing low and intermediate representations of richer shape feature compositions (LeCun et al., 1998; Riesenhuber and Poggio, 1999; Mutch and Lowe, 2008; Tabernik et al., 2014). The orientation-curvature representation of V4 cells reported by Pasupathy and Connor (1999); Connor et al. (2007) has been investigated in the models reported in Cadieu et al.

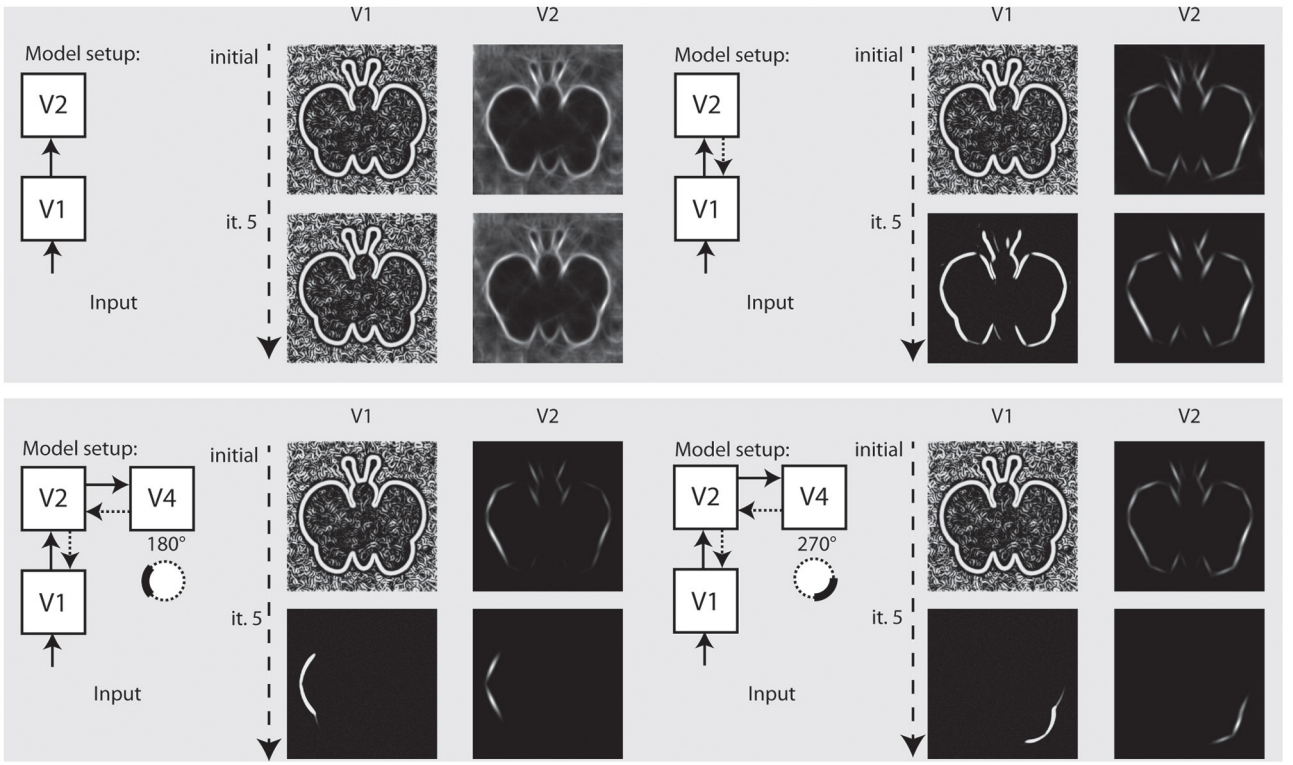


FIGURE 8 | Representation of curved boundary segments and effect of feedback. Top left: Model setup without feedback connections. The initial representation at model V1 and V2 undergo no change. Top right: Feedback from V2 causes a refinement of elongated structure within a few iterations. Bottom left: Feedback from V4 allows the accentuation of boundary segments with distinct curvature strength and direction. Here, a curvature segment as found on the left part of the shape is highly emphasized by feedback. Bottom right: Same as left, but with selectivity for another segment of the shape. Note that while boundaries with the same orientation are present in the stimulus, only the one with matching curvature is emphasized.

(2007); Rodríguez-Sánchez and Tsotsos (2012); Hatori and Sakai (2012). We share the principles of the hierarchical organization of processing and the emergence of rich orientation-curvature sensitivity in our proposal. Initial processing utilizes orientation sensitive filters to extract local oriented contrast. Unlike the previous models we incorporate a stage of boundary grouping at the interface between low and intermediate levels of representation. Such grouping operations integrate oriented contrast responses that are arranged in the local neighborhood of a target location. The local responses are enhanced by evaluating a support function that measures feature compatibility [(Neumann and Mingolla, 2001) for an overview and taxonomy of grouping schemes]. The measure of compatibility, or relatability, depends on the lateral integration that utilizes oriented weighting functions for contrast features arranged along a model shape outline, e.g., circular arcs with different radii (Parent and Zucker, 1989). Such a scheme thus implicitly incorporates curvature as a local contour feature. In order to make this explicit, different contour radii and signs of curvature (for individual orientations) have been considered in Rodríguez-Sánchez and Tsotsos (2012). Rather then implementing this curvature selectivity in a hard-wired scheme of local oriented filter conjunctions, we propose that this selectivity is generated via bottom-up and top-down filter mechanisms organized in a hierarchy. In this architecture the responses from model V2 contour groupings (based on different radii) are integrated by model V4 curvature sensitive cells with coarse bipartite odd-symmetric receptive fields (similar to simple cell profiles, but at much larger spatial scale). The sign of curvature is distinguished by cells of opposite polarity that mutually compete for each orientation. As a consequence responses are generated preferentially in cases where a single dominant curvature is present while responses are suppressed for straight contours which feed curvature cells symmetrically. The curvature radius is represented through a family of differently scaled integration sizes of such model V4 cells. Each of these cells have a specific peak selectivity. In the simulations we used three different sizes for each curvature sign. In order to make those cell responses selective to the feature specificity but mainly invariant to luminance contrast we suggested that each V4 cell response competes against the responses of other curvature selective cells in a local pool that interact via a mechanism of shunting inhibition. This leads to normalization of responses just like in those mechanisms proposed to account for various non-linearities at different stages in cortical processing, e.g., for context related contour responses in V1 (Carandini and Heeger, 1994; Carandini et al., 1999), attention selection (Carandini and Heeger, 2012), and higher level cognitive functions (Louie et al., 2011). Since the curvature sensitive model V4 cells, in turn, send feedback

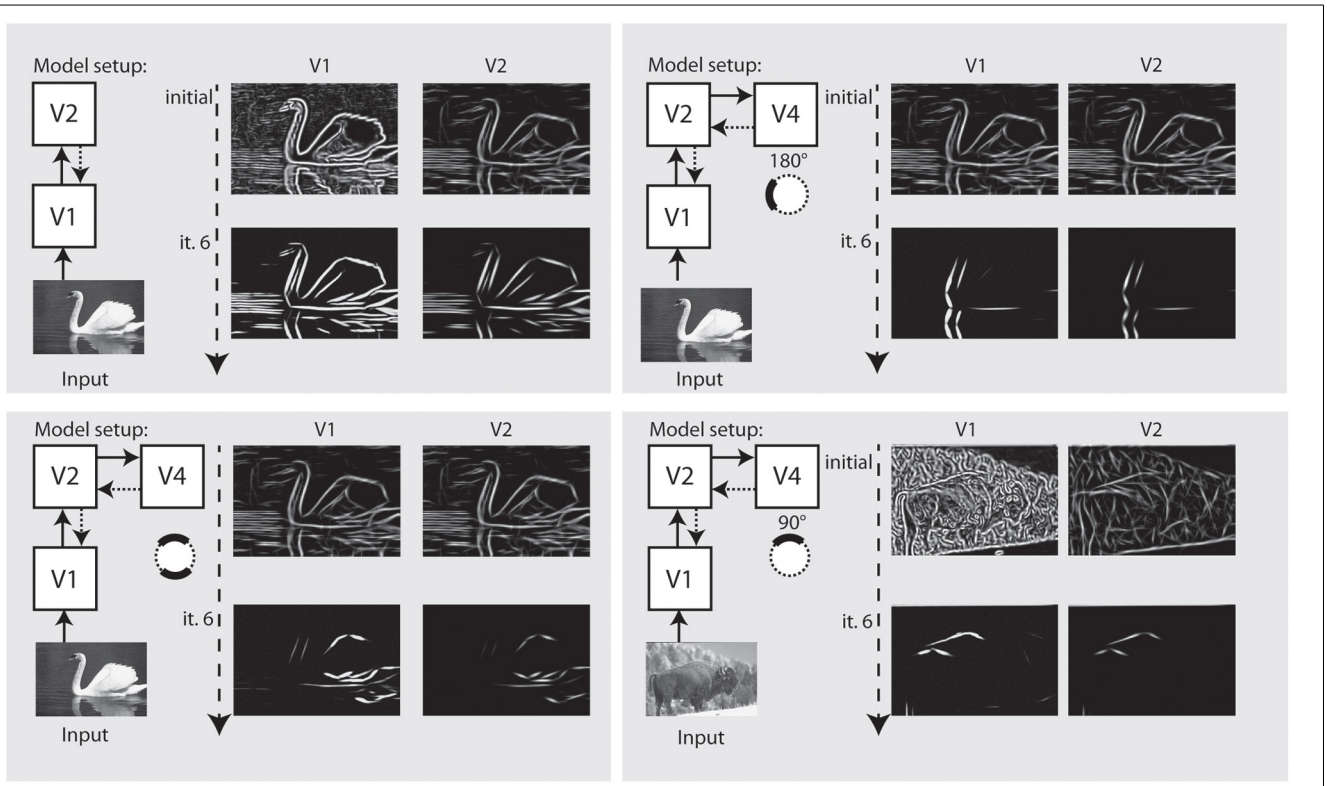


FIGURE 9 | Refinement and modulation of shape contours in a real world example. Left: Within 5 iterations, the outline of the animal is very well visible at V1 and at V2 stage. **Right:** With modulatory feedback from model V4, various parts of the animal like those contours with a certain curvature and orientation can be emphasized.

to their input contour representations in model V2 and filter response in model V1 those corresponding input activations will be enhanced. The amplitude of responses in distributed boundary representations will be amplified as an emergent net effect such that local salient curvature features in a shape outline will be amplified to yield distributed component feature representations of figural shapes.

These local boundary and curvature representations also feed mechanisms of border ownership assignment at the level of the model V4/IT complex. Such mechanisms have been investigated before in e.g., Zhou et al. (2000); O'Herron and von der Heydt (2011). Our computational framework belongs to the group of feedback models for border ownership encoding (see the overview of the current state in Williford and von der Heydt, 2013, see discussion below). We adopted this generic scheme by integrating responses from curvature selective cells with the compatible sign of curvature. In such a way the ownership configuration favors contributions from coarsely presented convex components. If a shape with multiple convex and concave segments is present then the ownership cells with opponent direction selectivities compete in order to arrive at a disambiguated assignment of surface belongingness. This makes the testable prediction that bumpy outlines should lead to slightly longer ownership disambiguation than for smooth convex shapes since the disambiguation will take more time when initially opposite assignment hypotheses coexist.

An additional investigation was argued to be of importance in the work proposed here. Several experimental investigations have reported that cells in extra-striate cortex selectively respond to corner junctions. For example, Ito and Komatsu (2004) (compare also Hegdé and Van Essen, 2000) reported that cells in area V2 selectively respond as to generate representations of sharp corners, or angles, selective for a particular opening angle. Similarly, Pasupathy and Connor (1999); Yau et al. (2013) show that area V4 cells respond to sharp shape corners with a sub-population of cells preferring sharp corners with different orientation and opening angles while another sub-population prefers smooth rounded corners. While the previous hierarchical models can account for the response selectivity for any of these generic corner types the perceptual representation of sharp localized features that allow, e.g., to distinguish between sharp and rounded corners remain unanswered. Sharp corners of any opening angle would be indistinguishable from the smooth variants of these corners given the increasing smoothing and subsampling of the visual representation while proceeding in the hierarchy. Our model argues in favor of a distributed representation: While shape sensitive cells at an intermediate level represent the salient shape protrusions (as in V4) the localized detail of an outline is represented at a higher spatial resolution in lower-level representations, e.g., in V1, V2, V3. In our model we suggest representations of smooth boundaries with different curvatures represented by groupings in model V2 while

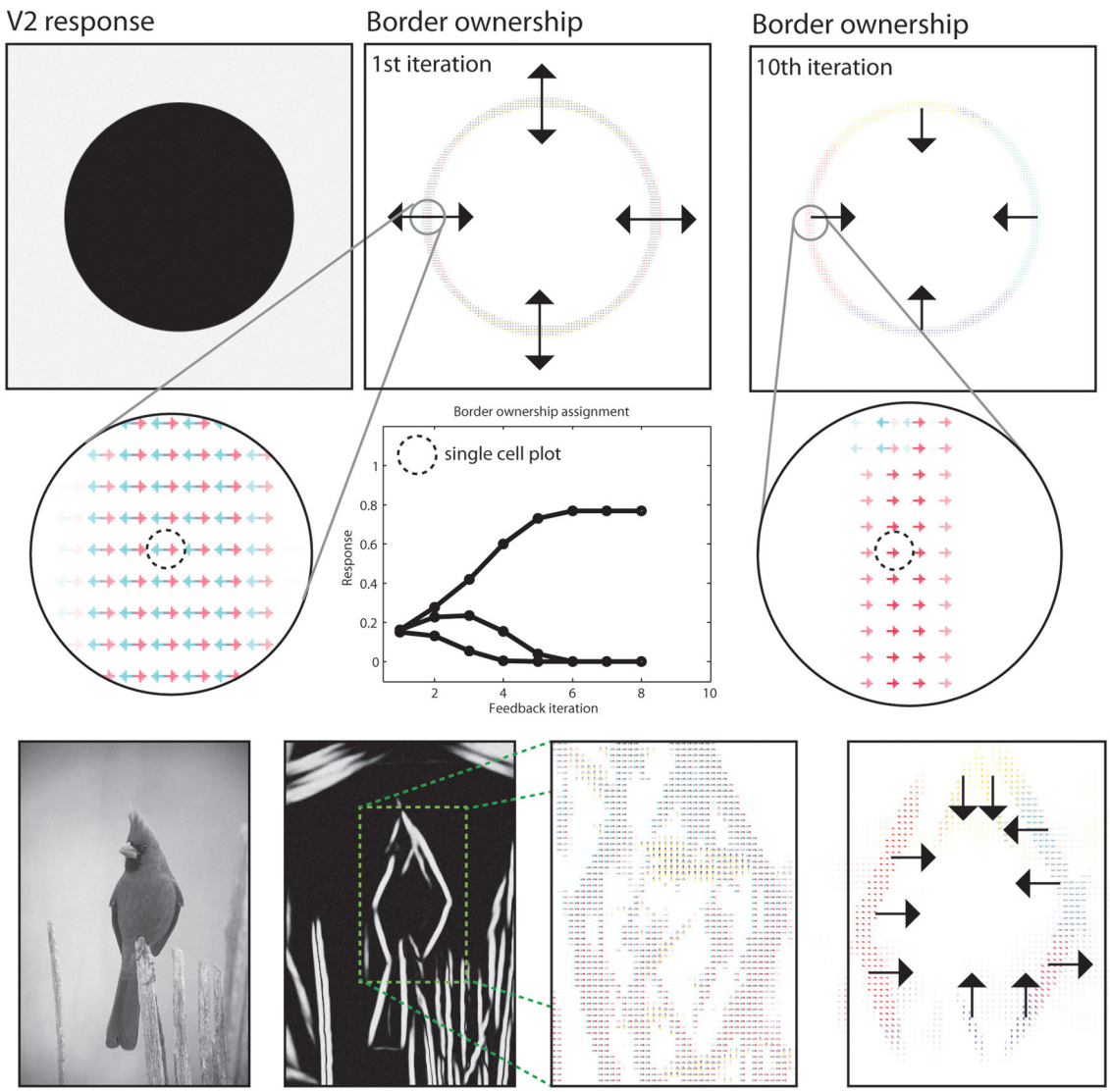


FIGURE 10 | Results for border ownership assignment. First two rows: Cells at model V2 indicate the direction of figure side at positions where boundaries exist. Initially, four hypotheses exist for possible figure direction. These are refined in model V2, where only two hypotheses remain after the orientation of the boundary is represented. Contextual integration in model V4 then provides correct estimates with modulatory feedback. Subsequent normalization and mutual competition leaves only one hypotheses for border ownership direction. See text for details on the time phases of border ownership assignment. Third row: Demonstration of the boundary assignment for a natural image. The initial responses are improved after a few iteration steps.

sharp corners are implicitly represented by convergent V1 input in local representations in model V2/V3. We assume that responses of cells in the model V2/V3 complex mutually compete such that their energy provides a measure to normalize individual responses. These provide convergent input to curvature selective contour cells in model V4 which, in turn, send feedback signals to their input sites at preceding stages. Since they are driven by either smooth or sharp contour arrangements the interaction of bottom-up sensory and top-down context-driven signals leads to selective enhancement of the particular corner configuration in the present stimulus. The specific details of the interaction between such counter-stream signal flows are discussed below.

4.3. FEEDBACK AS PREDICTION MECHANISM TO LINK SHAPE COMPONENTS

The hierarchical model architecture proposed here is composed of multiple model areas each of which is represented by a three-stage columnar cascade model. In a nutshell, the model cascade consists of (i) an initial stage of input filtering, (ii) a stage of activity modulation of filter outputs by top-down or lateral re-entrant signals, and (iii) a stage of center-surround interaction of target cells against an inhibitory pool of cells leading to activity normalization to generate the net output response of the model area. These three stages can be roughly mapped onto compartments of cortical area subdivisions (as suggested in Self et al., 2012). The filtering stage of the driving feedforward input signals is specific

to the particular (model) area under consideration. At the output stage, the activity normalization is computed by a mechanism of shunting inhibition, like the non-linear divisive mechanisms proposed in Carandini and Heeger (1994); Carandini et al. (1999); Kouh and Poggio (2008); Carandini and Heeger (2012). The feedback signal is generated at higher-level cortical stages or parallel processing pathways and is thought to provide context information that is re-entered at the stage earlier in the processing hierarchy (Grossberg, 1980; Edelman, 1993).

The functional role feedback signals play still remains controversial. Different proposals how feedback signals interact and combine with the driving feedforward stream have been discussed in the literature which have received different support from the experimental literature (Markov et al., 2013). One such framework proposes that the goal of computation is to reduce the residual error between the different signal streams in order to approach the sensory prediction generated by higher stages of processing (Ullman, 1995; Bastos et al., 2012). This idea is rooted in the Bayesian theory of predictor-corrector mechanisms which yields to the Kalman optimal filter realization under some restricting assumptions (Rao and Ballard, 1999). We follow an alternative route in which the feedback mechanism is modulatory in nature. Unlike predictive coding which tried to drive the difference between driving signals and the prediction to zero bottom-up input signals are amplified by matching feedback signals. This leads to a gain enhancement for those cell responses where a matching top-down predictive signal template has been generated. This feedback signal amplifies the sensory signal such that the subsequent competition between neurons yields a competitive advantage for the enhanced response patterns [biased competition; (Girard and Bullier, 1989; Desimone, 1998; Roelfsema et al., 2002; Reynolds and Heeger, 2009)]. The modulation mechanism is reminiscent of the linking mechanism suggested by Eckhorn et al. (1990); Eckhorn (1999) to account for activity synchronization in networks of spiking neurons. We have recently demonstrated (Brosch and Neumann, 2014) that such mechanism of convergent bottom-up feedforward and top-down feedback signal correlation accounts for the signal amplification as measured at the level of cortical pyramidal cells (Larkum, 2013).

In the shape processing architecture described here the modulatory feedback serves the role of a predictor (Spratling, 2008). For example, bottom-up input in oriented contrast is integrated by mechanisms of contour grouping and integration to generate continuous boundary representations. This is similar in spirit as the recent investigation of Piëch et al. (2013) who emphasized how context information at higher cortical stages influence more local feature representation at lower levels. Here, the same principle is replicated over different stages of model cortical processing. Contour representations after grouping in model V2 and junction configurations in model V3 send their output activations to curvature sensitive cells in model V4 where the activities are integrated. These cells, in turn, send their feedback to the input populations of neurons that have generated their input. The computational logic is that the curvature responses provide a template of context-related information about the local presence of oriented shape features. The modulatory feedback amplifies those inputs that are consistent with the curvature feature representation. The mutual competition of responses in a pool of cells at the lower level leads to a suppression of inputs that do not contribute to the present curvature feature. In all, a distributed representation of shape information is created that contains coarse-grained configurational information about stimulus shape and, at the same time, the spatially localized detail needed to distinguish between sharp and smooth corners. Similarly, the action of feedback sent from ownership sensitive cells (in the V4/IT complex of the model) to curvature sensitive and grouping cells in model V2 and V4 also provides context information for the assignment of configurational information. Here, the ownership assignment is based on the consolidation of evidence which convex shape elements make to establish a closed shape region in the visual field. This context is delivered via feedback to their input that represents fragments of shape components (irrespective of the sign of curvature) and also to the grouping representations. Those shapes that finally receive assigned direction of border ownership, and thus figure-ground direction, will enhance the associated inputs at the intermediate level orientation-curvature representations.

In all, the hierarchical processing scheme proposed here relies on extensive bidirectional flow of information in which the feedback signals that represent context-sensitive templates are gated by feedforward driving input signals. Such a modulating feedback driven gain control mechanism relates to mechanisms proposed by Roelfsema and colleagues (Lamme and Roelfsema, 2000; Roelfsema et al., 2002; Roelfsema, 2006) in which spatial detail is generated by feature-driven low-level processes and representations and subsequently associated with coarse-grained context information provided by intermediate and higher-levels of cortical computation. The mechanisms implemented in the proposed model are consistent with theoretical predictions from computational constraints visual perception imposes on the underlying architecture (Tsotsos, 1988). The advantages in computational complexity have been calculated for principles such as hierarchical organization, localized receptive field computations, and dedicated (distributed) maps of feature representation and their combination. Feedback has been suggested to steer an attentional beam by selecting a spatial region and their computational resources (Tsotsos, 2005). In the proposed architecture feedback also selectively enhances representations of features by increasing their gain which are coherent with the predictions generated at higher-level stages with more condensed coding of shape and figural properties. Also we emphasize that this provides a key to enhance (and make accessible) localized shape features, such as sharp edges, as part of a shape configuration that is represented on a coarser scale.

4.4. MODEL LIMITATIONS AND FURTHER EXTENSIONS

The proposed model architecture emphasized the computational role of feedforward and feedback mechanisms in order to generate a hierarchical distributed representation of shape information. For that reason, we focused on the representational aspects as steady-state solutions of an otherwise dynamic interaction between neuronal populations and representations distributed over several model areas. We did not, so far, investigate the temporal response phases observed for shape sensitive cells in V4

(Yau et al., 2013). The work of Roelfsema and colleagues has shown that different response phases exist that can be reliably assigned to different mechanisms in processing, namely for feature detection, figure-ground segregation, and attention (Roelfsema et al., 2007). We have demonstrated that such separate but temporally overlapping phases can be accounted for by a recurrent network of mutually interacting neuronal sites. The network model has been composed of the same components like the present model architecture (Raudies and Neumann, 2010). It would thus be interesting to reveal whether similar temporal phases can be identified for model V4 cells that may give rise to identify different signatures indicative of contributions from delayed neuronal mechanisms that are involved in the computation of figural shape information.

Different signal streams (particularly in the feedforward sweep of feature processing) operate on different temporal scales. Several lines of evidence suggest that the dorsal and the ventral streams of processing do not operate entirely in isolation but mutually interact at different levels (Felleman and Van Essen, 1991; Markov et al., 2013). Also different response characteristics of cells may define different temporal routes of fast and slow processing (Born, 2001) that may help fusing information from different pathways. Here, we did not take into account such interactions based on different temporal effectivenesses. However, other model investigations capitalized on combining information from different channels to improve the selectivity of representation. For example, edge detection and grouping (in the ventral pathway) could be enhanced through mutually inhibitory gain control (which is similar as the normalization stage described here) generated by representations in the dorsal pathway. Since the dorsal representation is created by magno-cellular responses, such inhibition arrives already early to shape the selectivity of shape representations in the ventral path that is mainly driven by parvo-cellular responses (Shi et al., 2013). Similarly, interactions between the motion and form pathway have been suggested to help disambiguating localized features that give rise to occlusion cues which, in turn, support the disambiguation of object representation in the motion representation (Bayerl and Neumann, 2007; Beck and Neumann, 2010). Such detailed mechanisms would further enhance the proposed model architecture in refining the selectivities at different levels of low and intermediate representation.

As already pointed out above, the focus here is on the processing of 2D shape representations. In Cadieu et al. (2007) the authors have highlighted that their specific model investigation on shape representation in V4 is part of a larger hierarchically organized architecture for object recognition (Riesenhuber and Poggio, 1999; Serre and Poggio, 2010). Since their model principles relied on purely feedforward processing the insights provided in the work presented here might also shed some light on the mutual interactions between different processes on an even larger scale of object recognition processes. In addition, it would be interesting to find out how the representation of 3D surface patches (Yamane et al., 2008) seamlessly fit into a model computational architecture of recurrent shape computation.

In the presented coverage our model does not respond to contours elicited by contrasts of spatial luminance statistics caused by differently textured regions. However, the core mechanisms, including initial filtering, modulatory feedback and competitive interaction for normalization, are like those proposed in the current contribution. A model that focuses on the processing of such boundaries has been developed in Thielscher and Neumann (2003). It is thus very likely that the recent model architecture proposed here can be extended with processing stages capable to process texture define boundaries as well without changing the basic architecture and computational principles. Also not considered in the current version is a multi-scale approach. We acknowledge the theoretical justification of hierarchical multi-stage processing to build up a pyramid-like structure (Tsotsos, 2005). Incorporating this representational diversity would allow the processing of a wider range of curvature configurations in shape outlines. In addition, this would support a more robust segregation of border ownership on the basis of convexities in the figural outline. We have focused our efforts on the specification of a hierarchically organized network architecture that utilized bottom-up and top-down convergent processing flows. In order to keep the computational efforts and the simulation times within reasonable bounds we restricted our description to single scale components at the different model stages within the hierarchy. A more extended realization of components is certainly desired but left for future investigations.

Intermediate level representations involve cells with receptive fields that recruit multiple sub-field components (Mineault et al., 2012; Yau et al., 2013). The model of Cadieu et al. (2007) accounts for this by sequentially fitting the subunits of intermediate level receptive field models to match the response profiles of V4 responses measured experimentally. This yields a sampling structure of statistically significant inputs in a feature space that contributes a significant amount of feature input to generate the final response of a shape selective cell. So far, in our modeling we sampled the spatial and the feature domains regularly. This of course demands high representational as well as computational resources. Consequently, it would be of interest to see how an irregularly sampled 4D space-feature domain (with orientation and curvature features) can be embedded into the scheme of shape representation proposed here.

ACKNOWLEDGMENTS

We thank our two anonymous reviewers for their valuable comments and constructive criticism that helped to improve the manuscript. This work has been supported by a grant from the Transregional Collaborative Research Center SFB/TRR62 "A Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

REFERENCES

Anzai, A., Peng, X., and Van Essen, D. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nat. Neurosci.* 10, 1313–1321. doi: 10.1038/nn1975

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Bayerl, P., and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Comput.* 16, 2041–2066. doi: 10.1162/0899766041732404

Bayerl, P., and Neumann, H. (2007). Disambiguating visual motion by form-motion interaction – a computational model. *Int. J. Comp. Vis.* 72, 27–45. doi: 10.1007/s11263-006-8891-8

Baylis, G. C., and Driver, J. (2001). Shape coding in IT cells generalizes over contrast and mirror reversal but not figureground reversal. *Nat. Neurosci.* 4, 937–942. doi: 10.1038/nn0901-937

Beck, C., and Neumann, H. (2010). Interactions of motion and form in visual cortex - a Neura model. *J. Physiol. Paris* 104, 61–70. doi: 10.1016/j.jphysparis.2009.11.005

Beck, C., and Neumann, H. (2011). Combining feature selection and integration. A neural model for mt motion selectivity. *PLoS ONE* 6:e21254. doi: 10.1371/journal.pone.0021254

Ben-Shahar, O., and Zucker, S. (2004). Geometrical computations explain projection patterns of long–range horizontal connections in visual cortex. *Neural Comput.* 16, 445–476. doi: 10.1162/089976604772744866

Born, R. T. (2001). Visual processing: parallel–er and Parallel–er. *Curr. Biol.* 11, R566–R568. doi: 10.1016/S0960-9822(01)00345-1

Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* 17, 2112–2127.

Brosch, T., and Neumann, H. (2014). Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations. *Neural Netw.* 54, 11–16. doi: 10.1016/j.neunet.2014.02.005

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., and Poggio, T. (2007). A model of V4 shape selectivity and invariance. *J. Neurophysiol.* 98, 1733–1750. doi: 10.1152/jn.01265.2006

Caplovitz, G., and Tse, P. (2007). V3A processes contour curvature as a trackable feature for the perception of rotational motion. *Cereb. Cortex* 17, 1179–1189. doi: 10.1093/cercor/bhl029

Carandini, M., and Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science* 264, 1333–1336. doi: 10.1126/science.8191289

Carandini, M., and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Carandini, M., Heeger, D. J., and Movshon, J. A. (1999). "Linearity and gain control in V1 simple cells," in *Cerebral Cortex Volume 13 Models of Cortical Circuits*, eds P. S. Ulinski, E. G. Jones, and A. Peters (New York, NY: Kluwer Academic; Plenum Publishers), 401–443.

Connor, C. E., Brincat, S. L., and Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.* 17, 140–147. doi: 10.1016/j.conb.2007.03.002

De Pasquale, R., and Murray Sherman, S. (2013). A modulatory effect of the feedback from higher visual areas to V1 in the mouse. *J. Neurophysiol.* 109, 2618–2631. doi: 10.1152/jn.01083.2012

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond. B* 353, 1245–1255. doi: 10.1098/rstb.1998.0280

Eckhorn, R. (1999). Neural mechanisms of visual feature binding investigated with microelectrodes and models. *Vis. Cogn.* 6, 231–265. doi: 10.1080/135062899394975

Eckhorn, R., Reitboeck, H. J., Arndt, M., and Dicke, P. (1990). Feature linking via synchronization among distributed assemblies: simulations of results from cat visual cortex. *Neural Comput.* 2, 293–307. doi: 10.1162/neco.1990.2.3.293

Edelman, G. (1993). Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron* 10, 115–125. doi: 10.1016/0896-6273(93)90304-A

Felleman, D. J., and Van Essen, D. C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. *J. Neurophysiol.* 57, 889–920.

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Field, D. J., Hayes, A., and Hass, R. F. (1993). Contour integration by the human visual system: evidence for a local association field. *Vision Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Fowlkes, C. C., Martin, D. R., and Malik, J. (2007). Local figure-ground cues are valid for natural images. *J. Vis.* 7, 1–9. doi: 10.1167/7.8.2

Fukushima, K. (1980). Neocognitron: a self–organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Netw.* 1, 119–130. doi: 10.1016/0893-6080(88)90014-7

Gilbert, C. D., and Li, W. (2013). Top–down influences on visual processing. *Nat. Rev. Neurosci.* 14, 350–363. doi: 10.1038/nrn3476

Girard, P., and Bullier, J. (1989). Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. *J. Neurophysiol.* 62, 1287–1302.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychol. Rev.* 87, 1–51. doi: 10.1037/0033-295X.87.1.1

Grossberg, S., and Mingolla, E. (1985). Neural dynamics of perceptual grouping: textual, boundaries, and emergent segmentations. *Percept. Psychophys.* 38, 141–171. doi: 10.3758/BF03198851

Hansen, T., and Neumann, H. (2004). Neural mechanisms for the robust representation of junctions. *Neural Comput.* 16, 1013–1037. doi: 10.1162/089976604773135087

Hatori, Y., and Sakai, K. (2012). "Surface-based construction of curvature selectivity from the integration of local orientations," in *19th International Conference on Neural Information Processing (ICONIP)*, Vol. 7665, eds H. Tingwen, Z. Zhigang, L. Chuandong, and L. ChiSing (Berlin; Heidelberg: Springer), 425–433. doi: 10.1007/978-3-642-34487-9_52

Hegdé, J., and Van Essen, D. (2000). Selectivity for complex shapes in primate visual area V2. *J. Neurosci.* 20, 1–6.

Heitger, F., von der Heydt, R., Peterhans, E., Rosenthaler, L., and Kübler, O. (1998). Simulation of neural contour mechanisms: representing anomalous contours. *Image Vis. Comput.* 16, 407–421. doi: 10.1016/S0262-8856(97)00083-8

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hupé, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., and Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature* 394, 784–787. doi: 10.1038/29537

Ito, M., and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.* 24, 3313–3324. doi: 10.1523/JNEUROSCI.4364-03.2004

Jehee, J. F. M., Lamme, V. A. F., and Roelfsema, P. R. (2006). Boundary assignment in a recurrent network architecture. *Vision Res.* 47, 1153–1165. doi: 10.1016/j.visres.2006.12.018

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolozione omogenea. *Rivista Psicologia* 49, 7–30.

Kouh, M., and Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput.* 20, 1427–1451. doi: 10.1162/neco.2008.02-07-466

Lamme, V., and Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X

Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151. doi: 10.1016/j.tins.2012.11.006

Layher, G., Giese, M., and Neumann, H. (2014). Learning representations of animated motion sequences - a neural model. *Top. Cogn. Sci.* 6, 170–182. doi: 10.1111/tops.12075

Layton, O. W., Mingolla, E., and Yazdanbakhsh, A. (2012). Dynamic coding of border–ownership in visual cortex. *J. Vis.* 12, 1–21. doi: 10.1167/12.13.8

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient–based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563. doi: 10.1016/S0960-9822(95)00108-4

Louie, K., Grattan, L. E., and Glimcher, P. W. (2011). Reward value–based gain control: divisive normalization in parietal cortex. *J. Neurosci.* 31, 10627–10639. doi: 10.1523/JNEUROSCI.1237-11.2011

Markov, N., Ercsey-Ravasz, M., Van Essen, D., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Science* 342, 1238406. doi: 10.1126/science.1238406

Mineault, P., Khawaja, F., Butts, D., and Pack, C. (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proc. Natl. Acad. Sci. U.S.A.* 109, E972–E980. doi: 10.1073/pnas.1115685109

Mokhtarian, F. (1995). Silhouette–based isolated object recognition through curvature scale space. *IEEE Trans. PAMI* 17, 539–544. doi: 10.1109/34.391387

Mokhtarian, F., and Mackworth, A. (1986). Scale–based description and recognition of planar curves and two–dimensional shapes. *IEEE Trans. PAMI* 8, 34–43. doi: 10.1109/TPAMI.1986.4767750

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comp. Vis.* 80, 45–57. doi: 10.1007/s11263-007-0118-0

Neumann, H., and Mingolla, E. (2001). "Computational neural models of spatial integration in perceptual grouping," in *From Fragments to Objects: Grouping and Segmentation in Vision*, Chapter 12, eds T. Shipley and P. Kellman (Amsterdam: Elsevier), 353–400.

Neumann, H., and Sepp, W. (1999). Recurrent v1-v2 interaction in early visual boundary processing. *Biol. Cybern.* 81, 425–444. doi: 10.1007/s004220050573

Neumann, H., Yazdanbakhsh, A., and Mingolla, E. (2007). Seeing surfaces: the brain's vision of the world. *Phys. Life Rev.* 4, 189–222. doi: 10.1016/j.plrev.2007.09.001

O'Herron, P., and von der Heydt, R. (2011). Representation of object continuity in the visual cortex. *J. Vis.* 11, 1–9. doi: 10.1167/11.2.12

O'Herron, P., and von der Heydt, R. (2013). Remapping of border ownership in the visual cortex. *J. Neurosci.* 33, 1964–1974. doi: 10.1523/JNEUROSCI.2797-12.2013

Parent, P., and Zucker, S. (1989). Trace inference, curvature consistency, and curve detection. *IEEE Trans. PAMI* 11, 823–839. doi: 10.1109/34.31445

Pasupathy, A., and Connor, C. (1999). Responses to contour features in macaque area V4. *J. Neurophysiol.* 82, 2490–2502.

Piëch, V., Li, W., Reeke, G. N., and Gilbert, C. D. (2013). Network model of top–down influences on local gain and contextual interactions in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 110, E4108–4117. doi: 10.1073/pnas.1317019110

Qiu, F. T., and von der Heydt, R. (2005). Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron* 47, 155–166. doi: 10.1016/j.neuron.2005.05.028

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Raudies, F., and Neumann, H. (2010). A model of neural mechanisms in monocular transparent motion perception. *J. Physiol. Paris* 104, 71–83. doi: 10.1016/j.jphysparis.2009.11.010

Reynolds, J. H., and Heeger, D. J. (2009). The normalization model of attention. *Neuron* 61, 168–185. doi: 10.1016/j.neuron.2009.01.002

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3, 1199–1204. doi: 10.1038/81479

Rodríguez-Sánchez, A. J., and Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. *PLoS ONE* 7:e42058. doi: 10.1371/journal.pone.0042058

Roelfsema, P., Lamme, V., Spekreijse, H., and Bosch, H. (2002). Figure–ground segregation in a recurrent network architecture. *J. Cogn. Neurosci.* 14, 525–537. doi: 10.1162/08989290260045756

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Ann. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Roelfsema, P. R., Tolboom, M., and Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron* 56, 785–792. doi: 10.1016/j.neuron.2007.10.006

Salin, P.-A., and Bullier, J. (1995). Corticocortical connections in the visual system: structure and function. *Physiol. Rev.* 75, 107–154.

Schwartz, E., Desimone, R., Albright, T., and Gross, C. (1983). Shape Recognition and inferior temporal neurons. *Proc. Natl. Acad. Sci. U.S.A.* 80, 5776–5778. doi: 10.1073/pnas.80.18.5776

Self, M. W., Kooijmans, R. N., Supér, H., Lamme, V. A., and Roelfsema, P. R. (2012). Different glutamate receptors convey feedforward and recurrent processing in macaque V1. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11031–11036. doi: 10.1073/pnas.1119527109

Serre, T., and Poggio, T. (2010). A neuromorphic approach to computer vision. *Commun. ACM* 53, 54. doi: 10.1145/1831407.1831425

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *TPAMI* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Shi, X., Wang, B., and Tsotsos, J. K. (2013). "Early recurrence improves edge detection," in *Proceedings of the 24th British Machine Vision Conference* (Bristol, UK).

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci.* 2:4. doi: 10.3389/neuro.10.004.2008

Tabernik, D., Kristan, M., Boben, M., and Leonardis, A. (2014). "Using discriminative analysis for improving hierarchical compositional models," in *Proc. 19th Computer Vision Winter Workshop*, ed Z. Kukelova (Krtiny).

Thielscher, A., and Neumann, H. (2003). Neural mechanisms of cortico–cortical interaction in texture boundary detection: a modeling approach. *Neuroscience* 122, 921–939. doi: 10.1016/j.neuroscience.2003.08.050

Tsotsos, J. K. (1988). "Computational processes in human vision: an interdisciplinary perspective," in *How Does Human Vision Beat the Computational Complexity of Visual Perception?* ed Z. Pylyshyn (Norwood, NJ: Ablex Press), 286–338.

Tsotsos, J. K. (2005). "Neurobiology of attention," in *Computational Foundations for Attentive Processes*, eds L. Itti, G. Rees, and J. K. Tsotsos (Amsterdam, NL: Elsevier), 3–7.

Ullman, S. (1995). Sequence-seeking and counter streams: a computational model for bi-directional information flow in the visual cortex. *Cereb. Cortex* 5, 1–11. doi: 10.1093/cercor/5.1.1

Ungerleider, L. G., and Haxby, J. V. (1994). 'what' and 'where' in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165. doi: 10.1016/0959-4388(94)90066-3

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science* 224, 1260–1262. doi: 10.1126/science.6539501

Weidenbacher, U., and Neumann, H. (2009). Extraction of surface–related features in a recurrent model of V1–V2 interactions. *PLoS ONE* 4:e5909. doi: 10.1371/journal.pone.0005909

Williford, J., and von der Heydt, R. (2013). Border–ownership coding. *Scholarpedia* 8, 30040. doi: 10.4249/scholarpedia.30040

Wilson, H., and Wilkinson, F. (1998). Detection of global structure in Glass patterns: implications for form vision. *Vision Res.* 38, 2933–2947. doi: 10.1016/S0042-6989(98)00109-6

Yamane, Y., Carlson, E., Bowman, K., Wang, Z., and Connor, C. (2008). A neural code for three–dimensional object shape in macque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360. doi: 10.1038/nn.2202

Yau, J., Pasupathy, A., Brincat, S., and Connor, C. (2013). Curvature processing dynamics in macaque area V4. *Cereb. Cortex* 23, 198–209.

Zhaoping, L. (1998). A neural model of contour integration in the primary visual cortex. *Neural Comput.* 10, 903–940. doi: 10.1162/089976698300017557

Zhou, H., Friedman, H. S., and Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611. doi: 10.1113/jphysiol.2002.033555

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 March 2014; accepted: 22 July 2014; published online: 11 August 2014.
*Citation: Tschechne S and Neumann H (2014) Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation. Front. Comput. Neurosci. **8**:93. doi: 10.3389/fncom.2014.00093*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Tschechne and Neumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 30 July 2014
doi: 10.3389/fncom.2014.00080

Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective COSFIRE models

George Azzopardi* and Nicolai Petkov

Intelligent Systems, Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, Netherlands

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Bart Ter Haar Romeny, Eindhoven University of Technology, Netherlands
Mario Vento, University of Salerno, Italy

****Correspondence:***
George Azzopardi, Intelligent Systems, Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, P.O. Box 800, 9700 AV Groningen, Netherlands
e-mail: g.azzopardi@rug.nl

The remarkable abilities of the primate visual system have inspired the construction of computational models of some visual neurons. We propose a trainable hierarchical object recognition model, which we call *S*-COSFIRE (*S* stands for *Shape* and COSFIRE stands for *Combination Of Shifted FIlter REsponses*) and use it to localize and recognize objects of interests embedded in complex scenes. It is inspired by the visual processing in the ventral stream (V1/V2 → V4 → TEO). Recognition and localization of objects embedded in complex scenes is important for many computer vision applications. Most existing methods require prior segmentation of the objects from the background which on its turn requires recognition. An *S*-COSFIRE filter is automatically configured to be selective for an arrangement of contour-based features that belong to a prototype shape specified by an example. The configuration comprises selecting relevant vertex detectors and determining certain blur and shift parameters. The response is computed as the weighted geometric mean of the blurred and shifted responses of the selected vertex detectors. *S*-COSFIRE filters share similar properties with some neurons in inferotemporal cortex, which provided inspiration for this work. We demonstrate the effectiveness of *S*-COSFIRE filters in two applications: letter and keyword spotting in handwritten manuscripts and object spotting in complex scenes for the computer vision system of a domestic robot. *S*-COSFIRE filters are effective to recognize and localize (deformable) objects in images of complex scenes without requiring prior segmentation. They are versatile trainable shape detectors, conceptually simple and easy to implement. The presented hierarchical shape representation contributes to a better understanding of the brain and to more robust computer vision algorithms.

Keywords: hierarchical representation, object recognition, shape, ventral stream, vision and scene understanding, robotics, handwriting analysis

1. INTRODUCTION

Shape is perceptually the most important visual characteristic of an object. Although there is no formal definition—as with most perceptual related concepts—it is understood that the two-dimensional shape of an object is characterized by the relative spatial positions of a collection of contour-based features.

Let us consider, for instance, the square in **Figure 1A**, which we refer to as a reference or prototype object. From the point of view of visual perception the incomplete object in **Figure 1B** is very similar to the prototype even though it is composed of only 25% of the contour pixels of the reference object. On the contrary, the closed polygon in **Figure 1C**, which has the bottom half equivalent to that of the prototype is perceptually less similar to it. Furthermore, there is little perceptual similarity between the prototype and its scrambled contour parts shown in **Figure 1D**.

As a matter of fact, there is neurophysiological evidence that objects, such as faces, are recognized by detecting certain features that are spatially arranged in a certain way (Kobatake and Tanaka, 1994). By means of single-cell recordings in adult monkeys it was, for instance, found that a neuron in inferotemporal cortex gives similar responses for the two images shown in **Figures 2A,B**. The icon presented in **Figure 2B** is a simplified version of the monkey's face shown in **Figure 2A**. It only consists of a circle that surrounds a horizontally-aligned pair of spots on top of a horizontal bar. Removing one of these features, **Figures 2C,D**, causes the concerned cell to give very small response.

Another neurophysiological study (Brincat and Connor, 2004) reveals that some neurons in inferotemporal cortex integrate information about the curvatures, orientations, and positions of multiple (typically 2–4) simple contour elements, such as angles or curved contour segments. In that study the authors argue that their findings are in line with other studies that support parts-based shape representation theories (Marr and Nishihara, 1978; Riesenhuber and Poggio, 1999; Mel and Fiser, 2000; Edelman and Intrator, 2003), and suggest that non-linear integration in the inferotemporal cortex might help to extend sparseness of shape representation along the ventral stream.

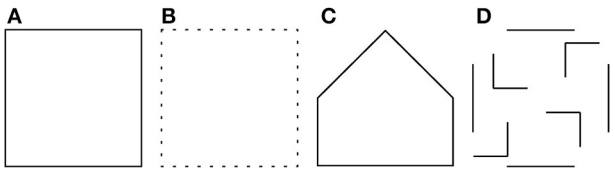


FIGURE 1 | (A) A prototype shape. **(B)** A test pattern that has only 25% similarity (computed by template matching) to the prototype is perceptually more similar to the prototype than the polygon in **(C)** and the set of contour parts in **(D)**, both of which have 50% similarity (computed by template matching) to the prototype.

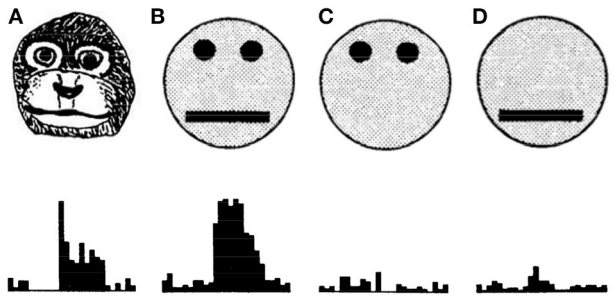


FIGURE 2 | (A–D) A set of stimuli used in an electrophysiological study Kobatake and Tanaka (1994) to test the selectivity of a neuron in inferotemporal cortex. **(Bottom)** The activity of the concerned neuron for the corresponding stimuli. The neuron gives high response only when the stimulus contains a detailed or simplified representation of the face boundary that surrounds a pair of eyes on top of a mouth. If any of these features is missing, the neuron gives negligible response.

Tsotsos (1990) showed that hierarchical architectures are more appropriate for object detection in contrast to unbounded visual search which is known to be NP-complete. This has led to the proposal of a number of hierarchical models (Mel and Fiser, 2000; Scalzo and Piater, 2005; DiCarlo and Cox, 2007; Rodríguez-Sánchez and Tsotsos, 2012). Existing approaches that consider the spatial relationship of features include the so-called standard model (Serre et al., 2007), some probabilistic techniques, such as the generative constellation model (Fergus et al., 2003; Fei-Fei et al., 2007) and a hierarchical model of object categories (Fidler and Leonardis, 2007; Fidler et al., 2008). These approaches rely on summation of the responses of elementary feature detectors and may find the images in **Figures 1C,D** quite similar to the prototype in **Figure 1A**. For instance, such a technique may consider a circle with a horizontal line within it as a face even though the representations of the eyes are missing, **Figures 2C,D**.

We introduce a hierarchical object detection technique which is motivated by the shape selectivity of some neurons in inferotemporal cortex. The principal idea is to construct a shape-selective filter that combines the responses of some simpler filters that detect some partial features of the concerned shape in specific positions that are characteristic of that shape. We call this approach to the construction of filters Combination Of Shifted Filter REsponses (COSFIRE). We successfully applied this approach to the construction of line and edge detectors (Azzopardi and Petkov, 2012; Azzopardi et al., 2014) and simple contour-related features, such as vascular bifurcations (Azzopardi and Petkov, 2013b). In Azzopardi and Petkov (2013b) we demonstrated how the collective responses of multiple COSFIRE filters to segmented patterns, such as handwritten digits, can be used to form a shape descriptor with high discrimination ability. That descriptor, however, does not take into account the relative spatial arrangement of the concerned features. Similar to other shape descriptors (Belongie et al., 2002; Grigorescu and Petkov, 2003; Ghosh and Petkov, 2005; Latecki et al., 2005; Lauer et al., 2007; Ling and Jacobs, 2007; Goh, 2008; Almazan et al., 2012) that approach works well with segmented objects, but it is not effective for the detection of objects embedded in complex scenes. In order to distinguish the two types of filter, we refer to the composite shape-selective filter that we propose in this paper as *S*-COSFIRE and to the filter proposed in Azzopardi and Petkov (2013b) as *V*-COSFIRE (*S* and *V* stand for shape and vertex, respectively).

There are three aspects in which the *S*-COSFIRE filters that we propose differ from other hierarchical models that also consider the spatial geometric arrangement of parts. *First*, our model is implemented in a filter that gives a scalar response (between 0 and 1) for each position in the image. The higher the value the more similar the shape around the concerned location is to the prototype shape. An *S*-COSFIRE filter can be thought of a model of a shape-selective neuron in inferotemporal cortex of the type studied in Kobatake and Tanaka (1994); Brincat and Connor (2004), which fires only when a specific arrangement of contour-based features is present in its receptive field. It addresses object recognition and localization as a joint problem, which is in line with how Marr (1982) defined the sense of seeing: "... to know what is where by looking." In contrast, the other methods referred to above use multiple prototypes and consider several responses from different feature detectors to form a mixture of probability distributions or a vector of responses. For these methods, the geometrical spatial arrangement of the concerned prototype defining parts is achieved by training a supervised classifier and subsequently the similarity between a test pattern and a prototype is computed by a distance metric. Moreover, they suffer from insufficient robustness to localization because they treat this matter at a region level (sliding window) rather than at a pixel level.

Second, since the omission of an object part can radically change shape perception, we regard every feature (and its relative position) that forms part of a prototype shape as essential. This aspect is implemented as an AND-type operation of an *S*-COSFIRE filter. It is in contrast to other models that rely on summation, and therefore achieve a response even when any of the prototype-defining features is missing. These models may thus match objects that are perceptually different.

Third, while the *S*-COSFIRE approach that we present achieves invariance to rotation, scaling, and reflection by simply manipulating some model parameters, the other techniques can only achieve invariance to such geometric transformations by extending the training set with example objects that are rotated, scaled and/or reflected versions of a prototype.

The rest of the paper is organized as follows: in section 2 we present the proposed hierarchical *S*-COSFIRE model. In

section 3, we demonstrate its effectiveness in two applications: keyword spotting in handwritten manuscripts and vision for a home tidying pickup robot. Section 4 contains a discussion on the properties of the *S*-COSFIRE filters and finally we draw conclusions in section 5.

2. METHODS

The following example illustrates the main idea of the proposed method. We consider the triangle, shown in **Figure 3A**, as a shape of interest and we call it *prototype*. We use this prototype to automatically configure an *S*-COSFIRE filter that will respond to shapes that are identical with or similar to this prototype.

A shape-selective *S*-COSFIRE filter takes input from simpler filters; here filters that are selective for vertices. We use vertex-selective COSFIRE filters of the type proposed in Azzopardi and Petkov (2013b) to detect the vertices of the prototype shape. Such a filter, which we refer to it as *V*-COSFIRE, combines the responses of line detectors, the areas of support of which are indicated by the small ellipses in **Figure 3A**.

The response of an *S*-COSFIRE filter is computed by combining the responses of the concerned *V*-COSFIRE filters in the centers of the corresponding circles by weighted geometric mean. The preferred orientations and the preferred apertures of these filters together with the locations at which we take their responses are determined by analysing the responses of a set of *V*-COSFIRE filters to the prototype shape. Consequently, the *S*-COSFIRE filter will be selective for the given spatial arrangement of vertices of specific orientations and apertures. Taking the responses of *V*-COSFIRE filters at different locations around a point can be implemented by shifting the responses appropriately before using them for the pixel-wise evaluation of a multivariate function which gives the *S*-COSFIRE filter output.

2.1. DETECTION OF VERTEX FEATURES BY *V*-COSFIRE FILTERS

We denote by $r_{V_{f_i}}(x, y)$ the response of a *V*-COSFIRE filter V_{f_i} that is selective for a vertex f_i. We threshold these responses at a given fraction t_1 $(0 \leq t_1 \leq 1)$ of the maximum response across all image coordinates (x, y) and denote these thresholded responses by $|r_{V_{f_i}}(x, y)|_{t_1}$. We use the publicly available Matlab implementation[1] of *V*-COSFIRE filters. Such a filter uses as input the responses of given channels of a bank[2] of Gabor filters. For further technical details about the properties of *V*-COSFIRE filters we refer to Azzopardi and Petkov (2013b).

We use a bank of *V*-COSFIRE filters that are selective for vertices of different orientations (in intervals of $\pi/6$ radians) and different apertures (in intervals of $\pi/6$ radians), **Figure 3B**. For the considered prototype the strongest responses are obtained by three *V*-COSFIRE filters that are selective for vertices of the types f_{13}, f_{17}, and f_{21}, shown in **Figure 3B**. The corresponding locations, (x_1, y_1), (x_2, y_2), (x_3, y_3), at which they obtain the maximum responses are indicated in **Figure 3C**.

2.2. CONFIGURATION OF AN *S*-COSFIRE FILTER

An *S*-COSFIRE filter uses as input the responses of selected *V*-COSFIRE filters $V_{f_{j_i}}$, $i = 1 \ldots n$, each selective for some vertex f_{j_i}, around a certain position (ρ_i, ϕ_i) with respect to the center of the *S*-COSFIRE filter. A 3-tuple $(V_{f_{j_i}}, \rho_i, \phi_i)$ that consists of a *V*-COSFIRE filter specification $V_{f_{j_i}}$ and two scalar values (ρ_i, ϕ_i) characterizes the properties of a vertex that is present in the given prototype shape: $V_{f_{j_i}}$ represents a *V*-COSFIRE filter that is selective for a vertex f_{j_i} and (ρ_i, ϕ_i) are the polar coordinates of the location at which its response is taken with respect to the center of the *S*-COSFIRE filter. In the following we explain how we obtain the parameter values of such vertices around a given point of interest.

For each location in the input image of the prototype shape we take the maximum value of all responses achieved by the bank of *V*-COSFIRE filters mentioned above. The positions that have values greater than those of their corresponding 8-neighbors are chosen as the points that have local maximum responses. For each such point (x_i, y_i) we determine the polar coordinates (ρ_i, ϕ_i) with respect to the center of the *S*-COSFIRE filter, **Figure 3C**.

[1]The Matlab implementation of a *V*-COSFIRE filter can be downloaded from http://matlabserver.cs.rug.nl/

[2]Here we use a bank of Gabor filters with five wavelengths $\lambda = \{4, 4\sqrt{2}, 8, 8\sqrt{2}, 16\}$ and six equidistant orientations $\theta \in \{0, \frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \frac{5\pi}{6}\}$

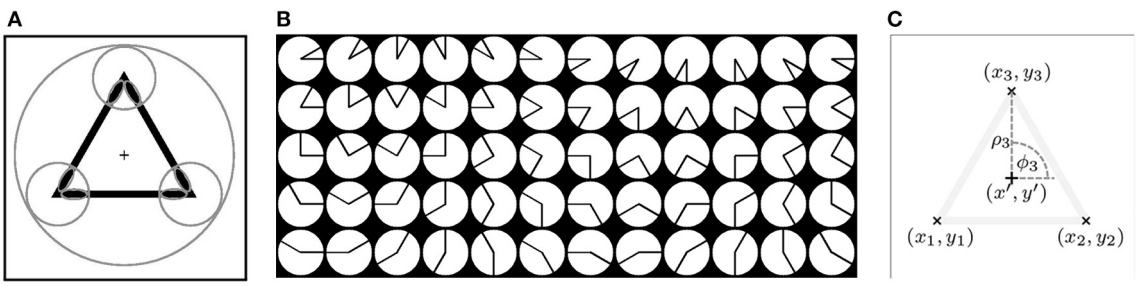


FIGURE 3 | (A) The triangle is the prototype shape of interest and the "+" marker indicates the center of the user-specified large circle. The small circles indicate the supports of three vertex detectors that are identified as relevant for the concerned prototype shape. The small ellipses represent the supports of line detectors that are selective for the contour parts of the corresponding vertices. **(B)** A data set of 60 synthetic vertices, $f_1, \ldots, f_{60}$ (left-to-right, top-to-bottom). A *V*-COSFIRE filter V_{f_k} is selective for a vertex f_k. **(C)** Configuration of an *S*-COSFIRE filter. The "×" markers indicate the locations, (x_1, y_1), (x_2, y_2), (x_3, y_3), where the corresponding three *V*-COSFIRE filters, $V_{f_{13}}$, $V_{f_{17}}$, $V_{f_{21}}$, achieve the maximum responses. These locations correspond to the three vertices of the prototype shape, which is rendered here with low contrast. The Cartesian coordinates of each point (x_i, y_i) are converted into the polar coordinates (ρ_i, ϕ_i) with respect to the given point of interest (x', y'), indicated by the "+" marker.

Then we determine the V-COSFIRE filters, the responses of which are greater than a fraction $t_2 = 0.75$ of the maximum response $r_{V_{f_i}}(x, y)$ for all $i \in \{1, \ldots n_f\}$ where n_f is the number of V-COSFIRE filters used across all locations in the input image. Thus, multiple V-COSFIRE filters can be significantly activated for the same location (ρ_i, ϕ_i). The selected points characterize the dominant vertices in the given prototype shape of interest.

We denote by $S_{\mathsf{S}} = \left\{(V_{f_{j_i}}, \rho_i, \phi_i) \mid i = 1 \ldots n_f\right\}$ the set of parameter value combinations, which describes the properties and locations of a number of vertices. The subscript S stands for the prototype shape of interest. Every tuple in set S_{S} specifies the parameters of some vertex in prototype S. For the prototype shape of interest in **Figure 3A**, the selection method described above results in three vertices with parameter values specified by the tuples in the following set: $S_{\mathsf{S}} = \left\{(V_{f_{j_1=21}}, \rho_1 = 50, \phi_1 = \pi/2),\ (V_{f_{j_2=13}}, \rho_2 = 50, \phi_2 = 7\pi/6),\ (V_{f_{j_3=17}}, \rho_3 = 50, \phi_3 = 5\pi/3)\right\}$.

2.3. BLURRING AND SHIFTING *V*-COSFIRE RESPONSES

The above configuration results in an S-COSFIRE filter that is selective for a preferred spatial arrangement of three vertices forming an equilateral triangle. Next, we use the responses of the V-COSFIRE filters that are selective for the corresponding vertices to compute the output of the S-COSFIRE filter as follows.

First, we *blur* the responses of the V-COSFIRE filters in order to allow for some tolerance in the position of the respective vertices. This increases the generalization ability of the S-COSFIRE filter under construction. We define the blurring operation as the computation of maximum value of the weighted thresholded responses of a V-COSFIRE filter. For weighting we use a Gaussian function $G_\sigma(x, y)$, the standard deviation σ of which is a linear function of the distance ρ from the center of the S-COSFIRE filter: $\sigma = \sigma_0 + \alpha\rho$ where σ_0 and α are constants. The choice of this linear function is inspired by the visual system of the brain for which we provide more detail in section 4. For $\alpha > 0$, which we use, the tolerance to the position of the respective vertices increases with an increasing distance ρ from the support center of the concerned S-COSFIRE filter.

Second, we *shift* the blurred responses of each V-COSFIRE filter by a distance ρ_i in the direction opposite to ϕ_i. With this shifting the concerned V-COSFIRE filter responses, which are located at different positions (ρ_i, ϕ_i) meet at the support center of the S-COSFIRE filter. The output of the S-COSFIRE filter can then be evaluated as a pixel-wise multivariate function of the shifted and blurred responses of V-COSFIRE filter responses. In polar coordinates, the shift vector is specified by $(\rho_i, \phi_i + \pi)$, and in Cartesian coordinates, it is $(\Delta x_i, \Delta y_i)$ where $\Delta x_i = -\rho_i \cos\phi_i$, and $\Delta y_i = -\rho_i \sin\phi_i$. We denote by $s_{V_{f_{j_i}}, \rho_i, \phi_i}(x, y)$, the blurred and shifted thresholded response of a V-COSFIRE filter that is specified by the i-th tuple $(V_{f_{j_i}}, \rho_i, \phi_i)$ in the set S_{S}:

$$s_{V_{f_{j_i}}, \rho_i, \phi_i}(x, y) \stackrel{\text{def}}{=} \max_{x', y'} \left\{ \left| r_{V_{f_{j_i}}}(x - x' - \Delta x_i, y - y' - \Delta y_i) \right|_{t_1} G_\sigma(x', y') \right\},$$
$$\text{where } -3\sigma \le x', y' \le 3\sigma \qquad (1)$$

Figure 4 illustrates the blurring and shifting operations for this S-COSFIRE filter, applied to the image shown in **Figure 3A**.

We define the response $r_{S_{\mathsf{S}}}(x, y)$ of an S-COSFIRE filter as the weighted geometric mean of the blurred and shifted thresholded responses of the selected V-COSFIRE filters $s_{V_{f_{j_i}}, \rho_i, \phi_i}(x, y)$:

$$r_{S_{\mathsf{S}}}(x, y) \stackrel{\text{def}}{=} \left| \left(\prod_{i=1}^{|S_{\mathsf{S}}|} \left(s_{V_{f_{j_i}}, \rho_i, \phi_i}(x, y) \right)^{\omega_i} \right)^{1 / \sum_{i=1}^{|S_{\mathsf{S}}|} \omega_i} \right|_{t_3},$$
$$\omega_i = \exp^{-\frac{\rho_i^2}{2\sigma'^2}},\ 0 \le t_3 \le 1 \qquad (2)$$

where $|.|_{t_3}$ stands for thresholding the response at a fraction t_3 of its maximum across all image coordinates (x, y). For $1/\sigma' = 0$, the computation of the S-COSFIRE filter is equivalent to the standard geometric mean, where the s-quantities have the same contribution. Otherwise, for $1/\sigma' > 0$, the input contribution of s-quantities decreases with an increasing value of the corresponding parameter ρ. In our experiments we use a value of the standard deviation σ' that is computed as a function of the maximum value of the given set of ρ values: $\sigma' = (-\rho_{\max}{}^2 / 2 \ln 0.5)^{1/2}$, where $\rho_{\max} = \max_{i \in \{1 \ldots |S_{\mathsf{S}}|\}}\{\rho_i\}$. We make this choice in order to achieve a maximum value $\omega = 1$ of the weights in the center (for $\rho = 0$), and a minimum value $\omega = 0.5$ in the periphery (for $\rho = \rho_{\max}$).

Figure 4D shows the output of an S-COSFIRE filter which is defined as the weighted geometric mean of three blurred and shifted response images obtained by the three concerned V-COSFIRE filters. Note that this filter responds in the middle of a spatial arrangement of three vertices that is identical with or similar to that of the prototype shape S, which was used for the configuration of the S-COSFIRE filter. In this example, the S-COSFIRE filter reacts strongly in a given point that is surrounded by three vertices each having an aperture of $\pi/3$ radians: one northward-pointing, another one south-west-pointing and a south-east-pointing vertex to the north, south-west, and south-east of that point, respectively. Besides the complete triangle that was used for configuration, the concerned filter also detects the Kanizsa-type illusory triangle. This is in line with neurophysiological and psychophysical evidence, in that the visual system is capable of detecting a shape with illusory contours, based on its visible salient parts. A thorough review of this phenomenon is provided in Roelfsema (2006).

2.4. TOLERANCE TO GEOMETRIC TRANSFORMATIONS

The proposed S-COSFIRE filters are tolerant to rotations, scales and reflections. Similar to a V-COSFIRE filter, such a tolerance is achieved by manipulating the values of some parameters rather than by configuring separate filters by rotated, scaled, and reflected versions of the prototype shape of interest.

2.5. TOLERANCE TO ROTATION

Using the set S_{S} that defines the concerned S-COSFIRE filter, we form a new set $\Re_\psi(S_{\mathsf{S}})$ that defines a new filter, which is

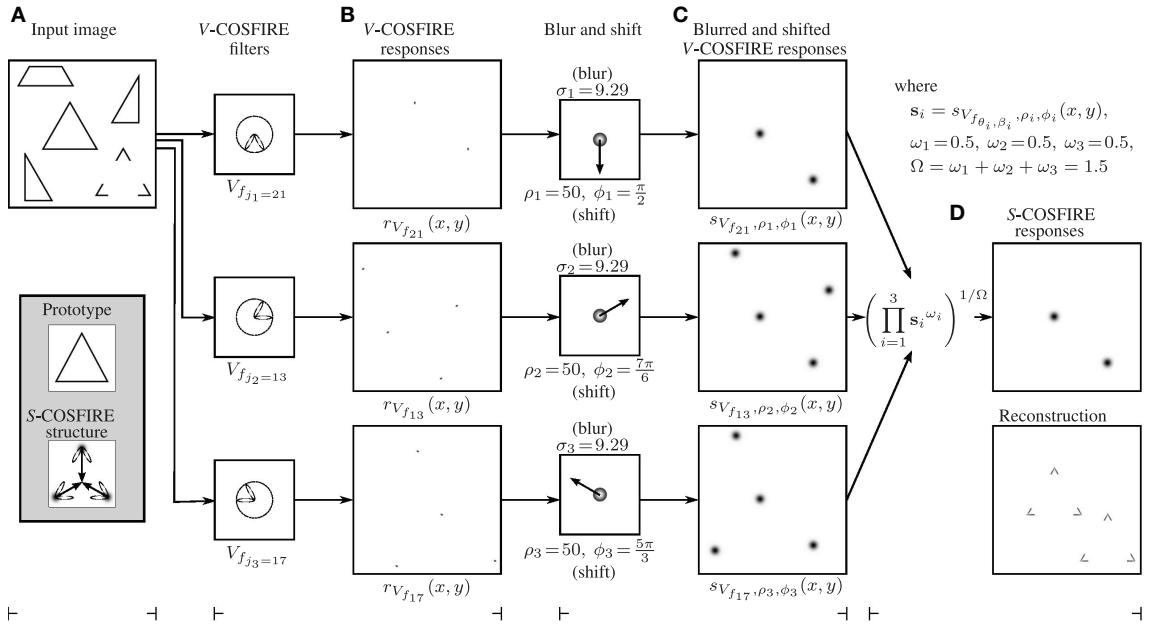


FIGURE 4 | (A) Input image (of size 512×512 pixels). The enframed inlay images show (top) the enlarged prototype shape of interest, which is identical to the equilateral triangle in the input image and (bottom) the structure of the *S*-COSFIRE filter that is configured by this prototype. The ellipses illustrate the wavelengths and orientations of the Gabor filters that are used by the *V*-COSFIRE filters, and the dark blobs are intensity maps for blurring (Gaussian) functions. The blurred responses are then shifted by the corresponding vectors. **(B)** The *V*-COSFIRE filters that are automatically identified from the prototype shape and the corresponding response images to the input image. **(C)** We then blur (here we use $\sigma_0 = 0.1$ and $\alpha = 0.0853$ to compute σ_i) the thresholded (here at $t_1 = 0$) response $|r_{V_{f_j}}(x,y)|_{t_1}$ of each concerned *V*-COSFIRE filter and subsequently shift the resulting blurred response images by corresponding polar-coordinate vectors $(\rho_i, \phi_i + \pi)$. **(D)** We use weighted geometric mean (here $\sigma' = 91.44$) of all the blurred and shifted *V*-COSFIRE filters to compute (top) the output of the *S*-COSFIRE filter and show (bottom) the reconstruction of the detected features. The reconstruction is achieved by superimposing the Gabor filter responses that give input to the *S*-COSFIRE filter. The two local maxima in the output of the *S*-COSFIRE filter correspond to the triangle and to the perceived one in the input image. For better clarity we use inverted gray-level rendering to show the images in the right of the columns **(B–D)**.

selective for a version of the prototype shape S that is rotated by an angle ψ:

$$\Re_\psi(S_S) \stackrel{\text{def}}{=} \left\{ (\Re_\psi(V_{f_{j_i}}), \rho_i, \phi_i + \psi) \mid \forall\, (V_{f_{j_i}}, \rho_i, \phi_i) \in S_S \right\} \quad (3)$$

For each tuple $(V_{f_{j_i}}, \rho_i, \phi_i)$ in the original filter S_S that describes a certain vertex of the prototype shape, we provide a counterpart tuple $(\Re_\psi(V_{f_{j_i}}), \rho_i, \phi_i + \psi)$ in the new set $\Re_\psi(S_S)$. The set $\Re_\psi(V_{f_{j_i}})$ defines[3] a *V*-COSFIRE filter that is selective for vertex f_{j_i} that is also rotated by an angle ψ. The orientation of the concerned vertex and its polar angle position ϕ_i with respect to the support center of the *S*-COSFIRE filter are off-set by an angle ψ relative to the values of the corresponding parameters of the original vertex.

A rotation-invariant response is achieved by taking the maximum value of the responses of filters that are obtained with different values of the parameter ψ:

$$\hat{r}_{S_S}(x,y) \stackrel{\text{def}}{=} \max_{\psi \in \Psi} \left\{ r_{\Re_\psi(S_S)}(x,y) \right\} \quad (4)$$

where Ψ is a set of n_ψ equidistant orientations defined as $\Psi = \left\{ \frac{2\pi}{n_\psi} i \mid 0 \le i < n_\psi \right\}$.

2.6. TOLERANCE TO SCALING

Tolerance to scaling is achieved in a similar way. Using the set S_S that defines the concerned *S*-COSFIRE filter, we form a new set $T_\upsilon(S_S)$ that defines a new filter, which is selective for a version of the prototype shape S that is scaled in size by a factor υ:

$$T_\upsilon(S_S) \stackrel{\text{def}}{=} \left\{ (T_\upsilon(V_{f_{j_i}}), \upsilon\rho_i, \phi_i) \mid \forall\, (V_{f_{j_i}}, \rho_i, \phi_i) \in S_S \right\} \quad (5)$$

For each tuple $(V_{f_{j_i}}, \rho_i, \phi_i)$ in the original *S*-COSFIRE filter S_S that describes a certain vertex of the prototype shape, we provide a counterpart tuple $(T_\upsilon(V_{f_{j_i}}), \upsilon\rho_i, \phi_i)$ in the new set $T_\upsilon(S_S)$. The set $T_\upsilon(V_{f_{j_i}})$ defines[1] a V-COSFIRE filter that responds to a version of the vertex f_{j_i} scaled by the factor υ. The size of the concerned vertex and its distance to the center of the filter are scaled by the factor υ relative to the original values of the corresponding parameters.

A scale-invariant response is achieved by taking the maximum value of the responses of filters that are obtained with different values of the parameter υ:

[3]We refer to Azzopardi and Petkov (2013b) for the technical details about the invariance that is achieved by a *V*-COSFIRE filter.

$$\tilde{r}_{S_S}(x,y) \stackrel{\text{def}}{=} \max_{\upsilon \in \Upsilon}\{r_{T_\upsilon(S_S)}(x,y)\} \quad (6)$$

where Υ is a set of υ values equidistant on a logarithmic scale defined as $\Upsilon = \{2^{\frac{i}{2}} \mid i \in \mathbb{Z}\}$.

2.7. REFLECTION INVARIANCE

As to reflection invariance we first form a new set $\acute{S}_S$ from the set S_S as follows:

$$\acute{S}_S \stackrel{\text{def}}{=} \{(\acute{V}_{f_{j_i}}, \rho_i, \pi - \phi_i) \mid \forall\,(V_{f_{j_i}}, \rho_i, \phi_i) \in S_S\} \quad (7)$$

The set $\acute{V}_{f_{j_i}}$ defines[1] a new V-COSFIRE filter that is selective for the corresponding vertex f_{j_i} reflected about the y–axis. Similarly, the new S-COSFIRE filter $\acute{S}_S$ is selective for a reflected version of the prototype shape S also about the $y-$axis. A reflection-invariant response is achieved by taking the maximum value of the responses of the filters S_S and $\acute{S}_S$:

$$\acute{r}_{S_S}(x,y) \stackrel{\text{def}}{=} \max\{r_{S_S}(x,y), r_{\acute{S}_S}(x,y)\} \quad (8)$$

2.8. COMBINED TOLERANCE TO ROTATION, SCALING, AND REFLECTION

An S-COSFIRE filter achieves tolerance to all the above geometric transformations by taking the maximum value of the rotation- and scale-tolerant responses of the filters S_S and $\acute{S}_S$ that are obtained with different values of the parameters ψ and υ:

$$\bar{r}_{S_S}(x,y) \stackrel{\text{def}}{=} \max_{\psi \in \Psi, \upsilon \in \Upsilon}\left\{\hat{r}_{\Re_\psi(T_\upsilon(S_S))}(x,y), \hat{r}_{\Re_\psi(T_\upsilon(\acute{S}_S))}(x,y)\right\} \quad (9)$$

3. APPLICATIONS

In the following we demonstrate the effectiveness of the proposed S-COSFIRE filters by applying them in two practical applications: the spotting of keywords in handwritten manuscripts and the spotting of objects in complex scenes for the computer vision system of a domestic robot.

3.1. SPOTTING KEYWORDS IN HANDWRITTEN MANUSCRIPTS

The automatic recognition of keywords in handwritten manuscripts is an application that has been extensively investigated for several decades (Plamondon and Srihari, 2000; Frinken et al., 2012). Despite this effort the problem has not been solved yet.

As a demonstration, in **Figure 5** we show how to detect the keyword "Germany" in two handwritten manuscripts. We use the keyword prototype "Germany" that is shown enframed in **Figure 5A** to configure an S-COSFIRE filter that receives input from 13 V-COSFIRE filters, **Figure 5E**. **Figures 5C,D** show the responses of the concerned S-COSFIRE filter ($t_1 = 0.1, t_2 = 0.75$, $t_3 = 0.1$, $\sigma_0 = 0.67$, and $\alpha = 0.1$.) to the two manuscript images[4] in **Figures 5A,B**. It spots all the six instances of the keyword "Germany" and does not produce any false positives.

The S-COSFIRE filters that are selective for specific words may correspond to neurons or networks of neurons in a certain area in the posterior lateral-occipital cortex. This area receives input from V4 and is selective for combinations of vertices. It has been shown to play a role in the recognition of words and has been named Visual Word Form Area (Szwed et al., 2011).

3.2. VISION FOR A HOME TIDYING PICKUP ROBOT

Daily service robots that perform routine tasks are becoming popular as household appliances. Such tedious tasks include, but are not limited to, vacuum cleaning, setting up and cleaning up a dinner table, tidying up toys, and organizing closets. The design of domestic robots is a growing research area (Bandera et al., 2012; Jiang et al., 2012).

[4]The images in **Figures 5A,B** are extracted from the files named b01-049.png and b01-044.png, respectively, in the IAM offline database.

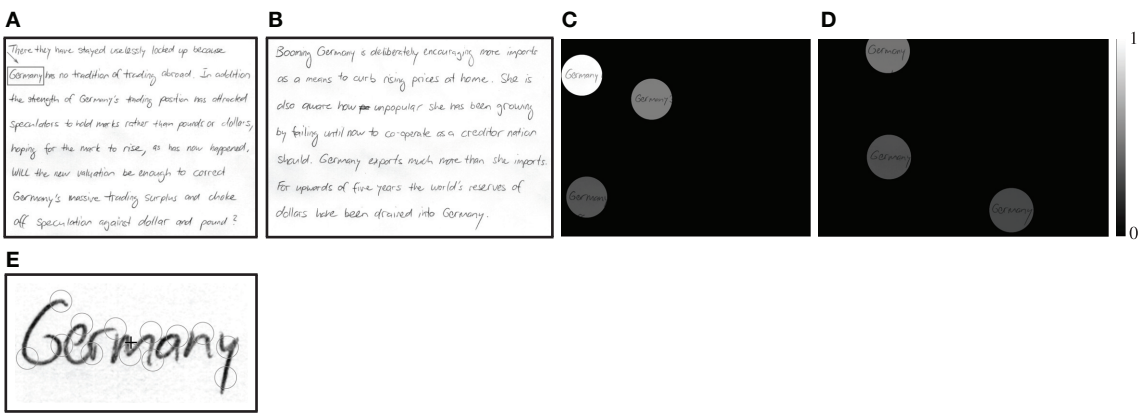


FIGURE 5 | An example of spotting the keyword "Germany" in (A,B) two handwritten manuscripts taken from the IAM offline database (Marti and Bunke, 2002). The indicated keyword "Germany" in **(A)** is used as a prototype to configure an S-COSFIRE filter. **(E)** The circles indicate the support areas of 13 V-COSFIRE filters that are used to provide input to the concerned S-COSFIRE filter with the "+" marker indicating its support center. **(C,D)** Normalized responses of this filter to the images in **(A,B)** rendered by shading of the spotted words. All six instances are detected. The strongest response is achieved for the word that was used for the configuration of the S-COSFIRE filter.

We demonstrate how the *S*-COSFIRE filters that we propose can be used by a personal robot to visually recognize objects of interest in indoor environments. As an illustration we consider a task for a tidying pickup robot to detect shoes in different rooms of a home that match the prototype shoe shown in **Figure 6A**.

We use a segmented prototype image of the shoe to configure an *S*-COSFIRE filter. The concerned *S*-COSFIRE filter receives input from three *V*-COSFIRE filters that are selective for different parts of the shoe. These parts are automatically chosen by the system from a circular local neighborhood of a point of interest that is indicated by a "+" marker. In practice, the concerned point of interest and the radius of the corresponding local neighborhood are manually specified by the user. The radii of the three circles are automatically computed in such a way that the circles touch each other. For the configuration of the concerned *V*-COSFIRE filters we use a bank of Gabor energy filters[5] with one wavelength ($\lambda = 4$) and 16 equidistant orientations $\left(\theta = \left\{\frac{\pi}{8}i \mid 0 \ldots 15\right\}\right)$, and we threshold the responses with $t_1 = 0.3$. Within each of the three circles, we consider a number of concentric circles, the radii of which increment in intervals of 4 pixels starting from 0. For the concerned three *V*-COSFIRE filters as well as the *S*-COSFIRE filter we use the same values of parameters α ($\alpha = 0.67$) and σ_0 ($\sigma_0 = 0.1$) in order to allow the same tolerance in the position of the involved edges and curvatures.

We created a data set that we call RUG-Shoes of 60 color images (of size 256 × 342 pixels) by taking pictures in different rooms of the same house. Of these images, 39 contain a pair of shoes of interest, another nine contain a single shoe and the remaining 12 do not contain any shoes. The distance above ground of the digital camera was varied between 50 cm and 1 m. All pictures of shoes were taken from the side view of the corresponding shoes. The shoes were, however, arranged in different orientations and their distances from the camera varied by at most 25% as compared to the distance which we used to take the image of the prototype shoe. We made the RUG-Shoes data set publicly available[6].

We use the configured *S*-COSFIRE filter to detect shoes in the data set of 60 images. We first convert every color image to grayscale and subsequently apply the concerned *S*-COSFIRE filter in reflection-, scale- $\left(\upsilon \in \left\{\frac{3}{4}, 1, \frac{5}{4}\right\}\right)$ and partially rotation-invariant $\left(\psi \in \left\{-\frac{\pi}{8}, 0, \frac{\pi}{8}\right\}\right)$ mode. The Gabor energy filters that we use to provide inputs to the *V*-COSFIRE filters are applied with isotropic suppression (Grigorescu et al., 2004) in order to reduce responses to texture. We threshold the responses of the concerned *S*-COSFIRE filter with t_3=0.1 and for each image we consider only the highest two responses. We obtain a perfect detection and recognition performance for all the 60 images in the RUG-Shoes data set. This means that we detect all the shoes in the given images with no false positives. **Figure 6B** illustrates the detection of some shoes in two of the images.

4. DISCUSSION

The trainable *S*-COSFIRE filters that we propose are part of a hierarchical object recognition approach that shares similarity with the ventral stream of visual cortex. In the first layer we detect lines and edges by Gabor filters, which are inspired by the function of orientation-selective cells in primary visual cortex (Daugman, 1985). Their responses are projected to a second layer and used by *V*-COSFIRE filters that detect vertices and curved contour segments. In our previous work (Azzopardi and Petkov, 2013b), we showed that such filters give responses that are qualitatively similar to a class of cells in area V4 in visual cortex. Finally, in a third layer we have *S*-COSFIRE filters that combine the

[5]The response of a Gabor energy filter is computed as the L2-norm of the responses of a symmetric and anti-symmetric Gabor filters.

[6]The RUG-Shoes data set can be downloaded from http://matlabserver.cs.rug.nl/

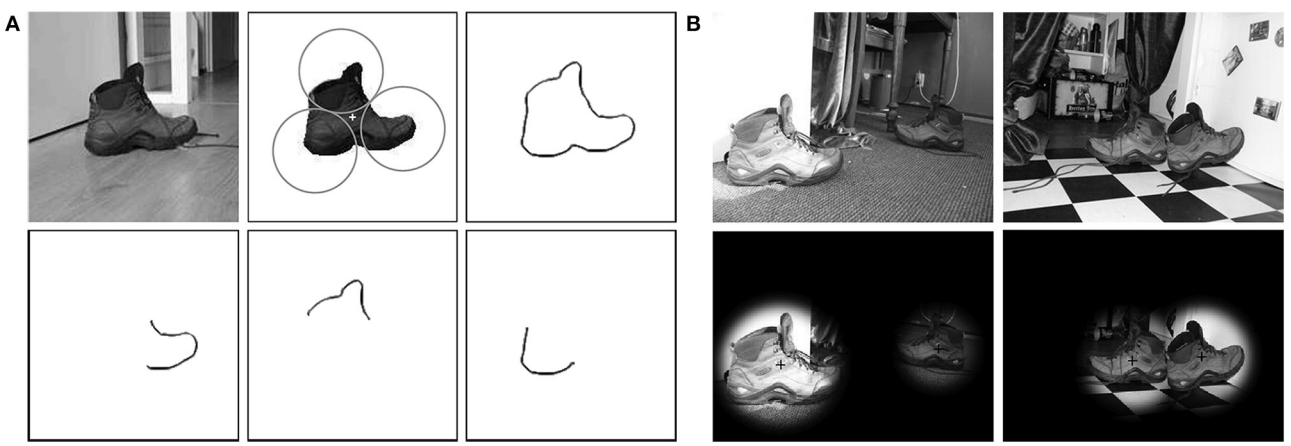


FIGURE 6 | Detection of shoes in complex scenes. (A) A protoype shoe used for the configuration of an *S*-COSFIRE filter. The circles represent the non-overlapping supports of three *V*-COSFIRE filters, and the "+" marker indicates the center of support of the concerned *S*-COSFIRE filter. (Top right) The superimposed (inverted) thresholded responses ($t_1 = 0.3$) of a bank of Gabor energy filters with one wavelength ($\lambda = 4$) and 16 orientations in intervals of $\pi/8$. (Bottom) Reconstructions of the local patterns for which the three resulting *V*-COSFIRE filters are selective. **(B)** Detection results to two input images (of size 256 × 342 pixels) from the RUG-Shoes data set with filenames (a) Shoes03_1.jpg, (c) Shoes17_2.jpg, (e) Shoes58_2.jpg, and (g) Shoes38_1.jpg.

responses of certain *V*-COSFIRE filters. Such a filter is selective for a given spatial configuration of vertices and curved contour segments that defines a simple to moderately complex shape. *S*-COSFIRE filters share similar properties with shape-selective neurons in inferotemporal cortex, which provided inspiration for this work.

This hierarchical object recognition approach is, however, not restricted to three layers. The addition of further layers may be more appropriate for prototype objects of higher deformation complexity. For instance, let us consider a prototype shape of a simplistic human-body figure that is composed of a head, a pair of eyes, a nose, a mouth, two arms, two hands, a torso, two legs, and two feet. We may configure an *S*-COSFIRE filter to be selective for the entire body with its center being at the center of mass of the body. Such a filter receives input from *V*-COSFIRE filters that are selective for distinct body parts. With this type of configuration the tolerance in the position of the body parts is computed with the same function that depends on the distance from the center of the *S*-COSFIRE filter. However, we know that certain body parts may require more tolerance or may be more correlated than others. For instance, the positions of the eyes, the nose and the mouth depend more on the position of the head than on the position of the legs. By taking this aspect in consideration it would be better to construct a hierarchical filter in the following way: configure an *S*-COSFIRE filter to be selective for the spatial arrangement of the head components (eyes, nose, and mouth), an *S*-COSFIRE filter for a hand and an arm, another one for a foot and a leg and a fourth one for the torso. Then, the responses of these four *S*-COSFIRE filters may be used as inputs to another, more complex *S*-COSFIRE filter.

The configuration of an *S*-COSFIRE filter determines which responses of which *V*-COSFIRE filters need to be multiplied in order to obtain the output of the filter. The number of *V*-COSFIRE filters used is a model parameter that is specified by the user. This value depends on the shape complexity of the concerned prototype (as represented by the number of vertex features). The selectivity of an *S*-COSFIRE filter increases with an increasing number of *V*-COSFIRE filters. The sizes of the *V*-COSFIRE supports and their position are automatically determined in such a way that they do not overlap each other. In future work, we will incorporate a learning mechanism in the configuration stage. It will use multiple prototype examples of the object of interest (instead of only one prototype that we use here) and negative examples (e.g., other objects and scenes). It will learn the optimal number of *V*-COSFIRE filters as well as the size and position of their support in order to maximize selectivity and generalization abilities.

An *S*-COSFIRE filter achieves a response when all parts of a shape of interest are present in a specific spatial arrangement around a given point in an image. The rigidity of this geometrical configuration may vary according to the application at hand. The standard deviation of a blurring (Gaussian) function that we use to allow for some tolerance depend on the distance from the center of the concerned *S*-COSFIRE filter: it grows linearly with a rate that is defined by the parameter α. Small values of α are more appropriate for the selectivity of rigid objects. Generalization ability increases with an increasing value of α. This mechanism is inspired by neurophysiological evidence that the average diameter of receptive fields of some neurons in visual cortex increases with the eccentricity (Gattass et al., 1988).

The specific type of function that we use to combine the responses of costituent (*V*-COSFIRE) filters for the considered applications is a weighted geometric mean. This output function, which is also used to compute a *V*-COSFIRE filter response, proved to give better results than various forms of addition. Furthermore, there is psychophysical evidence that human visual processing of shape is likely performed by a non-linear neural operation that multiplies afferent responses (Gheorghiu and Kingdom, 2009). In future work, we plan to experiment with functions other than (weighted) geometric mean.

The application of the home tidying robot in section 3.2 demonstrates the benefits of the rotation, scale and reflection invariances that we use. With one *S*-COSFIRE filter that is configured by a single prototype, the filter is able to achieve responses to different views of the object used for training. While this ability implies more operations, the computational cost does not grow linearly with the number of considered views. This is attributable to the fact that the responses of the bank of Gabor filters at the bottom layer can be shared among the involved *V*-COSFIRE filters, irrespective of the view. We refer the reader to Azzopardi and Petkov (2013a,b) for the technical details. The majority of the new operations required due to the invariances are shifting computations, which have very low computational cost. In practice, the shoe-selective filter used in section 3.2 takes 3.5 s to process an image (256 × 342 pixels) with no invariances, and less than 5 s with rotation-, scale-, and reflection-invariance.

The proposed *S*-COSFIRE filters are particularly useful due to their versatility and selectivity, in that an *S*-COSFIRE filter can be configured to be selective for any given deformable object and used to detect other objects embedded in complex scenes that are perceptually similar to it. This effectiveness is attributable to taking into account the mutual spatial positions of the responses of certain *V*-COSFIRE filters that are selective for simpler object parts.

5. CONCLUSIONS

The *S*-COSFIRE filters that we propose are highly effective to detect and recognize deformable objects that are embedded in complex scenes without prior segmentation. This effectiveness is due to the deployment of both the presence of certain object-characteristic features and their mutual spatial arrangement. They are versatile shape detectors as they can be trained to be selective for any given visual pattern of interest.

An *S*-COSFIRE filter is conceptually simple and easy to implement: the filter output is computed as the weighted geometric mean of blurred and shifted responses of simpler *V*-COSFIRE filters.

REFERENCES

Almazan, J., Fornes, A., and Valveny, E. (2012). A non-rigid appearance model for shape description and recognition. *Pattern Recogn.* 45, 3105–3113. doi: 10.1016/j.patcog.2012.01.010

Azzopardi, G., and Petkov, N. (2012). A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model. *Biol. Cybernet.* 106, 177–189. doi: 10.1007/s00422-012-0486-6

Azzopardi, G., and Petkov, N. (2013a). Automatic detection of vascular bifurcations in segmented retinal images using trainable COSFIRE filters. *Pattern Recogn. Lett.* 34, 922–933. doi: 10.1016/j.patrec.2012.11.002

Azzopardi, G., and Petkov, N. (2013b). Trainable COSFIRE Filters for Keypoint Detection and Pattern Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 490–503. doi: 10.1109/TPAMI.2012.106

Azzopardi, G., Rodriguez Sanchez, A., Piater, J., and Petkov, N. (2014). A push-pull CORF model of a simple cell with antiphase inhibition improves SNR and contour detection. *PLoS ONE* 9:e98424. doi: 10.1371/journal.pone.0098424

Bandera, J. P., Rodriguez, J. A., Molina-Tanco, L., and Bandera, A. (2012). A survey of vision-based architectures for robot learning by imitation. *Int. J. Human. Robot.* 9:1250006. doi: 10.1142/S0219843612500065

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 509–522. doi: 10.1109/34.993558

Brincat, S. L., and Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* 7, 880–886. doi: 10.1038/nn1278

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Optic. Soc. Am. Optic. Image Sci. Vis.* 2, 1160–1169. doi: 10.1364/JOSAA.2.001160

DiCarlo, J., and Cox, D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

Edelman, S., and Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cogn. Sci.* 27, 73–109. doi: 10.1207/s15516709cog2701_3

Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Understand.* 106, 59–70. doi: 10.1016/j.cviu.2005.09.012. 2nd International Workshop on Generative-Model Based Vision, Washington, DC, 2005.

Fergus, R., Perona, P., and Zisserman, A. (2003). "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, Technical Committee on Pattern Analysis and Machine Intelligence (TCPAMI)* (Madison, WI), 264–271.

Fidler, S., Boben, M., and Leonardis, A. (2008). "Similarity-based cross-layered hierarchical representation for object categorization," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1–12 (Anchorage, AK), 525–532. doi: 10.1109/CVPR.2008.4587409

Fidler, S., and Leonardis, A. (2007). "Towards scalable representations of object categories: learning a hierarchy of parts," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1–8 (Minneapolis, MN), 2295–2302. doi: 10.1109/CVPR.2007.383269

Frinken, V., Fischer, A., Manmatha, R., and Bunke, H. (2012). A novel word spotting method based on recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 211–224. doi: 10.1109/TPAMI.2011.113

Gattass, R., Sousa, A. P., and Gross, C. G. (1988). Visuotopic organization and extent of v3 and v4 of the macaque. *J. Neurosci.* 8, 1831–1845.

Gheorghiu, E., and Kingdom, F. A. A. (2009). Multiplication in curvature processing. *J. Vis.* 9, 1–17. doi: 10.1167/9.2.23

Ghosh, A., and Petkov, N. (2005). Robustness of shape descriptors to incomplete contour representations. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1793–1804. doi: 10.1109/TPAMI.2005.225

Goh, W. B. (2008). Strategies for shape matching using skeletons. *Comput. Vis. Image Understand.* 110, 326–345. doi: 10.1016/j.cviu.2007.09.013

Grigorescu, C., and Petkov, N. (2003). Distance sets for shape filters and shape recognition. *IEEE Trans. Image Process.* 12, 1274–1286. doi: 10.1109/TIP.2003.816010

Grigorescu, C., Petkov, N., and Westenberg, M. A. (2004). Contour and boundary detection improved by surround suppression of texture edges. *Image Vis. Comput.* 22, 609–622. doi: 10.1016/j.imavis.2003.12.004

Jiang, Y., Lim, M., Zheng, C., and Saxena, A. (2012). Learning to place new objects in a scene. *Int. J. Robot. Res.* 31, 1021–1043. doi: 10.1177/0278364912438781

Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral-cortex. *J. Neurophysiol.* 71, 856–867.

Latecki, L., Lakaemper, R., and Wolter, D. (2005). Optimal partial shape similarity. *Image Vis. Comput.* 23, 227–236. doi: 10.1016/j.imavis.2004.06.015. 11th International Conference on Discrete Geometry for Computer Imagery, Italian Inst Philosoph Studies, Naples, Italy, Nov 19-21, 2003.

Lauer, F., Suen, C. Y., and Bloch, G. (2007). A trainable feature extractor for handwritten digit recognition. *Pattern Recogn.* 40, 1816–1824. doi: 10.1016/j.patcog.2006.10.011

Ling, H., and Jacobs, D. W. (2007). Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 286–299. doi: 10.1109/TPAMI.2007.41

Marr, D. (1982). *Vision: A Computational Investigation Into The Human Respresentation and Processing of Visual Information*. New York, NY: Freeman.

Marr, D., and Nishihara, H. K. (1978). Representation and recognition of spatial-organization of 3-dimensional shapes. *Proc. R. Soc. London B Biol. Sci.* 200, 269–294. doi: 10.1098/rspb.1978.0020

Marti, U.-V., and Bunke, H. (2002). The IAM-database: an english sentence database for offline handwriting recognition. *Int. J. Doc. Anal. Recogn.* 5, 39–46. doi: 10.1007/s100320200071

Mel, B. W., and Fiser, J. (2000). Minimizing binding errors using learned conjunctive features (vol 12, pg 247, 1999). *Neural Comput.* 12, 731–762. doi: 10.1162/089976600300015574

Plamondon, R., and Srihari, S. (2000). On-line and off-line handwriting recognition: a comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 63–84. doi: 10.1109/34.824821

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rodríguez-Sánchez, A. J., and Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2d shape. *PLoS ONE* 7:e42058. doi: 10.1371/journal.pone.0042058

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Scalzo, F., and Piater, J. (2005). "Statistical learning of visual feature hierarchies," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops* (San Diego, CA), 44.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Szwed, M., Dehaene, S., Kleinschmidt, A., Eger, E., Valabregue, R., Amadon, A., et al. (2011). Specialization for written words over objects in the visual cortex. *Neuroimage* 56, 330–344. doi: 10.1016/j.neuroimage.2011.01.073

Tsotsos, J. (1990). Analyzing vision at the complexity level. *Behav. Brain Sci.* 13, 423–444. doi: 10.1017/S0140525X00079577

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 April 2014; accepted: 09 July 2014; published online: 30 July 2014.
Citation: Azzopardi G and Petkov N (2014) Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective COSFIRE models. Front. Comput. Neurosci. ***8****:80. doi: 10.3389/fncom.2014.00080*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Azzopardi and Petkov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 15 January 2015
doi: 10.3389/fncom.2014.00168

Visual dictionaries as intermediate features in the human brain

Kandan Ramakrishnan[1]*, H. Steven Scholte[2], Iris I. A. Groen[2], Arnold W. M. Smeulders[1] and Sennay Ghebreab[1,2]

[1] *Intelligent Systems Lab Amsterdam, Institute of Informatics, University of Amsterdam, Amsterdam, Netherlands*
[2] *Cognitive Neuroscience Group, Department of Psychology, University of Amsterdam, Amsterdam, Netherlands*

Edited by:
Mazyar Fallah, York University, Canada

Reviewed by:
Marcel Van Gerven, Donders Institute for Brain, Cognition and Behaviour, Netherlands
Tianming Liu, Uga, USA
Daniel Leeds, Fordham University, USA

***Correspondence:**
Kandan Ramakrishnan, Intelligent Sensory Information Systems Group, ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands
e-mail: k.ramakrishnan@uva.nl

The human visual system is assumed to transform low level visual features to object and scene representations via features of intermediate complexity. How the brain computationally represents intermediate features is still unclear. To further elucidate this, we compared the biologically plausible HMAX model and Bag of Words (BoW) model from computer vision. Both these computational models use visual dictionaries, candidate features of intermediate complexity, to represent visual scenes, and the models have been proven effective in automatic object and scene recognition. These models however differ in the computation of visual dictionaries and pooling techniques. We investigated where in the brain and to what extent human fMRI responses to short video can be accounted for by multiple hierarchical levels of the HMAX and BoW models. Brain activity of 20 subjects obtained while viewing a short video clip was analyzed voxel-wise using a distance-based variation partitioning method. Results revealed that both HMAX and BoW explain a significant amount of brain activity in early visual regions V1, V2, and V3. However, BoW exhibits more consistency across subjects in accounting for brain activity compared to HMAX. Furthermore, visual dictionary representations by HMAX and BoW explain significantly some brain activity in higher areas which are believed to process intermediate features. Overall our results indicate that, although both HMAX and BoW account for activity in the human visual system, the BoW seems to more faithfully represent neural responses in low and intermediate level visual areas of the brain.

Keywords: visual perception, fMRI, low and intermediate features, HMAX, bag of words, representation similarity analysis

1. INTRODUCTION

The human visual system transforms low-level features in the visual input into high-level concepts such as objects and scene categories. Visual recognition has been typically viewed as a bottom-up hierarchy in which information is processed sequentially with increasing complexities, where lower-level cortical processors, such as the primary visual cortex, are at the bottom of the processing hierarchy and higher-level cortical processors, such as the inferotemporal cortex (IT), are at the top, where recognition is facilitated (Bar, 2003). Much is known about the computation in the earliest processing stages, which involve the retina, lateral geniculate nucleus (LGN) and primary visual cortex (V1). These areas extract simple local features such as blobs, oriented lines, edges and color from the visual input. However, there remain many questions on how such low-level features are transformed into high-level object and scene percepts.

One possibility is that the human visual system transforms low-level features into object and scene representations via an intermediate step (Riesenhuber and Poggio, 1999). After extraction of low-level features in areas such as V1, moderately complex features are created in areas V4 and the adjacent region TO. Then partial or complete object views are represented in anterior regions of inferotemporal (IT) cortex (Tanaka, 1997). It has been suggested that such intermediate features along the ventral visual pathway are important for object and scene representation (Logothetis and Sheinberg, 1996).

Previous studies have provided some evidence of what intermediate features might entail. In Tanaka (1996) it has been shown that cells in the V4/IT region respond selectively to complex features such as simple patterns and shapes. Similarly, Hung et al. (2012) identified contour selectivity for individual neurons in the primate visual cortex and found that most contour-selective neurons in V4 and IT each encoded some subset of the parameter space and that a small collection of the contour-selective units were sufficient to capture the overall appearance of an object. Together these findings suggest that intermediate features capture object information encoded within the human ventral pathway.

In an attempt to answer the question of intermediate features underlying neural object representation, Leeds et al. (2013) compared five different computational models of visual representation against human brain activity to object stimuli. They found that the Bag of Words (BoW) model was most strongly correlated with brain activity associated with midlevel perception. These results were based on fMRI data from 5 subjects. Recently Yamins et al. (2014) used a wider set of models including HMAX and BoW against neural responses from two monkeys in IT and V4.

HMAX (Riesenhuber and Poggio, 1999) and BoW (Csurka et al., 2004) models represent scenes in a hierarchical manner transforming low level features to high level concepts. HMAX is a model for the initial feedforward stage of object recognition in the ventral visual pathway. It extends the idea of simple cells (detecting oriented edges) and complex cells (detecting oriented edges with spatial invariance) by forming a hierarchy in which alternate template matching and max pooling operations progressively build up feature selectivity and invariance to position and scale. HMAX is thus a simple and elegant model used by many neuroscientists to describe feedforward visual processing. In computer vision, different algorithms are used for object and scene representation. The commonly used model in computer vision is BoW which performs very well on large TRECvid (Smeaton et al., 2006) and PASCAL (Everingham et al., 2010) datasets, in some cases even approaching human classification performance (Parikh and Zitnick, 2010). The key idea behind this model is to quantize local Scale Invariant Feature Transform (SIFT) features (Lowe, 2004) into visual words (Jurie and Bill, 2005), features of intermediate complexity, and then to represent an image by a histogram of visual words. To further understand the nature of intermediate features underlying scene perception, we test these two computational models against human brain activity while subjects view a movie of natural scenes.

Although HMAX and BoW are different models they both rely on the concept of visual dictionaries to represent scenes. In HMAX after the initial convolution and pooling stage, template patches are learnt from responses of the pooling layer (from a dataset of images) which are used as visual dictionaries. In the BoW model, clustering of SIFT features forms the visual dictionary. In both the models, visual dictionaries are medium size image patches that are informative and at the same time distinctive. They can be thought of, as features of intermediate complexity. This comparison of different computational approaches to visual dictionaries might provide further insight about the representation of intermediate features in the human brain.

In this work we test two layers of the HMAX and BoW models against human brain activity. We show 20 subjects a 11-min video of dynamic natural scenes and record their fMRI activity while watching the video. We use dynamic scenes instead of static scenes because they are more realistic, and because they may evoke brain responses that allow for a better acquisition of neural processes in the visual areas of the brain (Hasson et al., 2004). Furthermore, the use of a relatively large pool of subjects allows us to compare computational models in terms of their consistency in explaining brain activity. The fMRI data is compared to HMAX and BoW models. For the HMAX model we test how Gabor and visual dictionary representation of an image explain brain activity. Similarly for BoW, we test how SIFT and visual dictionary explain brain activity. If the models are good representations of intermediate features in the human brain, they should account for brain activity across multiple subjects.

Testing hierarchical models of vision against brain activity is challenging for two reasons. First, computational and neural representations of visual stimuli are of different nature but both very high-dimensional. Second the different hierarchical levels of the models need to be dissociated properly in order to determine how brain activity is accounted by each of the individual hiearchical levels of the model. This cannot be done easily in standard multivariate neuroimaging analysis. We address the first challenge by using dissimilarity matrices (Kriegeskorte et al., 2008) that capture computational and neural distances between any pair of stimuli. The second is resolved by applying a novel technique, variation partitioning (Peres-Neto et al., 2006) on the dissimilarity matrices. This enables us to compute the unique contributions of the hierarchical layers of HMAX and BoW models in explaining neural activity. Distance based variation partitioning has been successfully used in ecological and evolutionary studies, and will be applied here to fMRI data. This will enable us to establish correspondence between computational vision models, their different hierarchical layers and fMRI brain activity.

2. MATERIALS AND METHODS

2.1. COMPUTATIONAL MODELS

We use HMAX and BoW computational models to represent images at the different hierarchical levels. For the HMAX model we compute the Gabor representation at the first level and the visual dictionary representation at the second level (**Figure 1**). Similarly for the BoW model we compute the SIFT representation and visual dictionary representation. It is important to note that HMAX and BoW models refer to the entire hierarchical model combining low level feature and visual dictionary.

2.1.1. HMAX model

We use the HMAX model (Mutch and Lowe, 2008), where features are computed hierarchically in layers: an initial image layer and four subsequent layers, each built from the previous layer by alternating template matching and max pooling operations as seen in **Figure 1**. In the first step, the greyscale version of an image is downsampled and a image pyramid of 10 scales is created. Gabor filters of four orientations are convolved over the image at different positions and scales in the next step, the S1 layer. Then in the C1 layer, the Gabor responses are maximally pooled over $10 \times 10 \times 2$ regions of the responses from the previous layer (the max filter is a pyramid). The Gabor representation of an image I is denoted by the vector $\mathbf{f}_{gabor}$.

In the next step, template matching is performed between the patch of C1 units centered at every position/scale and each of P prototype patches. These $P = 4096$ prototype patches are learned as done in Mutch and Lowe (2008) by randomly sampling patches from the C1 layer. We use images from the PASCAL VOC 2007 dataset (Smeaton et al., 2006) to sample the prototypes for the dictionary. In the last layer, a P dimensional feature is created by maximally pooling over all scales and orientations to one of the models P patches from the visual dictionary. This results in a visual dictionary representation of image I denoted by the vector $\mathbf{f}_{vdhmax} = [h_1 \ldots h_P]$ where each dimension h_p represents the max response of the dictionary elements convolved over the output of the C1 layer.

2.1.2. BoW model

The first step in the BoW model (**Figure 1**) is extraction of SIFT descriptors (Lowe, 2004) from the image. SIFT combines a scale

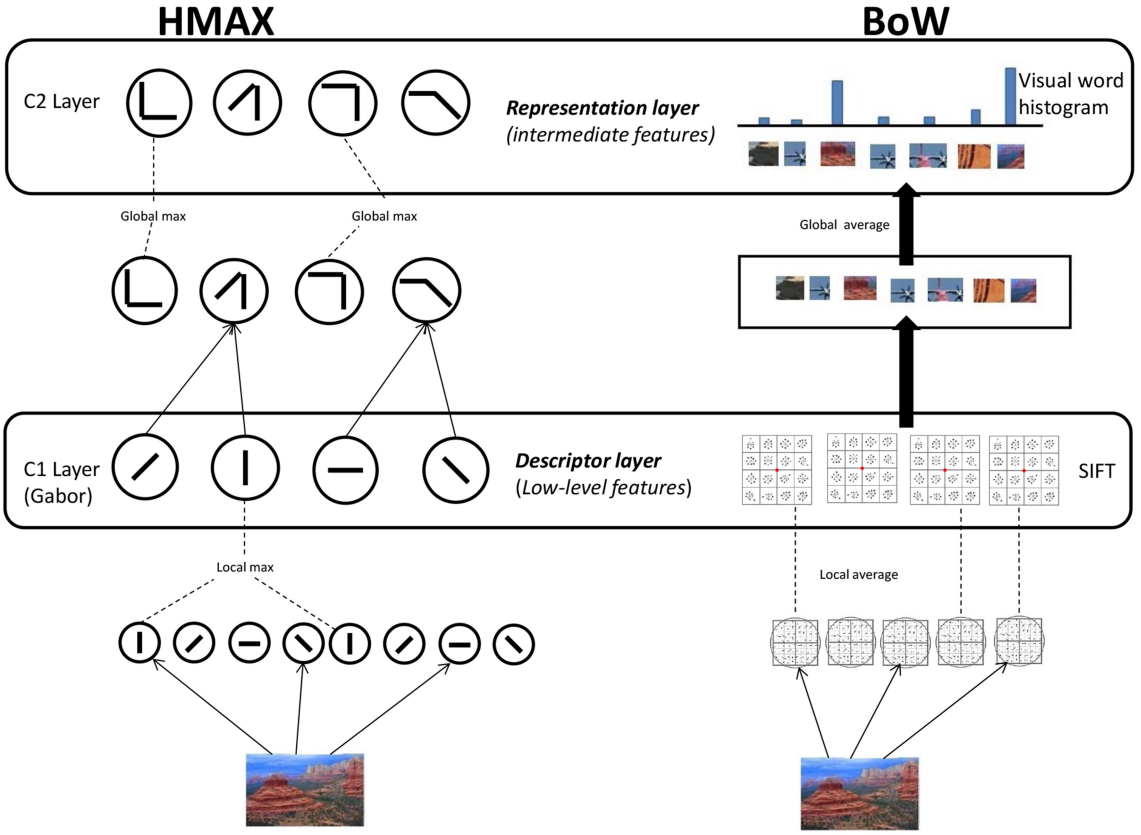


FIGURE 1 | Computational models. HMAX model: Gabor filters of 4 orientations and 10 different scales are convolved in the S1 Layer. The responses are pooled to form the C1 Layer. For the S2 Layer random samples of the pooled responses from the C1 layer to the PASCAL dataset of images is used to form the visual dictionary of dimension 4096. These template patches are detected from the responses to form the S2 Layer. A global max pooling operation is done for the final C2 Layer which is of dimension 4096. **BoW model:** SIFT features are extracted densely over the image. Visual dictionary of dimension 4000 is learnt by kmeans clustering on SIFT features extracted PASCAL dataset images. Each SIFT descriptor from an image is encoded to the nearest element of visual dictionary. Average pooling is done to form the 4000 dimension visual dictionary representation.

invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations and orientations weighted by the gradient magnitude. The quantization of gradient locations and orientations makes the descriptor robust to small geometric distortions and small errors in the region detection. SIFT feature is a 128 dimensional vector which is computed densely over the image. Here the SIFT representation of an image I is obtained by concatenating all the SIFT features over the image. It is denoted by the vector $\mathbf{f}_{sift}$.

Secondly, a dictionary of visual words (Csurka et al., 2004) is learned from a set of scenes independent of the scenes in the stimuli video. We use k-means clustering to identify cluster centers $\mathbf{c}_m = \mathbf{c}_1, \ldots, \mathbf{c}_M$ in SIFT space, where $m = 1, \ldots, M$ denotes the number of visual words. We use the PASCAL VOC 2007 (Smeaton et al., 2006) dataset to create a codebook of dimension $M = 4000$.

The SIFT features of a new image are quantized (assigned to the nearest visual word) to a element in the visual dictionary and the image is represented by counting the occurrences of all words. This results for image I in the visual dictionary representation $\mathbf{f}_{vdbow} = [h_1 \ldots h_M]$ where each bin h_m indicates the frequency(number of times) the visual word $\mathbf{c}_m$ is present in the image.

2.2. REPRESENTATIONAL DISSIMILARITY MATRICES

A representational dissimilarity matrix (Kriegeskorte et al., 2008) (RDM) F is computed separately for each of the image representations. The elements in this matrix are the Euclidean distance between the representations of pairs of images. Thus, F_{gabor}, F_{vdhmax}, F_{sift}, and F_{vdbow} are dissimilarity matrices for the different representations respectively. **Figure 2A** shows the 290×290 dissimilarity matrices for 290 images (frames) from the video stimulus used in this study.

2.3. STIMULI

An 11-min video track consisting of about 20 different dynamic scenes was used for this study. The scenes were taken from the movie Koyaanisqatsi: Life Out of Balance and consisted primarily of slow motion and time-lapse footage of cities and many natural landscapes across the United States as in **Figure 2B**.

The movie Life Out of Balance was chosen as a stimulus because it contained all kinds of scenes we encounter in our daily

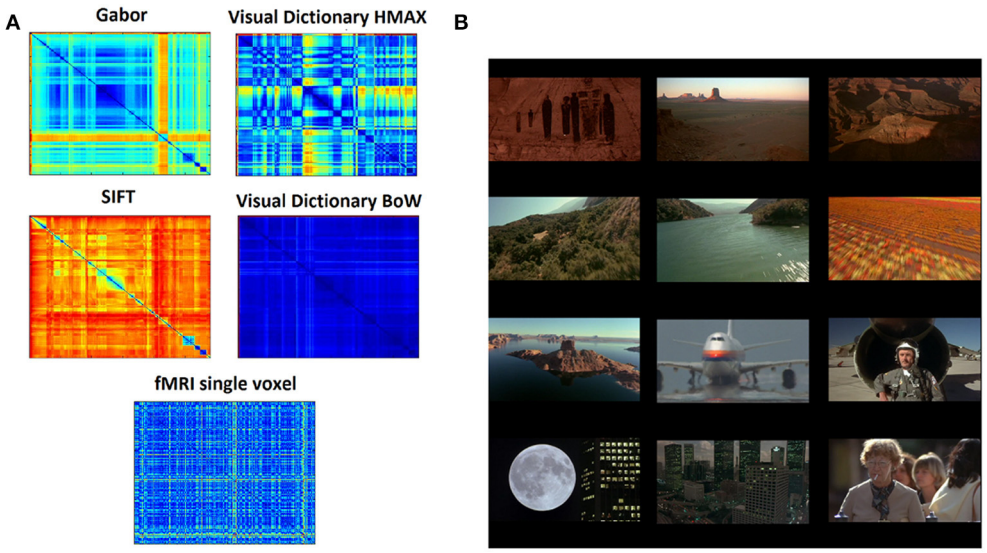


FIGURE 2 | Data and representations (A) Dissimilarity matrices computed for the different hierarchical levels of HMAX and BoW using pairwise distance between all the scenes from the stimuli resulting in a 290 × 290 matrix. Similarly the dissimilarity matrix computed for the fMRI brain responses where each element is the distance in multivariate voxel responses to any image pair resulting in a 290 × 290 matrix. **(B)** Example images (frames) from the 11 min video stimuli that was used in the fMRI study. There are totally 290 scenes representing a wide variety of scenes, ranging from natural to man-made.

live with no human emotional content or specific storyline, from natural (e.g., forest) to more man made scenes (e.g., streets). In addition the movie exhibits different motion elements such as zooming, scaling, luminance etc. In this respect, the movie is rich in its underlying low-level properties such as spatial frequency and color.

2.4. SUBJECTS

The fMRI data of the video stimuli was collected for over 500 subjects, from which 20 were randomly sampled for this study. Subjects were not assigned with any specific tasks when watching. They watched the video track passively one time each. The experiment was approved by the ethical committee of the University of Amsterdam and all participants gave written informed consent prior to participation. They were rewarded for participation with either study credits or financial compensation.

2.5. fMRI

We recorded 290 volumes of BOLD-MRI (GE-EPI, 192^2 mm, 42 slices, voxel size of 3 × 3 × 3.3, TR 2200 ms, TE 27.63 ms, SENSE 2, FA 90°C) using a 3T Philips Achieve scanner with a 32 channel headcoil. A high-resolution T1-weighted image (TR, 8.141 ms; TE, 3.74 ms; FOV, 256 × 256 × 160 mm) was collected for registration purposes. Stimuli were backward-projected onto a screen that was viewed through a mirror attached to the head-coil.

2.6. fMRI PREPROCESSING

FEAT (fMRI Expert Analysis Tool) version 5.0, part of FSL (Jenkinson et al., 2012) was used to analyze the fMRI data. Preprocessing steps included slice-time correction, motion correction, high-pass filtering in the temporal domain ($\sigma = 100$ s), spatially filtered with a FWHM of 5 mm and prewhitened (Woolrich et al., 2001). Data was transformed using an ICA and we subsequently, automatically identified artifacts using the FIX algorithm (Salimi-Khorshidi et al., 2014). Structural images were coregistered to the functional images and transformed to MNI standard space (Montreal Neurological Institute) using FLIRT (FMRIB's Linear Image Registration Tool; FSL). The resulting normalization parameters were applied to the functional images. The data was transformed into standard space for cross-participant analyses, so that the same voxels and features were used across subjects.

These 290 image frames and volumes were used to establish a relation between the two computational models and BOLD responses. Although in this approach the haemodynamic response might be influenced by other image frames, we expect this influence to be limited because the video is slowly changing without any abrupt variations. In addition, BOLD responses are intrinsically slow and develop over a period of up to 20 s. Still they summate linearly reasonably well (Buckner, 1998) and also match the timecourse in typical scenes which develop over multiple seconds. This also probably explains the power of BOLD-MRI in decoding the content of movies (Nishimoto et al., 2011) and indicates it is possible to compare different models of information processing on the basis of MRI volumes.

2.7. VARIATION PARTITIONING

A 3 × 3 × 3 searchlight cube is centered at each voxel in the brain and BOLD responses within the cube to each of the 290 still images compared against each other. This results for each subject and for each voxel in a 290 × 290 dissimilarity matrix Y. Each element in the Y matrix is the pairwise distance of the 27 dimensional (from the searchlight cube) multivariate voxel responses to any image pair. As a distance measure Cityblock is taken. We now

perform variation partitioning voxel-wise (each voxel described by its searchlight cube) for all the voxels across all subjects.

Variation partitioning (Peres-Neto et al., 2006) for the HMAX model is done by a series of multiple regression, producing fractions of explained variation R^2_{gabor} (unique to gabor representation), $R^2_{gaborvdhmax}$ (common to both gabor and visual dictionary representation) and R^2_{vdhmax} (unique to visual dictionary). First the multiple regression of Y against F_{gabor} and F_{vdhmax} together is computed, where Y denotes the fMRI dissimilarity matrix, and F_{gabor} and F_{vdhmax} the Gabor and visual dictionary dissimilarity matrices respectively. The corresponding R^2_{hmax} measures the total fraction of explained variation, which is the sum of the fractions of variation R^2_{gabor}, $R^2_{gaborvdhmax}$, and R^2_{vdhmax}. Then the multiple regression of Y against F_{gabor} is computed. The corresponding $R^2_{gabor+gaborvdhmax}$ measure is the sum of the fractions R^2_{gabor} and $R^2_{gaborvdhmax}$. In the next step, the multiple regression of Y against F_{vdhmax} is obtained, with corresponding $R^2_{vdhmax+gaborvdhmax}$ being the sum of the fractions of variation $R^2_{gaborvdhmax}$ and R^2_{vdhmax}. The fraction of variation uniquely explained by the Gabor dissimilarity matrix is computed by substraction: $R^2_{gabor} = R^2_{hmax} - R^2_{vdhmax+gaborvdhmax}$. Similarly, variation uniquely explained by visual dictionary dissimilarity matrix is: $R^2_{vdhmax} = R^2_{hmax} - R^2_{gabor+gaborvdhmax}$. The residual fraction may be computed by: $1 - (R^2_{gabor} + R^2_{gaborvdhmax} + R^2_{vdhmax})$.

Exactly the same steps of computation are taken to determine the fraction of variation uniquely explained by the SIFT dissimilarity matrix, the fraction explained by BoW visual dictionary dissimilarity matrix, and by the combination of both the SIFT and visual dictionary dissimilarity matrices as shown in **Figure 3**.

We also compare the models at their respective hierarchical levels. At the first level, Gabor and SIFT dissimilarity matrices are used to explain brain activity Y. Similarly at the level of visual dictionaries, we compare how HMAX and BoW visual dictionary dissimilarity matrices explain Y.

Note that these R^2 statistics are the canonical equivalent of the regression coefficient of determination, R^2 (Peres-Neto et al., 2006). They can interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

A permutation test (1000 times) determines the statistical significance (p value) of the fractions that we obtain for each voxel by variation partitioning. To account for the multiple comparison problem, we perform cluster size correction and only report here clusters of voxels that survive the statistical thresholding at $p < 0.05$ and have a minimum cluster size of 25 voxels. We determine the minimum cluster size by calculating the probability of a false positive from the frequency count of cluster sizes within the entire volume, using a Monte Carlo simulation (Ward, 2000).

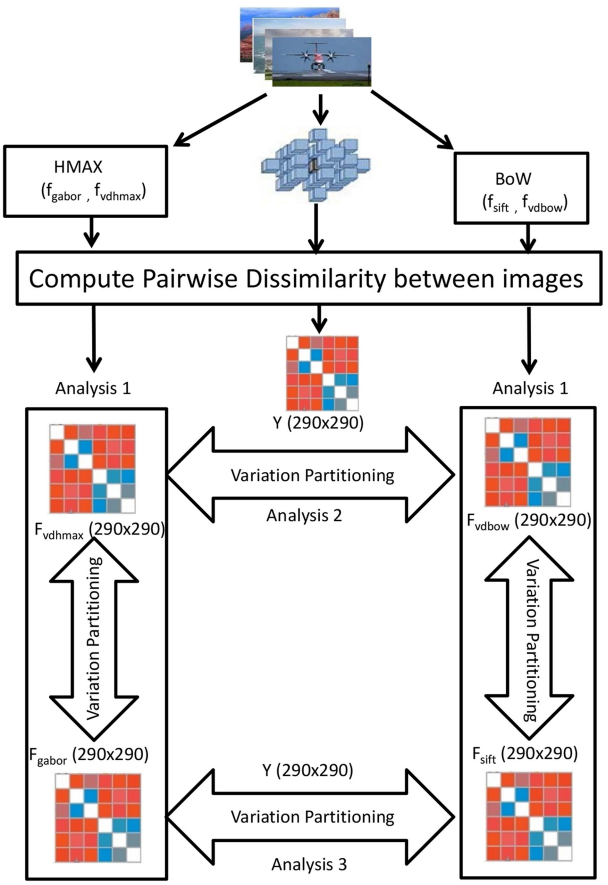


FIGURE 3 | Visualization of the variation partitioning on the RDMs obtained from the 290 images of the ID1000 stimuli. For the HMAX model we obtain a 290 × 290 Gabor dissimilarity matrix(F_{gabor}) and visual dictionary dissimilarity matrix(F_{vdhmax}) using pairwise image distances. Similarly for BoW, we obtain 290 × 290 SIFT dissimilarity matrix(F_{sift}) and visual dictionary matrix(F_{vdbow}). Then variation partitioning is applied at each of the hierarchical level and across the hierarchical levels on the 290 × 290 fMRI dissimilarity matrix(Y).

3. RESULTS

3.1. COMPARING FULL MODELS : INTERSUBJECT CONSISTENCY

Using distance-based variation partitioning for each subject we dissociate the explained variation of the HMAX model into unique contributions of Gabor R^2_{gabor} and visual dictionary representation R^2_{vdhmax}. The total explained variation by HMAX model is given by the combination of R^2_{gabor} and R^2_{vdhmax}. We do the same for the BoW model, based on SIFT R^2_{sift} and visual dictionary representation R^2_{vdbow}. HMAX and BoW models refer to the entire hierarchical model combining low level feature and visual dictionary. Cluster size correction ($p < 0.05$ and minimal cluster size of 25 voxels) was performed to solve for the multiple comparison problem.

To test whether our results are consistent across subjects, for each voxel we counted the number subjects for which brain activity was explained significantly by the HMAX and BoW models. A spatial version of the chi-square statistic (Rogerson, 1999) was subsequently applied to determine whether the observed frequency at a particular voxel deviated significantly from the expected value (the average number of subjects across all voxels).

Figure 4A shows how consistently across subjects, HMAX and BoW models account for brain activity. We observe that

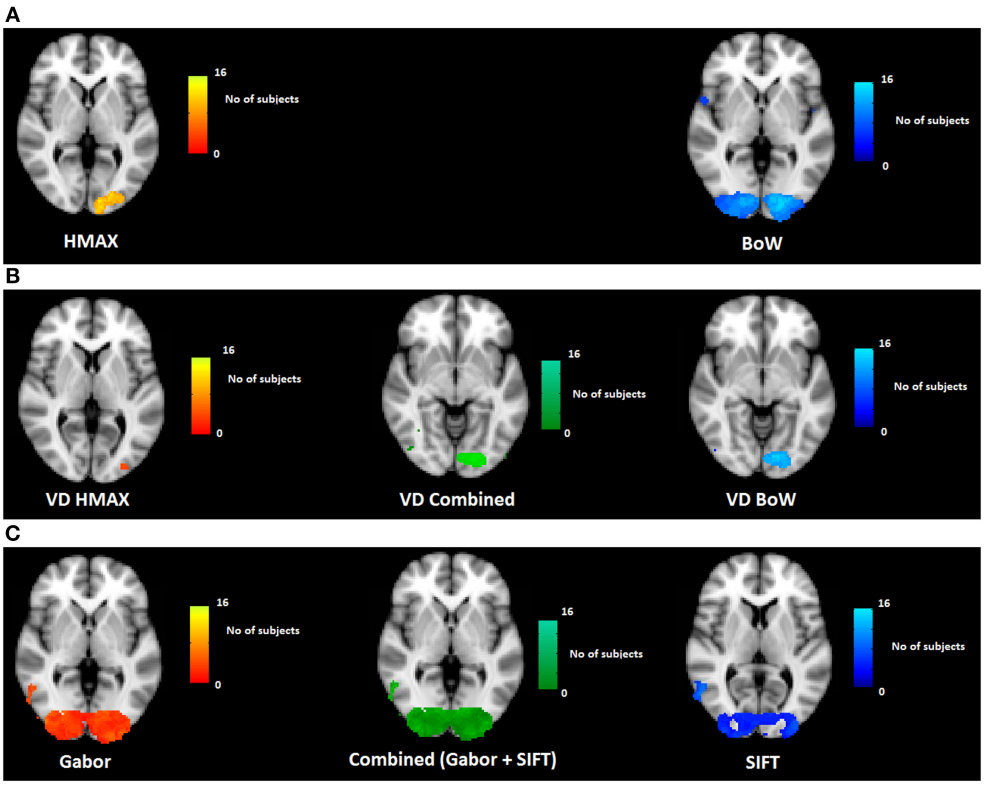


FIGURE 4 | Visualization of across subject consistency at each voxel for the complete HMAX and BoW models and their individual components. To find consistency across subjects, first significant voxel clusters are determined subject wise and then a spatial frequency count is performed on detected clusters across subjects. **(A)** Across subjects consistency for HMAX and BoW model based on voxel clusters and the spatial chi-square statistic (Analysis 1 in **Figure 3**). **(B)** Across subjects consistency for visual dictionaries from HMAX (VD HMAX), BoW (VD BoW) and their combination (Analysis 2 in **Figure 3**). **(C)** Across subjects consistency for Gabor, SIFT and their combination (Analysis 3 in **Figure 3**).

the HMAX model explains brain activity in areas V2 and V3 consistently across subjects. In these areas the HMAX model explains brain activity in overlapping voxels for 16 out of 20 subjects. In contrast, the BoW model accounts for brain activity across wider and bilateral regions including V1, V2 and V3. Most consistency is found at the left V3 and V4 regions, where for 14 out of 20 subjects, the BoW model was relevant in explaining brain activity. This difference in the number of subjects is not significant however the extent of the voxels is much more for BoW than HMAX.

Both HMAX and BoW models use low level features (Gabor filters and histogram of orientations) as their first step of computation. This is explicitly modeled and tested in our study (low-level feature representation in **Figure 1**). This explains why low level visual regions such as V1 and V2 emerge in our results. Interestingly, however, the BoW model also accounts for brain activity in regions higher up in the visual system such as V4 and LO (lateral occipital cortex). These regions are hypothesized to process intermediate features. This suggests that while both models appropriately represent low-level features, the transformation of these features to intermediate features is better modeled by BoW. Figure 1 in Supplementary section shows for each individual subject the explained variation of the two representational levels in both the models.

We observe that for the HMAX model the combination of hierarchies provide 5% of additional explanation compared to the maximum explaining hierarchical level. The two levels of the BoW together additionally account for 8% of the variation in brain activity. A t-test on the two distributions of additional explained variations show a significant difference ($p < 0.0001$). Thus, in both models, but more strongly in BoW, the aggregation of low level features into visual dictionaries describes brain activity, not captured by individual hierarchical levels. Thus, the aggregation of low level features into visual dictionaries provide additional value to account for brain activity. The hierarchical levels in BoW contribute slightly more to the explained brain activity as compared to the hierarchical levels from HMAX.

3.2. COMPARING VISUAL DICTIONARIES : INTERSUBJECT CONSISTENCY

We tested the two visual dictionary representations against each other. As before, we use variation partitioning on the visual dictionary dissimilarity matrices from HMAX and BoW to explain Y. For each voxel we counted the number of subjects for which brain activity was explained significantly by the visual dictionary from HMAX and BoW models. A spatial version of the chi-square statistic (Rogerson, 1999) was applied to determine whether the observed frequency at a particular voxel deviated

significantly from the expected value (the average count across all voxels).

Figure 4B shows the across subject consistency of visual dictionaries from HMAX and BoW models ($p < 0.05$, cluster size correction). We observe for the HMAX visual dictionary representation that consistency across subjects occurs in few voxels in area V4. In contrast the visual dictionary representation of the BoW model explains brain activity in areas V3 and V4 for 14 out of 20 subjects. The combination of visual dictionary representation explain brain activity for 14 out of 20 subjects in areas V3 and V4.

The visual dictionary representation from the BoW model has a much higher across subject consistency than the HMAX model. In addition the results of the combined model are similar to those of the BoW visual dictionary representations, suggesting that the HMAX visual word representation adds little to the BoW representation in terms of accounting for brain activity. Moreover, the BoW visual dictionary representation is localized in an area V4 that is hypothesized to compute intermediate features. Altogether, these results suggest that the BoW model provides a better representation for visual dictionaries, compared to the HMAX model. Single subject results confirming consistency across subjects can be found in the Supplementary section.

3.3. COMPARING LOW-LEVEL FEATURE REPRESENTATIONS : INTERSUBJECT CONSISTENCY

We tested Gabor and SIFT representation against each other. As before, we use variation partitioning on Gabor and SIFT representations to explain Y. For each voxel we counted the number subjects for which brain activity was explained significantly. The spatial version of the chi-square statistic was applied to determine whether the observed frequency at a particular voxel deviated significantly from the expected value (the average count across all voxels).

Figure 4C shows the across subject consistency of Gabor and SIFT representations ($p < 0.05$, cluster size correction). We observe that the Gabor representation explains brain activity in early visual areas for a large number of voxels such as V1, V2, and V3. The Gabor representation also explains brain activity consistently across subjects in the higher brain areas such as LO and precentral gyrus for 10 out of the 20 subjects. Similarly for the SIFT representation we observe that it explains brain activity in the lower visual areas such as V1, V2 and also higher areas of the brain such as LO across 9 out of 20 subjects. Overall Gabor and SIFT representations account for brain activity in similar areas of the brain. It is expected that Gabor and SIFT explain brain responses in early visual areas since both rely on edge filters. However, it is interesting to observe that they also explain brain activity in the higher areas of the brain.

We also observe areas where Gabor and SIFT together explain neural response consistently across subjects. The combination of Gabor and SIFT representations explain brain activity in 14 out of 20 subjects in the early visual area V1. The combination also explains brain activity in higher areas of the brain such as V4 and LO. This suggests that Gabor and SIFT representation have complementary low-level gradient information. Taken together, Gabor and SIFT provide a better computational basis for V1 representation.

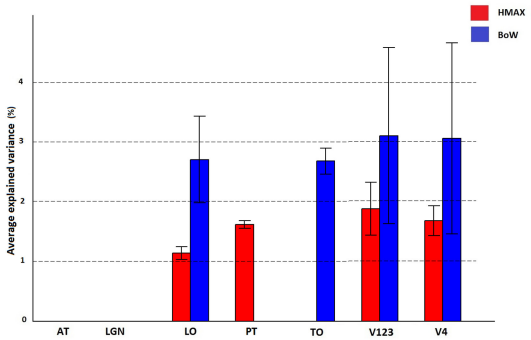


FIGURE 5 | Visualization of brain activity explained by HMAX and BoW model across all subjects in the selected ROIs. The explained variation of the significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

3.4. CROSS SUBJECT ROI ANALYSIS

A region of interest analysis was conducted to explicitly test the sensitivity of different brain regions to the models and their individual components. **Figure 5** shows how HMAX and BoW explain brain activity in 6 brain regions (out of the 25 brain areas analyzed). These ROIs are obtained based on the Jülich MNI 2 mm atlas. We show the explained variation for each model averaged across subjects and the voxels within each ROI (Note that this doesn't show single subject variation across ROIs). We observe that there is significant explained variation in areas TO (temporal occipital), LO, explained variationV123 and V4. The representations do not account for brain activity in areas such as LGN (lateral geniculate nucleus) and AT (anterior temporal). In all the regions the BoW model has a higher average explained variation than the HMAX model The difference in explained variation is significant ($p < 0.0001$). **Table 1** shows the number of voxels in each ROI obtained across subjects that exhibited significant brain activity and the maximum explained variation across subjects. We observe that the HMAX and BoW models explain more brain activity in early visual areas compared to the other areas.

Figure 6 shows how visual dictionaries from HMAX and BoW explain brain activity in the 6 brain regions (out of the 25 brain areas analyzed). It can be seen that there is significant explained variation in areas TO (temporal occipital), LO, V123 and V4. The average explained variation is slightly higher in the TO regions compared to V123. In all the regions the visual dictionary from BoW model has a higher average explained variation compared to the visual dictionary from the HMAX model ($p < 0.0001$). Also the combination of visual dictionaries from HMAX and BoW do not significantly increase the explained variation and is similar to the explained variation from BoW. **Table 2** shows that the visual dictionary from both the models explains a large number of voxels in LO and V4, however the visual dictionary from BoW has highest explained variations in LO and TO compared to HMAX. Also, we do not notice any brain activity in brain regions

Table 1 | Number of significant voxels and maximum explained variation for HMAX and Bow models in each ROI.

	No of significant voxels		Max explained variation	
	HMAX	BoW	HMAX	BoW
AT	0	0	0	0
LGN	0	0	0	0
LO	43	387	2	4
TO	0	31	0	4
V123	701	3671	3	5
V4	53	1141	2	5

The significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

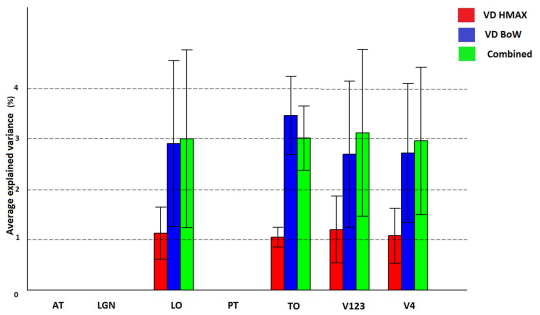


FIGURE 6 | Visualization of brain activity explained uniquely and the combination of visual dictionaries from HMAX, BoW across all subjects in the selected ROIs. The explained variation of the significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

Table 2 | Number of significant voxels and maximum explained variation for visual dictionaries from each ROI.

	No of significant voxels			Max explained variation		
	VD HMAX	VD BoW	Combined	VD HMAX	VD BoW	Combined
AT	0	0	0	0	0	0
LGN	0	0	0	0	0	0
LO	1427	1255	1402	3	9	10
TO	178	143	164	2	11	8
V123	4818	5705	5716	3	6	6
V4	1682	1878	1910	2	6	6

The significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

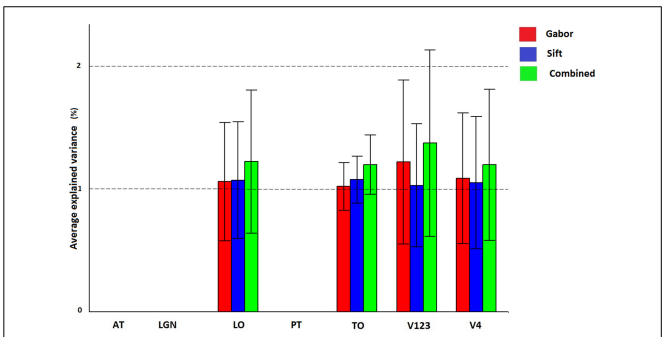


FIGURE 7 | Visualization of brain activity explained uniquely and the combination of Gabor and SIFT across all subjects in the selected ROIs. The explained variation of the significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

such as parahippocampal gyrus, retrosplenial corted and medial temporal lobe for either HMAX and BoW models.

Figure 7 and **Table 3** show how Gabor and SIFT explain brain activity in the 6 brain regions (out of the 25 brain areas analyzed). We observe that there is significant explained variation in areas TO (temporal occipital), LO, V123 and V4. For Gabor, the average explained variation is slightly higher in the V123 region compared to the other areas. Here we also observe that the Gabor and SIFT representations are not significantly different from each other and also the combination explains brain acitivity to the same extent.

Overall these results suggest that individually, the BoW visual dictionary is a better computational representation of neural responses (as measured by percent explained variation and consistency across subjects) than the visual dictionary from HMAX, which provides little additional information over the BoW visual dictionary.

Table 3 | Number of significant voxels and maximum explained variation for each ROI.

	No of significant voxels			Max explained variation		
	Gabor	SIFT	Combined	Gabor	SIFT	Combined
AT	0	0	0	0	0	0
LGN	0	0	0	0	0	0
LO	1394	1283	1427	2	2	2
TO	183	152	186	2	2	2
V123	5531	4418	5718	3	2	3
V4	1850	1535	1912	2	2	3

The significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

4. DISCUSSION

The success of models such as HMAX and BoW can be attributed to their use of features of intermediate complexity. The BoW model in particular has proven capable of learning to distinguish visual objects from only five hundred labeled examples (for each category of twenty different categories) in a fully automatic fashion and with good recognition rates (Salimi-Khorshidi et al., 2014). Many variations of this model exists (Jégou et al., 2012), and the recognition performance on a wide range of visual scenes and objects, improves steadily year by year (Salimi-Khorshidi et al., 2014). The HMAX model is a biologically plausible model for object recognition in the visual cortex which follows the hierarchical feedforward nature of the human brain. Both the models are candidate computational models of intermediate visual processing in the brain.

Our results show that in early visual brain areas such as V1, V2, and V3 there are regions in which brain activity is explained consistently across subjects by both the HMAX and BoW models. These models rely on gradient information. In the HMAX model, Gabor filters similar to the receptive fields in the V1 region of the brain are at the basis of visual representation. Similarly in the BoW model, the Scale Invariant Feature Transform (SIFT) features are the low level representation based on multi-scale and multi-orientation gradient features (Lowe, 2004). Although SIFT features originate from computer vision, their inspiration goes back to Hubel and Wiesel's (1962) simple and complex receptive fields, and Fukushima's (1980) Neocognitron model. SIFT features thus have an embedding in the visual system, much like Gabor filters have. In light of this, the sensitivity in brain areas V1, V2, and V3 to representations of the HMAX and BoW models is natural and in part due to low-level features. Interestingly we also observe that SIFT and Gabor representations explain brain activity in higher regions of the brain. This indicates that neurons in higher level visual areas process low level features pooled over local patches of the image for feedforward or feedback processing within visual cortex.

Brain areas higher up in the processing hierarchy appear to be particularly sensitive to visual dictionaries. Visual dictionaries are medium size image patches that are informative and distinctive at the same time, allowing for sparse and intermediate representations of objects and scenes. In computer vision visual dictionaries have proven to be very effective for object and scene classification (Jégou et al., 2012). The brain may compute visual dictionaries as higher-level visual building blocks composed of slightly larger receptive fields, and use visual dictionaries as intermediate features to arrive at a higher-level representation of visual input. We observe that both HMAX and BoW visual dictionaries explain some brain activity in higher level visual regions, with the BoW visual dictionary representation outperforming the HMAX model both in terms of explained variance and consistency.

HMAX and BoW both use low level features that are pooled differently in the various stages of processing. First HMAX pools Gabor features by a local max operator whereas BoW creates a histogram of orientations (SIFT). Then, BoW uses a learning technique (k-means clustering) on all the SIFT features from the image to form the visual dictionary. On the other hand HMAX uses random samples of Gabor features pooled over patches as its visual dictionary. This difference in aggregating low level features might explain why BoW provides a better computational representation of images.

BoW visual dictionaries may facilitate scene gist perception, which occurs rapidly and early in visual processing. While there is evidence that simple low-level regularities such as spatial frequency (Schyns and Oliva, 1994), color (Oliva, 2000) and local edge aligment (Loschky and Larson, 2008) underly scene gist representation, it is hitherto unknown whether and how mid-level features facilitate scene gist perception. BoW summarizes SIFT features computed over the entire image. It has been observed that such patterns of orientations and scales are believed to be used by V4 and IT (Oliva and Torralba, 2006). This is in accordance with our observation that the localization of BoW visual dictionary representations occur in V4 and areas anterior to V4 in the brain.

Our findings are in line with a recent study by Leeds et al. (2013). They compared multiple vision models against MRI brain activity in response to image stimuli. Leeds et al. conclude that the BoW model explains most brain activity in intermediate areas of the brain. For this model, they report that the correlation of the BoW model varies from 0.1 to 0.15 across the 5 subjects. In our study, we obtain similar results for the BoW model, and with an average explained variation across subjects of around 5% (with explained variations varying across subjects). Similarities and consistencies between our results and results in Leeds et al. (2013) further suggest that BoW computation might provide a suitable basis for intermediate features in the brain. Yamins et al. (2014) observe explained variance of up 25% for both HMAX and BoW models, and up to 50% for their HMO model (4-layer Convolutional neural network model) in brain areas IT and V4. The discrepancy between these results and our findings in terms of the magnitude of explained brain activity can be in part attributed to the use of high signal-to-noise ratio measurements in Yamins et al. (2014), such as electrophysiological data from monkeys. The neural sensitivity to convolutional neural network model is nevertheless promising. We will include deep neural networks in future work to understand how it performs on video stimuli.

Our study aims to understand if intermediate features used in the brain are connected to how computational models of vision use such intermediate features. Our findings suggest that visual dictionaries used in HMAX and BoW account for brain activity consistently across subjects. The result does not imply that visual dictionaries as computed by HMAX or BoW are actually used by the brain to represent scenes but it does suggest visual dictionaries might capture aspects of intermediate features. The results from this work are similar to previous work and provides new interesting insights into the nature of intermediate features in the brain. We have also provided a novel framework which allows us to dissociate the different levels of a hierarchical model, and individually understand their contribution to explain brain activity.

ACKNOWLEDGMENT

This research was supported by the Dutch national public-private research program COMMIT.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fncom.2014.00168/abstract

REFERENCES

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Buckner, R. L. (1998). Event-related fMRI and the hemodynamic response. *Hum. Brain Mapp.* 6, 373–377. doi: 10.1002/(SICI)1097-0193(1998)6:5/6<373::AID-HBM8>3.0.CO;2-P

Csurka, G., Christopher, D., Lixin, F., Jutta, W., and Cédric, B. (2004). "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, Vol. 1 (Prague) 1–2.

Everingham, M., Gool, L. J. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* 88, 303–338. doi: 10.1007/s11263-009-0275-4

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640. doi: 10.1126/science.1089506

Hubel, D., and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hung, C.-C., Carlson, E. T., and Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron* 74, 1099–1113. doi: 10.1016/j.neuron.2012.04.029

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., and Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1704–1716. doi: 10.1109/TPAMI.2011.235

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Jurie, F., and Bill, T. (2005). "Creating efficient codebooks for visual recognition," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference*, Vol. 1 (Beijing), 604–610.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 1–28. doi: 10.3389/neuro.06.004.2008

Leeds, D. D., Seibert, D. A., Pyles, J. A., and Tarr, M. J. (2013). Comparing visual representations across human fmri and computational vision. *J. Vis.* 13:25. doi: 10.1167/13.13.25

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Loschky, L., and Larson, A. (2008). Localized information is necessary for scene categorization, including the natural/man-made distinction. *J. Vis.* 8, 19. doi: 10.1167/8.1.4

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80, 45–57. doi: 10.1007/s11263-007-0118-0

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing 416 visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Oliva, A. (2000). Diagnostic colors mediate scene recognition. *Cognit. Psychol.* 41, 176–210. doi: 10.1006/cogp.1999.0728

Oliva, A., and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Vis. Percept. Prog. Brain Res.* 155, 23–36. doi: 10.1016/S0079-6123(06)55002-2

Parikh, D., and Zitnick, C. L. (2010). "The role of features, algorithms and data in visual recognition," in *CVPR* (San Francisco, CA: IEEE), 2328–2335.

Peres-Neto, P. R., Legendre, P., Dray, S., and Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87, 2614–2625. doi: 10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 3, 1199–1204. doi: 10.1038/81479

Rogerson, P. A. (1999). The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. *Geogr. Anal.* 31, 130–147. doi: 10.1111/j.1538-4632.1999.tb00973.x

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* 90, 449–468. doi: 10.1016/j.neuroimage.2013.11.046

Schyns, P., and Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychol. Sci.* 5, 195–200. doi: 10.1111/j.1467-9280.1994.tb00500.x

Smeaton, A. F., Over, P., and Kraaij, W. (2006). "Evaluation campaigns and trecvid," in *Multimedia Information Retrieval*, eds J. Z. Wang, N. Boujemaa, and Y. Chen (Santa Barbara, CA: ACM), 321–330.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. *Curr. Opin. Neurobiol.* 7, 523–529. doi: 10.1016/S0959-4388(97)80032-3

Ward, B. D. (2000). Simultaneous inference for FMRI data. *AFNI 3dDeconvolve Documentation, Medical College of Wisconsin*.

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* 14, 1370–1386. doi: 10.1006/nimg.2001.0931

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 June 2014; accepted: 05 December 2014; published online: 15 January 2015.

Citation: Ramakrishnan K, Scholte HS, Groen IIA, Smeulders AWM and Ghebreab S (2015) Visual dictionaries as intermediate features in the human brain. Front. Comput. Neurosci. ***8****:168. doi: 10.3389/fncom.2014.00168*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2015 Ramakrishnan, Scholte, Groen, Smeulders and Ghebreab. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 24 September 2014
doi: 10.3389/fncom.2014.00112

Complex cells decrease errors for the Müller-Lyer illusion in a model of the visual ventral stream

***Astrid Zeman*[1,2,3]*, *Oliver Obst*[2] *and Kevin R. Brooks*[3,4]**

[1] *Department of Cognitive Science, ARC Centre of Excellence in Cognition and its Disorders (CCD), Macquarie University, Sydney, NSW, Australia*
[2] *Digital Productivity and Services Flagship (DPAS), Commonwealth Scientific and Industrial Research Organisation, Marsfield, NSW, Australia*
[3] *Perception in Action Research Centre, Macquarie University, Sydney, NSW, Australia*
[4] *Department of Psychology, Macquarie University, Sydney, NSW, Australia*

Edited by:
Mazyar Fallah, York University, Canada

Reviewed by:
Ján Antolík, Centre National de la Recherche Scientifique, France
Victor De Lafuente, Universidad Nacional Autónoma de México, Mexico

***Correspondence:**
Astrid Zeman, Macquarie University, Level 3 Australian Hearing Hub, 16 University Avenue, Sydney, NSW 2109, Australia
e-mail: astrid.zeman@mq.edu.au

To improve robustness in object recognition, many artificial visual systems imitate the way in which the human visual cortex encodes object information as a hierarchical set of features. These systems are usually evaluated in terms of their ability to accurately categorize well-defined, unambiguous objects and scenes. In the real world, however, not all objects and scenes are presented clearly, with well-defined labels and interpretations. Visual illusions demonstrate a disparity between perception and objective reality, allowing psychophysicists to methodically manipulate stimuli and study our interpretation of the environment. One prominent effect, the Müller-Lyer illusion, is demonstrated when the perceived length of a line is contracted (or expanded) by the addition of arrowheads (or arrow-tails) to its ends. HMAX, a benchmark object recognition system, consistently produces a bias when classifying Müller-Lyer images. HMAX is a hierarchical, artificial neural network that imitates the "simple" and "complex" cell layers found in the visual ventral stream. In this study, we perform two experiments to explore the Müller-Lyer illusion in HMAX, asking: (1) How do simple vs. complex cell operations within HMAX affect illusory bias and precision? (2) How does varying the position of the figures in the input image affect classification using HMAX? In our first experiment, we assessed classification after traversing each layer of HMAX and found that in general, kernel operations performed by simple cells increase bias and uncertainty while max-pooling operations executed by complex cells decrease bias and uncertainty. In our second experiment, we increased variation in the positions of figures in the input images that reduced bias and uncertainty in HMAX. Our findings suggest that the Müller-Lyer illusion is exacerbated by the vulnerability of simple cell operations to positional fluctuations, but ameliorated by the robustness of complex cell responses to such variance.

Keywords: Müller-Lyer, illusion, HMAX, hierarchical, computational, model, visual, cortex

1. INTRODUCTION

Much of what is known today about our visual perception has been discovered through visual illusions. Visual illusions allow us to study the difference between objective reality and our interpretation of the visual information that we receive. Recently it has been shown that computational vision models that imitate neural mechanisms found in the ventral visual stream can exhibit human-like illusory biases (Zeman et al., 2013) . To the extent that the models are accurate reflections of human physiology, these results can be used to further elucidate some of the neural mechanisms behind particular illusions.

In this paper, we focus on the Müller-Lyer Illusion (MLI), which is a geometrical size illusion where a line with arrowheads appears contracted and a line with arrow-tails appears elongated (Müller-Lyer, 1889) (see **Figure 1**). The strength of the illusion can be affected by the fin angle (Dewar, 1967), shaft length (Fellows, 1967; Brigell and Uhlarik, 1979), inspection time (Coren and Porac, 1984; Predebon, 1997), observer age (Restle and Decker, 1977), the distance between the fins and the shaft (Fellows, 1967) and many other factors. The illusion classically appears in a four-wing form but can also manifest with other shapes, such as circles or squares, replacing the fins at the shaft ends. Even with the shafts completely removed, the MLI is still evident.

Here, we employ an underused method to explore the Müller-Lyer illusion and its potential causes using an Artificial Neural Network (ANN). To date, few studies have used ANNs to explore visual illusions (Ogawa et al., 1999; Bertulis and Bulatov, 2001; Corney and Lotto, 2007). In some cases, these artificial neural networks were not built to emulate their biological counterparts, but rather to demonstrate statistical correlations in the input. One such example is the model used by Corney and Lotto (2007), consisting of only one hidden layer with four homogenous neurons, which few would consider to be even a crude representation of visual cortex. The work presented by Ogawa et al. (1999) used a network with three hidden layers of "orientational neurons,"

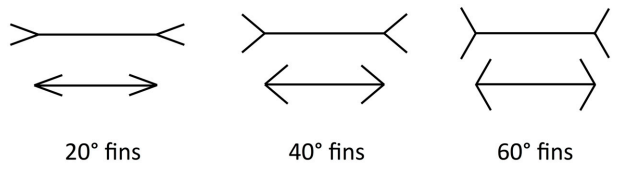


FIGURE 1 | The ML illusion in classical four-wing form. Horizontal lines are the same length in all cases. The ML effect is stronger for more acute angles (Left) and weaker for more obtuse angles (Right).

"rotational neurons" and "line unifying neurons." This network could roughly correspond to one layer of simple cells that provide orientation filters and one layer of complex cells that combine their output. However, this study presented no quantitative data and lacked a detailed description of the model, such as the size or connectivity of their network. Bertulis and Bulatov (2001) created a computer model to replicate the spatial filtering properties of simple cells and the combination of these units' outputs by complex cells in visual cortical area V1. Although they compared human and model data for the Müller-Lyer Illusion, their model centered only on the filtering properties of early visual neurons. These models do not adequately represent the multi-layered system that would best describe the relevant neural structures. Neuroimaging studies have shown areas V1, V2, V4, and IT are recruited when viewing the MLI (Weidner and Fink, 2007; Weidner et al., 2010) and hence the inclusion of operations from such visual ventral stream subdivisions is desirable. Therefore, studying the MLI in a computational model known to mimic these areas would provide a more biologically representative result.

In a previous report, we studied the MLI in a benchmark model of the ventral visual stream that imitates these cortical areas (Zeman et al., 2013). Following from our hypothesis that the MLI could occur in a model that imitates the structure and function of visual ventral areas, we demonstrated its manifestation in a biologically plausible artificial neural network. Although the models listed above are capable of reproducing the MLI, we believe our work provides a significant advance, being one of the first studies to model a visual illusion in a simulated replica of the ventral visual stream. In addition, our study contrasts with those above by employing techniques to train the model on multiple images before running a classification task and comparing the task of interest to a control. This allows us to separate the inner workings of the model from the input in the form of training images.

The model we recruit, HMAX (Serre et al., 2005), is a feed-forward, multi-layer, artificial neural network with layers corresponding to simple and complex cells found in visual cortex. Like visual cortex, the layers of HMAX alternate between simple and complex cells, creating a hierarchy of representations that correspond to increasing levels of abstraction as you traverse each layer. The simple and complex cells in the model are designed to match their physiological counterparts, as established by single cell recordings in visual cortex (Hubel and Wiesel, 1959). Here, we briefly describe single and complex cell functions and provide further detail on these later in Section 2.1. In short, simple cells extract low-level features, such as edges, an example of which would be Gabor filters that are often used to model V1 operations. The outputs of simple cells are pooled together by complex cells that extract combined or high-level features, such as lines of one particular orientation that cover a variety of positions within a visual field. Within HMAX, the max pooling function is used to imitate complex cell operations, giving the model its trademark name. In general, low-level features extracted by simple cells are shared across a variety of input images. High-level features are less common across image categories. The high-level features output by complex cells are more stable, invariant and robust to slight changes in the input.

HMAX has been extensively studied in its ability to match and predict physiological and psychological data (Serre and Poggio, 2010). Like many object recognition models, HMAX has been frequently tested using well-defined, unambiguous objects and scenes but has not been thoroughly assessed in its ability to handle visual illusions. Our previous demonstration of the MLI within HMAX showed not only a general illusory bias, but also a greater effect with more acute fin angles, corresponding to the pattern of errors shown by humans. Our replication of the MLI in this model allowed us to rule out some of the necessary causes for the illusion. There are a number of theories that attempt to explain the MLI (Gregory, 1963; Segall et al., 1966; Ginsburg, 1978; Coren and Porac, 1984; Müller-Lyer, 1896a,b; Bertulis and Bulatov, 2001; Howe and Purves, 2005; Brown and Friston, 2012) and here we discuss two. One common hypothesis is the "carpentered-world" theory—that images in our environment influence our perception of the MLI (Gregory, 1963; Segall et al., 1966). To interpret and maneuver within our visual environment, we apply a size-constancy scaling rule that allows us to infer the actual size of objects from the image that falls on our retina. While arrowhead images usually correspond to the near, exterior corners of cuboids, arrow-tail configurations are associated with more distant features, such as the right-angled corners of a room. If the expected distance of the features is used to scale our perception of size, when a line with arrowheads is compared to a line with arrow-tails that is physically equal in length, the more proximal arrowhead line is perceived as being smaller. Another common theory is based upon visual filtering mechanisms (Ginsburg, 1978). By applying a low spatial frequency filter to a Müller-Lyer image, the overall object (shaft plus fins) will appear elongated or contracted. Therefore, it could simply be a reliance on low spatial frequency information that causes the MLI. In our previous study, we were able to replicate the MLI in HMAX, allowing us to establish that exposure to 3-dimensional "carpentered world" scenes (Gregory, 1963) is not necessary to explain the MLI, as the model had no representation of distance and hence involved no size constancy scaling for depth. We also demonstrated that the illusion was not a result of reliance upon low spatial frequency filters, as information from a broad range of spatial frequency filters was used for classification.

In the current study, we set out to investigate the conditions under which the Müller-Lyer illusion manifests in HMAX and what factors influence the magnitude and precision of the effect. In particular, we address the following questions: (1) How do simple vs. complex cell operations within HMAX affect illusory bias and precision? (2) How would increasing the positional

variance of the input affect classification in HMAX? Our principal motivation is to discover how HMAX processes Müller-Lyer images and transforms them layer to layer. Following from this, we aim to find ways to reduce errors associated with classifying Müller-Lyer images, leading to improvements in biologically inspired computational models. We are particularly interested in how hierarchical feature representation could potentially lead to improvements in the fidelity of visual perception both in terms of accuracy (bias) and precision (discrimination thresholds).

2. MATERIALS AND METHODS

2.1. COMPUTATIONAL MODEL : HMAX

To explore where and how the illusion manifests, we first examined the architecture of HMAX: a multi-layer, feed-forward, artificial neural network (Serre et al., 2005; Mutch and Lowe, 2008; Mutch et al., 2010). Input is fed into an image layer that forms a multi-scale representation of the original image. Processing then flows sequentially through four more stages, where alternate layers perform either template matching or max pooling (defined below). HMAX operations approximate the processing of neurons in cat striate cortex, as established by single cell recordings (Hubel and Wiesel, 1959). Simple cells are modeled using template matching, responding with higher intensity to specific stimuli, while complex cell properties are simulated using max pooling, where the maximum response is taken from a pool of cells that share common features, such as size or shape.

Image information travels unidirectionally through four layers of alternating simple ("S") and complex ("C") layers of HMAX that are labeled S1, C1, S2, and C2. When the final C2 level is reached, output is compressed into a 1D vector representation that is sent to a linear classifier for final categorization. While previous versions of HMAX employed a support vector machine (SVM), in this paper we used the GPU-based version of HMAX (Mutch et al., 2010) that uses a linear classifier to perform final classification. The task for the classifier was to distinguish Long (i.e., top shaft longer) from Short (top shaft shorter) stimulus categories under a range of conditions, where the top or bottom line length varied by a known positive or negative extent. **Figure 2** summarizes the layers and operations in the model. Precise details are included in the original papers (Serre et al., 2005; Mutch and Lowe, 2008; Mutch et al., 2010).

2.2. STIMULI: TRAINING AND TEST SETS(CONTROL AND MÜLLER-LYER)

To carry out our procedure, we generated three separate image sets: a training (cross fin) set, a control test set (CTL) and an illusion test set (ML). All images were 256 × 256 pixels in size, with black 2 × 2 pixel lines drawn onto a white background (see **Figure 3**). Each image contained two horizontal lines ("shafts") with various fins appended. Each different image set was defined by the type of fins appended to the ends of the shafts. The fin type determines whether an illusory bias will be induced or not. Unlike the ML set, the cross fin and control test sets do not induce any illusions of line length in humans (Glazebrook et al., 2005; Zeman et al., 2013).

Within each two-line stimulus, the length of the top line was either "long" (L), or "short" (S), compared to the bottom line. The horizontal shaft length of the longest line was independently randomized between 120 and 240 pixels. The shorter line was varied by a negative extent randomly between 2 and 62 pixels for the training set, or by a known negative extent between 10 and 60 pixels for the test sets. The positions of each unified figure (shaft plus fins) were independently randomly jittered in the vertical direction between 0 and 30 pixels and in the horizontal direction between −30 and 30 pixels from center. The vertical position of the top line was randomized between 58 and 88 pixels from the top of the image while the bottom line's vertical position was randomized between 168 and 198 pixels. Top and bottom fin lengths randomized independently between 15 and 40 pixels. Fin lengths, line lengths and line positions remained consistent across all image sets. The parameters that varied between sets were fin angle, the direction of fins and the set size. If an image was generated that had any overlapping lines, for example, arrowheads touching or intersecting, these images were excluded from the sets.

Training images contained two horizontal lines with cross fins appended to the ends of the shafts (see Row 1, **Figure 3**). Fin angles were randomized independently for the top and bottom lines between 10 and 90°. Five hundred images per category (long and short) were used for training.

Two sets of test images were used, one as a control test set (CTL) and one as an illusion test set (ML). The CTL set used for parameterization contained left facing arrows for the top line and right facing arrows for the bottom line (see Row 2, **Figure 3**). CTL fin angles were randomized between 10 and 80° (the angles between top and bottom lines was the same). For parameterization, we used 200 images per category (totaling 400 images for both long and short) to test for overall accuracy levels with a randomized line length difference between 2 and 62 pixels. To establish performance levels for the control set, we tested 200 images per pixel condition for each category i.e., 200 images at 10, 20, 30, 40, 50, and 60 pixel increment differences for both short and long.

The ML set was used to infer performance levels for images known to induce an illusory bias in humans. In this ML set, all top lines contained arrow-tails and all bottom lines contained arrowheads (see Row 3, **Figure 3**). Fin angles for ML images were fixed at 20 and at 40° in two separate conditions. At the C2 layer, we tested 200 images for each pixel condition within each category (totaling 1200 images for the short category at 10, 20, 30, 40, 50, and 60 pixel length increments and 1200 for the long category). For all other layers (Input, S1, C1, and S2), we tested 100 images per pixel condition within each category. In each case we took the average of 10 runs, randomizing the order of training images. Classification results for the input, S1 and C1 levels are based on deterministic operations, without dependence on the weights developed during training. In these cases, randomizing the order of training images has no effect on classification results. To produce variation for these conditions, we generated additional test images that were randomized within the parameters specified above (with identical position ranges, fin angles, fin lengths, etc).

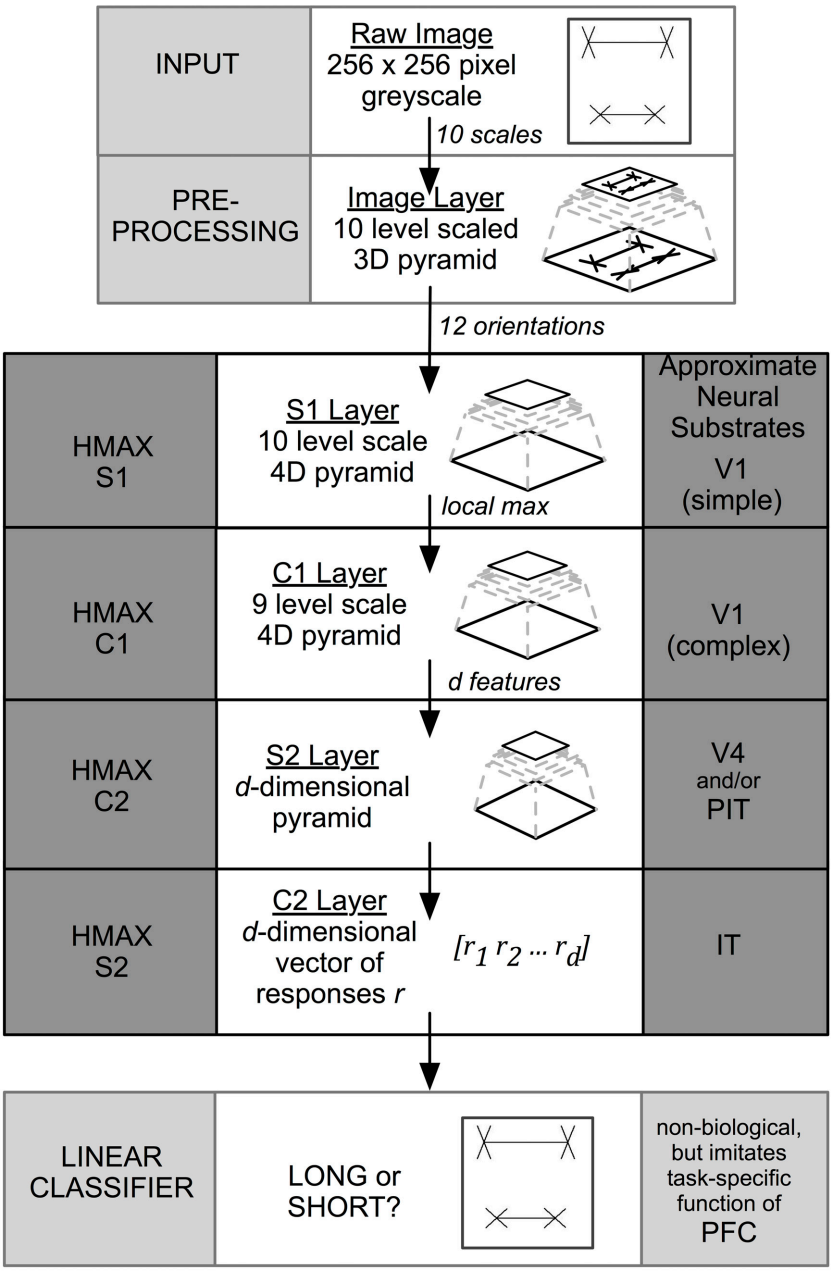


FIGURE 2 | HMAX architecture. The input to the system is a 256 × 256 pixel image. The output is a binary classification. HMAX consists of alternating template matching (S layers) and feature pooling (C layers). The neural substrate approximations are taken from Serre et al. (2005).

2.3. PROCEDURE: LEARNING, PARAMETERIZATION, ILLUSION CLASSIFICATION

Our method, established in Zeman et al. (2013), was carried out in three stages:

1. Training. Given a set of training images, a fixed-size network adjusted its internal weights to find the most informative features using unsupervised learning.
2. Test Phase 1: Parameterization. Using the CTL set, we ensured that the classifier was able to distinguish long from short images at an acceptable level of classification performance (above 85% correct), before testing with illusory stimuli. If performance fell below this level, we increased the size of the network and retrained (step 1).
3. Test Phase 2: Illusion classification. Using the ML set, we established the discrimination thresholds and the magnitude of the illusion that manifested in the model.

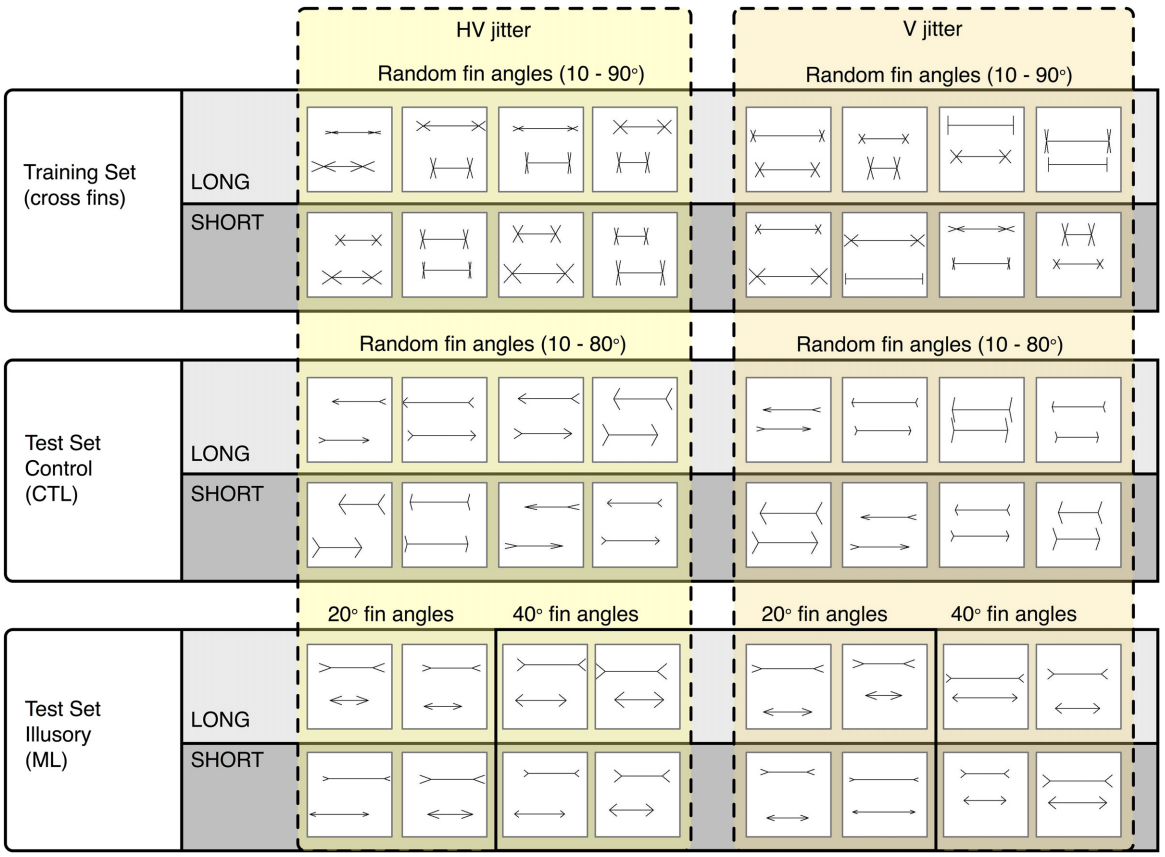


FIGURE 3 | Representative sample of images categorized as LONG or SHORT. The Cross fin set (Top row) was used for training. The Control CTL set (Middle row) and Illusory ML set were both used for testing. Images are grouped into those that were jittered both horizontally and vertically (Left group) and those that were jittered only vertically (Right group).

3. RESULTS

3.1. EXPERIMENT I: CLASSIFICATION OF ML IMAGES AFTER EACH LEVEL OF HMAX

The aim of this experiment was to assess how simple and complex cell operations contribute toward bringing about the MLI. To this end, we examined the inner workings of HMAX, looking at classification performance for illusory images at each level of the architecture. We used a linear classifier to perform classification after each subsequent layer of HMAX, (which included processing of all previous layers required to reach that stage). Therefore, we ran classification on the Input only, on S1 (after information arrived from Input), on C1 (after information traversed through Input and S1 layers) and so on.

We first tested classification performance on our control images, which exceeded 85% when the size of the S2 layer was 1000 nodes. Using this network configuration, we tested classification on 20 and 40° ML images at the C2 level. We then tested classification at each layer of HMAX using the same illusory set.

When plotted in terms of the percentage of stimuli classified as "long" as a function of the difference in line length (top–bottom) for each separate data set (i.e., control images, illusory images with 20° fins and with 40° fins), we observed a sigmoidal psychometric function, characteristic of human performance in equivalent psychophysical tasks. The data were characterized by a cumulative Gaussian, with the parameters of the best-fitting function determined using a least-squares procedure. **Figure 4** illustrates an example data set. When Gaussian curves did not fit significantly better than a horizontal line at 50% (chance responding) in an extra sum of squares F-test, the results were discarded (2 runs out of a total of 52). This allowed us to determine the Point of Subjective Equality (PSE) the line length difference for which stimuli were equally likely to be classified as long or short (50%), represented by the mean of the cumulative Gaussian. Here, PSEs are taken as a measure of accuracy, representing the magnitude of the Müller-Lyer Illusion manifested in the model. We also established the Just Noticeable Difference (JND) for each data set. The JND represents perceptual precision—the level of certainty of judgments for a stimulus type, and is indicated by the semi-interquartile difference of the Gaussian curve (the standard deviation multiplied by 0.6745). A higher JND represents greater uncertainty, and hence lower precision.

As can be seen in our results (see **Figure 5A**), the model produces a pattern of PSEs for illusory images consistent with human bias. We see a larger bias for more acute angles (20°) vs. less acute angles (40°), a pattern that is also consistent with human perception. This constitutes a replication of our previous findings

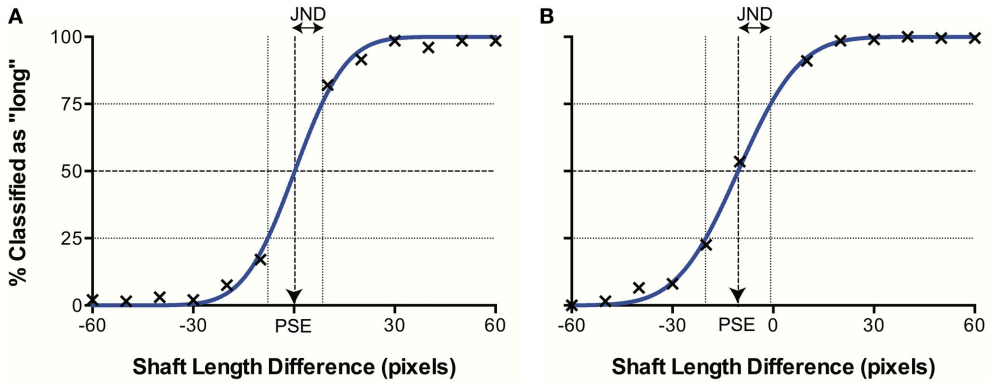


FIGURE 4 | Example data sets from (A) CTL and (B) ML (40°) conditions. The best fitting curve (blue) allows derivation of the point of subjective equality (PSE) where classification is at 50%, and the just noticeable difference (JND), corresponding to the semi-interquartile difference.

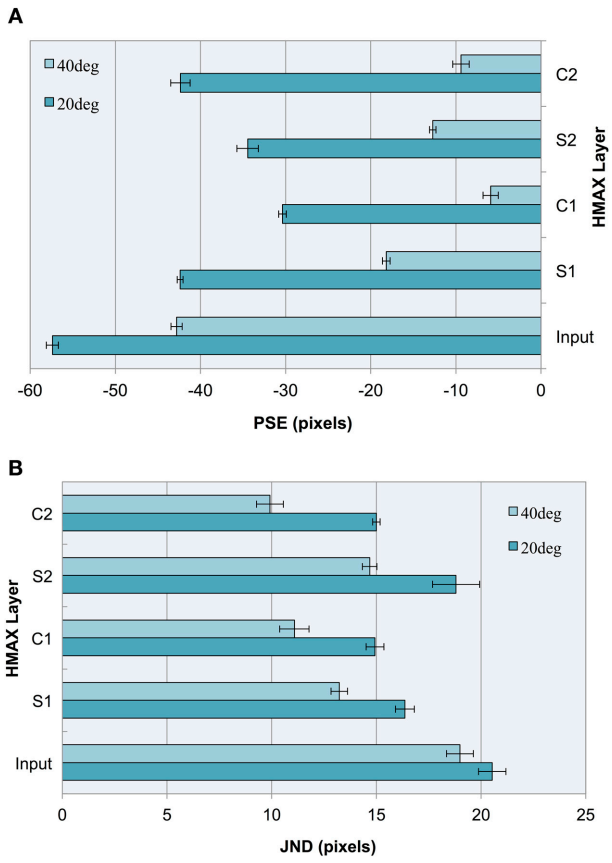


FIGURE 5 | Experiment I results as a function of the HMAX layer for images with 20° and 40° fins. Error bars represent ± 1 s.e.m. across multiple runs. **(A)** Accuracy (PSEs). **(B)** Precision (JNDs).

We observe that the illusion is present at the input level, suggesting that underlying statistical information may be present in our training images, despite careful design to remove bounding box cues and low spatial frequency information. The influence of image-source statistics on the Müller-Lyer illusion has already been studied using real-world environmental images and an input layer bias is to be expected (Howe and Purves, 2005). Because the aim of our study is to explore the Müller-Lyer within a biologically plausible model of the visual ventral stream, we are more interested in how the network would process the input. Our novel contribution, therefore, is to focus on how such information is transformed in terms of changes in accuracy and precision layer to layer as we traverse the cortical hierarchy within the HMAX network.

Observing the PSE for each HMAX layer after a linear classifier is applied, this experiment demonstrates three key findings:

1. Running a linear classifier on the raw images revealed a bias at the input level that would represent statistical influences such as those proposed by Howe and Purves (2005). However, each layer of the HMAX architecture counteracts this bias producing a reduction in PSE magnitude after every S and C layer is traversed, when compared to the input layer.
2. In the majority of cases (87.5% of the time), illusory bias and uncertainty is reduced after complex cell operations have been applied. A reduction in uncertainty and bias can be seen when comparing the PSE and JND for S1 vs. C1 layers, for both 20 and 40° fin angles in the illusion set. Going from S2 to C2, PSE is reduced for 40° angles but not for 20° angles in the ML set, whereas JND is reduced for all cases.
3. When simple cell operations follow complex, illusory bias and uncertainty is increased. At the S2 layer, we see an increase in PSE and in JND for both 20° and 40° ML images.

(Zeman et al., 2013) using a linear classifier, as opposed to a support vector machine (SVM), confirming that these findings are robust to the specific method of classification. These two trends are observable not only at the final C2 layer but at all levels of the architecture.

The observations concerning accuracy data are echoed for precision. In **Figure 5B**, we see a higher JND (lower precision) for images with more acute fin angles at all levels of HMAX architecture. Looking at each layer of the architecture, we see lower JNDs (higher precision) at each level of HMAX compared to the

input alone. We also observe higher precision (smaller JNDs) following processing by complex cells, but lower precision when the output from these layers is passed through a simple cell layer. In the case of results for precision, these observations held without exception.

The contrast between results following processing by simple cell and complex cell layers encourages examination of the principal differences between the operations performed by these cells. The major distinction between S-layer and C-layer operations concerns the response to variance in the image. Unlike simple cells, whose outputs are susceptible to image variations such as fluctuations in the locations of features, complex cells' filtering properties allow them to respond similarly to stimuli despite considerable positional variance. When initially designing the training stimuli for HMAX, we wanted the system to build higher-level representations of short and long independent of line position, exact line length and of features appended to the shaft ends. This would require an engagement of complex cell functionality and less reliance on simple cell properties. To this end, we varied these parameters randomly in a controlled fashion to reduce reliance on trivial image details. If one of our training parameters were to be restricted, the architecture would be less able to build such robust concepts of short and long. Given that complex cells are designed to pool information across simple cells with similar response properties and fire regardless of small changes in the afferent information, decreasing the variance in one of our training parameters would underutilize C cell properties and the short and long concepts within HMAX would become less flexible. This is likely to reduce the overall categorization performance of the computational model. More specifically, we hypothesize that restricting positional jitter to only one dimension would decrease accuracy and precision with which HMAX categorizes Müller-Lyer images. If this hypothesis holds true, we would demonstrate that greater positional variance reduces illusory bias and uncertainty. To seek further support for this proposition, we remove horizontal positional jitter from all stimuli in our second experiment.

3.2. EXPERIMENT II: HMAX CLASSIFICATION OF ML IMAGES WITH REDUCED VARIANCE

In our previous experiment, we observed a reduction in the level of bias after complex cell operations and hypothesized that introducing greater variance in the input would further reduce bias levels. To test this, we measured classification performance for HMAX layer C2 under two conditions: (1) Using our default horizontal and vertical jitter (HV) and (2) Under conditions of decreased positional jitter (V). We reduced the positional jitter in our training and test images from two-dimensional jitter in both the horizontal and vertical dimensions to one-dimensional, vertical jitter. While the top and bottom lines and their attached fins in our training and test sets remained independently jittered vertically (between 0 and 60 pixels), we removed all horizontal jitter, instead centering each stimulus. The vertical position of the top line was randomized between 48 and 108 pixels from the top of the image while the bottom line's vertical position was randomized between 148 and 208 pixels. We thus maintained a maximal 60 pixel jitter difference per line while limiting jitter to only one dimension.

In an initial parameterization stage, we first tested performance using the CTL set, and found an overall classification score of 91.5% with an S2 size of 1000 nodes. The results of control and illusion image classification for our default jitter condition and for reduced positional jitter is shown in **Figure 6**. In terms of accuracy measurements (**Figure 6A**), it can be seen that for ML images PSEs are more extreme for V jitter only, compared to HV jitter. These results provide support for our hypothesis, demonstrating an increase in the magnitude of the Müller-Lyer effect for both 20 and 40° illusory conditions when reducing positional jitter, and hence image variance. As in before, the pattern of results for accuracy is echoed in terms of precision measurements (**Figure 6B**). Following the trend from our previous experiment, we see lower JND values for more obtuse angles compared to more acute angles. Comparing JND results for HV jitter with those for V jitter, we see that the classifier has higher precision when distinguishing short from long lines in the HV condition. In summary, decreasing the amount of positional variance in our stimuli increases bias and reduces the level of certainty in making decisions.

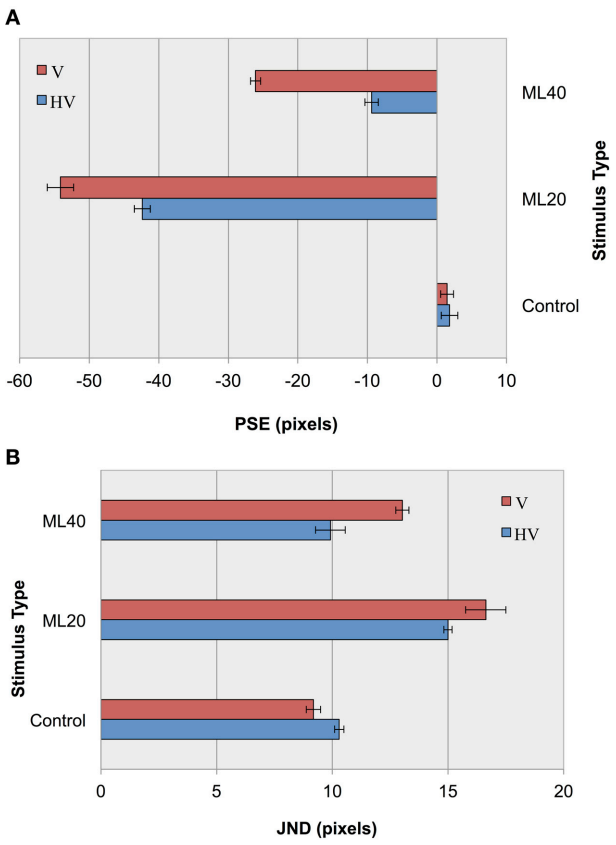


FIGURE 6 | Experiment II results as a function of jitter type for control images, and Müller-Lyer images with 20° and 40° fins. (A) Accuracy (PSEs). **(B)** Precision (JNDs).

4. DISCUSSION

Our aim for this study was to investigate the conditions under which the Müller-Lyer illusion manifests in HMAX and the factors that could influence the magnitude of the effect. Our primary motivation was to explore how hierarchical feature representation within HMAX affects classification performance. We ran two experiments performing binary image classification using HMAX. Images contained two horizontal lines that were jittered independently. Various configurations of fins were appended to the line shafts to create separate training and test images. Our first experiment compared the effects of operations performed by simple vs. complex cells by applying a linear classifier after each layer of HMAX when distinguishing long from short MLI images. Our second experiment examined HMAX classification of MLI images with decreased positional jitter.

The main finding from our first experiment is that the addition of any simple or complex cell layers reduces bias, compared to classification directly made on the input images. Illusory bias changes from layer to layer within a simple-complex cell architecture, with increases in MLI magnitude as information passes through simple layers. In most cases, the effect decreases as information passes through complex layers. The pattern of results for accuracy is replicated when measurements of precision are considered. All levels of HMAX show improved precision compared to classified input images, with further JND reductions caused by complex cell layers, and increases caused by simple cell layers. Proposing that the C layers' property of invariant responding may underlie their ability to increase accuracy and precision, we hypothesized that decreasing variance in the input images and re-training the network would increase the MLI. We chose to decrease the positional variance by removing horizontal jitter and including only vertical jitter for the stimuli in our second experiment. Consistent with our hypothesis, experiment 2 showed an increase in illusion magnitude for both 20 and 40° angles.

In this paper and in our previous study, we focused solely on the ML illusion in its classical four-wing form. It would also be possible to study other variants of the Müller-Lyer and other illusory figures to test more generally for the susceptibility of hierarchical artificial neural networks. Some variants of the Müller-Lyer to be tested could include changing the fins to circles (the "dumbbell" version) or ovals (the "spectacle" version) (Parker and Newbigging, 1963). Other monocular line length or distance judgment illusions occurring within the visual ventral stream may also manifest in similar hierarchical architectures, for example, the Oppel-Kundt illusion (Oppel, 1854/1855; Kundt, 1863).

Some illusions are moderated by the angle at which the stimulus is presented (de Lafuente and Ruiz, 2004). This raises the question whether illusory bias and uncertainty changes in classifying Müller-Lyer images that are presented diagonally, rotated by a number degrees to the left or to the right. Simple cells in HMAX consist of linear oriented filters, and are present in multiple orientations. The max pooling operations combine input from these and provide an output that is invariant to rotation. As a result, we would predict no difference in results when processing versions of the Müller-Lyer illusion in HMAX rotated at any arbitrary angle. This prediction is also consistent with human studies. While a number of illusions demonstrate an increase in magnitude when presented in a tilted condition, there is no difference in magnitude for the MLI (Prinzmetal and Beck, 2001).

In our last study, we recruited a previous version of HMAX known as FHLib, a Multi-scale Feature Hierarchy Library (Mutch and Lowe, 2008). In the current study, a more recent, GPU-based version of HMAX, known as CNS: Cortical Network Simulator (Mutch et al., 2010) was used. The main difference between these architectures was a linear classifier replacing the SVM in the final layer of the more recent code. The network setup between architectures was identical: one image layer followed by four layers of alternating S and C layers. Both had the same levels of inhibition (50% of cells in S1 and C1). The image layer contained 10 scales, each level $2^{1/4}$ smaller than the previous. Compared to our previous study, we were able to replicate similar levels of bias despite a change in the classifier, demonstrating that our result is robust and dependent upon properties of the HMAX hierarchical architecture, rather than the small differences between the implementation of these two related models.

Reflecting upon the implication of our results for other models, we would predict that those that have a similar hierarchical architecture would exhibit similar trends. That is, comparable networks would demonstrate increased bias with decreased precision when categorizing MLI images with less variance. Considering models that only contain filtering operations (akin to layers of simple cells) we would observe an illusory effect that may also be exacerbated compared to those with more complex operations, with low accuracy and precision. Examples of would include the model of Bertulis and Bulatov (2001).

The reduction of bias in computer vision systems has significant ramifications for applications such as automated driving, flight control and landing, target detection and camera surveillance. Correct judgment of distances and object dimensions in these systems could affect target accuracy and reduce the potential for crashes and errors. Our hypothesis that increasing positional variance in the stimuli would reduce the magnitude of illusory bias could be extended to include other forms of variance, such as image rotation, articulation or deformation, hence examining the generality of this proposal. Furthermore, it would be informative to test the generality of the results presented in this study in other computational models. If a general effect could be confirmed, then we would advise the implementation of many forms of input variance during training to improve their judgment capabilities, providing more accurate and precise information.

Our work not only has implications for the field of computer science, but also for psychology. Computational models allow manipulations of parameters that are impossible or impracticable to perform in human subjects, such as isolating the contributions of different neural structures to the effect. Artificial architectures allow us to make predictions about overall human performance as well as how performance changes from layer to layer within the visual system. Considering that this model not only provides an overall system performance (C2 output), but also supplies information at multiple levels of the architecture that correspond approximately to identifiable neural substrates, it may be possible to test the model's predictions with neuroimaging data. Using functional magnetic resonance imaging (fMRI), we could obtain

blood-oxygen-level dependent (BOLD) signals at different levels of the visual cortices of observers viewing the MLI compared to a control condition (using a similar method to that described by Weidner and Fink, 2007). Then by applying a classifier to these signals, we could map this information to changes in model bias and quantify how well the model matches human brain data. This forms a possible direction for future research.

FUNDING

Astrid Zeman is supported by a CSIRO Top-Up Scholarship and the Australian Postgraduate Award (APA) provided by the Australian Federal Government. Astrid Zeman is also supported by the Australian Research Council Centre of Excellence for Cognition and its Disorders (CE110001021) http://www.ccd.edu.au.

ACKNOWLEDGMENTS

We thank Dr. Kiley Seymour and Dr. Jason Friedman from the Cognitive Science Department at Macquarie University (Dr. Friedman now at Tel Aviv University, Israel) for helpful discussions. We thank Jim Mutch for making HMAX publicly available via GNU General Public License. We would like to thank Assistant Professor Sennay Ghebreab from the Intelligent Systems Lab at the University of Amsterdam for valuable input and suggestions. Lastly, we thank the reviewers for their helpful feedback, particularly with suggestions that we have incorporated into our discussion.

REFERENCES

Bertulis, A., and Bulatov, A. (2001). Distortions of length perception in human vision. *Biomedicine* 1, 3–26.

Brigell, M., and Uhlarik, J. (1979). The relational determination of length illusions and length aftereffects. *Perception* 8, 187–197. doi: 10.1068/p080187

Brown, H., and Friston, K. J. (2012). Free-energy and illusions: the cornsweet effect. *Front. Psychol.* 3:43. doi: 10.3389/fpsyg.2012.00043

Coren, S., and Porac, C. (1984). Structural and cognitive components in the Müller-Lyer illusion assessed via cyclopean presentation. *Percept. Psychophys.* 35, 313–318. doi: 10.3758/BF03206334

Corney, D., and Lotto, R. B. (2007). What are lightness illusions and why do we see them? *PLoS Comput. Biol.* 3:e180. doi: 10.1371/journal.pcbi.0030180

de Lafuente, V., and Ruiz, O. (2004). The orientation dependence of the Hermann grid illusion. *Exp. Brain Res.* 154, 255–260. doi: 10.1007/s00221-003-1700-5

Dewar, R. (1967). Stimulus determinants of the practice decrement of the Müller-Lyer illusion. *Can. J. Psychol.* 21, 504–520. doi: 10.1037/h0083002

Fellows, B. J. (1967). Reversal of the Müller-Lyer illusion with changes in the length of the inter-fins line. *Q. J. Exp. Psychol.* 19, 208–214. doi: 10.1080/14640746708400094

Ginsburg, A. (1978). *Visual Information Processing Based on Spatial Filters Constrained by Biological Data*. Ph.D. thesis, Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Cambridge.

Glazebrook, C. M., Dhillon, V. P., Keetch, K. M., Lyons, J., Amazeen, E., Weeks, D. J., et al. (2005). Perception–action and the Müller-Lyer illusion: amplitude or endpoint bias? *Exp. Brain Res.* 160, 71–78. doi: 10.1007/s00221-004-1986-y

Gregory, R. L. (1963). Distortion of visual space as inappropriate constancy scaling. *Nature* 199, 678–680. doi: 10.1038/199678a0

Howe, C. Q., and Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image–source relationships. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1234–1239. doi: 10.1073/pnas.0409314102

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591.

Kundt, A. (1863). Untersuchungen über Augenmaß und optische Täuschungen. *Poggen-dorffs Annalen der Physik und Chemie* 120, 118–158. doi: 10.1002/andp.18631960909

Müller-Lyer, F. C. (1889). Optische Urteilstäuschungen. *Archiv für Anatomie und Physiologie* 2, 263–270.

Müller-Lyer, F. C. (1896a). Über Kontrast und Konfluxion. (Zweiter Artikel). *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* 10, 421–431.

Müller-Lyer, F. C. (1896b). Zur Lehre von den optischen Tauschungen über Kontrast und Konfluxion. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* 9, 1–16.

Mutch, J., Knoblich, U., and Poggio, T. (2010). *Cns: A gpu-Based Framework for Simulating Cortically-Organized Networks*. Cambridge, MA: Technical Report MIT-CSAIL-TR-2010-013 / CBCL-286, Massachusetts Institute of Technology.

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comp. Vis.* 80, 45–57. doi: 10.1007/s11263-007-0118-0

Ogawa, T., Minohara, T., Kanaka, I., and Kosugi, Y. (1999). "A neural network model for realizing geometric illusions based on acute-angled expansion," in *6th International Conference on Neural Information Processing (ICONIP '99) Proceedings*, Vol. 2 (Perth, WA: IEEE), 550–555.

Oppel, J. (1854/1855). Über geometrischoptische Täuschungen (Zweite Nachlese). *Jahres-Bericht des Physikalischen Vereins zu Frankfurt am Main* 37–47.

Parker, N. I., and Newbigging, P. L. (1963). Magnitude and decrement of the Muller-Lyer illusion as a function of pre-training. *Can. J. Psychol* 17, 134–140. doi: 10.1037/h0083262

Predebon, J. (1997). Decrement of the Brentano Müller-Lyer illusion as a function of inspection time. *Perception* 27, 183–192. doi: 10.1068/p270183

Prinzmetal, W., and Beck, D. M. (2001). The tilt-constancy theory of visual illusions. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 206–217. doi: 10.1037/0096-1523.27.1.206

Restle, F., and Decker, J. (1977). Size of the Mueller-Lyer illusion as a function of its dimensions: theory and data. *Percept. Psychophys.* 21, 489–503. doi: 10.3758/BF03198729

Segall, M. H., Campbell, D. T., and Herskovits, M. J. (1966). *The Influence of Culture on Visual Perception*. New York, NY: The Bobbs-Merrill Company, Inc.

Serre, T., and Poggio, T. (2010). A neuromorphic approach to computer vision. *Commun. ACM* 53, 54–61. doi: 10.1145/1831407.1831425

Serre, T., Wolf, L., and Poggio, T. (2005). "Object recognition with features inspired by visual cortex," in *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)* (San Diego, CA: IEEE Computer Society Press), 886–893. doi: 10.1109/CVPR.2005.254

Weidner, R., Boers, F., Mathiak, K., Dammers, J., and Fink, G. R. (2010). The temporal dynamics of the Müller-Lyer illusion. *Cereb. Cortex* 20, 1586–1595. doi: 10.1093/cercor/bhp217

Weidner, R., and Fink, G. R. (2007). The neural mechanisms underlying the Müller-Lyer illusion and its interaction with visuospatial judgments. *Cereb. Cortex* 17, 878–884. doi: 10.1093/cercor/bhk042

Zeman, A., Obst, O., Brooks, K. R., and Rich, A. N. (2013). The Müller-Lyer illusion in a computational model of biological object recognition. *PLoS ONE* 8:e56126. doi: 10.1371/journal.pone.0056126

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2014; accepted: 29 August 2014; published online: 24 September 2014.

*Citation: Zeman A, Obst O and Brooks KR (2014) Complex cells decrease errors for the Müller-Lyer illusion in a model of the visual ventral stream. Front. Comput. Neurosci. **8**:112. doi: 10.3389/fncom.2014.00112*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Zeman, Obst and Brooks. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 September 2014
doi: 10.3389/fncom.2014.00106

Exploration of complex visual feature spaces for object perception

***Daniel D. Leeds*[1,2]*, *John A. Pyles*[2,3] *and Michael J. Tarr*[2,3]**

[1] *Computer and Information Science Department, Fordham University, Bronx, NY, USA*
[2] *Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA*
[3] *Psychology Department, Carnegie Mellon University, Pittsburgh, PA, USA*

Edited by:
Mazyar Fallah, York University, Canada

Reviewed by:
Cees Van Leeuwen, Katholieke Universiteit Leuven, Belgium
Paul Downing, University of Wales, UK

***Correspondence:**
Daniel D. Leeds, Computer and Information Science Department, Fordham University, 441 East Fordham Road, Room 328A, 340 John Mulcahy Hall, Bronx, NY 10458, USA
e-mail: dleeds@fordham.edu

The mid- and high-level visual properties supporting object perception in the ventral visual pathway are poorly understood. In the absence of well-specified theory, many groups have adopted a data-driven approach in which they progressively interrogate neural units to establish each unit's selectivity. Such methods are challenging in that they require search through a wide space of feature models and stimuli using a limited number of samples. To more rapidly identify higher-level features underlying human cortical object perception, we implemented a novel functional magnetic resonance imaging method in which visual stimuli are selected in real-time based on BOLD responses to recently shown stimuli. This work was inspired by earlier primate physiology work, in which neural selectivity for mid-level features in IT was characterized using a simple parametric approach (Hung et al., 2012). To extend such work to human neuroimaging, we used natural and synthetic object stimuli embedded in feature spaces constructed on the basis of the complex visual properties of the objects themselves. During fMRI scanning, we employed a real-time search method to control continuous stimulus selection within each image space. This search was designed to maximize neural responses across a predetermined 1 cm^3 brain region within ventral cortex. To assess the value of this method for understanding object encoding, we examined both the behavior of the method itself and the complex visual properties the method identified as reliably activating selected brain regions. We observed: (1) Regions selective for both holistic and component object features and for a variety of surface properties; (2) Object stimulus pairs near one another in feature space that produce responses at the opposite extremes of the measured activity range. Together, these results suggest that real-time fMRI methods may yield more widely informative measures of selectivity within the broad classes of visual features associated with cortical object representation.

Keywords: neuroimaging, object recognition, computational modeling, intermediate feature representation, real-time stimulus selection

1. INTRODUCTION

Object recognition associates visual inputs—beginning with an array of light intensities falling on our retinas—with semantic categories, for example, "cow," "car," or "face." Inspired by the architecture of the ventral occipito-temporal pathway of the human brain, models that attempt to implement or account for this process assume a feedforward architecture in which the features of representation progressively increase in complexity as information moves up the hierarchy (Riesenhuber and Poggio, 1999). The top layers of such a hierarchy are typically construed as high-level *object representations* that correspond to and allow the assignment of category-level labels. Critically, within such models, there is the presupposition of one or more levels of *intermediate* features that, while less complex than entire objects, nonetheless capture important—and compositional—object-level visual properties (Ullman et al., 2002). Yet, despite significant interest and study of biological vision, the nature of such putative intermediate features remains frustratingly elusive. To begin to address this gap, we explored the intermediate visual properties encoded within human visual cortex along the ventral pathway.

The majority of what we *have* learned about intermediate representation within the ventral cortex has come from primate neurophysiology studies. In a pioneering study, Tanaka (1996) explored the minimal visual stimulus sufficient to drive a given cortical neuron at a level equivalent to the complete object. He found that individual neurons in area TE were selective for a wide variety of simple patterns and that these patterns bore some resemblance to image features embedded within the objects initially used to elicit a response. Tanaka hypothesized that this pattern-specific selectivity has a columnar structure that maps out a high-dimensional feature space for representing visual objects. In more recent neurophysiological work, Yamane et al. (2008) and Hung et al. (2012) used a somewhat different search procedure to

identify the contour selectivity of individual neurons in primate inferotemporal cortex (IT). Using a highly-constrained, parameterized stimulus space based on 2D contours, they found that most contour-selective neurons in IT encoded a subset of that parameter space. Importantly, each 2D contour within this space mapped to specific 3D surface properties—thus, collections of these contour-selective units should be sufficient to capture the 3D appearance of an object or part. At the same time, recent primate physiology and human fMRI studies have begun to address the issue of intermediate representations. For example, op de Beeck et al. (2001) and op de Beeck et al. (2008) demonstrated that the pattern of responses to complex synthetic stimuli in object-selective cortex is associated with perceived shape similarity and, in particular, that this intermediate region of visual cortex is sensitive to shape features such as curved vs. straight.

Also within the domain of human neuroscience, Kay et al. (2008) explored how responses—as measured by fMRI—of voxels coarsely coding for orientation and scale within human V1, V2, and V3 can be combined to reconstruct complex images. Although this work offers a demonstration of how human neuroimaging methods may support more fine-grained analyses (and inspiration for further investigation), it does not inform us regarding the nature of intermediate features. In particular, models of the featural properties of V1 and V2 are common, so Kay et al.'s study largely demonstrates that such models hold even at the voxel/millions-of-neurons level without explicating any new properties or principles for these visual areas. Put another way, Kay et al. decoded features within a well-understood parameter space in which it is already agreed that the particular brain regions in question encode information about the orientations and scales of local edges. In contrast, the core problem in identifying the features of intermediate-level object representation is that the parameter space is extremely large and highly underspecified, therefore it is difficult to find effective prior models that will fit the data. In this context, the proposal of Ullman et al. (2002) that intermediate features can be construed as image fragments of varying scale and location—leaving the contents of said fragments entirely unspecified—is still one of the strongest models of intermediate-level representation. In particular, this model predicts which task-relevant object information is likely to be encoded within the human ventral pathway (Harel et al., 2007).

Note that the large majority of models applied to biological object recognition have made weak assumptions regarding the nature of intermediate features (with the notable exception being Hummel and Biederman (1992) who made very strong assumptions as to the core features used in object representation; unfortunately, such strong assumptions worked against the generality of the model). For example, many models employ variants of Gabor filterbanks, detecting local edges in visual stimuli, to explain selectivities in primary visual cortex (V1) (Hubel and Wiesel, 1968). Extending this approach, both Kay et al. (2008) and Serre et al. (2007) propose hierarchies of linear and non-linear spatial pooling computations, with Gabor filters at the base, to model higher-level vision. In this vein, perhaps the most well-specified hierarchical model is "HMAX" (Cadieu et al., 2007) and its variants (Serre et al., 2007). While these models partially predict neural selectivity in the mid-level ventral stream (V4) for simple synthetic stimuli (Cadieu et al., 2007), HMAX imperfectly clusters images of real-world objects relative to the clusterings obtained from primate neurophysiology or human fMRI (Kriegeskorte et al., 2008).

To address the question of the intermediate-level features underlying neural object processing, we adopted two different models of visual representation. First, we explored a visual parameter space defined by "SIFT" (Lowe, 2004)—a method drawn from computer vision that we have, previously, established as effective in explaining some of the variance observed in the neural processing of objects (Leeds et al., 2013). Second, we explored a novel visual parameter space defined by collections of 3D components—akin to Biederman's approach (Hummel and Biederman, 1992)—"Fribble" objects (Williams and Simons, 2000). Both of these representational choices arise from a diverse set of linear and non-linear operations across image properties and, as such, can be thought of as proxies for more detailed models of visual representation within biological systems (see Leeds et al., 2013).

Using these two models, we collected fMRI data from human observers performing a simple object processing task using real-world objects characterized by coordinates in SIFT space or synthetic objects characterized by coordinates in Fribble space. That is, stimuli were projected onto one of two types of feature spaces, constructed to reflect the SIFT and Fribble models of object representation. During scanning, specific stimuli from these spaces were sequentially selected in *real-time* based on an algorithmic search of each feature space for images (and their corresponding image features) that produced *maximal* BOLD activity in a preselected brain region of interest (ROI) within the ventral visual pathway.

These novel methods allowed us to evaluate principles of object representation within human visual cortex. In particular, beyond the specifically-observed organizational structure of cortex, we found some evidence for "local inhibition," in which cortical activity was reduced for viewing object images that varied slightly from preferred images for a given brain region. This finding expands on similar observations seen for earlier stages of visual processing (Hubel and Wiesel, 1968; Wang et al., 2012). With respect to topographic organization for objects, we observed that the object images producing the highest responses for a given ROI were often distributed across multiple areas of the visual feature space, potentially reflecting multiple neural populations with distinct selectivities encoded within small regions of visual cortex. Finally, across both real-world objects and Fribbles, we obtained some evidence for selectivity to local contours and textural surface properties.

Next we describe the novel methods that were integral to the execution of our study. In particular, we addressed two challenges. First, the potential space of object images, even given the reductions afforded by adopting SIFT or Fribble space, is massive. It was incumbent on us to implement a computationally-efficient image search strategy for stimulus selection. Second, because our goal was the real-time selection of stimuli, we developed a time-efficient means for measuring and processing BOLD signals.

2. METHODS

2.1. STIMULUS SELECTION METHOD

We developed methods for the dynamic selection of stimuli, choosing new images to display based on the BOLD response to previous images within a given pre-selected brain region. This search chooses each new stimulus by considering a space of visual properties and probing locations in this space (corresponding to stimuli with particular visual properties) in order to efficiently identify those locations that are likely to elicit maximal activity from the brain region under study (**Figure 1**). Each stimulus i that could be displayed is assigned a point in space p_i based on its visual properties. The measured response of the brain region to this stimulus r_i is understood as:

$$r_i = f(p_i) + \eta \quad (1)$$

That is, a function f of the stimulus' visual properties as encoded by its location in the representational space plus a noise term η, drawn from a zero-centered Gaussian distribution. The process of displaying an image, recording the ensuing cortical activity via fMRI, and isolating the response of the brain region of interest using the preprocessing program we model as performing an evaluation under noise of the function describing the region's response. For simplicity's sake, we perform stimulus selection assuming our chosen brain region has a selectivity function f that reaches a maximum at a certain point in the visual space and falls off with increasing Euclidean distance from this point. Under these assumptions, we use a modified version of the simplex simulated annealing Matlab code available from Donckels (2012), implementing the algorithm from Cardoso et al. (1996). An idealized example of what a search run might look like based on this algorithm is shown in **Figure 1B**. For each group, we performed searches in each of two scan sessions, starting at distinct points in the feature space for each session to probe the consistency of search results across different initial simplex settings. Further details are provided by Leeds (2013) and Cardoso et al. (1996).

2.2. STIMULUS DISPLAY

All stimuli were presented using MATLAB (2012) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) controlled by an Apple Macintosh and were displayed on a BOLD screen (Cambridge Research, Inc.) 24 inch MR compatible LCD display placed at the head end of the bore. Subjects viewed the images through a mirror attached to the head coil with object stimuli subtending a visual angle of approximately 8.3° × 8.3°.

2.3. fMRI PROCEDURES

Subjects were scanned using a 3 T Siemens Verio MRI scanner with a 32-channel head coil. Functional images were acquired with a $T2^*$-weighted echo-planar imaging (EPI) pulse sequence (31 oblique axial slices, in-plane resolution 2 × 2 mm, 3 mm slice thickness, no gap, sequential descending acquisition, repetition time $TR = 2000$ ms, echo time $TE = 29$ ms, flip angle = 72°, GRAPPA = 2, matrix size = 96 × 96, field of view FOV = 192 mm). An MP-RAGE sequence (1 × 1 × 1 mm, 176 sagittal slices, $TR = 1870$, $TI = 1100$, $FA = 8°$, GRAPPA = 2) was used for anatomical imaging.

2.4. EXPERIMENTAL DESIGN

For each subject, our study was divided into an initial "reference" scanning session and two "real-time" scanning sessions (**Figure 2A**). In the reference session we gathered cortical responses to four classes of object stimuli to identify cortical regions selective for each separate stimulus class. As described in Sections 2.6 and 2.7, two different stimulus sets, comprised of four visually-similar object classes, were used to explore visual feature selectivity: real-world objects and synthetic "Fribble" objects; each subject viewed stimuli from only one set. In the real-time scan sessions we searched for stimuli producing the maximal response from each of the four brain regions, dynamically choosing new stimuli based on each region's responses to recently shown stimuli.

Runs in the reference scan session followed a slow event-related design. Each stimulus was displayed in the center of the

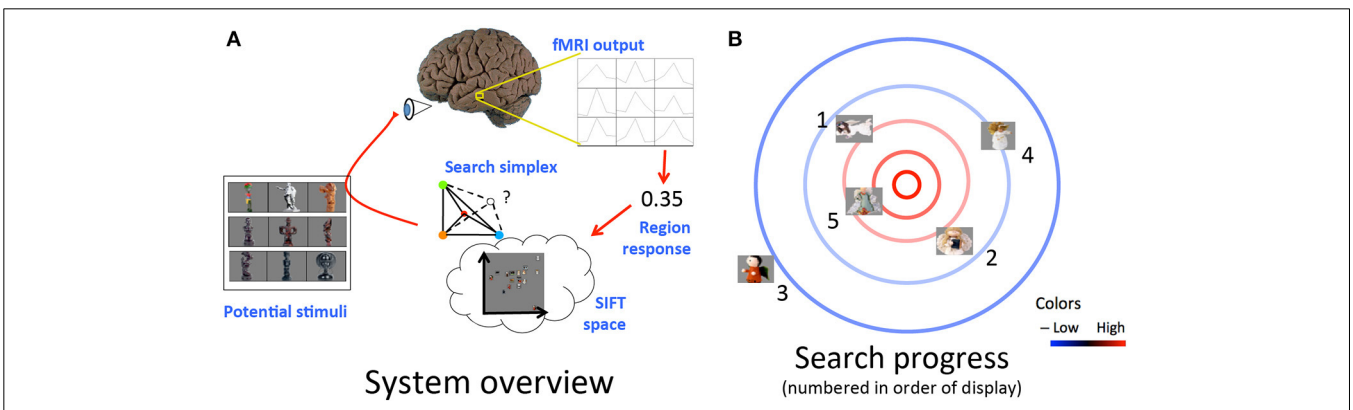


FIGURE 1 | (A) Schematic of loop from stimulus display to measurement and extraction of cortical region response to selection of next stimulus. **(B)** Example progression of desired stimulus search. Cortical response is highest toward the center of the space (red contours) and lowest toward the edges of the space (blue contours). Stimuli displayed in order listed. Cortical responses to initial stimuli, e.g., those numbered 1, 2, and 3, influence selection of further stimuli closer to maximal response region in visual space, e.g., those numbered 4 and 5.

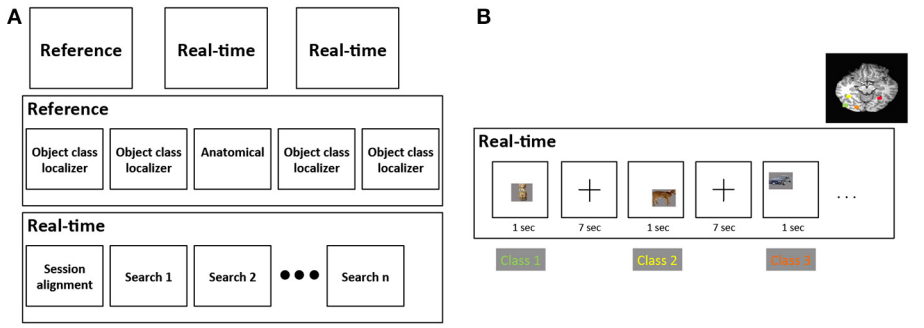


FIGURE 2 | (A) Structure of the three scanning sessions performed for each subject. *First row* depicts the three sessions, *second row* depicts the runs for the reference session, and *third row* depicts the runs for each real-time session. **(B)** An example of the alternation among four stimulus class searches in a real-time search run. These four classes are comprised of mammals, human-forms, cars, and containers, and correspond to four colored brain regions shown on the upper-right of the figure.

screen for 2 s followed by a blank 53% gray screen shown for a time period randomly selected to be between 500 and 3000 ms, followed by a centered fixation cross that remained displayed until the end of each 10 s trial, at which point the next trial began. As such, the SOA between consecutive stimulus displays was fixed at 10 s. Subjects were instructed to press a button when the fixation cross appeared. The fixation onset detection task was used to engage subject attention throughout the experiment. No other task was required of subjects, meaning that the scan assessed object perception under passive viewing conditions.

The stimuli were presented in four 3-min runs, spread across the 1-h scanning sessions and arranged to minimize potential adaptation and priming effects. Each run contained 36 object pictures, 9 objects from each of the four classes, ordered to alternate among the four classes. Stimulus order was randomized across runs. Over the course of the experiment, each subject viewed each picture four times; averaging across multiple repetitions was performed for each stimulus, described below, to reduce trial-by-trial noise. We determined from data gathered in Leeds et al. (2013) that relatively little information is gained by averaging over more than four repetitions.

To provide anatomical information, a T1-weighted structural MRI was performed between runs within the reference scan session.

In each of the two 1.5-h real-time scan sessions, the image selectivities of four distinct brain regions within ventral cortex were explored. For each brain region, a distinct search was performed using stimuli drawn from a single class of visual objects. Stimuli were presented for each search in 8.5-min "search" runs (4–8 runs were used per subject depending on other factors). Each stimulus was selected by the real-time search program based on responses of a pre-selected region of interest (ROI) to stimuli previously shown from the same object category. Each run contained 60 object pictures, 15 objects from each object class, ordered to alternate through the four classes—that is, `search 1` → `search 2` → `search 3` → `search 4` → `search 1` ⋯—as illustrated in **Figure 2B**. Alternation among distinct searches employing visually-distinct classes was advantageous in decreasing the risk of cortical adaptation that would have been present if multiple similar stimuli had been shown in direct succession. The focus of each search within an object class also limited visual variability across stimuli within that search. This also enabled the remaining sources of variability to be more intuitively identified and more readily associated with their influence on the magnitude of cortical activity. Note that the overt task during search runs varied depending on the stimuli shown. Task details are provided in Sections 2.6.4 and 2.7.4.

Each real-time session began with a 318-s functional scan performed with a viewing task to engage subject attention. The first functional volume scanned for this task was used to align the ROI masks (defined in Sections 2.6.5 and 2.7.5) selected from the reference session for a given subject to the subject's brain's position in the current session. This alignment corrects for changes in head position between the reference and the real-time scan sessions that might result in the brain, and its associated ROIs, moving to different locations in the scan volume. The remaining data volumes from this beginning task were ignored in that this task was designed simple to occupy the attention of the subject while computing inter-session brain alignment to be used for the remainder of the session.

2.5. PREPROCESSING

During analyses of the reference scan session, functional scans were coregistered to the anatomical image and motion corrected using AFNI (Pittman, 2011). Highpass filtering was implemented in AFNI by removing sinusoidal trends with periods of half and full length of each run (338 s) as well as polynomial trends of orders one through three. The data then were normalized so that each voxel's time course was zero-mean and unit variance (Just et al., 2010). To allow multivariate analysis to exploit information present at high spatial frequencies, no spatial smoothing was performed (Swisher et al., 2010).

During real-time scan sessions, functional volumes were motion corrected using AFNI. Polynomial trends of orders one through three were removed. The data then were normalized for each voxel by subtracting the average and dividing by the standard deviation, obtained from the currently analyzed response

and from the previous "reference" scan session, respectively, to approximate zero-mean and unit variance (Just et al., 2010). The standard deviation was determined from 1 h of recorded signal from the reference scan session to gain a more reliable estimate of signal variability in each voxel. Due to variations in baseline signal magnitude across and within scans, each voxel's mean signal value required updating based on activity in each block (the time covering the responses for two consecutive trials). To allow multivariate analysis to exploit information present at high spatial frequencies, no spatial smoothing was performed (Swisher et al., 2010).

Matlab was used to perform further processing on the fMRI time courses for the voxels in the cortical region of interest for the associated search. For each stimulus presentation, the measured response of each voxel consisted of five data samples starting 2 s/1 TR after onset. Each five-sample response was consolidated into a weighted sum by computing the dot product of the response and the average hemodynamic response function (HRF) for the associated region. The HRF was determined based on data from the reference scan session. The pattern of voxel responses across the region was consolidated further into a single scalar response value by computing a similar weighted sum. Like the HRF, the voxel weights were determined from reference scan data. The weights corresponded to the most common multi-voxel pattern observed in the region during the earlier scan; that is, the first principal component of the set of multi-voxel patterns. This projection of recorded real-time responses onto the first principal component treats the activity across the region of interest as a single locally-distributed code, emphasizing voxels whose contributions to this code are most significant and de-emphasizing those voxels with typically weak contributions to the average pattern.

During the alignment run of each real-time session, AFNI was used to compute an alignment transformation between the initial functional volume of the localizer and the first functional volume recorded during the reference scan session. The transformation computed between the first real-time volume and the first reference volume was applied in reverse to each voxel in the four ROIs determined from the reference scan.

2.6. REAL-WORLD OBJECTS EMBEDDED IN SIFT SPACE

We pursued two methods to search for visual feature selectivity. In our first method, we focused on the perception of real-world objects with visual features represented by the scale invariant feature transform (SIFT, Lowe, 2004).

2.6.1. Subjects

Ten subjects (four female, age range 19–31) from the Carnegie Mellon University community participated, provided written informed consent, and were monetarily compensated for their participation. All procedures were approved by the Institutional Review Board of Carnegie Mellon University.

2.6.2. Stimuli

Stimulus images were drawn from a picture set comprised of 400 distinct color object photos displayed on 53% gray backgrounds (**Figure 3A**). The photographic images were taken from the Hemera Photo Objects dataset from Hemera Technologies (2000–2003). The number of distinct exemplars in each object class varied from 68 to 150 object images. Note that our use of real-world images of objects rather than the hand-drawn or computer-generated stimuli employed in past studies of intermediate-level visual coding (e.g., Cadieu et al., 2007; Yamane et al., 2008) was intended to more accurately capture a broad set of naturally-occurring visual features.

2.6.3. Defining SIFT space

Our real-world stimuli were organized into a Euclidean space (**Figure 3B**) that was constructed to reflect a scale invariant feature transform (SIFT) representation of object images (Lowe, 2004). Leeds et al. (2013) found that a SIFT-based representation of visual objects was the best match among several machine vision models in accounting for the neural encoding of objects in mid-level visual areas along the ventral visual pathway. The SIFT measure groups stimuli according to a distance matrix for object pairs (Leeds et al., 2013). In our present work, we defined a Euclidean space based on the distance matrix using Matlab's implementation of metric multidimensional scaling (MDS, Seber, 1984). MDS finds a space in which the original pairwise distances

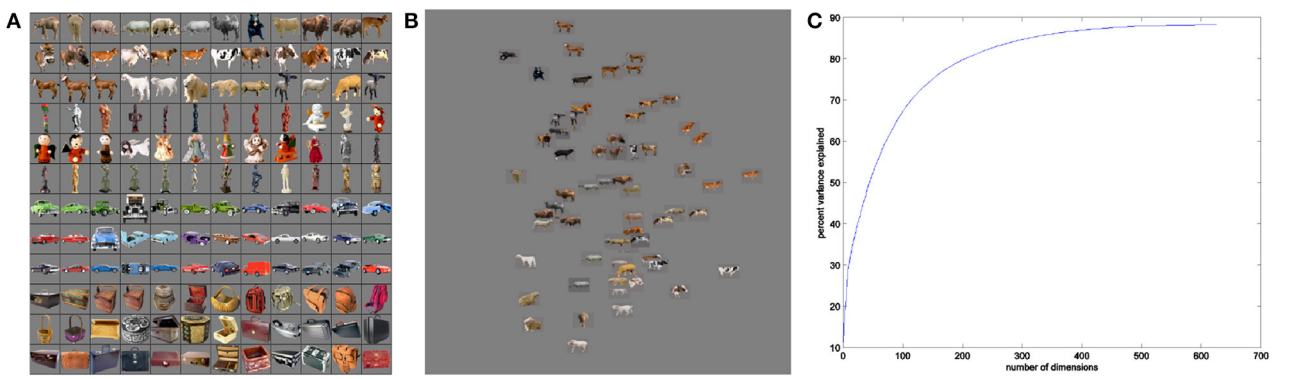


FIGURE 3 | Example real-world objects (A) and corresponding SIFT feature space (B). Real-world object images were selected from four object classes—mammals, human-forms, cars, and containers. Feature space shows example stimuli projected onto first two dimensions of space. **(C)** Percent variance explained using first *n* dimensions of MDS feature space for SIFT.

between data points—that is, SIFT distances between stimuli—are maximally preserved for any given n dimensions. This focus on maintaining the SIFT-defined visual similarity groupings among stimuli—using MDS—was motivated by the observations of Kriegeskorte et al. (2008) and Edelman and Shahbazi (2012), both of whom argued for the value of studying representational similarities to understand cortical vision.

The specific Euclidean space used in our study was derived from a SIFT-based distance matrix for 1600 Hemera photo objects, containing the 500 stimuli available for display across the real-time searches, as well as 1100 additional stimuli included to further capture visual diversity across the appearances of real-world objects (nb. ideally, the object space would be better covered by many more than 1600 objects, however, we necessarily had to restrict the total number of objects in order to limit the computation time required to generate large distance matrices). This distance matrix was computed using a "bag of words" method (Nowak et al., 2006; Leeds et al., 2013):

(1) Several SIFT feature vectors were computed for each stimulus
(2) 128 visual words were defined to describe typical feature vectors appearing in the 1600 photo objects
(3) Each vector for each stimulus was assigned to its closest "word"
(4) Each stimulus was represented by a histogram counting the number of times each word occurs in the image
(5) Stimulus pairs were compared using the Kullback-Leibler divergence (Kullback and Leibler, 1951) between their corresponding histograms

MDS was then used to generate a Euclidean space into which all stimulus images were projected. The real-time searches for each object class operated within the same MDS space. This method produced an MDS space containing over 600 dimensions. Unfortunately, as the number of dimensions in a search space increases, the sparsity of data in the space will increase exponentially. As such, any conclusions regarding the underlying selectivity function will become increasingly more uncertain absent further search constraints. To address this challenge, we constrained our real-time searches to use only the four most-representative dimensions from the MDS space.

2.6.4. Experimental design

Search runs in the real-time scan sessions employed a one-back location task to engage subject attention throughout the experiment. Each stimulus was displayed centered on one of nine locations on the screen for 1 s followed by a centered fixation cross that remained until the end of each 8 s trial, at which point the next trial began. Subjects were instructed to press a button when the image shown in this subsequent trial was centered on the same location as the image shown in the previous trial. The specific nine locations were defined by centering the stimulus at +2.5, 0, or −2.5° horizontally and/or vertically displaced from the screen center. From one trial to the next, the stimulus center shifted with a 30% probability.

2.6.5. Selection of regions of interest (ROIs)

Reference scan data was used to select ROIs for further study in real-time scan sessions.

Class localizer: For each stimulus class S, selectivity s_c was assessed for each voxel by computing:

$$s_c = \frac{\langle r_c \rangle - \langle r_{\bar{c}} \rangle}{\sigma(r_c)} \quad (2)$$

where $\langle r_c \rangle$ is the mean response for stimuli within the given class c, $\langle r_{\bar{c}} \rangle$ is the mean response for stimuli outside of the class $\bar{c}$, and $\sigma(r_c)$ is the standard deviation of responses within the class[1]. We identified clusters of voxels with the highest relative responses for the given class using a manually-selected threshold and clustering through AFNI.

SIFT localizer: The representational dissimilarity matrix-searchlight method described in Leeds et al. (2013) was used to determine brain regions with neural representations of objects similar to the representation of the same objects by SIFT. Thresholds were adjusted by hand to find contiguous clusters with high voxel sphere z values.

Selection of ROIs: Visual inspection was used to find overlaps between the class-selective and SIFT-representational regions. For each class, a 125 voxel cube ROI was selected based on the observed overlap in a location in the ventral visual stream. The use of relatively small—one cubic centimeter—cortical regions enables exploration of local selectivities for complex visual properties. Analyses were successfully pursued on similar spatial scales in Leeds et al. (2013), using 123-voxel searchlights.

2.7. FRIBBLE OBJECTS EMBEDDED IN FRIBBLE SPACE

Our second attempt to search for visual feature selectivity focused on the perception of synthetic novel objects—Fribbles—in which visual features were parameterized as interchangeable 3D components (Williams and Simons, 2000).

2.7.1. Subjects

Ten subjects (six female, age range 21–43) from the Carnegie Mellon University community participated, provided written informed consent, and were monitarily compensated for their participation. All procedures were approved by the Institutional Review Board of Carnegie Mellon University.

2.7.2. Stimuli

Stimulus images were generated based on a library of synthetic Fribbles (Williams and Simons, 2000; Tarr, 2013), and were displayed on 54% gray backgrounds as in Section 2.6.2. Fribbles are animal-like objects composed of colored, textured geometric shapes. They are divided into classes, each defined by a specific body form and a set of four locations for attached appendages. In the library, each appendage has three potential shapes, e.g., a circle, star, or square head for the first class in **Figure 4A**, with

[1]The measured response of each voxel for each stimulus repetition consisted of five data samples starting 2 s after stimulus onset, corresponding to the 10 s between stimuli. Each five-sample response was consolidated into a single value—the average of the middle three samples of the response (Just et al., 2010)—intended to estimate the peak response.

potentially variable corresponding textures. These stimuli provide a careful control on the varying properties displayed to subjects, in contrast to the more natural, but less parameterized real-world objects.

2.7.3. Defining Fribble space

We organized our Fribble stimuli into Euclidean spaces. In the space for a given Fribble class, movement along an axis corresponded to morphing the shape of an associated appendage. For example, for the purple-bodied Fribble class, the axes were assigned to (1) the tan head, (2) the green tail tip, and (3) the brown legs, with the legs grouped and morphed together as a single appendage type. Valid locations on each axis spanned from −1 to 1 representing two end-point shapes for the associated appendage, (e.g., a circle head or a star head). Appendage appearance at intermediate locations was computed through the morphing program Norrkross MorphX (Wennerberg, 2009) based on the two end-point shapes. Example morphs can be seen in the Fribble space visualization in **Figure 4B**.

For each Fribble class, stimuli were generated for each of 7 locations—the end-points −1 and 1 as well as coordinates −0.66, −0.33, 0, 0.33, and 0.66—on each of 3 axes, i.e., $7^3 = 343$ locations. A separate space was searched for each class of Fribble objects.

2.7.4. Experimental design

Search runs in the real-time scan sessions employed a dimness detection task to engage subject attention throughout the experiment. Each stimulus was displayed in the center of the screen for 1 s followed by a centered fixation cross that remained displayed until the end of each 8 s trial, at which point the next trial began. On any trial there was a 10% chance the stimulus would be displayed as a darker version of itself—namely, the stimulus' red, green, and blue color values each would be decreased by 50 (max intensity 256). Subjects were instructed to press a button when the image appeared to be "dim or dark." For the Fribble stimuli, the dimness detection task was used to address concerns we had regarding the one-back location task used with real-world object stimuli. In particular, the fact that subjects necessarily had to hold two objects in memory simultaneously in order to perform the one-back location task may have "blurred" our ability to assess the neural representation of single objects on any given trial. This confound may have limited the strength of real-world object search results. Thus, our change to the dimness detection task.

2.7.5. Selection of Fribble class regions of interest

We employed the representational dissimilarity matrix-searchlight procedure of Leeds et al. (2013) to identify cortical areas whose visual representations are well characterized by each simple Fribble space. ROIs were selected manually from these areas for study during the real-time scan sessions. In these regions, we could search effectively for complex featural selectivities using the associated Fribble space.

3. RESULTS

Our study was designed to explore complex visual properties utilized for object perception by the ventral pathway of the brain. We studied the distribution of recorded ROI responses in our novel visual feature spaces, defined and explored separately for real-world objects and for Fribble objects.

3.1. VISUALIZING FEATURE SPACES

To search for those visual properties selectively activating different cortical regions within the ventral pathway we constructed two types of visual feature spaces. Each of these spaces—Euclidean in nature—represented an array of complex visual properties through the spatial grouping of image stimuli that were considered similar according to the defining visual metric, as in Sections 2.6.3 and 2.7.3.

Of note, each space contains a low number of dimensions—four dimensions for SIFT and three dimensions for each Fribble class—to allow the searches for visual selectivity to converge in the limited number of simplex steps that can be evaluated in real-time over the course of a scanning session. These low dimensional spaces also permit visualization of search activity over each scan session and visualization of general ROI response intensities across the continuum of visual properties represented by a given space. We display this information through colored scatter plots. For example, representing each stimulus as a point in feature space, **Figure 5** shows the locations in SIFT-based space selected, or "visited," by the search for human-form images evoking high activity in the pre-selected SIFT/"human-form" region of subject S3, and shows the regional response to each of the displayed

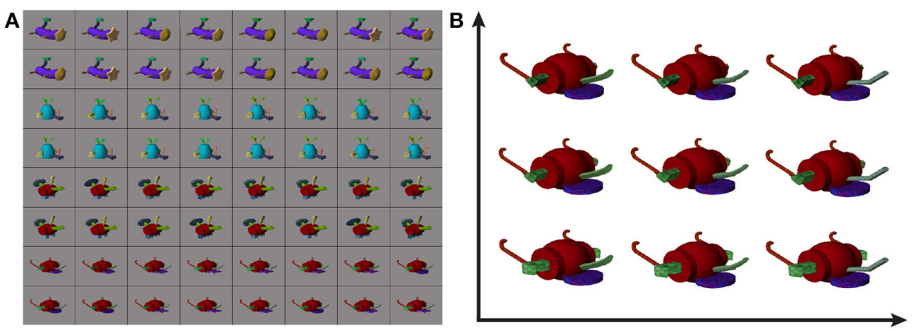


FIGURE 4 | Example Fribble objects (A) and example corresponding Fribble feature space (B). Fribble images were selected from four synthesized classes, shown in rows 1/2, 3/4, 5/6, and 7/8, respectively. Feature space shows stimuli projected onto first two dimensions of space.

stimuli. The four dimensions of SIFT-based space are projected onto its first two and second two dimensions in **Figures 5A,B**, respectively. Stimuli visited during the first and second real-time sessions are shown as circles and diamonds, respectively, centered at the stimuli's corresponding coordinates in the space. (Black dots correspond to the locations of all stimuli in the human-form class that were available for selection by the search program.) The magnitude of the average ROI response to a given visited stimulus is reflected in the color of its corresponding shape. For stimuli visited three or more times, colors span dark blue–blue–dark red–red for low through high average responses.

3.2. REAL-TIME SEARCH BEHAVIOR

In real-time scanning sessions, dynamic stimulus selection was pursued to more effectively explore each space of visual properties in limited scan time and to quickly identify visual properties producing the strongest activity from cortical regions in the ventral object perception pathway. Because the methods for real-time search are novel, we assess and confirm their expected performance in addition to studying the visual properties discovered by these methods. In particular, we expected each search in visual feature space to show the following two properties:

1. Convergence onto one, or a few, location(s) in the associated visual space producing greatest cortical response, corresponding to the regional selectivity.
2. Consistency in stimuli found to be preferred by the ROI, despite differing search starting points in visual feature space in the two scanning sessions for each subject.

However, because of the novelty of our methods—and thus our limited knowledge about optimal search parameters—and because of the limited number of stimulus display trials available, convergence occurred for only 10% of searches of real-world objects and 25% of searches of Fribble objects, as judged by a measure of convergence significance devised for our study (Section S1). We focus our ensuing analyses on the convergent and consistent searches. We anticipate further methodological development stemming from our present study will improve search convergence in future studies.

3.3. SELECTION OF BRAIN REGIONS OF INTEREST

Both for subjects viewing real-world objects and subjects viewing Fribble objects, ROIs containing cubes of 125 voxels were manually selected for each of four stimulus classes searched (**Figure 6**). Beyond incorporating voxels most highlighted by reference scan analyses reviewed above, the four regions for each subject were selected to be non-overlapping and to lie within the ventral pathway, with a preference for more anterior voxels, presumably involved in higher levels of object perception. With this selection approach in mind, consideration of the anatomical locations of the chosen ROIs provides perspective on the span of areas using SIFT-like and "Fribble-morph-like" representational structures across subjects, and the distribution of areas most strongly encoding each of the four studied object classes across subjects. We also gain perspective on the range of brain areas across subjects and searches studied for complex visual selectivities.

ROIs used for real-world object searches are distributed around and adjacent to the fusiform cortex, while ROIs used for Fribble object searches are distributed more broadly across the ventral pathway.

3.4. COMPLEX VISUAL SELECTIVITIES

We examine cortical responses observed for stimuli displayed in searches, selected for convergence and consistency, to determine visual properties significant to ROI representations of visual objects. In particular, we study the frequently-visited stimuli, ranked by ROI responses, to intuit important visual properties for each ROI and use the scatter plot introduced in Section 3.1 to visualize cortical activity across visual space, as well as to observe search behavior. The adaptive trajectory of each real-time search further reflects ROI selectivities. In the following two sections,

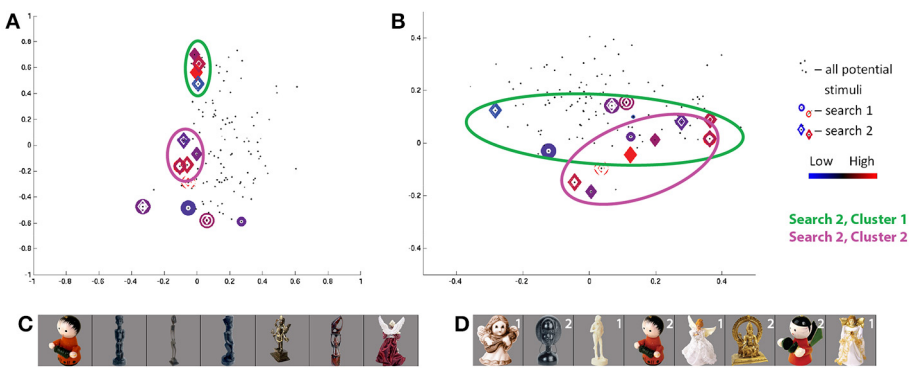


FIGURE 5 | Search results for S3, class 2 (human-forms), shown in (A) first and second SIFT space dimensions and (B) third and fourth dimensions. Location of all potential stimuli in space shown as black dots. Results from real-time scan session 1 are circles, results from real-time scan session 2 are diamonds. For stimuli "visited" (i.e., selected by the search) three or more times, colors span dark blue–blue–red–dark red for low through high responses. First and second clusters of points visited in second search session are highlighted by green and pink circles, respectively. Note axes for **(A)** are from −1 to 1 and for **(B)** are from −0.5 to 0.5. **(C,D)** Stimuli visited three or more times in search session 1 **(C)** and search session 2 **(D)**, sorted in order of decreasing ROI response, averaged across all trials for each image. Stimuli from second search are labeled in white according to location in cluster 1 or 2.

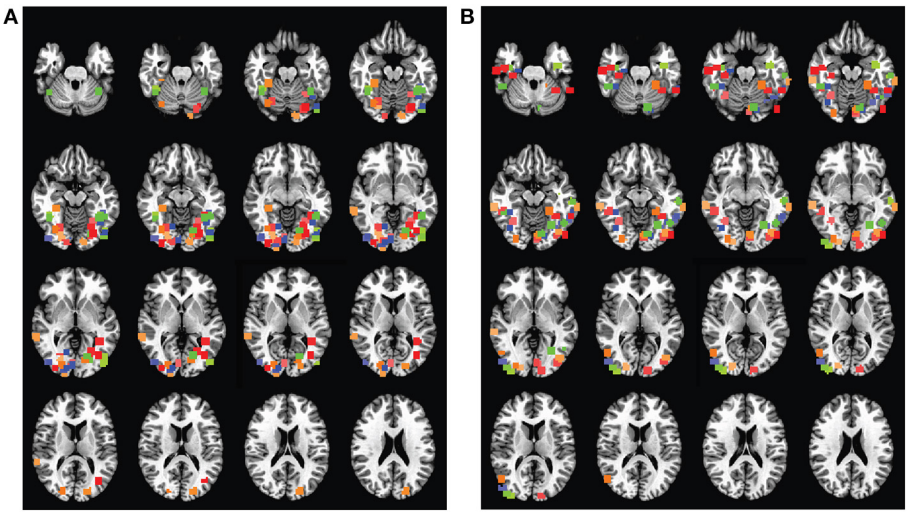


FIGURE 6 | Class-selective regions for 10 subjects in (A) real-world objects search and (B) Fribble objects search, projected onto the Talairach brain. Colors are associated as follows (listed as real-world/Fribble, respectively): blue for mammals/purple curved tube object, green for human-forms/blue-bodied–yellow-winged object, orange for cars/bipedal–metal-tipped-tail object, red for containers/wheelbarrow object, overlaps shown as single color. Each subject is assigned a shade of each of the four colors.

we use the feature space for real world objects and then the feature spaces for Fribble objects as powerful new tools for characterizing and understanding cortical responses to complex visual properties.

3.4.1. Real-world objects search

Examination of points frequently visited by each search and the responses of the corresponding brain regions revealed (1) multiple distinct selectivities within search of single ROIs, (2) marked change in cortical response resulting from slight deviations in visual properties/slight changes in location in visual space, and (3) several intuitive classes of visual properties used by the ventral pathway—including surface texture as well as two- and three-dimensional shape.

We examine the results of the two "most-converged" searches in detail below, and summarize results for all other searches with above-threshold convergence.

The class 2/human-forms search in the second session for subject S3 was one of the most convergent. Projecting the visited stimuli along the first two dimensions in SIFT-based space in **Figure 5A**, and focusing on frequently-visited stimuli, we see two clusters, circled in green and pink. The images visually are split into two groups[2]: one group containing light/generally-narrow-shapes and the second group containing less-light/wide-shapes, as shown in **Figure 5D**. Notably, stimuli evoking high and low responses appear in both clusters, and similar-looking images can elicit opposite ROI activities—e.g., the two red characters. We consider this as potential evidence of local inhibition.

[2]For the interpretation of real-world objects results, grouping was done by visual inspection of a single linkage dendrogram constructed in the four-dimensional SIFT-based space.

The class 2 search in the first session for S3 shows a quite weak convergence measure. Unlike results for the second session, there is no concentration of focus around one (or two) spatial locations. Despite a very low consistency measure, there is evidence for a degree of consistency between session results. The stimuli evoking the strongest and weakest responses in the first session appear in the lower cluster of visited points in the second session. The red wingless character, again, elicits high response while the purple winged character in the first session and the red-green winged character in the second session, nearby in visual SIFT-based space, elicit low responses. The winged character in the first session is projected as a very small blue circle at $(-0.05, 0.02, 0.15, 0.10)$ in the SIFT space in **Figures 5A,B**. By starting from a separate location, the second search finds two ROI response maxima in SIFT space.

The class 2 search in the first session for S6 showed the greatest convergence measure across all searches. Projecting the visited stimuli along the SIFT dimensions in **Figure 7**, we see one cluster (of red and blue circles) around the coordinates $(-0.1, -0.15, 0, -0.15)$ and several outliers for the first session. The three stimuli in the cluster producing the highest responses (**Figure 7C**) may be linked by their wide circular head/halo, while the smallest-response stimulus is notably thin—potentially indicating response intensity as a wide/thin indicator. Notably, stimuli evoking high and low responses, coming from the two ends of the wide/thin spectrum, are nearby one another in the part of the SIFT space under study by the search—a potential example of the limits of four SIFT dimensions to capture magnitudes of all visual differences among real-world objects.

The class 2 search in the second session for S6 shows a quite weak convergence measure. Similarly, as the consistency measure is low, the stimuli frequently visited in the second session fail to overlap with similar feature space locations and

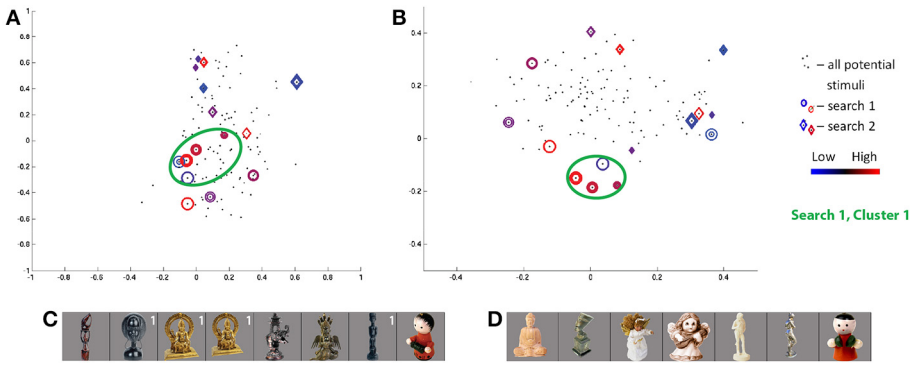


FIGURE 7 | Search results for S6, class 2 (human-forms), shown in (A) first and second SIFT space dimensions and (B) third and fourth dimensions. Colors and shapes used as in **Figure 5**. The stimuli visited three or more times in the first **(C)** and second **(D)** search sessions are shown sorted in order of decreasing ROI response, averaged across all trials for each image. Stimuli from first search are labeled in white if located in cluster 1.

"similar-looking"[3] stimuli frequently visited in the first session. Although a red character produces the minimum responses in each of the two searches (**Figure 7**), the two characters are located in distinct corners of the SIFT space (dark red diamond and blue circle in **Figure 7**).

Comparison of searches for S3 and S6, in **Figures 5**, **7**, respectively, shows a similar pattern of visited stimuli in the feature space. For both subjects, there is a focus close to the first dimensional axis, i.e., a vertical line of red and blue circles and diamonds along the first two dimensions; visited stimuli follow a V pattern in the second two dimensions. Furthermore, some of the highest ROI response stimuli appear (in red) at high locations along the second and third dimensions. Similarly, frequently-visited stimuli for S6 session 1 (dark blue circles) appear close to the the observed lower cluster for S3 session 2, though the cortical responses for the two subjects appear to differ. Comparing **Figures 5**, **7**, we also can confirm a degree of overlap between the images frequently shown for each subject. In both subjects, frequently visited stimuli seemed to show regional selectivity, and potentially differentiation, for narrow-versus-wide shapes.

Study of frequently visited stimuli in search sessions showing lower degrees of convergence reveals a mix of results, summarized in **Table 1**. Most searches identify one potential cluster producing marked high, and possibly marked low, responses from the ROI. A variety of visual properties are identified for different regions under study, from surface details to shapes of object parts. In one of the searches considered in the table, for subject S6, we note stimuli producing high and low cortical responses are close together in visual space.

Looking more broadly for evidence of local inhibition across both convergent and non-convergent searches, we measure the distance in feature space between stimuli producing the highest and lowest ROI responses, and compare it with the typical distribution of inter-stimulus distances in feature space in **Figure 8**.

[3]Similarity in appearance is not well-defined, as explored by our past work in Leeds et al. (2013). Generally, we limit our similarity judgements to identification of identical pictures, e.g., in **Figure 5**. Here, we occasionally use more rough intuition.

Table 1 | Summary of results for additional convergent and consistent searches.

Subject/Class/ Session	Local inhibition	# Cluster centers	Visual properties
S1/2/1	No	1, many outliers	Metallic surfaces, rectangular base (uncertain)
S5/2/2	No	1, many outliers	Sharp local angles defining internal holes or feathers
S7/2/1	No	1	High spatial frequencies on surface, shiny spots
S7/2/2	No	1	(uncertain)
S1/4/1	No	1	(uncertain)
S6/4/1	No	2	Cluster 1: same object in different colors; Cluster 2: multiple long edges (uncertain)
S6/4/2	Yes	1	Top handles, horizontal and vertical lines (uncertain)

"Local inhibition" marks the observation that stimuli close to one another in visual space evoke particularly high and low ROI responses. "(uncertain)" notes uncertainty about visual properties of frequently-visited stimuli clustered in feature space.

Stimuli were deemed to be close in space if their distance was more than a standard deviation below the average inter-stimulus pairwise distance among the stimuli in the class. Out of 80 searches performed, we observed nine in which nearby stimuli produced extremely different cortical responses.

A comparison of class 2 searches for S1, S3, and S6 reveals a similar pattern of stimulus responses in feature space. Qualitatively, the stimuli are arranged roughly linearly along the first two dimensions and show a more complex "V" pattern in the second two dimensions. Some of the highest ROI response stimuli appear at high locations along the second and third dimensions for S1 session 2, S3 session 2, and S6 session 2. Notably, the 4 human figures with largely-uniform white surfaces (**Figure 5D**)

constituting the first cluster for S2 from session 2, were also frequently displayed to S1 in session 2; 3 of the 4 figures are sorted in the same order based on ROI response size.

In contrast, comparison of class 2 searches for S5 with those of the subjects reported above, S1, S3, and S6, shows a great degree of difference in the pattern of frequently visited stimuli in feature space and in the pattern of cortical responses across space. This finding reflects the expected diversity of selectivities employed in perception of a given object class, e.g., human-forms.

3.4.2. Fribble objects search

Among subjects viewing Fribble objects, 20 selectivity searches converged and 7 searches showed consistency across search sessions. As in real-world object searches, examination of stimuli frequently visited by each search and the responses of the corresponding brain regions revealed (1) multiple distinct selectivities within search of single ROIs, (2) marked change in cortical response resulting from slight deviations in visual properties/slight changes in location in visual space, and (3) several perception approaches used by the ventral pathway—including focus on the form of one or multiple component "appendages" for a given Fribble object.

We examine in detail the results of two of the most convergent searches as well as the results of the two most inter-session consistent searches. We also summarize results for all other searches with above-threshold convergence and consistency.

The class 1/curved tube object search in the second session for S11 showed high convergence. Projecting the visited stimuli along the three Fribble-specific morph dimensions in **Figure 9A**, noting the third dimension is indicated by diagonal displacement, we see one cluster[4] (of red and blue diamonds) centered around (0, −0.33, 0.66). The cluster contains three of the four stimuli visited three or more times in the second session, as shown in **Figure 9C**. These stimuli show similar appearances in their three appendages. The outlying stimulus, while deviating in its more circular head and more flat-topped tail tip, retains the round leg shape of the three in-cluster stimuli. We observe Fribble ROIs sometimes are most selective for the shape of a subset of the component appendages, although clustering appears to indicate the head and tail-tip shape remain important for S11's ROI as well, as does cross-session comparison of results below.

The class 1 search in the first session for S11 shows quite weak convergence. Projecting the visited stimuli along the three Fribble-space dimensions (red and blue circles) shows the search spreading across the space. In several locations, pairs of near-adjacent stimuli were visited, as in the lower left, upper right, and center of **Figure 9A**. In each location, the stimuli evoked opposite strength responses from the ROI—the second and seventh highest responses are coupled, as are the first and sixth, and the third and seventh. Sensitivity to slight changes in visual features—potential local inhibition—thus is seen both for Fribble and real-world object perception.

The class 3/bipedal, metal-tipped tail object searches for S17 showed high cross-session consistency. Projecting the visited stimuli along the three Fribble-specific morph dimensions in

[4]For the interpretation of Fribble results, grouping was done by visual inspection of the three-dimensional scatter plots, e.g., **Figure 9A**.

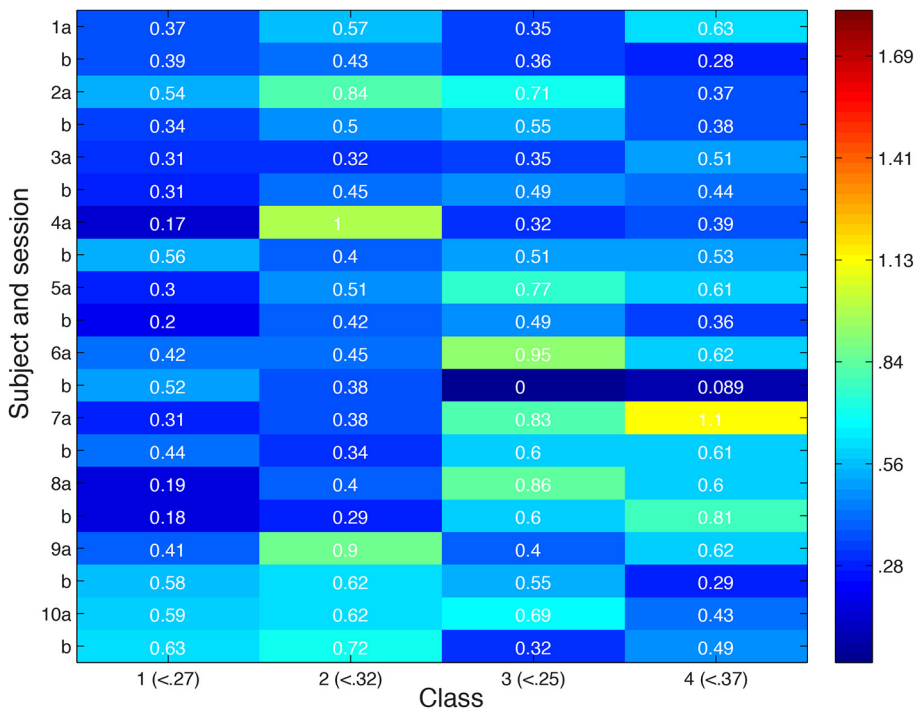


FIGURE 8 | Distance in feature space between stimulus pair evoking greatest difference in cortical responses or second greatest difference in cortical responses. Feature space distance color coded from red (high) to blue (low). For each class, number in parentheses indicates distance one standard deviation below average distance between randomly-selected stimuli. Search session 1 is represented by letter "a" and session 2 by letter "b."

Figure 10A, we see the first session focuses on the axis of dimension 1, the second session focuses on the axis of dimension 2, and both emphasize stimuli with $dim2 \approx 0$. The visited points for each session spread widely, albeit roughly confined to a single axis. Visually, in **Figures 10B,C**, these stimuli are grouped for their spiked feet ($dim2 = 0$), as well as for their tails appearing half-way between a circle and a cog shape (see **Figure 4A**) and their yellow "plumes" half-way between a round, patterned and angled, uniformly-shaded shape. The importance of spike-shaped feet indicated in both searches, even beyond the (0, 0, 0.66) cluster focus, may relate to prominance of edge detection in biological vision, expanding to the detection of sharp angles. As seen for other Fribble and real-world objects searches above, stimuli evoking the lowest and highest responses are notably clustered in the search space.

Visual comparison of searches and of regional responses for different subjects cannot be made across classes, as each Fribble space is defined by a different set of morph operations. Within class comparisons do not reveal strong consistent patterns across ROIs.

The class 4/wheelbarrow object search for S19 showed high convergence in both sessions. Furthermore, the two searches together showed the highest cross-session consistency across all subjects and object classes. Projecting the visited stimuli along the three Fribble-specific morph dimensions in **Figure 11A**, we see clustering along $dim1 = 0$ and $dim3 = -0.33$ for the first session (red and blue circles); dimension 2 location of the stimuli is more broadly-distributed, but limited to $dim2 \leq 0$. The stimuli at the center of the first session cluster, shown in **Figure 11B**, are linked by their purple tongue and green ear shapes. The ROI appears to be selective for the shape of a subset of component appendages, without regard for other elements of the object (i.e., the green nose). As observed throughout our search results, stimuli evoking high and low responses appear in the same cluster, sometimes adjacent to one another in space and appearing rather similar by visual inspection, indicating ROI sensitivity to slight changes in appearance.

Projecting the visited stimuli for the second session along the three Fribble dimensions (as red and blue diamonds) shows two clusters. The presence of multiple selectivity centers is consistent with observed ROI response properties for subjects viewing real-world objects, as well as Fribble subject S11 discussed above. The stimuli at the center of the larger second session cluster show a similar green ear and similar mid-extremes nose but a

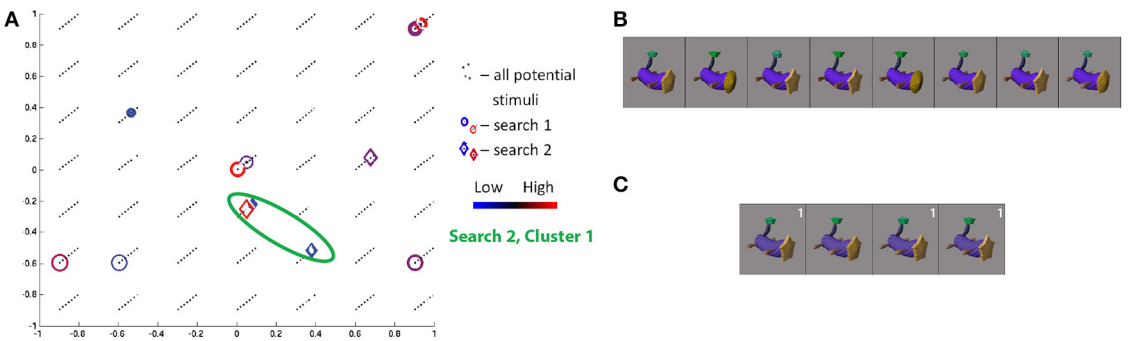


FIGURE 9 | Search results for S11, class 1, shown in three-dimensional Fribble space (A), with third dimension represented as diagonal offset. Positive third dimension results in displacement up and to the right. Location of all potential stimuli in space shown as black dots. Results from real-time scan session 1 are circles, results from real-time scan session 2 are diamonds. For stimuli visited three or more times, colors span dark blue–blue–red–dark red for low through high responses. Cluster of points visited in second search session is highlighted by green circle. The stimuli visited three or more times in the first **(B)** and second **(C)** search sessions are shown sorted in order of decreasing ROI response, averaged across all trials for each image. Stimuli from second search is labeled white if located in cluster 1.

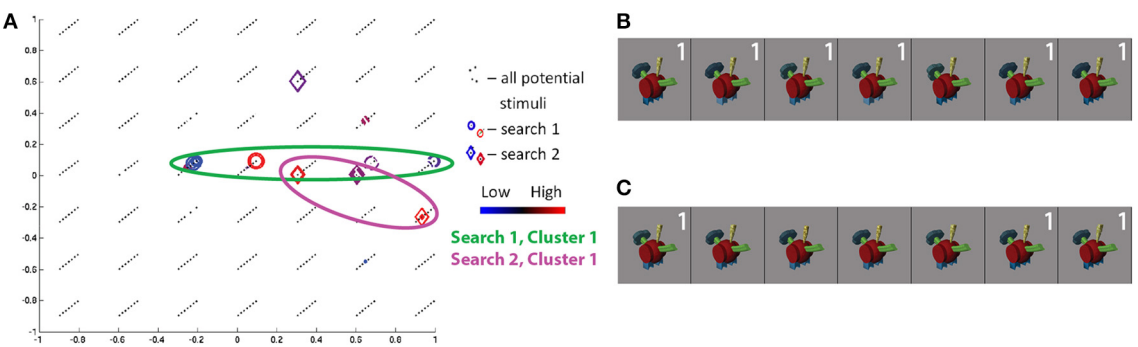


FIGURE 10 | Search results for S17, class 3, shown in three-dimensional Fribble space (A). Colors and shapes used as in **Figure 9**. The stimuli visited three or more times in the first **(B)** and second **(C)** search sessions are shown sorted in order of decreasing ROI response, averaged across all trials for each image. Stimuli from each search are labeled in white if located in cluster 1 for their respective searches.

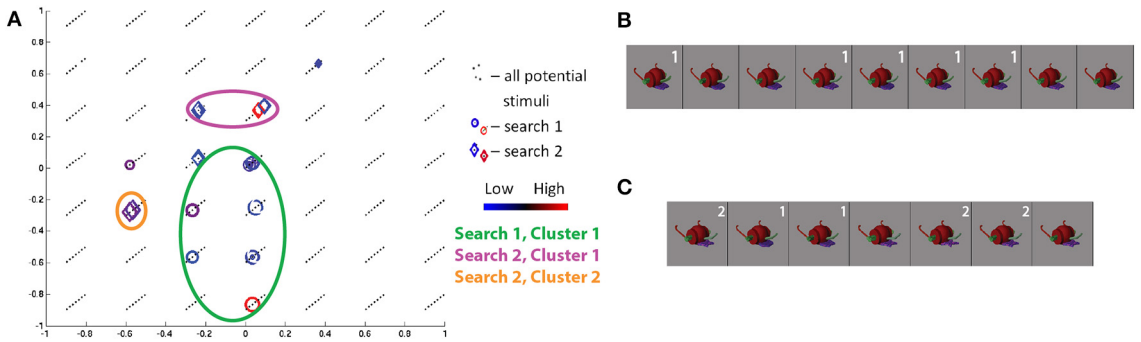


FIGURE 11 | Search results for S19, class 4, shown in three-dimensional Fribble space. Colors and shapes used as in **Figure 9**. The stimuli visited three or more times in the first **(B)** and second **(C)** search sessions are shown sorted in order of decreasing ROI response, averaged across all trials for each image. Stimuli from each search are labeled in white if located in a cluster for their respective search.

more star-shaped purple tongue. The two stimuli with the most-circular tongues form the second cluster. This second cluster has the highest consistency with two of the cluster outliers from the first session, i.e., the second and third most active stimuli for the first session. Stimuli evoking high and low responses appear in the same cluster, sometimes adjacent to one another in space and appearing rather similar by visual inspection.

Study of frequently visited stimuli in search sessions showing lower degrees of convergence reveals a mix of results, summarized in **Table 2**. Most searches identify one potential cluster producing marked high and low responses from the ROI. Most searches also show ROI selectivity for shapes of all three object appendages, each corresponding to a feature space dimension, though several searches indicate selectivity for only two appendages. In almost all the searches considered in the table, we note stimuli producing high and low cortical responses are close together in visual space.

Looking more broadly for evidence of local inhibition across both convergent and non-convergent searches, we measure the distance in feature space between stimuli producing the highest and lowest ROI responses, and compare it with the typical distribution of inter-stimulus distances in feature space in **Figure 12**. Stimuli were deemed to be close in space if their distance was less than 0.87. Notably, the minimum distance between a pair of stimulus points was 0.3. Out of eighty searches performed, we observed over 75% of searches in which nearby stimuli produced extremely different cortical responses. 50% of searches showed extremely different cortical responses for stimuli at most two minimum edit steps away in visual space, stepping between neighboring black dots in the scatter plots shown above.

In sum, searches in most ROIs discussed above cluster around a single location, indicating a single selectivity in visual space specific for all three component appendages in a given Fribble, though several searches find multiple clusters and some results show Fribble location along certain dimensions does not affect ROI response. Locations of clusters, and of high ROI responses, are roughly equally likely to be in the middle of the space (morphing between clear end-point shapes) or close to the extreme ends (showing clear end-point shapes like star heads or sharp-toed feet). For several (but not all) ROIs, stimuli close to one another in visual space evoked high and low cortical responses—indicating sensitivity to slight changes in visual properties.

Table 2 | Summary of results for additional convergent and consistent searches.

Subject/Class/ Session	Local inhibition	# Cluster centers	# Selectivity dims
S13/1/2	Yes	1	3
S16/1/2	Yes	1	3
S17/1/2	Yes	2	3
S16/2/1	Yes	1	3
S17/2/2	No	1	3
S16/3/1	Yes	2	2, 3
S18/3/1	Yes	1	2
S11/4/1	Yes	1	3
S18/4/1	Yes	1	3
S20/4/1	Yes	1	3

"Local inhibition" marks the observation that stimuli close to one another in visual space evoke particularly high and low ROI repsonses. "# selectivity dims" indicates whether clustering occurs in all dimensions (entry is 3) or only along a few dimension (entry can be 1 or 2).

3.5. LIMITATIONS OF USING SIFT MULTI-DIMENSIONAL SCALING SPACE

The use of a SIFT-based Euclidean space yielded relatively poor search performance across subjects and ROIs, despite the abilities of SIFT to capture representations of groups of visual objects in cortical regions associated with "intermediate-level" visual processing, discussed by Leeds et al. (2013). Significant convergence and consistency was observed more rarely than expected—certainly compared to those statistics in Fribble spaces—and visual inspection of frequently-visited stimuli frequently failed to provide intuition about visual properties of importance to the brain region under study.

Confining the SIFT representation to four dimensions, found through multi-dimensional scaling as discussed in Section 2.6.3, limited SIFT space's descriptive power over the broad span of visual properties encompassed by real-world objects. Use of

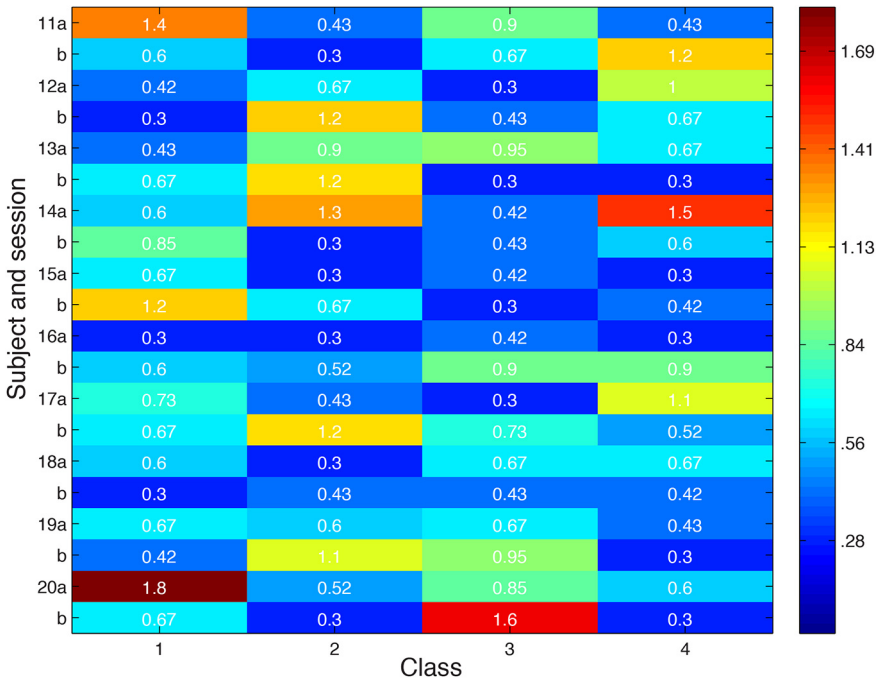


FIGURE 12 | Distance in feature space between stimulus pair evoking greatest difference in cortical responses or second greatest difference in cortical responses. Feature space distance color coded from red to blue. Distance one standard deviation below average distance between randomly-selected stimuli is 0.87. Search session 1 is represented by letter "a" and session 2 by letter "b."

a small number of dimensions was required to enable effective search over a limited number of scan trials. However, **Figure 3C** shows that at least 50 dimensions would be required to explain 50% of the variance in a SIFT-based pairwise distance matrix for 1600 images. Even among the 100 stimuli employed for each object class, the four dimensions used account for less than 50% of variance. The missing dimensions account for grouping pairwise distance patterns across large sets of images—therefore, more-careful selection of stimuli included in a given object class still renders four-dimensional SIFT space insufficiently-descriptive.

Intuitively, it is not surprising that there are more than four axes required to describe the visual world, even in the non-linear pooling space of SIFT. Indeed, the method employed in our present study employs 128 descriptors and 128 visual words (Leeds et al., 2013). Further study shows that tailoring SIFT space for each of the four object classes used in our sessions still requires over 10 dimensions each to account for 50% of variance. The exploration of selectivities for real-world objects using Euclidean space may well require more dimensions, and thus more trials or a more efficient real-time analysis approach. The number of dimensions may be kept small by identification of a better-fitted feature set, or by limitations on the stimuli. We pursue the latter through Fribble spaces, with notable improvement.

For the real-world object searches, our use of multi-dimensional scaling to define SIFT space may also have obscured observations of unifying properties for the stimuli producing high cortical activation. In particular, MDS identifies dimensions maximizing the preservation of pairwise distances between images. Within the first few dimensions, MDS allows groups of objects deemed similar within SIFT to be clustered together—such clustering of visually-similar objects is one of the key assumptions we rely on in our stimulus selection methods. However, this representation of the stimulus images may not capture more subtle variations *within* a cluster of visually-similar objects. For example, within the mammal class, dogs may form a cluster clearly distinct from cows, but this method does not guarantee that two sitting dogs will be closer together within the dog cluster than a sitting dog and a standing dog. Similarly, we would not expect dogs to be sorted according to ear length (or many other intuitive properties) along any vector in SIFT space, even though we would expect all dogs to be spatially far from rhinoceroses. In contrast, Fribble space is defined to better capture such nuanced and graded visual variability, and, perhaps as a consequence, reveals ROIs invariant to changes in some dimensions but selective to changes in others.

Looking forward, we note that there exist a wide range of alternative feature spaces that might be explored in future studies. For example, real-world objects might be rated on a large number of visual properties (or properties could be automatically extracted using unsupervised statistical learning over a large number of images, e.g., Chen et al., 2013), and PCA could then be used to determine a smaller number of dimensions capturing common patterns among these properties—an approach that is somewhat of a compromise between SIFT space and Fribble space. At the same time, acknowledging the limitations of the SIFT-based space, we feel that our experimental findings provide some insight into visual selectivity within selected cortical regions across multiple subjects.

4. DISCUSSION

Our overall goal was to better elucidate the complex visual properties used in visual object perception. In contrast to the field's understanding of early visual processing (e.g., edge detection in primary visual cortex), the intermediate-level visual features encoded in the ventral pathway are poorly understood. To address this gap, we relied on computational models of vision to build low-dimensional feature spaces as a framework in which to characterize neural activity across the high-dimensional world of visual objects. Whereas Hubel and Wiesel explored varying orientations and locations of edges to examine single neuron selectivity in primate V1 (Hubel and Wiesel, 1968), we explored these visual feature spaces to examine neural selectivity across 10 mm^3 brain regions in the human ventral pathway.

Uniquely, we employed real-time fMRI to determine neural selectivity—rapidly identifying those visual properties that evoked maximal responses within brain regions of interest in the context of limited scanning time. These real-time searches across visual feature space(s) provide new understanding of the complex visual properties encoded in mid- and high-level brain regions in the ventral pathway. First, we found that individual brain regions produced high responses for several sets of visual properties, that is, for two or three locations within a given feature space. Second, we found that many brain regions show a suppressed response for stimuli adjacent in feature space—and slightly varied in visual appearance—to those stimuli evoking strong neural responses. This observation indicates a form of a high-level "local inhibition"—a phenomenon often seen for simpler features encoded in earlier visual areas. Finally, a visual inspection of the stimulus images corresponding to the spatial selectivity centers of positive responses offers some intuition regarding what higher-level visual properties—for example, holistic object shape, part shape, and surface texture—are encoded in these specific areas of the brain.

Critically, an examination of the distribution of cortical responses for both visual feature spaces indicates repeating patterns across subjects and ROIs. In particular, for both the SIFT and Fribble spaces, a subset of searches show that stimuli eliciting extreme high or low responses cluster together, while stimuli eliciting responses more in the middle are spread further from cluster centers. This pattern of slightly differing stimuli producing extremely different neural responses is consistent with known visual coding principles within earlier stages of the ventral pathway. At the same time, this observation is not universal—within both the SIFT and Fribble spaces, some searches produce cortical response maxima that are distributed broadly across a given feature space, rather than concentrated in one location.

4.1. PROXIMITY OF DIFFERENTIAL RESPONSES

The proximity of stimuli evoking ROI responses of opposite extremes can be seen in the scatter plots in **Figures 9**, **11**[5]. Similar structure is apparent in the sorted stimulus images illustrated as the red figures in **Figure 5D**[6]. As mentioned earlier, these findings are consistent with the principle of local inhibition often observed within earlier processing stages of the visual system. For example, Hubel and Wiesel (1968) observed spatially adjacent "on" and "off" edge regions in visual stimuli exciting or inhibiting, respectively, the spiking of neurons in mammalian V1. In modern hierarchical models of the primate visual system, the first stage of processing is often held to reflect such early findings as realized as a series of Gabor filters (Serre et al., 2007; Kay et al., 2008). Even earlier within the visual system, prior to cortical coding, retinal ganglion cells are similarly known to have receptive fields characterized by concentric "on" and "off" rings in the image plane of any given stimulus (Rodiek and Stone, 1965). More broadly, multiple stages of alternating patterns of excitation and suppression are consistent with principles of successful neural coding models, in which lateral inhibition of representational units "located" adjacent to or nearby one another are found to be advantageous to a variety of visual tasks (Rolls and Milward, 2000; Jarrett et al., 2009). The sort of local competition observed in our study—that is, in alternative feature spaces—is conceptually plausible based on such models. Our findings indicate that local inhibition does indeed seem to occur in the complex representational spaces employed at more advanced stages of cortical visual object perception.

4.2. VISUAL INTUITIONS ABOUT FEATURE SELECTIVITY

Analysis of cortical activities over visual space provides further understanding of the presence of one or more selectivities for a given brain region and the presence of local inhibition within the defined visual space. However, intuition about the nature of preferred stimuli, and their underlying visual properties, is perhaps better obtained through visual inspection of those stimuli frequently visited by each search and evoking extreme cortical responses. For many real-world objects searches, it was not possible to identify unifying visual patterns of preferred stimuli. For some searches we did observe potentially consistent selected shape and surface properties. In particular, for Fribble object searches, executed in constrained visual spaces, unifying visual patterns for stimuli producing high cortical activity largely were holistic Fribble shapes. At the same time, there were no clear patterns across subjects regarding the preferred types of holistic shapes (which are dependent upon the shapes of the three component appendages of each Fribble class).

For both real-world and Fribble objects searches, visual inspection of the ordering of stimuli by ROI response, that is, as shown in **Figures 5**, **9**, fails to yield any specific insights. *A priori*, we would expect shape properties to smoothly transition as measured responses decreased. That we did not observe this transition may stem from the fact that our measurements reflect a mix of multiple coding units or noise in our fMRI data (despite averaging). For real-world objects, note that the construction of our four-dimensional search space using MDS

[5]Scatter plot examples only are given here in Fribble spaces as they are more easily evaluated visually in one two-dimensional plot.

[6]Stimulus examples only are given here for SIFT searches as similarities of the real-world object stimulus set are easier to see than they are for Fribbles that all look predominantly similar within a given class to the uninitiated reader.

may also limit our ability to detect the fine-grained organization of the stimuli, yet maintain the broader similarity groupings of these same images. At the same time—in light of our finding of evidence for local inhibition—we might alternatively expect that that similar-looking stimuli would appear at opposite ends of the line of sorted stimuli. Interestingly, such a visual disconnect between top-ranked stimuli for single ventral pathway neurons was observed by Woloszyn and Sheinberg (2012).

More broadly, frequently visited stimuli clustered together in SIFT space—evoking both extreme high and low responses, consistent with the observations above—can be united by coarse shape (e.g., width in **Figures 5D**, **7C**), surface properties (e.g., brightness in **Figure 5D** or texture for S1 class 2 in **Table 1**), and fine internal contours (e.g., sharp-edged holes for S5 class 2 in **Table 1**). This observed selectivity for shapes is consistent with the findings of Yamane et al. (2008) and Hung et al. (2012), who successfully identified two- and three-dimensional contour preferences for neurons in V4 and IT using uniform-gray blob stimuli. Unlike these prior studies, our work employs real-world stimuli and thus identifies classes of preferred shapes likely to be encountered in normal life experience. Observed selectivity for surface properties is a more novel finding, though Tanaka et al. (1991) observed such selectivities in primate IT neurons in the context of perception of object drawings. While some searches provided insights about cortically relevant visual properties, many searches performed for real-world objects revealed no clear patterns among stimuli evoking extreme cortical region responses, clustered together in SIFT-based space. This lack of clear patterns likely reflects the difficulty of capturing the diversity of real-world visual properties in a four dimensional space, as discussed in Section 3.5.

We also note that changes in the cortical representation of the stimuli due to repeated exposures across the three study sessions may have made interpretation of our results more difficult across our entire study. However, arguing against this possibility, our observation of stronger search performance for subjects viewing Fribble stimuli—novel images with significant similarity in appearance within each class (thereby increasing the likelihood of overlap in the neural representations of individual stimuli)—suggests that potential adaptation or learning effects did not constitute a significant problem.

4.3. SELECTIVITY TO VISUAL PARTS

Fribble objects, and corresponding "Fribble spaces," were more controlled in their span of visual properties than were the real-world stimuli. Frequently visited stimuli in each Fribble space can cluster around a three-dimensional coordinate. Each dimension corresponds to variations of a single component shape morphed between two options, such as a star/circle head or flat/curved feet, as in **Figure 4A**. Thus, clustering around a point indicates slight variations on three component shapes, with focus around a fixed holistic shape. However, across subjects, there was no clear pattern of preferred holistic Fribble shapes, nor of preferred shapes for any of the three varying component "appendages." For some searches, frequently visited Fribble stimuli evoking strong cortical responses varied along one axis or two axes while staying constant on the remaining one(s). Depending on the brain region being interrogated, one to three component shapes were able to account for such selectivity.

In interpreting this result, we note that it is possible that the appendage-based construction and variations of Fribble stimuli may have biased subjects to rely on perceptual strategies focused on object parts. Nonetheless, observations on cortical responses in these subjects may be supported by evidence for parts-based neural representations observed in subjects viewing less-structured real-world objects—for example, ROIs selective for rectangular statue bases or for bag handles in **Table 1**. One possibility is that more local selectivity for parts of an object, rather than the whole, may be associated with cortical areas particularly earlier in the ventral pathway—an organization that would be consistent with the focus of early and intermediate stages of vision on spatially-distinct parts of a viewed image, pooled together over increasingly broader parts of the image at higher stages of vision (Riesenhuber and Poggio, 1999).

5. CONCLUSIONS

Our study is one of the first to address head on the challenge of identifying intermediate-level feature representation in human ventral cortex. That is, although there is a great deal known about early visual coding and increasing knowledge regarding high-level visual representation (Huth et al., 2012), the field has been relatively silent (with the exception of Tanaka, 1996, Yamane et al., 2008, Hung et al., 2012) on how simple edge-like features are combined to encode more complex features such as parts, textures, and complex shapes. Here we explored this question in two ways. First, by advancing the application of a novel research methodology—real-time methods for rapidly measuring and processing the BOLD signal on a trial-by-trial basis. Second, by introducing a new research approach as applied to human neuroimaging—search methods for efficiently seeking the image or images that are most effective in driving specific brain regions. Although our overall findings are somewhat mixed regarding what we have learned about intermediate-level neural coding, we observed sufficient consistency—in particular, with respect to apparent high-level local inhibition—to suggest that as these methods mature, they offer a promising new direction for exploring the fine-grained neural representation of visual stimuli within the human brain.

ACKNOWLEDGMENTS

This research was funded by an NIH EUREKA Award #1R01MH084195-01, the Temporal Dynamics of Learning Center at UCSD (NSF Science of Learning Center SMA-1041755), and, in part by a grant from the Pennsylvania Department of Health's Commonwealth Universal Research Enhancement Program. Daniel D. Leeds was supported by an NSF IGERT Fellowship through the Center for the Neural Basis of Cognition, an R. K. Mellon Fellowship through Carnegie Mellon University, and the Program in Neural Computation Train Program (NIH Grant T90 DA022762).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fncom.2014.00106/abstract

REFERENCES

Brainard, D. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 443–446. doi: 10.1163/156856897X00357

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., and Poggio, T. (2007). A model of v4 shape selectivity and invariance. *J. Neurophysiol.* 98, 1733–1750. doi: 10.1152/jn.01265.2006

Cardoso, M., Salcedo, R., and de Azevedo, S. (1996). The simplex-simulated annealing approach to continuous non-linear optimization. *Comput. Chem. Eng.* 20, 1065–1080. doi: 10.1016/0098-1354(95)00221-9

Chen, X., Shrivastava, A., and Gupta, A. (2013). "Neil: extracting visual knowledge from web data," in *Computer Vision (ICCV), 2013 IEEE International Conference on* (Sydney), 1409–1416. doi: 10.1109/ICCV.2013.178

Donckels, B. (2012). *Global Optimization Algorithms for Matlab* (Online) (Accessed May 31, 2012). Available online at: http://biomath.ugent.be/∼brecht/downloads.html

Edelman, S., and Shahbazi, R. (2012). Renewing the respect for similarity. *Front. Comput. Neurosci.* 6:45. doi: 10.3389/fncom.2012.00045

Harel, A., Ullman, S., Epstein, B., and Bentin, S. (2007). Mutual information of image fragments predicts categorization in humans: neurophysiological and behavioral evidence. *Vision Res.* 47, 2010–2020. doi: 10.1016/j.visres.2007.04.004

Hemera Technologies, I. (2000–2003). *Hemera Photo Objects*. Vol. i–iii. Gatineau, QC.

Hubel, D., and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hummel, J., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Pschol. Rev.* 99, 480–517. doi: 10.1037/0033-295X.99.3.480

Hung, C., Carlson, E., and Connor, C. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron* 74, 1099–1113. doi: 10.1016/j.neuron.2012.04.029

Huth, A., Nishimoto, S., Vu, A., and Gallant, J. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). "What is the best multi-stage architecture for object recognition," in *IEEE International Conference on Computer Vision*, Vol. 1 (Kyoto).

Just, M., Cherkassky, V., Aryal, S., and Mitchell, T. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE* 5:e8622. doi: 10.1371/journal.pone.0008622

Kay, K., Naselaris, T., Prenger, R., and Gallant, J. (2008). Identifying natural images from human brain activity. *Nature* 452, 352–355. doi: 10.1038/nature06713

Kriegeskorte, N., Murr, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Kullback, S., and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Leeds, D. D. (2013). *Searching for The Visual Components of Object Perception*. Ph.D. thesis, Carnegie Mellon University.

Leeds, D. D., Seibert, D. A., Pyles, J. A., and Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *J. Vis.* 13:25. doi: 10.1167/13.13.25

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

MATLAB. (2012). *Version 8.0.0.783 (R2012b)*. Natick, MA: The MathWorks Inc.

Nowak, E., Jurie, F., and Triggs, B. (2006). "Sampling strategies for bag-of-features image classification," in *Computer Vision - ECCV 2006*, *Volume 3954 of Lecture Notes in Computer Science* (Graz), 490–503.

op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252. doi: 10.1038/nn767

op de Beeck, H. P., Torfs, K., and Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J. Neurosci.* 28, 10111–10123. doi: 10.1523/JNEUROSCI.2511-08.2008

Pelli, D. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442. doi: 10.1163/156856897X00366

Pittman, B. (2011). *Afni Main Page — Afni and Nifti Server for nimh/nih/phs/dhhs/usa/earth* (Online) (Accessed September 20, 2011). Available online at: http://afni.nimh.nih.gov/afni

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rodiek, R., and Stone, J. (1965). Analysis of receptive fields of cat retinal ganglion cells. *J. Neurophysiol.* 28, 833–849.

Rolls, E., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.* 12, 2547–2572. doi: 10.1162/089976600300014845

Seber, G. (1984). *Multivariate Observations*. Hoboken, NJ: John Wiley and Sons, Inc. doi: 10.1002/9780470316641

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6424–6429. doi: 10.1073/pnas.0700622104

Swisher, J., Gatenby, J., Gore, J., Wolfe, B., Moon, C.-H., Kim, S.-G., et al. (2010). Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. *J. Neurosci.* 30, 325–330. doi: 10.1523/JNEUROSCI.4811-09.2010

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Tanaka, K., Saito, H.-A., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.* 66, 170–189.

Tarr, M. (2013). *Novel Object – The Cnbc Wiki* (Online) (Accessed January 15, 2013). Available online at: http://wiki.cnbc.cmu.edu/Novel_Objects

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687.

Wang, H., Heeger, D., and Landy, M. (2012). Responses to second-order texture modulations undergo surround suppression. *Vision Res.* 60, 192–200. doi: 10.1016/j.visres.2012.03.008

Wennerberg, M. (2009). *Version 2.9.5*. Tynningo: Norrkross Software.

Williams, P., and Simons, D. (2000). Detecting changes in novel, complex three-dimensional objects. *Vis. Cogn.* 7, 297–322. doi: 10.1080/135062800394829

Woloszyn, L., and Sheinberg, D. (2012). Effects of long-term visual experience on responses of distinct class of single units in inferior temporal cortex. *Neuron* 74, 193–205. doi: 10.1016/j.neuron.2012.01.032

Yamane, Y., Carlson, E., Bowman, K., Wang, Z., and Connor, C. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360. doi: 10.1038/nn.2202

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 June 2014; accepted: 19 August 2014; published online: 12 September 2014.

Citation: Leeds DD, Pyles JA and Tarr MJ (2014) Exploration of complex visual feature spaces for object perception. Front. Comput. Neurosci. ***8****:106. doi: 10.3389/fncom.2014.00106*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Leeds, Pyles and Tarr. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 28 November 2014
doi: 10.3389/fncom.2014.00147

A neuromorphic system for video object recognition

Deepak Khosla*, Yang Chen and Kyungnam Kim

HRL Laboratories, LLC, Malibu, CA, USA

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Daniel L. Yamins, Massachussetts Institute of Technology, USA
Rangachar Kasturi, University of South Florida, USA

****Correspondence:***
Deepak Khosla, HRL Laboratories, LLC, 3011 Malibu Canyon Road, Malibu, CA 90265-4797, USA
e-mail: dkhosla@hrl.com

Automated video object recognition is a topic of emerging importance in both defense and civilian applications. This work describes an accurate and low-power neuromorphic architecture and system for real-time automated video object recognition. Our system, Neuormorphic Visual Understanding of Scenes (NEOVUS), is inspired by computational neuroscience models of feed-forward object detection and classification pipelines for processing visual data. The NEOVUS architecture is inspired by the ventral (*what*) and dorsal (*where*) streams of the mammalian visual pathway and integrates retinal processing, object detection based on form and motion modeling, and object classification based on convolutional neural networks. The object recognition performance and energy use of the NEOVUS was evaluated by the Defense Advanced Research Projects Agency (DARPA) under the Neovision2 program using three urban area video datasets collected from a mix of stationary and moving platforms. These datasets are challenging and include a large number of objects of different types in cluttered scenes, with varying illumination and occlusion conditions. In a systematic evaluation of five different teams by DARPA on these datasets, the NEOVUS demonstrated the best performance with high object recognition accuracy and the lowest energy consumption. Its energy use was three orders of magnitude lower than two independent state of the art baseline computer vision systems. The dynamic power requirement for the complete system mapped to commercial off-the-shelf (COTS) hardware that includes a 5.6 Megapixel color camera processed by object detection and classification algorithms at 30 frames per second was measured at 21.7 Watts (W), for an effective energy consumption of 5.45 nanoJoules (nJ) per bit of incoming video. These unprecedented results show that the NEOVUS has the potential to revolutionize automated video object recognition toward enabling practical low-power and mobile video processing applications.

Keywords: object detection, object classification, airborne, video image processing, neuromorphic, bio-inspired, low-power, real-time processing

INTRODUCTION

Unmanned platforms are becoming one of the major sources of data for intelligence and surveillance both on and off the battlefield. High resolution and wide field-of-view sensors are resulting in large volume of images and videos that then need to be processed and analyzed. Two problems arise from these emerging trends: (1) High bandwidth is required to send data from the platform to ground stations even with good compression and (2) High workload is imposed on analysts and end-users to process the data. One solution to these problems is to perform on-board automated image and video analysis (e.g., detect, recognize and track objects of interest) to enable better and timely situational awareness, reduce the amount of data to be streamed, and thus reduce the end-user workload. However, rapid and accurate situational awareness is virtually impossible on-board size- and power-constrained mobile platforms using current state of the art video-processing methods. Video data from these platforms typically includes a large number of objects with few pixels on target and occur under changing illumination, occlusion and clutter conditions. Conventional computer vision methods are generally engineered for specific and limited domains, lack robustness and are computationally expensive, making them unsuitable for onboard processing.

This work presents a real-time video object recognition system that is suitable for onboard processing, for example, on unmanned intelligence, surveillance and reconnaissance (ISR) platforms. This work was partially performed under the DARPA Neovision2 program (DARPA, 2011) whose goal was to enable an unattended visual object-recognition capability on unmanned reconnaissance platforms. Our system NEOVUS departs from traditional, domain-specific engineered video-processing capabilities and is instead inspired by neuromorphic algorithms that model visual information processing (Mishkin et al., 1983; Huang and Grossberg, 2010; LeCun et al., 2010). The goal of NEOVUS is to detect and classify objects of interest (e.g., car, truck, person and boat) in videos in real-time, while consuming significantly lower power than state of the art computer vision.

The NEOVUS combines two key capabilities: (1) Object detection that finds potential objects in the image and outlines a bounding box around them. It combines form-based detection using attention approaches to detect entities based on form (e.g., shape, color, intensity) and motion-based detection mediated

by spatial attention. (2) Object classification that then classifies the detected objects based on a feed-forward multi-layered convolutional neural network approach (ConvNet) (LeCun et al., 2010; Farabet et al., 2011). The combination of detection followed by classification provides object recognition results. The NEOVUS is implemented in COTS hardware to achieve real-time performance at low size, weight, and power (SWaP). Several components and capabilities of NEOVUS have been previously described by us Chen et al. (2011), Khosla et al. (2013a), Khosla et al. (2013b), Honda et al. (2013). This paper describes the complete system end-to-end, provides additional details of components and key capabilities, and describes in detail the results of DARPA evaluation on urban datasets.

The rest of the paper is organized as follows. Section Architecture describes the NEOVUS architecture. Section Algorithms describes the underlying algorithms for object detection and classification. Section Hardware Mapping describes mapping of the NEOVUS to COTS hardware. Section Results and Discussion describes object recognition performance and energy consumption results from evaluation of the NEOVUS. Finally, Section Conclusions provides conclusions and discusses some future directions of our work.

ARCHITECTURE

The NEOVUS is a neuromorphic object-recognition architecture and system that is inspired by the ventral (*what*) and dorsal (*where*) streams of the mammalian visual pathway (Mishkin et al., 1983). It is based on and consistent with neuroscience theories and models of mammalian pathways implicated in visual processing (Mishkin et al., 1983; Huang and Grossberg, 2010). The NEOVUS consists of three primary functions: retinal processing, object detection and object classification, and five components (I–V). **Figure 1** shows the architecture that combines retinal processing (I), object detection (II–IV) and object classification (V) for fast and accurate object recognition. **Table 1** below summarizes these functional components (I–V).

Data flow description and interaction between the components have been previously described (Khosla et al., 2013a). Unlike traditional computer vision approaches (e.g., Felzenszwalb et al., 2010; Kembhavi et al., 2011) that use engineered features and raster-scan processing, this architecture first detects objects and then classifies them based on a set of learned features. The object recognition results are then displayed to the end user for operation and decision-making. In an automated processing application, such results can be used to determine what data are to be transmitted to the end-user, fulfilling the requirements of bandwidth and workload reduction as outlined in the Introduction section. For example, one may wish to transmit sample images of only certain classes of objects (e.g., red car) when the system achieves certain confidence level in its classification results.

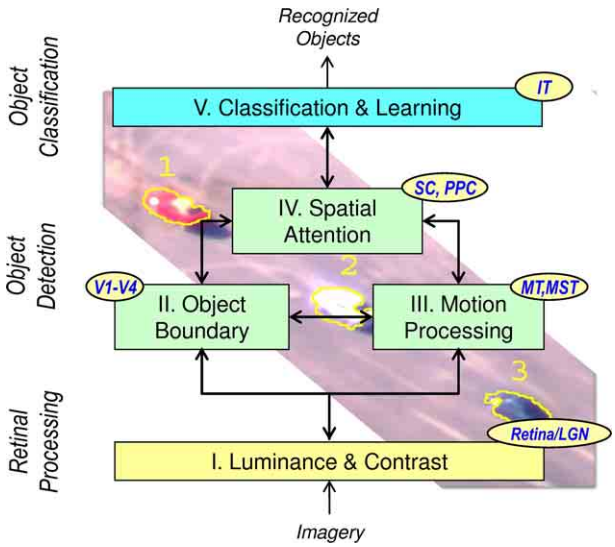


FIGURE 1 | NEOVUS is an integrated, visual cognition architecture that emulates mammalian dorsal and ventral visual functions. Five functional components I–V unify the state of the art in retinal processing, object detection and classification. (*LGN, Lateral Geniculate Nucleus; V1-V4, Visual Cortex areas; MT, Middle Temporal; MST, Medial Superior Temporal; SC, Superior Colliculus; PPC, Posterior Parietal Cortex; IT, Infero-Temporal Cortex*).

ALGORITHMS

OBJECT DETECTION

The purpose of object detection is to locate potential object regions in the video and pass them on to the object classification stage. This module detects motion and form by modeling interacting dorsal and ventral pathways based on the two stream models proposed by neuroscientists (Mishkin et al., 1983; Huang and Grossberg, 2010). It needs to operate with a high probability of detection even at reasonably high probability of false positives so as not to miss potential true objects. The false positives are then discarded during the classification stage as background or irrelevant objects. We now describe details of the object detection algorithm.

Static object detection

Static object detection emulates the attentional pathways that detect objects based on their form saliency with respect to the background. The algorithm is motivated by research on using spectral analysis for visual saliency (Hou and Zhang, 2007). **Figure 2** shows the block diagram of the major steps in our approach to detecting static objects using the RS method (Honda et al., 2013).

Targeted contrast enhancement (TCE). The original RS approach works only on gray-scale images. This works well on bright objects, but not on dark objects. TCE enhances gray-scale images for specified colors (e.g., black) which then enables us to easily detect objects with these colors. This idea can be extended to any arbitrary color for a potential object of interest. TCE is performed by using an un-normalized Gaussian function with mean value of the target color and variance incorporated into a single user-specified sensitivity parameter β. The color response for pixel value $V(c)$ and a select set of M target colors $T = \{T_i(c) : i \in \ldots M\}$ is:

$$R(V, T_i) = \exp\left(-\beta_i \sum_c (V(c) - T_i(c))^2\right),$$

Table 1 | Summary of the five functional components (I–V) in the NEOVUS architecture, their hypothesized brain structures, and functions.

Functional component	Brain structures	Function description
I. Light adaptation (Luminance and Contrast)	Retina-LGN	Light adaptation and contrast enhancement (Werblin and Dowling, 2010). This component was developed under a separate retinal camera hardware effort and is not described in this paper
II. Boundary processing	(P-retinal cells)—(LGN Parvo)-(V1 Interblob)-(V2 Interstripe)-V4; (M-retinal cells)—(LGN Parvo)-(V1 Blob)-(V2 Thin Stripe)-V4	Perceptual boundaries and surfaces (lightness and color) are hypothesized to form; complementary edge and surface processes define boundaries across noise occlusions, fill-in featural properties and aid in figure-ground separation (Mishkin et al., 1983; Elder and Zucker, 1998; Huang and Grossberg, 2010)
III. Motion processing	Toward posterior parietal cortex via V1, MT, and MST with feedback to V4	Computing object direction and speed further aids figure-ground separation and provides ability to detect objects even under occlusion (Sigala et al., 2005)
IV. Spatial attention	SC, LIP, and PPC	Core of the *where* stream; hypothesized formation of attention shrouds and final fused object detections (Mishkin et al., 1983; Huang and Grossberg, 2010)
V. Classification and learning	ITa, ITp, and PFC	Core of the *what* stream; hypothesized learning and recognizing object classes (Mishkin et al., 1983; Tanaka, 2000; Ranganath, 2006)



FIGURE 2 | Workflow of static target detection using form-based attention.

where c indexes the color channel and β is a sensitivity parameter for the similarity between that color channel and the target color. Usually β is set to 1.0, but can be a varying value for different colors depending on the application. The response R is 1 when the image matches the target color and falls off based on deviation from the target color values. Thus, TCE results in multiple gray-scale response maps, one for each target color. **Figure 3** illustrates the effect of TCE for a dark vehicle against dark background.

Anti-aliasing and down-sampling. In order to efficiently deal with large images, we down sample the response maps from previous step. Down-sampling can significantly reduce the computational load for subsequent processing. In addition, humans often fixate their attention to a specific scale when looking for objects. Down-sampling can achieve a similar effect by setting the scale for object detection. It effectively creates a pyramid of scales for each image from the previous step.

Saliency calculation. Each image from the previous step is now processed by the saliency algorithm. The resulting saliency map can be thought of as an approximate representation of human pre-attentive visual response in a bitmap format. We use Spectral Residual Saliency (RS) approach (Hou and Zhang, 2007) due to its speed and efficiency. The RS method exploits the inverse power law of natural images with the observation that the average of log-spectrums is locally smooth. This enables detecting salient objects based on the log-spectrum of individual images rather than *ensemble* of images. The key steps of this algorithm include:

- Convert the image into the Fourier domain;
- Calculate its phase and amplitude;
- Apply a high pass filter to the log amplitude;
- Transform back into the spatial domain using the modified amplitude and original phase.

These processing steps highlight the modes in the frequency domain, the idea being that objects are defined by boundaries constructed from ridges in the Fourier domain. The output is a set of saliency maps that are further processed to produce individual blobs bracketing objects.

Region detection. This step converts each saliency map into detections represented by their attributes (e.g., position, size, orientation, and score). A score is associated with each blob that indicates the confidence score based on its peak saliency value.

Post-processing and fusion. This final step uses various parameters, such as, object size, aspect ratio, and score to filter out false positives. In more general situation where there are multiple saliency maps or saliency maps of multiple scales, detections from these multiple maps are fused together with the results from motion to produce a final set of object detections, which will be discussed in Section Detection Fusion.



FIGURE 3 | An illustration of targeted contrast enhancement for color black ([0, 0, 0]) and $\beta = 55$. The low contrast vehicle in the shade of the tree is enhanced so that it clearly stands out from the shadows.

Moving object detection

Stationary platform. This module is used to detect foreground (moving) objects from a stationary platform, e.g., tower-mounted camera. The implementation in this work adopts our previous motion-based saliency technique that detects motion by modeling the background and comparing it to the input image (Kim et al., 2005). Key steps in our implementation are described below:

Codebook generation. The background model is constructed and updated by adopting our previous work on a codebook-based method (Kim et al., 2005). A codebook consists of one or more codewords that are formed by using samples from each pixel and clustering them based on a color distortion metric and a brightness ratio. A codeword typically contains an RGB-vector and other features such as the minimum and maximum brightness, occurrence frequency, the maximum negative run-length (MNRL) (Kim et al., 2005), and the first and last access times. These values are effectively used to evaluate feature differences, filter out non-background pixels, and recycle unused codewords.

Color model and filtering. The color model separates the color and brightness distortion evaluations to handle the problem of illumination changes, for example, shading and highlights. The color and brightness conditions are satisfied only when both the pure colors and brightness lie within acceptable bounds of the codeword. During training, a fat codebook encodes every incoming pixel in the training set. This fat codebook is then filtered to remove codewords that might contain moving foreground objects. We define MNRL as the maximum interval in time that the codeword has not recurred during the training period (Kim et al., 2005). A codeword with a large MNRL is eliminated from the codebook.

Foreground detection. To detect foreground in an incoming image frame, each sample pixel is matched against the corresponding background model. Unlike other Gaussian or kernel based methods, this step does not require probability calculations. The codebook method is fast as it only calculates the distance between the sample and the nearest cluster's mean in the codebook. The pixel is classified as foreground if no matching codeword is found; otherwise it is a background. This is followed by region extraction similar to that described in Section Static object detection.

Moving platform. When the sensor platform is moving, we leverage motion information in the video for object detection (Chen et al., 2011; Khosla et al., 2013a). Our approach illustrated in **Figure 4** works well even in the presence of jitter and vibration.

Feature extraction and matching. Our approach uses Scale Invariant Feature Transform (SIFT) (Lowe, 1999) features due to its attractive properties of invariance to scale, orientation, and affine distortions. SIFT descriptors represent the gradient orientation histograms and can be used to determine similarity between key points. The matching step matches the key points between a pair of images based on Euclidean distance between their SIFT descriptors and outputs a candidate set of matching points.

Ground-plane homography estimation. Some of the candidate matches from the previous step may be incorrect due to noise and inherent limitations of SIFT. To remove these false matches, we apply RANSAC (Fischler and Bolles, 1981) which is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers. We use RANSAC to find a ground-plane homography transform that best fits the candidate matches. This provides accurate matches and transformation (i.e., homography) between a pair of images.

Frame-to-reference registration. This step warps the frames into a global coordinate reference. Our approach uses a time window size of N frames, with the middle frame being the reference. Each frame is warped to the reference frame in the window based on the homography transformation. This enables all frames in a time window to be stabilized with respect to the center reference frame.

Difference accumulation. This step first computes the difference image between the reference frame and each frame registered to it. The difference image corresponds to independently moving objects against the ground plane. This moving pixel detection process accumulates the differences from several registered frames to increase confidence of detection. For example, with window size $N = 3$, the reference frame F_i, its previous frame F_{i-1} and next frame F_{i+1} are used to compute the frame differences.

Region extraction. The previous step effectively produces a motion saliency map, where higher values indicate more prominent motion due to object motion. We then apply the same region

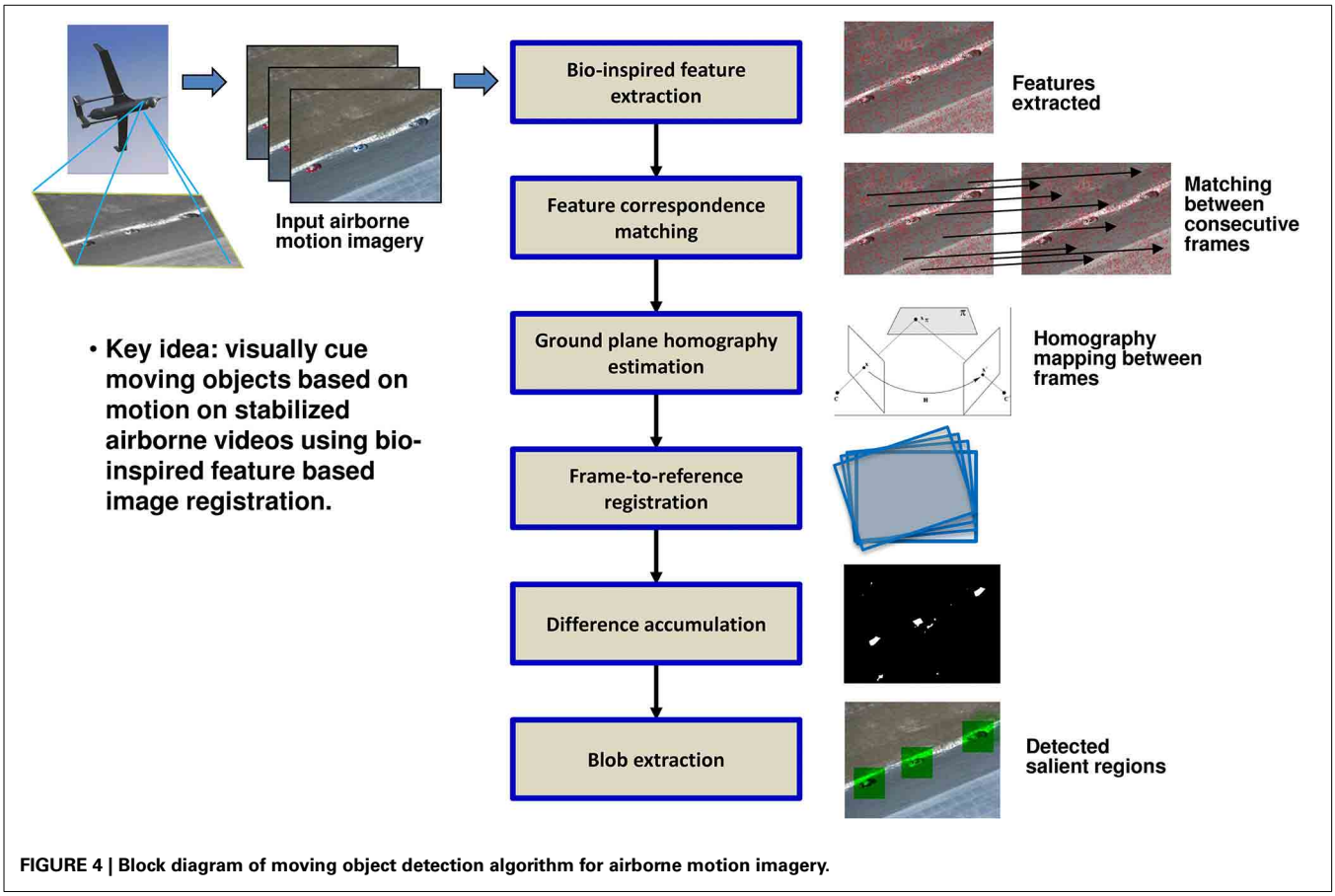


FIGURE 4 | Block diagram of moving object detection algorithm for airborne motion imagery.

extraction method as described in Section Static Object Detection to obtain a set of candidate detections.

Detection fusion

The final step in object detection is to combine the form and motion processing detections into a single detection set prior to object classification. Our approach performs fusion within the static object detection followed by its fusion with motion detection. Detections are all represented simply by the detection boxes in terms of their location, size, and score.

For fusion of detections from different saliency channels (e.g., intensity and TCE) and scales, we use several fusion stages. The first stage called "modal" fusion is motivated by observing that different saliency channels carry different importance in a particular application. For example, intensity channel is usually more important than any single channel from TCE, unless the application has a specific goal of finding objects of specific color (e.g., finding all red cars on the road). To account for this we shift the "mode" of detection scores (distribution) by adding an offset to the scores of detections from intensity channel such that detections from intensity channel have higher scores in general than detections from other channels, e.g., a dark channel. After that, fusion of the detections is accomplished by the unions of the detections from different channels, which are prioritized by their scores and will be further filtered in a resource-limited system before they reached the classification stage. The second fusion stage deals with overlapping detections. Detections in different scale can overlap or objects can split split-up in higher resolutions maps. In our implementation, we keep the enclosing and overlapping detections from lower-resolution. In case of detections from different color channels, we keep the detection with the higher score.

The next step is to merge motion detections with the results from fused detections from static object detection as described above. Here we use a variant of the second fusion scheme described above based on detection overlapping. Our experience is that motion detection is much more reliable than saliency based detection, therefore we keep all detections from motion detection. For the fused detections from static object detection, we only keep them if they do not have significant overlap with any detections from motion detection since such overlap indicates they are the same object. A typical example of the detection process from NEOVUS is shown in **Figure 5**.

OBJECT CLASSIFICATION

Convolutional neural networks (ConvNets) (LeCun et al., 2010) is a supervised deep-learning neural network with multiple layers of similarly structured convolutional feature extraction operations followed by a linear neural network (NN) classifier. ConvNets are an excellent model for image recognition because the structure allows automatic learning of image features as opposed to engineered ones used in traditional computer vision approaches

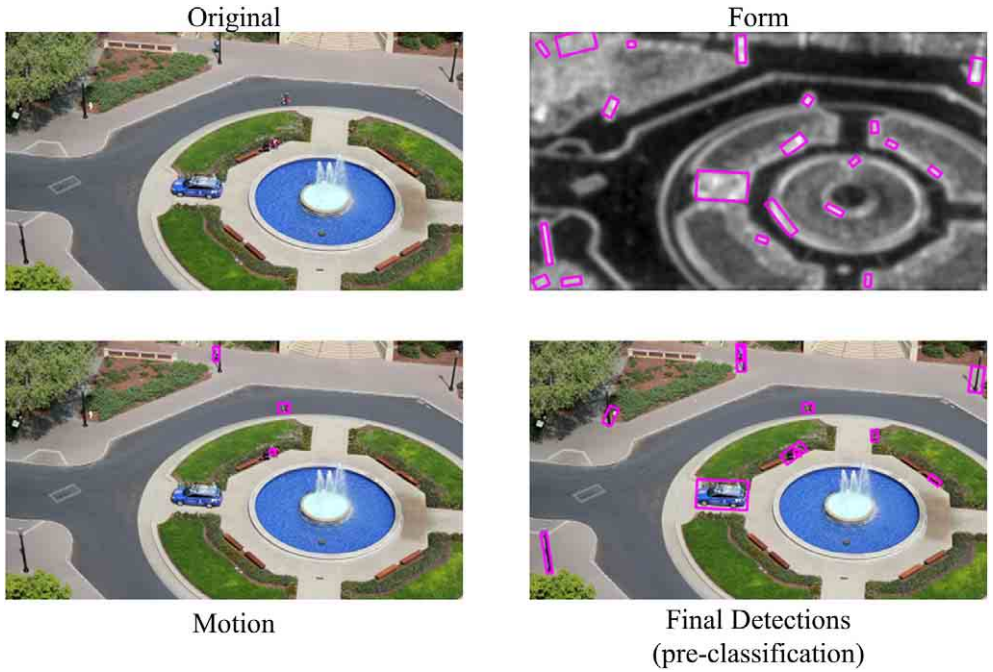


FIGURE 5 | NEOVUS object detection example by fusion of form and motion pathways. Input image **(top left)** is processed by separate and parallel pathways for form **(top right)** and motion **(bottom left)**. The detections for each pathway are fused **(bottom right)** and sent to the classification module.

(Felzenszwalb et al., 2010; Kembhavi et al., 2011). ConvNets typically consist of alternating layers of simple and complex cells as found in mammalian visual cortex. Simple cells perform template matching and complex cells pool these results to achieve invariance. A typical ConvNet has several of 3-layer convolution stages followed by a classifier stage which is a linear NN with one or more hidden layers. Each convolution stage has three layers: (1) a filter bank layer (convolutions) to simulate simple cells (2) non-linearity activation layer, and (3) feature pooling layer to simulate complex cells. The entire network can be trained using backpropagation with stochastic gradient descent (LeCun et al., 2010). Due to its feedforward nature (non-recursive) and uniform computation within each convolution stage, ConvNets are computationally very efficient. It has been implemented in NeuFlow data-flow processor (Farabet et al., 2011) on COTS field programmable gate arrays (FPGAs) to enable low-power, real-time operation. Our prototype hardware evaluation system described in Section Hardware Mapping used the NeuFlow approach.

Our ConvNet implementation follows the traditional architecture outlined above (LeCun et al., 2010) and has three stages. We first resize input RGB color images of candidate target region to 86 by 86 pixels regardless of their original aspect ratio. Then we convert the image to YUV color space, and further process the Y channel by local subtractive and divisive normalization (LeCun et al., 2010). The convolution layer of first stage has eight convolution filter kernels of 7 by 7 pixels, followed by point-wise sigmoid activation function (tanh()) and max-pooling in 4 by 4 pixels neighborhood and subsampling with a stride of 4 pixels, resulting in eight feature maps of 20 by 20 pixels each at the end of stage 1. The remaining stages are detailed in **Figure 6**. The output of the convolution layer at stage 3 is a 128-D vector, which is then fed to the tanh() non-linear activation layer followed by a fully-connected linear NN classifier. We chose the network parameters based on prior experience with similar datasets and validation.

To train our ConvNet, we use training videos provided by DARPA. Manually annotated video clips were used to train and a separate test set was used to quantify their performance (Section Results and Discussion). These videos contain hand-annotated ground truth for objects of interest in each dataset. The annotated image regions and their class labels are extracted and used to train the ConvNet. Depending on data sets and type of objects, we end up with a few hundreds to a few hundreds of thousands samples for the training step.

HARDWARE MAPPING

The NEOVUS hardware prototype was implemented on commercial COTS hardware (**Figure 7**) and can process both recorded and live video. For energy evaluation purposes, we processed live video from a 5.6 Megapixel RGB color camera at 30 frames per second. The camera connects to a frame grabber (via CameraLink) and a laptop computer (PCI-Express). The computer runs the object detection algorithm and sends image region corresponding to each detected object to a COTS FPGA board (Xilinx ML605 Virtex-6) for object classification. The FPGA board runs NeuFlow implementation of a trained ConvNet and sends the results to the laptop for subsequent display. The dynamic power of the complete system that includes the 5.6 Mpixel camera, object detection, and object classification components running at 30 frames per second was measured by

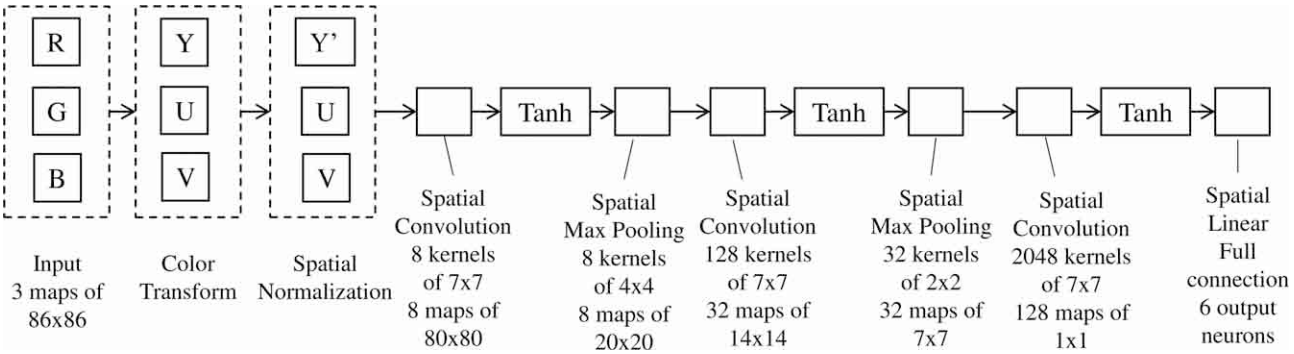


FIGURE 6 | ConvNet structure implemented in the NEOVUS.

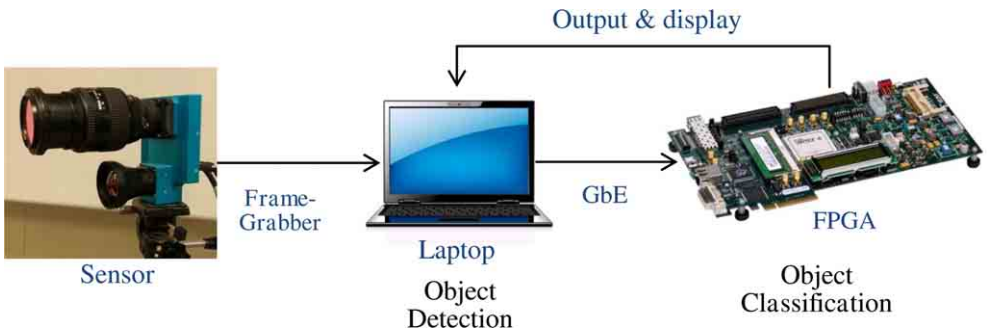


FIGURE 7 | Block Diagram of the NEOVUS hardware system. Video frames are captured from the sensor using a frame grabber card and sent to a laptop computer for buffering, display, and initial processing. Classification algorithms are executed on a COTS FPGA board that implements NeuFlow running a pre-trained ConvNet, which accepts image regions and returns a class label to the laptop computer for each region.

DARPA during a formal evaluation at 21.7 Watts (W). This translates to effective energy consumption of 5.45 nanoJoules (nJ) per bit of incoming video.

RESULTS AND DISCUSSION

In this section, we describe results of NEOVUS evaluation by DARPA on three urban area video datasets (Tower, Helicopter and TAILWIND, **Figure 8**) during summative testing conducted at the end of the Neovision2 program. The Tower dataset is filmed from a fixed camera on top of the Stanford University Hoover tower and the Helicopter dataset is filmed from a helicopter flying over Los Angeles. In both cases, the 1080p video imagery is converted to 8 bit PNG frames for analysis. A third dataset from DARPA TAILWIND (Tactical Aircraft to Increase Long Wave Infrared Nighttime Detection) program is captured from an airplane operating at two different altitudes. The Tower and Helicopter datasets are now available for download (iLab Neovision2 annotated video datasets, 2013a[1]). Each video dataset consists of variable number of object classes (**Figure 9**). The video frames typically have multiple objects per frame that may be occluded or even overlapping in some cases. The videos are processed through NEOVUS and its outputs in the form of object locations, bounding boxes, and class labels is logged for every frame. The NEOVUS logged results are evaluated using ground-truth and evaluation software. The ground-truth is created by human annotators at VideoMining Corporation who labeled 10 object classes (Boat, Car, Container, Cyclist, Helicopter, Person, Plane, Tractor-Trailer, Bus, and Truck) in these datasets. Each object is enclosed in an oriented rectangle that best encompasses the object. Ten percent of the datasets are annotated by two independent annotators and their outputs are compared to assess quality and consistency of annotations; significant differences between the two sets trigger a review of the annotation process to assure annotation quality. The ground-truth was created under DARPA supervision and no performer in the program had any control of that process. The evaluation software uses Normalized Multiple Object Thresholded Detection Accuracy (NMOTDA) score (Kasturi et al., 2009; iLab Neovision2 Performance Evaluation Protocol, 2013b[2]) to evaluate performance per sequence:

$$NMOTDA = 1 - \frac{\sum_{i=1}^{Nframes}\left(c_m m^{(t)} + c_f mf^{(t)}\right)}{\sum_{i=1}^{Nframes}\left(N_G^{(t)}\right)}$$

[1] Available online at: http://ilab.usc.edu/neo2/dataset/

[2] Available online at: http://ilab.usc.edu/neo2/dataset/neovision2-performance-evaluation-protocol.pdf



	Tower	Helicopter	TAILWIND
Resolution	1920 x 1088	1920 x 1080	3660 x 2748
Size (megapixel)	2	2	10
Size (megabits)	50	50	241
Ground sample distance (pixel/m)	30 - 40	25 – 40	4 – 8 High alt. 20 – 40 Low alt.
Frame rate (fps)	30	30	5
Bandwidth (Gb/s)	1.5	1.5	1.0

FIGURE 8 | Three video datasets (Tower, Helicopter, and TAILWIND) were used for the summative tests.

where $m^{(t)}$ = number of missed detections, $f^{(t)}$ = number of false positives, and $N_G^{(t)}$ = number of ground-truth objects in the tth frame. The summations are carried out over all evaluated frames. In Neovision2 evaluations, the cost functions c_m and c_f for missed detects and false positives, respectively were set to identity. This is a sequence-based measure which penalizes false detections, missed detections and object fragmentation. Note that maximizing NMOTDA for the sequence is the same as finding the optimal assignment of ground truth boxes to system output boxes at each frame image.

Figure 10 gives an example of NEOVUS results superimposed on a frame of Tower dataset. By aggregating the results, one can plot the ROC curves as P_d (probability of correct recognition) vs. FPPI (False Positives Per Image). **Figure 11** shows the NEOVUS results for four object classes on the Tower dataset. Overall, NEOVUS achieved excellent object recognition at about 90% per-frame accuracy and 0.1 FPPI.

In each data domain, multiple sequences are used for evaluation. The summary of the performance over the entire dataset, i.e., over all the video clips, is calculated using Weighted NMOTDA (WNMOTDA). WNMOTDA is calculated for each class separately and is given by,

$$WNMOTDA_i = \frac{\sum_{j=1}^{NVC} NMOTDA_{ij} * NGT_{ij}}{\sum_{j=1}^{NVC} NGT_{ij}}$$

where $WNMOTDA_i$ is the NMOTDA for class i calculated over all the video clips, $NMOTDA_{ij}$ is the NMOTDA measure calculated for class i in video clip j, NVC is the total number of video clips, and NGT_{ij} is the number of ground truth objects of class i in video clip j. Finally an average WNMOTDA score is generated for all object classes for each domain by ignoring the class labels. Before scoring, identical boxes are merged into one. Overlapped boxes (if Overlap_Ratio is over 20%) are merged into one and their union is used instead. All of these above calculations are part of the evaluation software.

The recognition performance and energy consumption results of summative testing were formally released by DARPA and are presented in **Figure 12** (DARPA, 2011). Five teams participated in the program and evaluation. Three neuromorphic teams (C, D, E) developed neuromorphic vision algorithms. Two baseline algorithms, based on the Deformable Part Model algorithm (Felzenszwalb et al., 2010), representing typical computer vision methods were implemented to serve as a benchmark for comparing against the neuromorphic algorithms. HRL's NEOVUS (Team C) was the best performer with high recognition accuracy (WNOMTDA) and the lowest energy use (nJ/bit) amongst all five teams. Particularly noteworthy is that the NEOVUS energy use was at least three orders of magnitude lower than the computer vision systems (Teams A and B, **Figure 12**). These unprecedented results show that neuromorphic

DARPA Object Classes

Object Class	Tower	Helicopter	TAILWIND
Car	✓	✓	✓
Truck	✓	✓	✓
Tractor-Trailer		✓	✓
Bus	✓	✓	✓
Container		✓	✓
Boat		✓	✓
Plane		✓	✓
Helicopter		✓	
Person	✓	✓	
Cyclist	✓	✓	



FIGURE 9 | A maximum of $N = 10$ object classes were to be recognized in the summative tests.

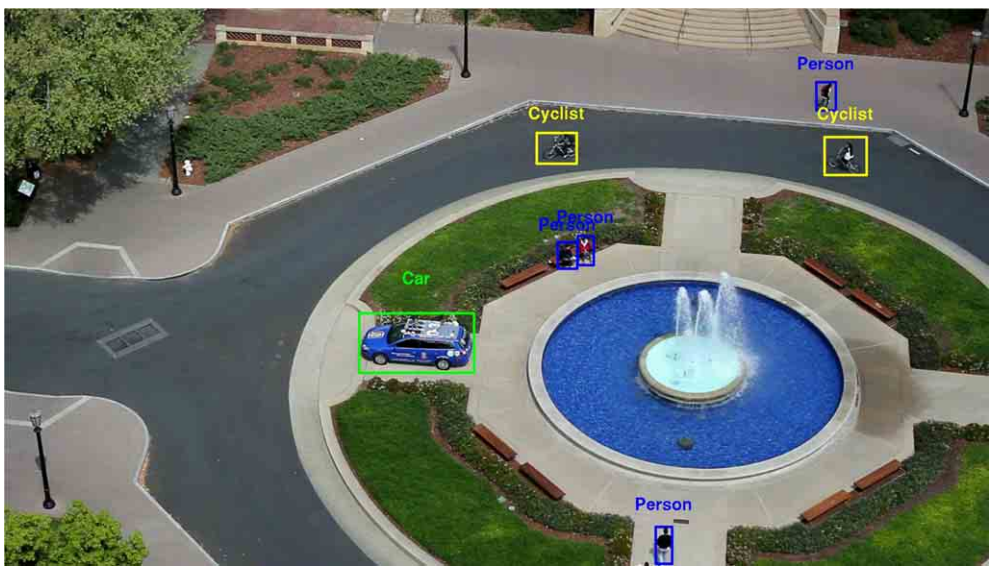


FIGURE 10 | Sample NEOVUS object classification results on one frame of Tower video dataset (object set: car, cyclist, person, bus, truck).

algorithms and hardware have the potential to revolutionize low-power situational awareness applications, e.g., on-board small unmanned aerial vehicles (UAV).

Our current low-power NEOVUS implementation is mature enough to be deployed on mobile platforms toward onboard processing. **Figure 13** supports this claim by analyzing platform payload available SWaP and extrapolating our measured energy use (5.45 nanoJoules per bit) to other camera resolutions. For example, NEOVUS could run onboard small UAV's, such as the AeroVironment Raven (Raven RQ-11, 2014[3]), in real-time for a 1-Megapixel camera. Larger UAVs, such as the Boeing-Insitu ScanEagle

[3]Available online at: http://www.avinc.com/uas/adc/raven/.

(Boeing Insitu ScanEagle, 2014[4]), can hold up to a 6-Megapixel camera and still support NEOVUS processing in real-time.

[4]Available online at: http://www.insitu.com/systems/scaneagle

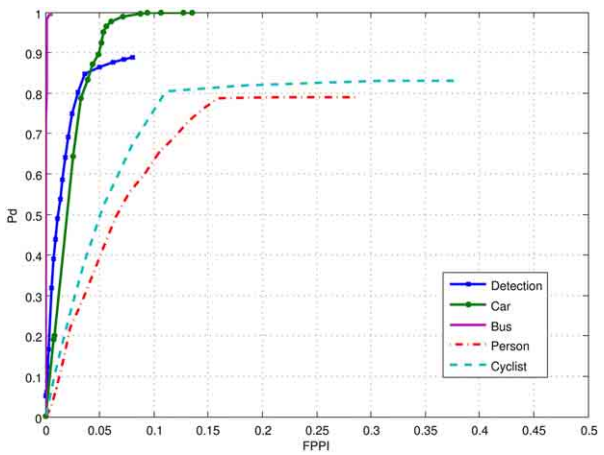


FIGURE 11 | Performance of the NEOVUS shown as ROCs in terms of P_d (probability of correct recognition) vs. False Positives Per Image (FPPI). The blue curve "Detection" corresponds to the overall detection performance of all objects regardless of class.

CONCLUSIONS

This work described a neuromorphic object recognition system inspired by neuroscience findings of object recognition pathways in the mammalian visual system. From a practical perspective, the NEOVUS is a proven collection of neuromorphic algorithms, software, and architecture for automated and accurate video object recognition at significantly lower power than state of the art computer vision. It processes video based on human-like search and classification models that are significantly different from computer vision brute-force raster-scan approaches. The NEOVUS was successfully evaluated on real-world urban video datasets and proven to accurately recognize objects at low-power. The successful hardware design and mapping of NEOVUS to COTS hardware can help enable potential autonomous and mobile applications. This onboard processing can reduce both the requirements for data bandwidth and analyst man power. While the NEOVUS hardware was geared for autonomous on-board processing, it is just as applicable to off-board processing of live or recorded images and videos. For off-board processing, the highly efficient algorithms used in NEOVUS can enable video analysis in faster than real-time even with COTS computer hardware.

The NEOVUS described in this work is a feed-forward, bottom-up object recognition architecture. However, models and algorithms for top-down attention and processing can be added to the current architecture with little modifications. For example, goal-directed search and classification, e.g., find and track all white trucks) can be added to our framework. Future work

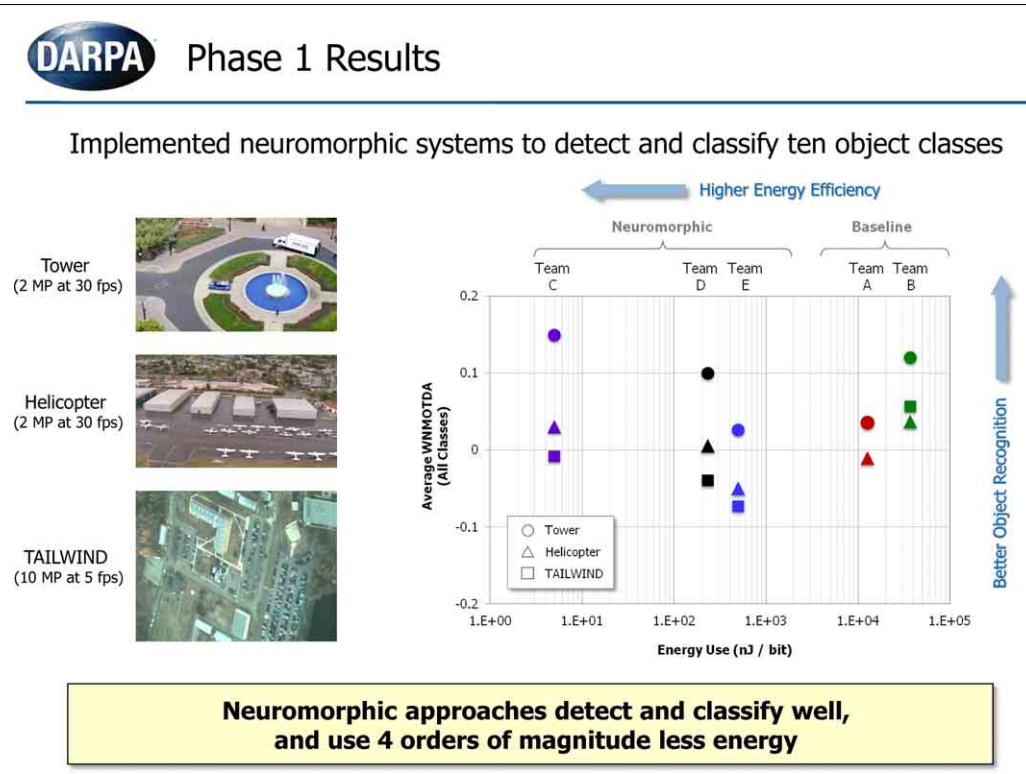


FIGURE 12 | Results of summative testing released by DARPA. Y-axis is object recognition performance expressed as WNMOTDA (higher is better). X-axis is energy use expressed as nJ/bit (lower is better). Teams A,B are baseline computer vision teams and teams C–E are neuromorphic teams. HRL's NEOVUS (Team C) was the best performer with high recognition accuracy and at least three orders of magnitude lower energy consumption than state of the art computer vision systems (A,B). (DARPA, 2011)

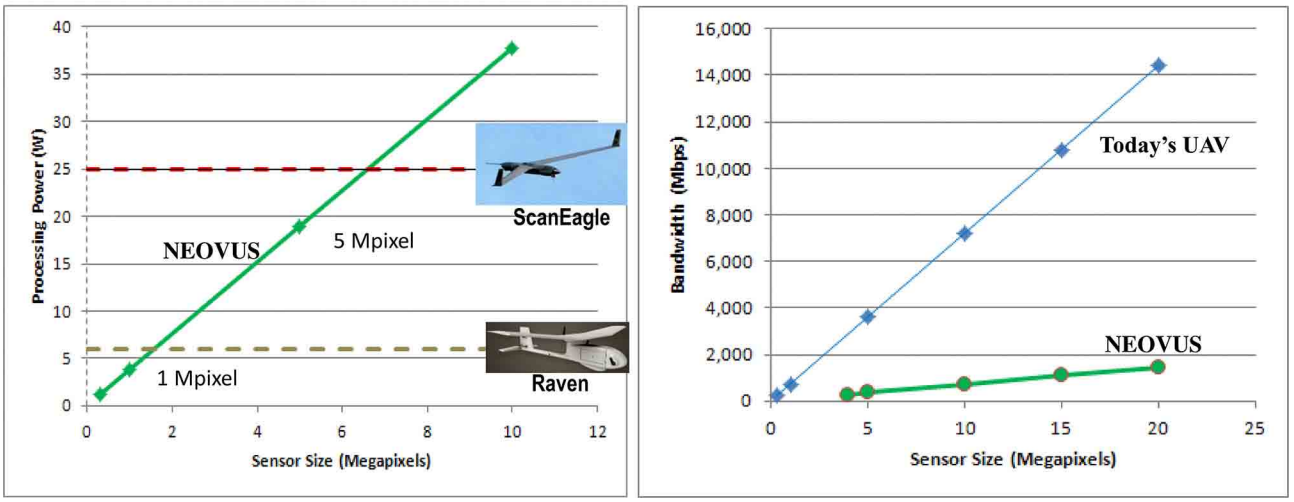


FIGURE 13 | Projected power consumption and bandwidth of the NEOVUS hardware system. The power consumption of the NEOVUS hardware system is low enough to even fit onto small UAVs. The bandwidth requirement of the system is far lower than that of a standard UAV surveillance system.

will include these top-down aspects and onboard evaluation of this capability. This is expected to yield the greatest level of improvement toward enabling practical systems.

ACKNOWLEDGMENTS

This work was partially supported by the Defense Advanced Research Projects Agency Neovision2 program (contract HR0011-10-C-0033). The views and conclusions contained in this document are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The authors would like to acknowledge Drs. Yann LeCun and Clement Farabet at the New York University (NYU) for their support on the classification algorithms and NeuFlow hardware mapping.

REFERENCES

Chen, Y., Khosla, D., Huber, D., Kim, K., and Cheng, S. Y. (2011). A neuromorphic approach to object detection and recognition in airborne videos with stabilization. *Adv. Visual Comput. Lect. Notes Comput. Sci.* 6939, 126–135. doi: 10.1007/978-3-642-24031-7_13

DARPA. (2011). *Neovision2 Evaluation Results. DISTAR Case Number* 18620, December, 2011.

Elder, J. H., and Zucker, S. W. (1998). Evidence for boundary-specific grouping. *Vis. Res.* 38, 143–152. doi: 10.1016/S0042-6989(97)00138-7

Farabet, C., Martini, B., Corda, B., Akselrod, P., Culurciello, E., and LeCun, Y. (2011). "NeuFlow: a runtime reconfigurable dataflow processor for vision," in *Proc. of the Fifth IEEE Workshop on Embedded Computer Vision (ECV'11 @ CVPR'11)*, IEEE, Colorado Springs, 2011.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Anal. Mach. Intell. IEEE Trans.* 32, 1627–1645. doi: 10.1109/TPAMI.2009.167

Fischler, M. A., and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM* 24, 381–395. doi: 10.1145/358669.358692

Honda, A. L., Chen, Y., and Khosla, D. (2013). Robust static and moving object detection via multi-scale attentional mechanisms. *Proc. SPIE* 8744, 87440S-1. doi: 10.1117/12.2016017

Hou, X., and Zhang, L. (2007). "Saliency detection: a spectral residual approach," in *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Minneapolis, MN).

Huang, T.-R., and Grossberg, S. (2010). Cortical dynamics of contextually cued attentive visual learning and search: spatial and object evidence accumulation. *Psychol. Rev.* 117, 1080–1112. doi: 10.1037/a0020664 doi: 10.1037/a0020664

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., et al. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. *IEEE Trans. Pattern Anal. Mach. Intell.* 31. 319–336. doi: 10.1109/TPAMI.2008.57

Kembhavi, A., Harwood, D., and Davis, L. S. (2011). Vehicle detection using partial least squares. *Pattern Anal. Mach. Intell. IEEE Trans.* 33, 1250–1265. doi: 10.1109/TPAMI.2010.182

Khosla, D., Chen, Y., Huber, D., and Van Buer, D. J. (2013b). Real-time low-power neuromorphic hardware for autonomous object recognition. *Proc. SPIE* 8713, 87131301. doi: 10.1117/12.2015944

Khosla, D., Chen, Y., Kim, K., Cheng, S. Y., Honda, A. L., and Zhang, L. (2013a). A neuromorphic system for object detection and classification. *Proc. SPIE* 8745, 87450X-1. doi: 10.1117/12.2016038

Kim, K., Chalidabhongse, T. H., Harwood, D., and Davis, L. (2005). Real-time Foreground-Background Segmentation using Codebook Model. *Real-time Imaging* 11, 167–256. doi: 10.1016/j.rti.2004.12.004

LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). "Convolutional networks and applications in vision," in *International Symposium on Circuits and Systems (ISCAS'10)*, IEEE, Paris.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proc. Int. Conf. Comput. Vis.* 2, 1150–1157. doi: 10.1109/ICCV.1999.790410

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* 6, 414–417. doi: 10.1016/0166-2236(83)90190-X

Ranganath, C. (2006). Working memory for visual objects: complimentary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience*139, 277–289. doi: 10.1016/j.neuroscience.2005.06.092

Sigala, R., Serre, T., Poggio, T., and Giese, M. (2005). "Learning features of intermediate complexity for the recognition of biological motion," in *Artificial Neural Networks: Biological Inspirations*, eds W. Duch, J. Kacprzyk, E. Oja, and S. Zadrożny (Berlin; Heidelberg: Springer), 241–246.

Tanaka, K. (2000). Mechanisms for visual object recognition studied in monkeys. *Spat. Vis.* 12, 147–163. doi: 10.1163/156856800741171

Werblin, F. S., and Dowling, J. (2010). "Evolution of retinal circuitry: from then to now," in *Handbook of Brain Microcircuits*, 1st Edn., eds G. Shepherd and S. Grillner (New York, NY: Oxford University Press), 200–207.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 April 2014; accepted: 29 October 2014; published online: 28 November 2014.

Citation: Khosla D, Chen Y and Kim K (2014) A neuromorphic system for video object recognition. Front. Comput. Neurosci. ***8****:147. doi: 10.3389/fncom.2014.00147*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Khosla, Chen and Kim. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Computational Neuroscience

REVIEW
published: 22 April 2015
doi: 10.3389/fncom.2015.00043

Shape representations in the primate dorsal visual stream

Tom Theys[1,2], *Maria C. Romero*[1], *Johannes van Loon*[2] *and Peter Janssen*[1]*

[1] *Laboratorium voor Neuro- en Psychofysiologie, Katholieke Universiteit Leuven, Leuven, Belgium,* [2] *Afdeling Experimentele Neurochirurgie en Neuroanatomie, Katholieke Universiteit Leuven, Leuven, Belgium*

The primate visual system extracts object shape information for object recognition in the ventral visual stream. Recent research has demonstrated that object shape is also processed in the dorsal visual stream, which is specialized for spatial vision and the planning of actions. A number of studies have investigated the coding of 2D shape in the anterior intraparietal area (AIP), one of the end-stage areas of the dorsal stream which has been implicated in the extraction of affordances for the purpose of grasping. These findings challenge the current understanding of area AIP as a critical stage in the dorsal stream for the extraction of object affordances. The representation of three-dimensional (3D) shape has been studied in two interconnected areas known to be critical for object grasping: area AIP and area F5a in the ventral premotor cortex (PMv), to which AIP projects. In both areas neurons respond selectively to 3D shape defined by binocular disparity, but the latency of the neural selectivity is approximately 10 ms longer in F5a compared to AIP, consistent with its higher position in the hierarchy of cortical areas. Furthermore, F5a neurons were more sensitive to small amplitudes of 3D curvature and could detect subtle differences in 3D structure more reliably than AIP neurons. In both areas, 3D-shape selective neurons were co-localized with neurons showing motor-related activity during object grasping in the dark, indicating a close convergence of visual and motor information on the same clusters of neurons.

Keywords: object, shape, visual cortex, macaque, depth, parietal cortex, dorsal stream

OPEN ACCESS

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Mazyar Fallah, York University, Canada
Charles Edward Connor, Johns Hopkins University School of Medicine, USA

***Correspondence:**
Peter Janssen, Laboratorium voor Neuro- en Psychofysiologie, Katholieke Universiteit Leuven, Herestraat 49, Bus 1021, B-3000 Leuven, Belgium
peter.janssen@med.kuleuven.be

Received: *28 March 2014*
Accepted: *20 March 2015*
Published: *22 April 2015*

Citation:
Theys T, Romero MC, van Loon J and Janssen P (2015) Shape representations in the primate dorsal visual stream. Front. Comput. Neurosci. 9:43. doi: 10.3389/fncom.2015.00043

Introduction

Visual object analysis in natural conditions is computationally demanding but critical for survival, hence the primate brain devotes considerable computing power to solve this problem. Lesion studies in monkeys (Ungerleider and Mishkin, 1982) and patients (Goodale et al., 1991) have demonstrated that the visual system beyond primary visual cortex consists of two subdivisions, a ventral stream directed toward the temporal cortex for object recognition and categorization, and a dorsal stream directed to the parietal cortex for spatial vision and the planning of actions (**Figure 1A**). Since primates not only recognize and categorize objects, but also grasp and manipulate those objects, it comes as no surprise that objects are processed in both the ventral and the dorsal visual stream.

The first recording experiments in the ventral stream, which is critical for object recognition, were published more than four decades ago (Gross et al., 1969), and the accumulated knowledge about the properties of individual neurons has spurred the development of a large number of computational models on object and shape analysis for object recognition (Riesenhuber and Poggio, 1999; Poggio and Ullman, 2013). However, neurophysiological evidence for the visual

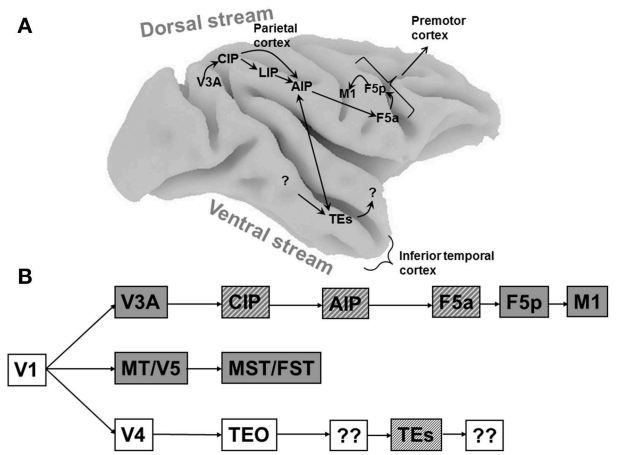


FIGURE 1 | Cortical areas processing object shape. (A) Overview of the macaque brain illustrating the locations of the areas involved in processing object shape, and the most important connections between these areas. Unidirectional arrows indicate the presumed flow of visual information along the dorsal stream, the bidirectional arrow between AIP and TEs indicates that the direction of information flow is unclear at present. Note that most connections in extrastriate cortex are bidirectional. CIP, caudal intraparietal area; LIP, lateral intraparietal area; AIP, anterior intraparietal area; F5a, anterior subsector of area F5; TEs, subsector of area TE in the anterior Superior Temporal Sulcus. **(B)** Schematic flow chart of visual 3D information. Dark boxes indicate areas of the dorsal visual stream, open boxes indicate ventral stream areas, hatched boxes indicate areas selective for higher-order disparity. Boxes with question marks indicate unknown areas.

analysis of objects in the dorsal stream has only recently emerged, and biologically-plausible models for the dorsal stream are scarce (Fagg and Arbib, 1998; Molina-Vilaplana et al., 2007). Robots have to interact with objects in an unpredictable environment, and artificial vision systems that operate based on principles used by the primate dorsal stream areas could undoubtedly advance the fields of computer vision and robotics (Kruger et al., 2013). To stimulate the interaction between neurophysiology and computational modeling, it is important to review recent progress in our understanding of the neural representation of object shape in the primate dorsal visual stream.

Objects contain both two-dimensional (2D: e.g., contour, color, texture) and three-dimensional (3D: e.g., orientation in depth and depth structure) information. Originating in primary visual cortex, at least three different pathways are sensitive to depth information (**Figure 1B**). The MT/V5–MST/FST pathway is primarily involved in the visual analysis of moving stimuli and ego-motion, and FST neurons are selective for three-dimensional shape defined by structure-from-motion (Mysore et al., 2010). The ventral pathway V4–TEO–TEs builds a very detailed representation of the depth structure of objects, and finally the V3A–CIP–AIP–F5a pathway analyses object shape for grasping and manipulation. These pathways should not be regarded as entirely separate entities, since numerous interactions between them exist at different levels in the hierarchy. Rather, each pathway has its own specialization and can function independently of the other pathways.

In this review we will focus on the properties of individual neurons in the parietal and frontal cortex, the hierarchy of cortical areas that links early visual areas to the motor system, and the relation between neuronal firing and behavior. Neurons in other parietal areas, such as area 5 in the medial bank of the IPS (Gardner et al., 2007) and area V6A in the medial parieto-occipital cortex (Fattori et al., 2012) also respond selectively to objects of different sizes and orientations. However, the role of these neurons in computing object shape to guide the preshaping of the hand during grasping is less clear at present. We will first discuss the coding of two-dimensional (2D) shape in areas LIP and AIP, the network of areas involved in processing three-dimensional (3D) shape investigated with fMRI, and finally the single-cell properties of neurons involved in 3D shape coding in the dorsal stream.

Two-Dimensional Shape Selectivity in the Dorsal Visual Stream

The first report of shape selectivity in the dorsal stream was a study by Sereno et al. (Sereno and Maunsell, 1998) in the lateral intraparietal area (LIP), an area in posterior parietal cortex (**Figure 1**) traditionally associated with eye movement planning and visual attention (Colby and Goldberg, 1999; Andersen and Buneo, 2002). In this study, many LIP neurons showed clear selectivity for simple two-dimensional (2D) shapes appearing in the receptive field (RF) in the absence of any eye movements. However, size and position invariance—two properties that are believed to be essential for genuine shape selectivity—were only tested in a small number of neurons. A more recent study (Janssen et al., 2008) confirmed the presence of shape-selective responses in LIP. However, a more systematic test of size and position invariance revealed that LIP neurons rarely exhibit these properties. In many cases shape-selective responses arose because of accidental interactions between the shape and the RF, such as a partial overlap. The RF structure of these LIP neurons was frequently inhomogeneous with multiple local maxima, and could even depend on the stimulus and the task: for example the RF tested with small shapes could be different from the RF tested with saccades. The lack of tolerance to changes in stimulus position in LIP neurons represented the first evidence that the shape representation in the dorsal stream is fundamentally distinct from the shape representation in the ventral visual stream, which is characterized by shape selectivity *and* tolerance of the shape preference to changes in stimulus position.

Just anterior to LIP lies area AIP (**Figure 1**), an area known to be critical for object grasping (Gallese et al., 1994; Murata et al., 2000; Baumann et al., 2009). Romero et al. (2012), recorded in area AIP using 2D images of familiar (e.g., fruits) and unfamiliar (tools) objects. Almost all AIP neurons showed significant selectivity to these images of objects, but subsequent testing with silhouettes and outline stimuli revealed that this selectivity was primarily based on the contours of the images. A follow-up study (Romero et al., 2013), demonstrated that for most AIP neurons, the presence of binocular disparity in these images was not necessary, and that a population of AIP neurons represents primarily

relatively simple stimulus features present in images of objects, such as aspect ratio and orientation.

The observation of neural selectivity for shape contours in AIP does not allow determining which shape features are being extracted by AIP neurons. For example, is the entire contour necessary or are parts of the shape contour (possibly corresponding to grasping affordances) sufficient to evoke AIP responses? Romero et al. (2014) used a systematic stimulus reduction approach, in which outline stimuli were fragmented into 4, 8, or 16 parts, the latter measuring merely 1–1.5°. Following previous studies in the ventral visual stream (Tanaka, 1996), the authors determined the minimal effective shape feature as the smallest fragment to which the neural response was at least 70% of the response to the intact outline. The example AIP neuron illustrated in **Figure 2** responded strongly to the outline of a key but not to the outline of a monkey hand. However, some of the smallest fragments in the test still elicited robust responses in this neuron. Hence although AIP is thought to be involved in extracting grasping affordances, the AIP responses were mainly driven by very simple shape features, but not by object parts that can be grasped (**Figure 2**). Similar to previous observations in neighboring area LIP, the fragment selectivity depended strongly on the spatial position of the stimulus, since even small position shifts (2.5°) evoked radically different responses and therefore a very different shape selectivity. Basic orientation selectivity or differences in eye movements could not explain the fragment responses. These results suggest that AIP neurons may not extract grasp affordances. Future studies should determine how the 2D-shape representation changes in ventral premotor areas.

A Network of Cortical Areas Sensitive to the Depth Structure of Objects

The depth structure of objects (i.e., flat, convex, or concave) can be specified by a large number of depth cues such as motion parallax, texture gradients, and shading. Many studies investigating the neural basis of 3D object vision have used random dot

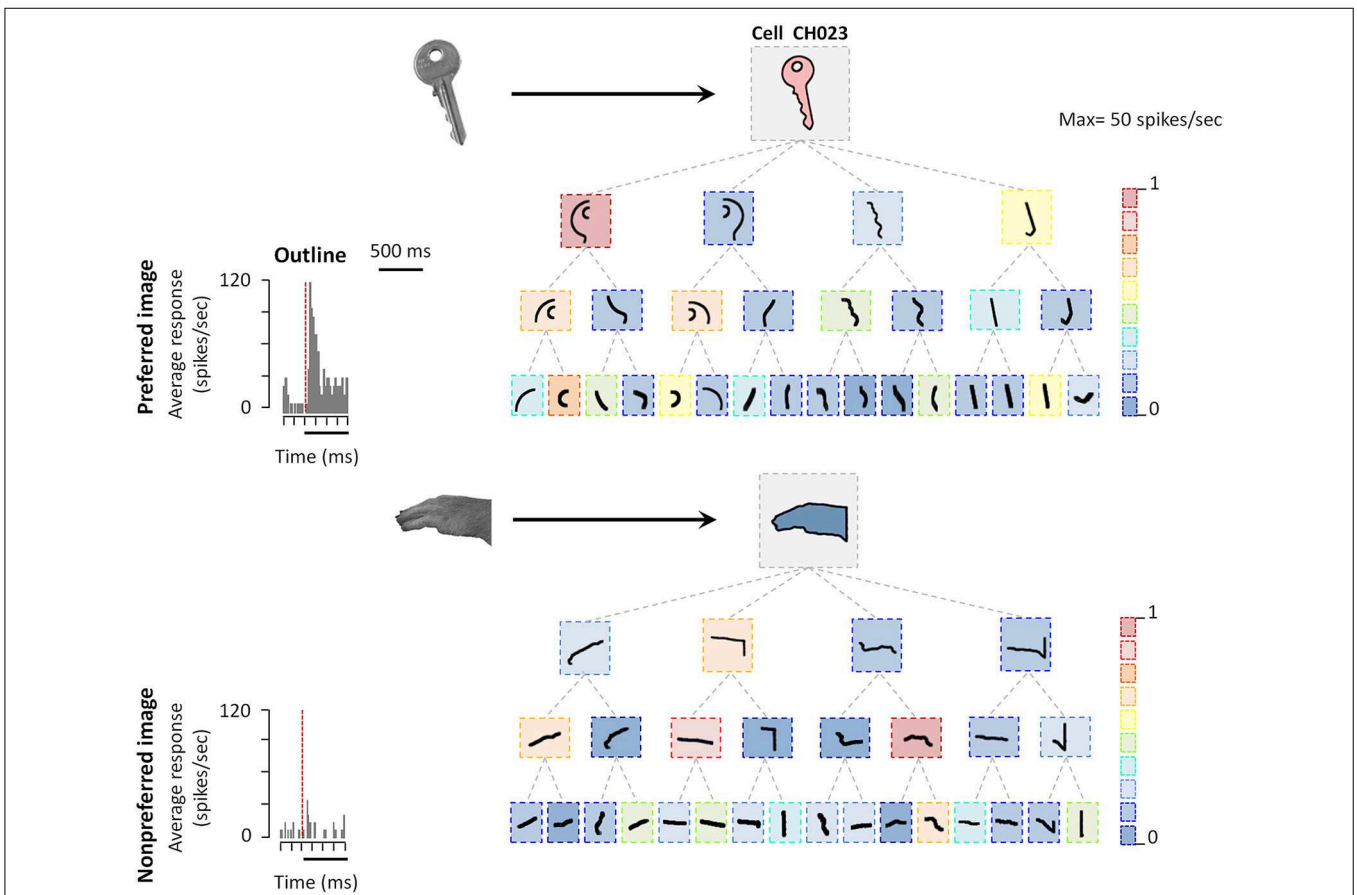


FIGURE 2 | Coding of shape features in AIP. Example AIP neuron responding to an image of an object and to fragments of image contours. In each box the stimulus is illustrated, the color of each box represents the normalized firing rate of the neuron to that stimulus (maximum response was 29 spikes/s). Top row: intact object contour, second row: the four fragments derived from subdividing the object contour into four fragments along the main axes of the shape (four-fragment stimuli). Third and fourth row: eight- and 16-fragment stimuli. Each contour fragment is connected to the stimulus from which it was derived. The original object images from which the contours were derived are illustrated on the left side (arrows pointing to their respective contour stimuli). All stimuli were presented at the center of the RF. Reproduced with permission from (Romero et al., 2014).

stereograms (**Figure 3**), in which depth information is exclusively defined by the gradients of binocular disparity, for obvious reasons. First of all, binocular disparity is the most powerful depth cue (Howard and Rogers, 1995): even when present in isolation, disparity evokes a very vivid percept of depth that is—in contrast to motion parallax or shading—unambiguous with respect to the sign of depth (near or far, convex or concave). Moreover, physiologists are particularly keen on this type of stimuli because one can easily determine whether a neuron (or even a cortical area in the case of fMRI) is responding to the depth from disparity in the stimulus and not to other stimulus features: simply presenting the stimulus to one eye only removes all depth information in the stimulus while preserving shape and texture (Janssen et al., 1999; Durand et al., 2007). In contrast, for other depth cues such as texture gradients, determining which aspect of the stimulus the neuron responds to requires numerous control stimuli. Finally, stereograms also allow precise behavioral control: monkeys and humans can be trained to discriminate depth in stereograms, and varying the percentage of correlation between the dots in the images presented to the left and the right eye (disparity coherence) furnishes a parametric manipulation of the strength of the depth stimulus which can then be related to behaviorial performance.

A series of functional imaging and single-cell studies in monkeys (Janssen et al., 2000a; Durand et al., 2007; Joly et al., 2009) as well as imaging experiments in humans (Georgieva et al., 2009) have suggested that many cortical areas located in both the dorsal and the ventral visual stream are sensitive to the depth structure of objects (**Figure 1A**). This network contains AIP which, as mentioned earlier, is known to be critical for grasping (Gallese et al., 1994; Murata et al., 2000), a region (TEs) in the inferior temporal cortex (ITC), the end-stage of the ventral visual stream and critical for object recognition, and a subsector of the ventral premotor cortex (area F5a, Joly et al., 2009). The observation that not only ventral stream but also dorsal stream areas are processing 3D object information came as no surprise, since many years before these imaging studies, investigations in a patient with a ventral stream lesion had already indicated that object grasping can be intact while object recognition is severely impaired (Goodale et al., 1991; Murata et al., 2000). The object analysis required for object grasping was presumably performed by her intact dorsal stream areas (Murata et al., 2000; James et al., 2003).

Single-Cell Studies in the Dorsal Visual Stream on the Visual Analysis of 3D Structure

fMRI can identify regions that are activated more by curved surfaces than by flat surfaces, but a detailed understanding of the neuronal selectivity in these areas requires invasive electrophysiological recordings of single neurons. Early in the hierarchy of the dorsal visual stream, the Caudal Intraparietal area (CIP) has been studied using inclined planar surfaces in which depth was defined by binocular disparity and/or texture gradients (Tsutsui et al., 2002). CIP neurons can signal the 3D-orientation

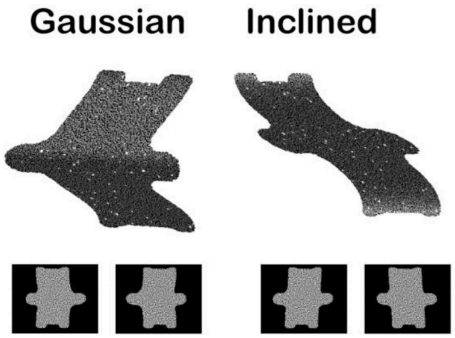


FIGURE 3 | Example random dot stereograms. The monocular images are illustrated below a 3D rendering of two depth stimuli (Gaussian depth profile and Inclined depth profile).

(the tilt) of large planar surfaces when either disparity or texture gradients are used as a depth cue (i.e., cue invariance). A more recent report suggests that CIP neurons can also be selective for disparity-defined concave and convex surfaces (Katsuyama et al., 2010). Rosenberg et al. (2013) showed that individual CIP neurons jointly encode the tilt and slant of large planar surfaces. In view of the anatomical connections of CIP, which run along the lateral bank of the IPS toward area AIP (Nakamura et al., 2001), and more recent preliminary monkey fMRI findings (Van Dromme and Janssen, unpublished observations), CIP could be an important—but not the only (Borra et al., 2008)—input area for AIP. However, since reversible inactivation of area CIP does not cause a grasping deficit (Tsutsui et al., 2001) but sometimes a perceptual deficit in the discrimination of tilt and slant, the role of area CIP in computing 3D object shape for object grasping remains largely unknown.

Previous studies in area AIP had reported object-selective responses in this area (Murata et al., 2000) but it was unclear whether these neurons encoded differences in 3D structure, 2D contour, orientation or any other feature that differed between the objects used in those experiments. Srivastava et al. (2009) recorded single-cell activity in the AIP of awake fixating rhesus monkeys using disparity-defined curved surfaces. A large proportion of AIP neurons responded selectively to concave and convex surfaces that had identical contours, as illustrated by the example neuron in **Figure 4**. This neuron fired vigorously when a convex surface was presented, but not at all when the surface was concave, a selectivity which could not be accounted for by the responses to the monocular presentations. Since this neuron preserved its selectivity across positions in depth (data not shown), the neuron must have responded to a change in binocular disparity along the surface of the stimulus, i.e., higher-order disparity selectivity. The same study observed that the neuronal properties in AIP were markedly different from the ones in TEs: AIP neurons fired much faster to the presentation of curved surfaces (a population latency of 60–70 ms in AIP compared to 90–100 ms in TEs), but appeared less sensitive to small differences in 3D structure, including the sign of curvature: while TEs neurons frequently showed similar responses to curved surfaces with different degrees of curvedness (provided they had the same sign, i.e.,

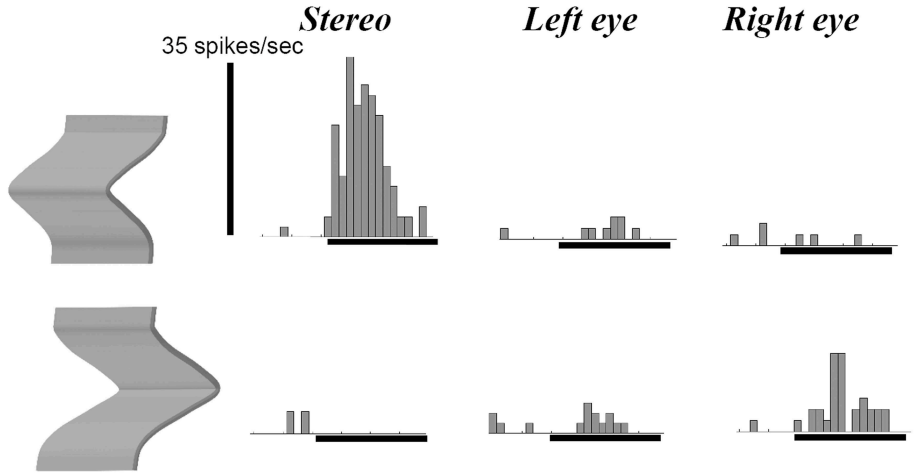


FIGURE 4 | Example neuron recorded in area AIP responding selectively to the depth structure of surfaces. The top row shows the peristimulus-time histograms (PSTHs) of the responses to a convex depth profile, the bottom row the responses to a concave depth profile. Stereo: binocular presentation with disparity. Left eye: monocular presentation to the left eye. Right eye: monocular presentation to the right eye. The horizontal line below the PSTHs indicates the duration of stimulus presentation (800 ms). This neuron preserved its selectivity across positions in depth (data not shown).

either convex or concave) and large response differences when the sign of curvature changed (even for very slightly curved surfaces), the response of AIP neurons declined monotonically as the degree of curvedness was decreased. These observations were the first demonstration that a distinct representation of the depth structure of objects exists in the dorsal visual stream. A follow-up study (Theys et al., 2012b), showed that the large majority of the AIP neurons are primarily sensitive to the disparity gradients on the boundary of the stimulus but largely ignore the depth structure information on the surface, which represents another difference with TEs neurons.

The ventral premotor cortex (PMv) represents the main target of area AIP (Borra et al., 2008). Reversible inactivation of PMv produces a grasping deficit that is highly similar to the one seen after AIP inactivation (Gallese et al., 1994; Fogassi et al., 2001). However, recent studies have provided functional and anatomical evidence suggesting that PMv is not a homogeneous area. A monkey fMRI study (Joly et al., 2009) showed that a subsector of PMv, area F5a, is more activated by curved surfaces than by flat surfaces at different positions in depth, similar to AIP. This activation was located in the depth of the more anterior part of the inferior ramus of the arcuate sulcus. Recent anatomical studies (Belmalih et al., 2009; Gerbella et al., 2011) described differences in the cytoarchitectonics and anatomical connectivity between area F5a and the surrounding subsectors of PMv: F5a does not project directly to primary motor cortex, but does so through its connections with F5p. Moreover, F5a is more strongly connected to the parietal (AIP) and prefrontal (areas 46 and 12) cortex. Based on the anatomical connectivity of F5a, Gerbella et al. (2011) coined the term "pre-premotor cortex" for area F5a, indicating that this subsector of PMv could represent a stage upstream from the more widely studied F5p and F5c sectors of PMv.

Guided by a monkey fMRI study (Joly et al., 2009), Theys et al. (2012a) targeted the F5a subsector with microelectrode recordings to investigate in detail to what extent this region differs functionally from the other subsectors of PMv. Accurately predicted by fMRI, neurons selective for disparity-defined curved surfaces were located in F5a but not in surrounding regions of PMv. The example neuron in **Figure 5** responded to a convex depth profile but not to a concave depth profile irrespective of the position in depth of the stimulus, and monocular responses could not account for the selectivity (data not shown). Remarkably in view of its anatomical location in the premotor cortex, the responses of these F5a neurons appeared very "visual," with robust increases in firing rate when visual stimuli (that the monkey could not grasp) appeared on a display, and with relatively short response latencies (70–80 ms) compared to 50–60 ms for AIP, which is consistent with the higher position in the cortical hierarchy of F5a compared to AIP. These strong visual responses to images presented on a display were unexpected, since previous studies (Graziano et al., 1997; Graziano and Gross, 1998) did not observe responses to images of objects presented on a display in PMv neurons with bimodal visuotactile responses. i.e., responses to tactile stimulation of the face or hand and to visual presentation of objects near the face or hand.

These observations raised the question whether F5a could still be considered part of PMv. To that end, (Theys et al., 2012a)—after having established higher-order disparity selectivity in a cluster of F5a neurons—recorded from the same neurons while the animal was grasping objects in the light (i.e., visually-guided grasping) and in the dark (i.e., memory-guided grasping). Surprisingly, almost all F5a neurons selective for disparity-defined depth structure were also active in the light when the monkey was grasping objects that did not resemble the random-dot

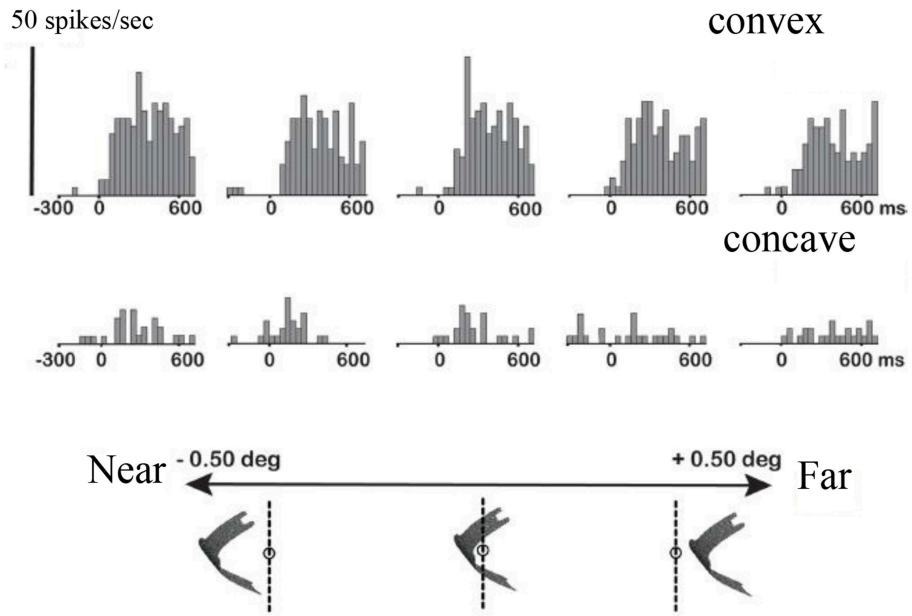


FIGURE 5 | Higher-order disparity selective F5a neuron. The PSTHs in the top row illustrate the responses of a single F5a neuron to a convex depth profile at five different positions in depth, the bottom row the responses to a concave depth profile at the same positions in depth. Below the PSTHs is a schematic illustration of the stimulus presentation indicating the positions in depth. Reproduced with permission from (Theys et al., 2012a).

stereograms. However, most of these F5a neurons were virtually silent when the animal was grasping the same objects in the dark. Hence, in contrast to earlier reports stating that all PMv neurons remain active during grasping in the dark (Raos et al., 2006), F5a neurons selective for the depth structure of objects are mostly visual-dominant. These results do not imply that all F5a neurons are visual-dominant, since multi-unit recordings of clusters of 3D-structure selective neurons revealed strong grasping-related activity in the dark. The implication of these findings is that (visual-dominant) 3D-structure selective F5a neurons are co-localized and strongly connected with visuomotor and motor-dominant neurons that remain active in the absence of visual information. Such functional clusters may form neural modules in which visual object information is mapped onto motor commands. The presence of activity during grasping in the dark implies that F5a is effectively a subsector of PMv, which differs from the other subsectors of PMv by the presence of visual-dominant responses during grasping and selectivity to the depth structure of objects.

The previous studies suggest that the hierarchy of dorsal stream cortical areas involved in 3D-shape processing is largely serial. Although most likely an oversimplification (e.g., the role of feedback connections is unknown and ignored here), the CIP–AIP–F5a serial chain of areas provides a unique opportunity to investigate how the 3D-shape representation changes along the dorsal pathway so that the underlying computations might be revealed. In a first attempt to address this question, Theys et al. (2013) recorded in F5a and AIP in the same animals during visual presentation of 3D surfaces, various approximations of these surfaces and during object grasping. The sensitivity for depth structure was measured by plotting the average responses of a population of neurons to curved surfaces with varying degrees of the disparity variation (from very convex, over almost flat to very concave surfaces). Although interindividual differences between the two animals were present, the sensitivity functions were virtually identical in F5a and AIP. Furthermore, testing F5a neurons with planar (i.e., least-square) and discrete approximations of the smoothly curved surfaces showed again strong similarities between AIP and F5a (**Figure 6**): the majority of neurons in AIP and F5a was also selective for discrete approximations of the convex and concave surfaces consisting of three separate planes in depth, but in F5a, the linear approximation evoked significantly less responses compared to the smoothly curved surfaces. Finally, AIP neurons encoding depth structure from disparity were also tested during object grasping, and similar to F5a, most of these AIP neurons were also strongly active during grasping. The only difference between the AIP and the F5a multi-unit activity consisted of stronger and faster responses in AIP during object fixation and higher activity in F5a during the hand movement epoch before object lift, suggesting that AIP neurons are mainly active during the visual analysis of the object whereas F5a neurons remain active throughout the trial. Overall, the representation of depth structure in premotor area F5a was highly similar to that in parietal areas AIP, which makes it difficult to identify the computations that take place between these two different stages in the dorsal stream shape hierarchy. Future studies may be able to document the differences in the object representation between F5a and AIP using different stimuli or tasks.

The strong correspondence between depth structure selectivity and grasping responses in AIP and F5a is consistent with

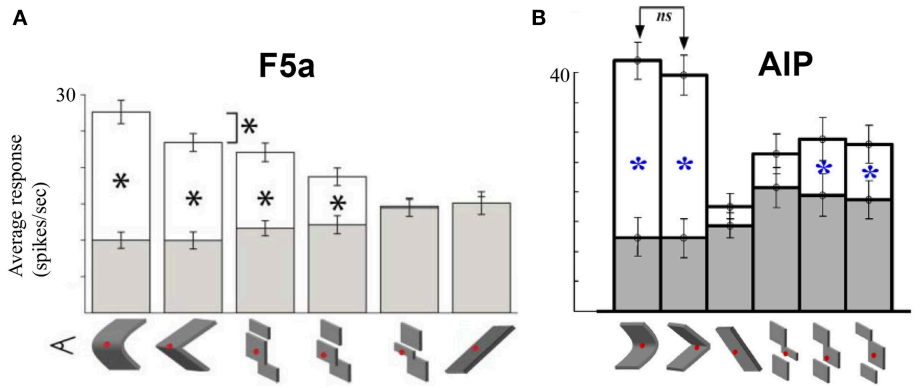


FIGURE 6 | Comparison between AIP and F5a. (A) Average responses of a population of F5a neurons to various approximations of the smoothly curved surfaces (on the left). White bars indicate the responses to the preferred depth profile, filled bars the responses to the non-preferred depth profile. Below the bar graph are schematic illustrations of the various approximations: linear, three different discrete approximations, and first-order (tilted plane). The asterisks indicate significant differences in response. **(B)** Average responses of a population of AIP neurons to the same stimuli. Same conventions as in **(A)**. Adapted with permission from (Srivastava et al., 2009).

the hypothesis that the pre-shaping of the hand during visually-guided object grasping relies on 3D-object information in these two areas. The question still remains why the primate brain would maintain two highly similar areas that communicate by means of (metabolically expensive) long-range connections. At this point the data suggest that the neural object representation in AIP does not have to become much more elaborate to guide the hand during grasping. Furthermore, F5a can directly interact with other premotor areas such as F5p, which AIP can also access directly or indirectly through either parietal area PFG or F5a. The interpretation of "motor" activity, i.e., activity during grasping in the dark, may be crucial in this respect. Both F5a and AIP contain large numbers of neurons that are active in the dark, in the absence of visual information, which are typically not selective for disparity-defined depth structure. Traditionally, this "motor" type of activity has been interpreted as related to action planning, since the activity remains high in the delay period when the animal is waiting for the go-signal to execute the grasping action. However, the activity in the dark in AIP may have a different status: AIP neurons may receive corollary information from premotor areas about the movement planning which can be integrated with visual information during visually-guided grasping (Rizzolatti and Luppino, 2001), an idea that has never been tested. Under this hypothesis, F5a motor activity is genuinely sub-serving action planning, whereas AIP motor activity is simply a corollary discharge reflecting premotor signals.

Conclusions

More than four decades of research have been devoted to investigations of the object representation in the ITC (for review, see Tanaka, 1996). ITC neurons respond selectively to shapes, and at the same time achieve selectivity invariance, i.e., these neurons exhibit tolerance of shape preference for stimulus transformations such as changes in retinal position, size, the visual cue defining the shape and occlusion. These neuronal properties are believed to be essential to support robust object recognition in an ever changing environment. Single-cell studies have demonstrated that neurons in TEs, a subsector of the ITC located anteriorly and therefore one of the end-stage areas of the ventral visual stream, also encode the 3D structure of surfaces (Janssen et al., 2000a,b) defined by binocular disparity and 3D orientation defined by disparity and texture (Liu et al., 2004), and TEs activity is causally related to the categorization of depth structure (Verhoef et al., 2012). Undoubtedly, a complex hierarchy of visual areas along the ventral visual stream supports the high-level 3D object representation culminating in TEs.

In the dorsal visual stream, the neural representations of depth structure can be traced from mid-level visual area CIP, which presumably receives input from early visual area V3A, until the motor system (F5a). Somewhere along this dorsal pathway, visual object representations are transformed into motor commands (grip type representations) that control the preshaping of the hand during object grasping. As outline above, inactivation studies have demonstrated that at least AIP and F5 are both important for motor control during grasping, but the role of other parietal areas such as PFG (Bonini et al., 2012) and V6A (Fattori et al., 2010) in grasping requires further study. Investigating the neural representation of object shape demands systematic stimulus manipulations (e.g., stimulus reduction), therefore visual object representations can be primarily studied in neurons that respond to images of objects (either 3D or 2D), as in AIP and F5a. Neurons in F5p and F5c, in contrast, respond selectively to real-world objects (Raos et al., 2006) but not to 3D images of objects (Theys et al., 2012a), most likely because these areas represent grip types, which are not activated by images of objects.

Numerous very basic questions remain to be addressed in future studies: for example, how do the RFs change along the dorsal pathway, which computations take place at different stages,

what is the role of feedback projections? As outlined earlier, it is also unclear at present how the ventral stream division of this 3D-shape network is organized. Our comprehension of this network will only increase when studies combine electrophysiological recordings, imaging and ultimately also computational modeling.

Acknowledgments

This work was supported by Geconcerteerde Onderzoeksacties (2010/19), Fonds voor Wetenschappelijk Onderzoek Vlaanderen Grants (G.0495.05, G.0713.09), Programmafinanciering (PFV/10/008), IUAP VII-11, and ERC-2010-StG-260607.

References

Andersen, R. A., and Buneo, C. A. (2002). Intentional maps in posterior parietal cortex. *Annu. Rev. Neurosci.* 25, 189–220. doi: 10.1146/annurev.neuro.25.112701.142922

Baumann, M. A., Fluet, M. C., and Scherberger, H. (2009). Context-specific grasp movement representation in the macaque anterior intraparietal area. *J. Neurosci.* 29, 6436–6448. doi: 10.1523/JNEUROSCI.5479-08.2009

Belmalih, A., Borra, E., Contini, M., Gerbella, M., Rozzi, S., and Luppino, G. (2009). Multimodal architectonic subdivision of the rostral part (area F5) of the macaque ventral premotor cortex. *J. Comp. Neurol.* 512, 183–217. doi: 10.1002/cne.21892

Bonini, L., Ugolotti, S. F., Bruni, S., Maranesi, M., Bimbi, M., Simone, L., et al. (2012). Selectivity for grip type and action goal in macaque inferior parietal and ventral premotor grasping neurons. *J. Neurophysiol.* 108, 1607–1619. doi: 10.1152/jn.01158.2011

Borra, E., Belmalih, A., Calzavara, R., Gerbella, M., Murata, A., Rozzi, S., et al. (2008). Cortical connections of the macaque anterior intraparietal (AIP) area. *Cereb. Cortex* 18, 1094–1111. doi: 10.1093/cercor/bhm146

Colby, C. L., and Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annu. Rev. Neurosci.* 22, 319–349. doi: 10.1146/annurev.neuro.22.1.319

Durand, J. B., Nelissen, K., Joly, O., Wardak, C., Todd, J. T., Norman, J. F., et al. (2007). Anterior regions of monkey parietal cortex process visual 3D shape. *Neuron* 55, 493–505. doi: 10.1016/j.neuron.2007.06.040

Fagg, A. H., and Arbib, M. A. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw.* 11, 1277–1303. doi: 10.1016/S0893-6080(98)00047-1

Fattori, P., Breveglieri, R., Raos, V., Bosco, A., and Galletti, C. (2012). Vision for action in the macaque medial posterior parietal cortex. *J. Neurosci.* 32, 3221–3234. doi: 10.1523/JNEUROSCI.5358-11.2012

Fattori, P., Raos, V., Breveglieri, R., Bosco, A., Marzocchi, N., and Galletti, C. (2010). The dorsomedial pathway is not just for reaching: grasping neurons in the medial parieto-occipital cortex of the macaque monkey. *J. Neurosci.* 30, 342–349. doi: 10.1523/JNEUROSCI.3800-09.2010

Fogassi, L., Gallese, V., Buccino, G., Craighero, L., Fadiga, L., and Rizzolatti, G. (2001). Cortical mechanism for the visual guidance of hand grasping movements in the monkey: a reversible inactivation study. *Brain* 124, 571–586. doi: 10.1093/brain/124.3.571

Gallese, V., Murata, A., Kaseda, M., Niki, N., and Sakata, H. (1994). Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport* 5, 1525–1529. doi: 10.1097/00001756-199407000-00029

Gardner, E. P., Babu, K. S., Ghosh, S., Sherwood, A., and Chen, J. (2007). Neurophysiology of prehension. III. Representation of object features in posterior parietal cortex of the macaque monkey. *J. Neurophysiol.* 98, 3708–3730. doi: 10.1152/jn.00609.2007

Georgieva, S., Peeters, R., Kolster, H., Todd, J. T., and Orban, G. A. (2009). The processing of three-dimensional shape from disparity in the human brain. *J. Neurosci.* 29, 727–742. doi: 10.1523/JNEUROSCI.4753-08.2009

Gerbella, M., Belmalih, A., Borra, E., Rozzi, S., and Luppino, G. (2011). Cortical connections of the anterior (F5a) subdivision of the macaque ventral premotor area F5. *Brain Struct. Funct.* 216, 43–65. doi: 10.1007/s00429-010-0293-6

Goodale, M. A., Milner, A. D., Jakobson, L. S., and Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature* 349, 154–156. doi: 10.1038/349154a0

Graziano, M. S., and Gross, C. G. (1998). Visual responses with and without fixation: neurons in premotor cortex encode spatial locations independently of eye position. *Exp. Brain Res.* 118, 373–380.

Graziano, M. S., Hu, X. T., and Gross, C. G. (1997). Visuospatial properties of ventral premotor cortex. *J. Neurophysiol.* 77, 2268–2292.

Gross, C. G., Bender, D. B., and Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science* 166, 1303–1306. doi: 10.1126/science.166.3910.1303

Howard, I. P., and Rogers, B. J. (1995). *Binocular Vision and Stereopsis*, Oxford University Press.

James, T. W., Culham, J., Humphrey, G. K., Milner, A. D., and Goodale, M. A. (2003). Ventral occipital lesions impair object recognition but not object-directed grasping: an fMRI study. *Brain* 126, 2463–2475. doi: 10.1093/brain/awg248

Janssen, P., Srivastava, S., Ombelet, S., and Orban, G. A. (2008). Coding of shape and position in macaque lateral intraparietal area. *J. Neurosci.* 28, 6679–6690. doi: 10.1523/JNEUROSCI.0499-08.2008

Janssen, P., Vogels, R., and Orban, G. A. (1999). Macaque inferior temporal neurons are selective for disparity-defined three-dimensional shapes. *Proc. Natl. Acad. Sci. U.S.A.* 96, 8217–8222. doi: 10.1073/pnas.96.14.8217

Janssen, P., Vogels, R., and Orban, G. A. (2000a). Selectivity for 3D shape that reveals distinct areas within macaque inferior temporal cortex. *Science* 288, 2054–2056. doi: 10.1126/science.288.5473.2054

Janssen, P., Vogels, R., and Orban, G. A. (2000b). Three-dimensional shape coding in inferior temporal cortex. *Neuron* 27, 385–397. doi: 10.1016/S0896-6273(00)00045-3

Joly, O., Vanduffel, W., and Orban, G. A. (2009). The monkey ventral premotor cortex processes 3D shape from disparity. *Neuroimage* 47, 262–272. doi: 10.1016/j.neuroimage.2009.04.043

Katsuyama, N., Yamashita, A., Sawada, K., Naganuma, T., Sakata, H., and Taira, M. (2010). Functional and histological properties of caudal intraparietal area of macaque monkey. *Neuroscience* 167, 1–10. doi: 10.1016/j.neuroscience.2010.01.028

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., et al. (2013). Deep hierarchies in the primate visual cortex: what can we learn for computer vision? *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1847–1871. doi: 10.1109/TPAMI.2012.272

Liu, Y., Vogels, R., and Orban, G. A. (2004). Convergence of depth from texture and depth from disparity in macaque inferior temporal cortex. *J. Neurosci.* 24, 3795–3800. doi: 10.1523/JNEUROSCI.0150-04.2004

Molina-Vilaplana, J., Feliu-Batlle, J., and Lopez-Coronado, J. (2007). A modular neural network architecture for step-wise learning of grasping tasks. *Neural Netw.* 20, 631–645. doi: 10.1016/j.neunet.2007.02.003

Murata, A., Gallese, V., Luppino, G., Kaseda, M., and Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *J. Neurophysiol.* 83, 2580–2601.

Mysore, S. G., Vogels, R., Raiguel, S. E., Todd, J. T., and Orban, G. A. (2010). The selectivity of neurons in the macaque fundus of the superior temporal area for three-dimensional structure from motion. *J. Neurosci.* 30, 15491–15508. doi: 10.1523/JNEUROSCI.0820-10.2010

Nakamura, H., Kuroda, T., Wakita, M., Kusunoki, M., Kato, A., Mikami, A., et al. (2001). From three-dimensional space vision to prehensile hand movements: the lateral intraparietal area links the area V3A and the anterior intraparietal area in macaques. *J. Neurosci.* 21, 8174–8187.

Poggio, T., and Ullman, S. (2013). Vision: are models of object recognition catching up with the brain? *Ann. N.Y. Acad. Sci.* 1305, 72–82. doi: 10.1111/nyas.12148

Raos, V., Umilta, M. A., Murata, A., Fogassi, L., and Gallese, V. (2006). Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey. *J. Neurophysiol.* 95, 709–729. doi: 10.1152/jn.00463.2005

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rizzolatti, G., and Luppino, G. (2001). The cortical motor system. *Neuron* 31, 889–901. doi: 10.1016/S0896-6273(01)00423-8

Romero, M. C., Pani, P., and Janssen, P. (2014). Coding of shape features in the macaque anterior intraparietal area. *J. Neurosci.* 34, 4006–4021. doi: 10.1523/JNEUROSCI.4095-13.2014

Romero, M. C., Van Dromme, I., and Janssen, P. (2012). Responses to two-dimensional shapes in the macaque anterior intraparietal area. *Eur. J. Neurosci.* 36, 2324–2334. doi: 10.1111/j.1460-9568.2012.08135.x

Romero, M. C., Van Dromme, I., and Janssen, P. (2013). The role of binocular disparity in stereoscopic images of objects in the macaque anterior intraparietal area. *PLoS ONE* 8:e55340. doi: 10.1371/journal.pone.0055340

Rosenberg, A., Cowan, N. J., and Angelaki, D. E. (2013). The visual representation of 3D object orientation in parietal cortex. *J. Neurosci.* 33, 19352–19361. doi: 10.1523/JNEUROSCI.3174-13.2013

Sereno, A. B., and Maunsell, J. H. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature* 395, 500–503. doi: 10.1038/26752

Srivastava, S., Orban, G. A., De Maziere, P. A., and Janssen, P. (2009). A distinct representation of three-dimensional shape in macaque anterior intraparietal area: fast, metric, and coarse. *J. Neurosci.* 29, 10613–10626. doi: 10.1523/JNEUROSCI.6016-08.2009

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Theys, T., Pani, P., van Loon, J., Goffin, J., and Janssen, P. (2012a). Selectivity for three-dimensional shape and grasping-related activity in the macaque ventral premotor cortex. *J. Neurosci.* 32, 12038–12050. doi: 10.1523/JNEUROSCI.1790-12.2012

Theys, T., Pani, P., van Loon, J., Goffin, J., and Janssen, P. (2013). Three-dimensional shape coding in grasping circuits: a comparison between the anterior intraparietal area and ventral premotor area F5a. *J. Cogn. Neurosci.* 25, 352–364. doi: 10.1162/jocn_a_00332

Theys, T., Srivastava, S., van Loon, J., Goffin, J., and Janssen, P. (2012b). Selectivity for three-dimensional contours and surfaces in the anterior intraparietal area. *J. Neurophysiol.* 107, 995–1008. doi: 10.1152/jn.00248.2011

Tsutsui, K., Jiang, M., Yara, K., Sakata, H., and Taira, M. (2001). Integration of perspective and disparity cues in surface-orientation-selective neurons of area CIP. *J. Neurophysiol.* 86, 2856–2867.

Tsutsui, K., Sakata, H., Naganuma, T., and Taira, M. (2002). Neural correlates for perception of 3D surface orientation from texture gradient. *Science* 298, 409–412. doi: 10.1126/science.1074128

Ungerleider, L. G., and Mishkin, M. (1982). *In Analysis of Visual Behavior*, eds D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (London; Cambridge, MA: MIT Press), 549–586.

Verhoef, B. E., Vogels, R., and Janssen, P. (2012). Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron* 73, 171–182. doi: 10.1016/j.neuron.2011.10.031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Theys, Romero, van Loon and Janssen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 27 October 2014
doi: 10.3389/fncom.2014.00132

Modeling the shape hierarchy for visually guided grasping

Omid Rezai[1], Ashley Kleinhans[2,3], Eduardo Matallanas[4], Ben Selby[1] and Bryan P. Tripp[1]**

[1] Department of Systems Design Engineering, Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, Canada
[2] Mobile Intelligent Autonomous Systems, Council for Scientific and Industrial Research, Pretoria, South Africa
[3] School of Mechanical and Industrial Engineering, University of Johannesburg, Johannesburg, South Africa
[4] ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Abdelmalik Moujahid, University of the Basque Country UPV/EHU, Spain
Ko Sakai, University of Tsukuba, Japan
Peter Janssen, Catholic University of Leuven, Belgium

***Correspondence:**
Omid Rezai and Bryan P. Tripp, Systems Design Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
e-mail: omid.srezai@uwaterloo.ca; bptripp@uwaterloo.ca

The monkey anterior intraparietal area (AIP) encodes visual information about three-dimensional object shape that is used to shape the hand for grasping. We modeled shape tuning in visual AIP neurons and its relationship with curvature and gradient information from the caudal intraparietal area (CIP). The main goal was to gain insight into the kinds of shape parameterizations that can account for AIP tuning and that are consistent with both the inputs to AIP and the role of AIP in grasping. We first experimented with superquadric shape parameters. We considered superquadrics because they occupy a role in robotics that is similar to AIP, in that superquadric fits are derived from visual input and used for grasp planning. We also experimented with an alternative shape parameterization that was based on an Isomap dimension reduction of spatial derivatives of depth (i.e., distance from the observer to the object surface). We considered an Isomap-based model because its parameters lacked discontinuities between similar shapes. When we matched the dimension of the Isomap to the number of superquadric parameters, the superquadric model fit the AIP data somewhat more closely. However, higher-dimensional Isomaps provided excellent fits. Also, we found that the Isomap parameters could be approximated much more accurately than superquadric parameters by feedforward neural networks with CIP-like inputs. We conclude that Isomaps, or perhaps alternative dimension reductions of visual inputs to AIP, provide a promising model of AIP electrophysiology data. Further work is needed to test whether such shape parameterizations actually provide an effective basis for grasp control.

Keywords: AIP, CIP, grasping, 3D shape, cosine tuning, superquadrics, Isomap

1. INTRODUCTION

The macaque anterior intraparietal area (AIP) receives input from the visual cortex, and is involved in visually guided grasping. A large fraction of neurons in this area encode information about three-dimensional object shapes from visual input (Murata et al., 2000; Sakaguchi et al., 2010). Responses are typically relatively invariant to object position in depth (Srivastava et al., 2009). The responses of some neurons are also invariant to other properties. For example, some are orientation-tuned but not highly sensitive to object shape (Murata et al., 2000). AIP has a strong recurrent connection with premotor area F5, which is involved in hand shaping for grasping (Rizzolatti et al., 1990; Luppino et al., 1999; Borra et al., 2008). Reversible inactivation of AIP leads to grasping impairment, specifically a mismatch between object shape and hand preshape (Gallese et al., 1994; Fogassi et al., 2001). AIP is therefore thought to provide visual information for grasp control (Jeannerod et al., 1995; Fagg and Arbib, 1998).

The focus of this paper is the pathway from V3 and V3A, to the caudal intraparietal area (CIP), to visual-dominant neurons in AIP (Nakamura et al., 2001; Tsutsui et al., 2002). This pathway makes binocular disparity information available for grasp control. Most V3 neurons are selective for binocular disparity (Adams and Zeki, 2001). V3 sends a major projection to V3A (Felleman et al., 1997), which is also strongly activated during binocular disparity processing (Tsao et al., 2003). Both V3 and V3A project to CIP (Katsuyama et al., 2010). CIP neurons are selective for depth gradients (Taira et al., 2000; Tsutsui et al., 2002; Rosenberg et al., 2013) and curvature (Katsuyama et al., 2010). Neurons in AIP receive disynaptic input from V3A via CIP (Nakamura et al., 2001; Borra et al., 2008). Visual-dominant AIP neurons are selective for 3D object shape (Srivastava et al., 2009; Sakaguchi et al., 2010) cued by binocular disparity, consistent with input from this pathway.

AIP also receives many other inputs that we do not model in the present study. The first of these is the premotor area F5, which together with AIP forms a circuit for grasp-related visuomotor transformations. AIP also receives input from the second somatosensory (SII) cortical region (Krubitzer et al., 1995; Fitzgerald et al., 2004; Gregoriou et al., 2006), which may provide tactile feedback and memory-based somatosensory expectations for grasping. Strong connections with other parietal areas are also identified, as well as with prefrontal areas 46 and 12. Area 12 is implicated in high level non-spatial processing including encoding of objects in working memory, suggesting that AIP may be influenced by visual memory of object features

(Borra et al., 2008). AIP also contains other neurons that fire in conjunction with motor plans in addition to or instead of visual input (Sakata et al., 1997; Murata et al., 2000; Taira et al., 2000). Interestingly, AIP also receives subcortical input (via the thalamus) from both the cerebellum and basal ganglia (Clower et al., 2005). Finally, AIP receives input from the inferotemporal cortex (IT), which is likely to provide additional visual information about shapes. Our present focus however is the visual input from CIP.

The main goal of this study is to model the neural spike code of object-selective visual-dominant AIP neurons. In particular, we wanted to know whether there are certain sets of shape parameters that are consistent with the responses of visual AIP neurons, and which can furthermore be estimated in a physiologically plausible way from the information available in CIP.

We therefore compared two ways of parameterizing shapes. First we considered the superquadric family of shapes, a continuum that includes cuboids, ellipsoids, spheres, octahedra, and cylinders, and which can also be extended in various ways to model more complex shapes (Solina and Bajcsy, 1990). We considered superquadrics because they play a role in robotic grasp control (Duncan et al., 2013) that seems to be similar to the role of AIP in primate grasp control, i.e., they represent shapes compactly as a basis for grasp planning. We also considered an alternative shape parameterization that is based on non-linear dimension reduction of the depth field. In particular, we used an Isomap (Tenenbaum et al., 2000). We considered Isomap parameters partly because they are continuous, i.e., similar shapes have similar parameters. This is consistent with datasets in which similar 3D stimuli elicit similar spike rate patterns in AIP (Theys et al., 2012, Figure 10; Srivastava et al., 2009, Figure 11C).

This study is one of the first to model the mapping from CIP to AIP. Oztop et al. (2006) modeled AIP as a hidden layer in a multilayer perceptron network that mapped visual depth onto hand configuration. The output layer of this model (corresponding to F5) was a self-organizing map of subnetworks that corresponded to different hand configurations. Prevete et al. (2011) developed a mixed neural and Gaussian-mixture model in which AIP received monocular infero-temporal input. This model did not include stereoscopic input from CIP. The FARS grasping model (Fagg and Arbib, 1998) did not address in detail how AIP activity arises from visual input. While past AIP models have been relatively abstract, here our goal is to fit published tuning curves from AIP recordings, and furthermore to do so using depth-related input from a model of CIP. As far as we are aware, there have not been previous attempts to model AIP tuning in terms of either superquadric parameters or non-linear dimension reduction of depth features.

2. MATERIALS AND METHODS

This study consists of three main parts. The first is a model of tuning for depth features in the caudal intraparietal area (CIP, see Section 2.1.1). The second is a model of tuning for three-dimensional shape features in the anterior intraparietal area (AIP, see Section 2.1.2). Finally, the third is an investigation of physiologically plausible feedforward mappings between CIP and AIP (see Section 2.5).

2.1. COSINE-TUNING MODELS OF NEUROPHYSIOLOGICAL DATA

We tested how well various tuning curves from the CIP and AIP electrophysiology literature could be approximated by cosine-tuned neuron models. In particular, given a vector x of stimulus variables, we modeled the net current, I, driving spiking activity in each neuron as

$$I = \tilde{\phi}^T x + b, \tag{1}$$

where b is a bias term and $\tilde{\phi}$ is parallel to the neuron's preferred direction in the space of stimulus parameters. Longer $\tilde{\phi}$ corresponds to higher sensitivity of the neuron to variations along its preferred direction.

We used a normalized version of the leaky-integrate-and-fire (LIF) spiking model. In this model, the membrane potential V has subthreshold dynamics $\tau_{RC}\dot{V} = -V + I$, where τ_{RC} is the membrane time constant and I is the driving current. The neuron spikes when $V >= 1$, after which V is held at 0 for a post-spike refractory time τ_{ref} before subthreshold integration begins again. These neurons have spike rate

$$r = \frac{1}{\tau_{ref} - \tau_{RC} \cdot \ln\left(1 - \frac{1}{I}\right)}. \tag{2}$$

Except where noted, τ_{RC} was included among the optimization parameters and constrained to the range $[0.02s, 0.2s]$. In some cases (where noted), when the basic cosine-LIF model (above) produced poor fits, we also added Gaussian background noise to I. Such background noise more realistically reflects the input to neurons *in vivo* (Carandini, 2004) and causes the LIF model to emit more realistic, irregular spike trains. It also has the potential to produce better tuning curve fits. The reason is that depending on the amplitude of the noise, the spike-rate function may be compressive [as in Equation (2)], sigmoidal, or nearly linear. In these cases we fixed $\tau_{ref} = 0.005s$ and $\tau_{RC} = 0.02s$, included the noise variance as an optimization parameter, and interpolated the spike rate from a lookup table based on simulations. Given a tuning curve from the electrophysiology literature and a list of hypothesized tuning variables, we found least-squares optimal parameters $\tilde{\phi}$ and b mainly, and either τ_{RC} or σ_{noise} (as noted in the corresponding sections), using Matlab's *lsqcurvefit* function. This function uses Matlab's trust-region-reflective algorithm, which is based partly on Coleman and Li (1994), to solve a non-linear curve-fitting problem in the sense of least-squares. We retried each optimization with at least 1000 random initial points in order to increase the probability of finding a global optimum.

We preferred cosine tuning models over more complex non-linear models for a number of reasons, including that they are simple and that cosine tuning is widespread in the cortex and elsewhere (Zhang and Sejnowski, 1999). (See more detailed rationale in the Discussion).

2.1.1. CIP Tuning

We approximated CIP responses in terms of depth and its first and second spatial derivatives. CIP has been proposed to encode these variables (Orban et al., 2006), and they have been the basis for several experimental studies of CIP responses (Sakata et al.,

1998; Taira et al., 2000; Tsutsui et al., 2001; Katsuyama et al., 2010; Rosenberg et al., 2013).

We fit cosine-tuned LIF neuron models to tuning curves from Tsutsui et al. (2002) and Rosenberg et al. (2013), and from Katsuyama et al. (2010), in which the stimuli varied in terms of first and second derivatives of depth, respectively. The stimuli in Katsuyama et al. (2010) consisted of curved surfaces with depth

$$z = \frac{1}{2}\left(K_1 x^2 + K_2 y^2\right). \tag{3}$$

K_1 and K_2 were varied to produce two levels of "curvedness,"

$$C = \sqrt{\frac{K_{max}^2 + K_{min}^2}{2}}$$

and a range of "shape indices"

$$SI = \frac{2}{\pi} \arctan \frac{K_{max} + K_{min}}{K_{max} - K_{min}},$$

where K_{max} and K_{min} are the larger and smaller curvatures along the x and y axes, respectively.

In terms of the depth z, the principal curvature along the x axis is

$$K_x = \frac{\partial^2 z/\partial x^2}{\left(1 + (\partial z/\partial x)^2\right)^{3/2}} \tag{4}$$

(de Vries et al., 1993). For these stimuli $\partial z/\partial x = 0$ at the center, and so $K_x = \partial^2 z/\partial x^2$.

2.1.2. AIP Tuning

Following Sakata et al. (1998) and Murata et al. (2000) and consistent with the role of AIP in grasping (Fagg and Arbib, 1998), we took the visual-dominant neurons in AIP to be responsive to three-dimensional shape. Available tuning curves (e.g., Murata et al., 2000) span small numbers of data points relative to the large space of shape variations that are relevant to hand pre-shaping. For this reason we fit models to various "augmented" tuning curves that matched published tuning curves for some shapes, and made assumptions about how these neurons might respond to other shapes (see **Figure 2**). These assumptions were based on additional data for separate AIP neurons (see below). Our augmented tuning curves spanned four of the shapes in Murata et al. (2000), specifically a sphere, cylinder, cube, and plate. Two other shapes (ring and cone) were omitted for simplicity, because they require additional superquadric shape parameters (see Section 2.2). The augmented tuning curves spanned four sizes and four orientations for each of the four shapes. Due to symmetries in the shapes, there were a total of 36 points in these tuning curves (see **Figure 1**). Four of these points corresponded to AIP data, and the rest (the augmented points) were extrapolated from the data.

We based the augmented points on additional data from other AIP neurons, including aggregate data. Murata et al. (2000) provide shape-tuning curves for six different object-type visual-dominant AIP neurons. We tested different augmented versions of these curves with various combinations of size and orientation

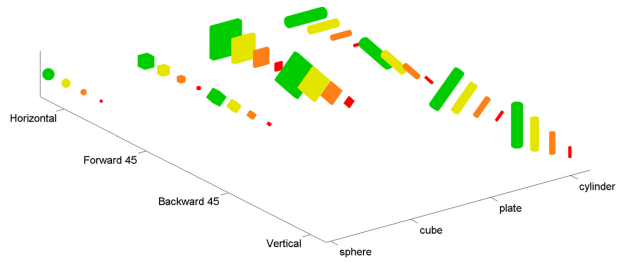


FIGURE 1 | The complete set of 36 shapes used in the augmented tuning curves. Four basic shapes (sphere, cube, plate, and cylinder) were adapted from Murata et al. (2000). In order to constrain the models more fully, and in particular to ensure that tuning curves included more points than there were parameters in our models, we augmented these basic shapes by adding copies with different sizes (shown with 4 different colors) and orientations (i.e., horizontal, vertical, tilted forward 45°, tilted backward 45°). Note that due to the symmetry of the basic shapes, some orientations are redundant (e.g., rotating a sphere does not create a distinguishable shape).

tuning (see **Figure 2**). Murata et al. (2000) reported (without plotting shape tuning for these neurons) that most object-type neurons were orientation selective, and that 16/26 were size-selective. Therefore, we created two augmented tuning curves for each of the six shape-tuning curves. Both were orientation-selective; one was size-selective and the other was size-invariant. For the size-selective tuning curves we assumed that spike rate increased monotonically with size (consistent with Murata et al., 2000, Figure 19; note that preference for intermediate sizes was reported only for motor-dominant neurons). We assumed that orientation tuning was roughly Gaussian and fairly narrow (consistent with Murata et al., 2000, Figure 18). Some AIP neurons are orientation selective with only mild selectivity across various elongated shapes (Sakata et al., 1998). Therefore, we created a final augmented tuning curve that was orientation selective but responded equally to cylinders and plates. **Figure 1** shows an example of an augmented tuning curve and its relationship to the data. This procedure made the tuning curve optimization more challenging. This was important because even our simple cosine-tuned neuron models had more parameters than the number of points in the published tuning curves (see Section 3). It also allowed us to make use of additional AIP data.

2.2. SUPERQUADRICS

We modeled AIP shape tuning both on the parameters of the superquadric family of shapes, and on an Isomap dimension reduction of depth features. The superquadric family is a continuum that includes cuboids, ellipsoids, spheres, octahedra, and cylinders as examples. Superquadrics are often used to approximate observed shapes as an intermediate step in robotic grasp control (Ikeuchi and Hebert, 1996; Biegelbauer and Vincze, 2007; Goldfeder et al., 2007; Huebner et al., 2008; Duncan et al., 2013). In this context, superquadric shape parameters are typically estimated from 3D point-cloud data using iterative non-linear optimization methods (Huebner et al., 2008).

Their role in robotics suggests that superquadrics are a plausible model of AIP shape tuning. Specifically, they can be parameterized from visual information and they contain

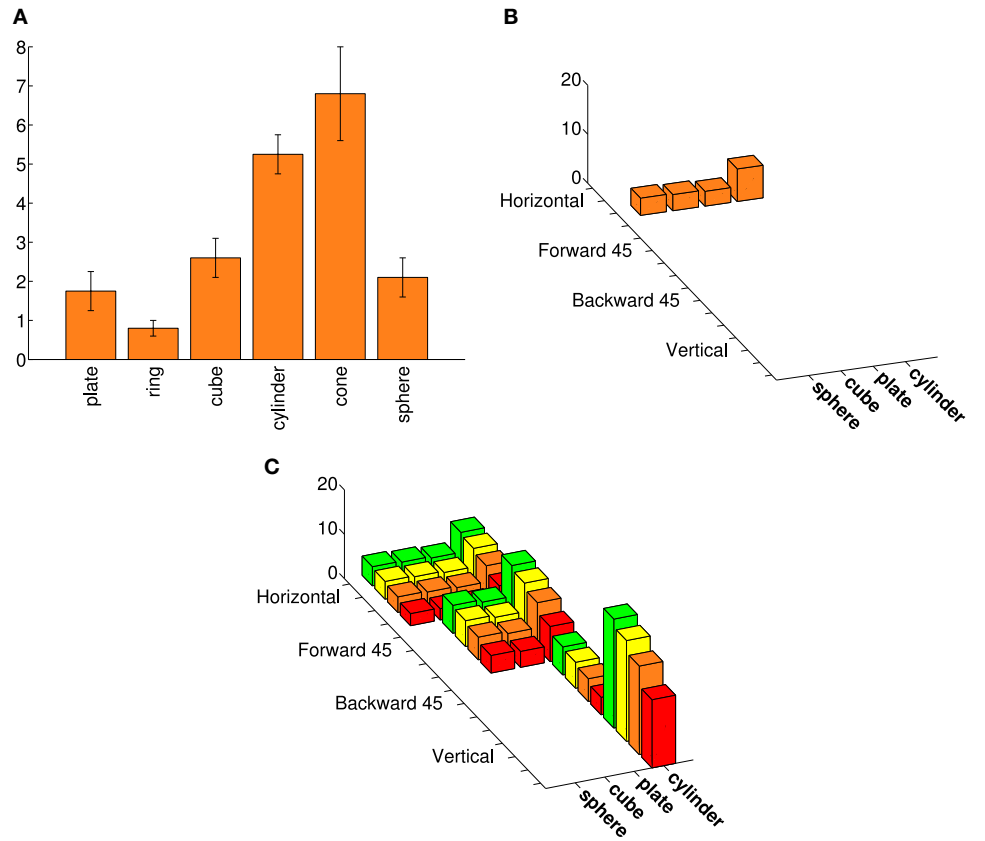


FIGURE 2 | An example of an augmented AIP tuning curve. (A) Tuning curve adapted from Murata et al. (2000), Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP, 2580-2601, with permission. (See their Figure 11.) **(B)** The four points from the same tuning curve that belong to the basic superquadric family (a ring and cone are excluded from the current study). The spike rates are plotted as 3D bars. **(C)** An augmented tuning curve that includes the points in **(B)**, as well as other rotations and scales. This augmented tuning curve is both size-tuned and orientation-tuned, as were the majority of object-type visual neurons in Murata et al. (2000). Another large minority were orientation-tuned but not size-tuned. As in **Figure 1**, the colors correspond to different sizes.

information about an object that is useful as a basis for grasp planning. One goal of the present study was to examine their physiological plausibility more closely, by fitting superquadric-tuned neuron models to AIP tuning curves. The surface of a superquadric shape is defined in *x*-*y*-*z* space as

$$\left(\frac{x}{A_1}\right)^{1/\epsilon_1} + \left(\frac{y}{A_2}\right)^{1/\epsilon_2} + \left(\frac{z}{A_3}\right)^{1/\epsilon_3} = 0,$$

where $A > 0$ are scale parameters and $\epsilon > 0$ are curvature parameters. Values of ϵ close to zero correspond to squared corners, while values close to one correspond to rounded corners. For example a sphere has $A_1 = A_2 = A_3$ and $\epsilon_1 = \epsilon_2 = \epsilon_3 = 1$. We also used another parameter, θ, that described the orientation of the superquadric. θ was composed of three angles, one per coordinate. The rotation of the superquadric is done applying the rotation matrix described in Equation 5.

We generated a database of 40,000 shapes that included spheres, cylinders, plates, and cubes as well as variations on these shapes with different scales in each dimension, and rotated versions of them. Our database contained roughly equal numbers of box-like, sphere-like, and cylinder-like shapes. For round edges we set $\epsilon = 1$. For squared edges we drew ϵ from an exponential distribution that was shifted slightly away from zero, $p = 10H(\epsilon - \eta)\exp(-(\epsilon - \eta)/0.1)$ with $\eta = 0.01$, where H is the Heaviside step function. The shift away from 0 (perfectly sharp corners) helped to avoid numerical problems. The objects had widths between 0.02 m and 0.12 m. We also allowed arbitrary rotations in three dimensions (except where symmetry made rotations redundant), so that each shape had a total of nine parameters.

This study considers only the basic superquadric family, which does not include all the shapes for which AIP responses have been reported. However, the basic family can also be extended in various ways to deal with more complex shapes. For exam-

$$R(\theta_1, \theta_2, \theta_3) = \begin{bmatrix} \cos(\theta_2)\cdot\cos(\theta_3) & \cos(\theta_1)\cdot\sin(\theta_3) + \sin(\theta_1)\cdot\sin(\theta_2)\cdot\cos(\theta_3) & \sin(\theta_1)\cdot\sin(\theta_3) - \cos(\theta_1)\cdot\sin(\theta_2)\cdot\cos(\theta_3) \\ -\cos(\theta_2)\cdot\sin(\theta_3) & \cos(\theta_1)\cdot\cos(\theta_3) - \sin(\theta_1)\cdot\sin(\theta_2)\cdot\sin(\theta_3) & \sin(\theta_1)\cdot\cos(\theta_3) + \cos(\theta_1)\cdot\sin(\theta_2)\cdot\sin(\theta_3) \\ \sin(\theta_2) & -\sin(\theta_1)\cdot\cos(\theta_2) & \cos(\theta_1)\cdot\cos(\theta_2) \end{bmatrix} \quad (5)$$

ple, hyperquadrics introduce asymmetry (Kumar et al., 1995), and trees of superquadrics can be used to approximate complex shapes with arbitrary precision (Goldfeder et al., 2007).

2.3. CREATION OF DEPTH MAPS

CIP receives input from V3 and V3A, which encode binocular disparity information (Anzai et al., 2011). Disparity is monotonically related to visual depth, or distance from observer to surface. As a simplified model of this input we created depth maps, i.e., grids of distances from a viewpoint to object surfaces. We created depth maps from the shapes in our superquadric database by finding intersections of the surfaces with rays at various visual angles from the view point. We used a 16 × 16 grid of visual angles. Grid spacing was closer near the center than in the periphery, in order to reflect higher visual acuity near the fovea and also to ensure that a few rays intersected with the smallest shapes (specifically, distances from the center were $a^{1.5}$, where a were evenly-spaced points). The grid covered ± 10° of visual angle in each direction. The object centers were at a depth of 0.75 m from the viewpoint. Depth at each grid point was found as the intersection of the superquadric surface with a line from the observation point (**Figure 3**).

2.4. ISOMAP SHAPE PARAMETERS

Within the superquadric family there is typically more than one set of parameters that can describe a given shape. For example, a tall box can either be parameterized as a tall box or a wide box on its end. This is not very problematic in robotics, because an iterative search for matching parameters finds one of these solutions. However, our goal was to model a feedforward mapping from depth (V3A) to shape parameters (AIP). In order to use the superquadric parameters as the basis for an AIP tuning we therefore needed the superquadric-to-depth function to be invertible.

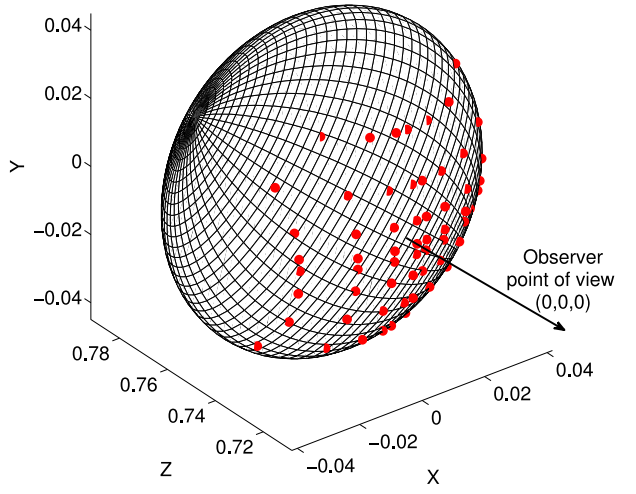


FIGURE 3 | Illustration of the depth map construction process. Each superquadric was centered at (0, 0, 0.75) relative to an observer at (0, 0, 0). Rays were traced between the observation point and a grid of points in the frontoparallel plane at $z = 0.75$, and intersections (red dots) were found with the superquadric surface. The depth map consisted of a grid of distances from (0, 0, 0) to these intersections.

We achieved this by restricting the ranges of angles. For example, for box-like shapes we restricted all angles to within $\pm\pi/4$. This resulted in a unique set of superquadric parameters for each shape. However, large discontinuities remained, in that some very similar shapes sometimes had very different parameters. For example, a tall box at an angle slightly less than $\pi/4$ has a depth map that is very similar to a wide box at angle just greater than $-\pi/4$ radians. Similar discontinuities seem to exist regardless of the angle convention. We anticipated that these discontinuities would impair feedforward mapping in a neural network, so we also explored an alternative low-dimensional shape parameterization.

In the alternative model, neurons were tuned to an Isomap (Tenenbaum et al., 2000) derived from depth data. Isomap is a non-linear dimension-reduction method in which samples are embedded in a lower-dimensional space in such a way that geodesic distances (i.e., distances along the shortest paths through edges between neighboring points) are maintained as well as possible. This method ensured that similar depth maps would be close together in the shape-parameter space, minimizing parameter discontinuities like those of the superquadric parameters. We constructed an Isomap of the first and second spatial derivatives of the depth maps in the horizontal and vertical directions.

We tested whether our augmented AIP tuning curves (above) were consistent with cosine tuning for these shape parameters. We also tested how well these shape parameters could be approximated by a neural network with CIP parameters as input.

2.5. NEURAL NETWORK MODELS OF CIP-TO-AIP MAP

In addition to fitting cosine-LIF models to neural tuning curves in CIP and AIP, we also developed feedforward networks to map from CIP variables to AIP variables. Our general approach was to decode shape parameters from the spike rates of CIP models.

We experimented with several different networks including neural engineering framework networks (Eliasmith and Anderson, 2003; Eliasmith et al., 2012), multilayer perceptrons trained with the back-propagation algorithm (Haykin, 1999) and convolutional networks (LeCun et al., 1998).

In each case the output units were linear. Linear decoding of the tuning parameters was of interest because decoding weights can be multiplied with preferred directions to give synaptic weights for any cosine tuning curve over the decoded variables (Eliasmith and Anderson, 2003). Specifically, suppose we have presynaptic rates $\mathbf{r}_{pre}$ and linearly decoded estimates $\hat{\mathbf{p}} = \Phi\mathbf{r}_{pre}$ of shape parameters $\mathbf{p}$, where Φ is a matrix of decoding weights. In this case the family of cosine tuning curves over $\hat{\mathbf{p}}$ is

$$r_{post} = G\left(\tilde{\phi}^T\hat{\mathbf{p}} + b\right), \tag{6}$$

where $\tilde{\phi}^T\hat{\mathbf{p}} + b$ is the driving current, $\tilde{\phi}$ is the neuron's preferred direction, G is a physiological model of the current-spike rate relationship, and b is a bias current. Such a tuning curve can then be obtained with synaptic weights (from all *presynaptic* neurons to a single *postsynaptic* neuron)

$$\mathbf{w}^T = \tilde{\phi}^T\Phi. \tag{7}$$

This allows us to draw general conclusions about how well our various models can account for AIP tuning, and how they would relate to future data.

Equations 6 and 7 are important components of the Neural Engineering Framework (Eliasmith and Anderson, 2003; Eliasmith et al., 2012), a method of developing large-scale neural circuit models.

3. RESULTS

3.1. CIP TUNING

Figure 4 shows an optimal fit of a cosine-tuned LIF model to a tuning curve from Katsuyama et al. (2010). Following their convention the spike rates are shown as a function of shape index, separately for the two curvedness levels. Inspection of the tuning curve revealed that it contained an expansive non-linearity, so we included Gaussian background noise in the model (as described in Section 2). To improve the fit further, in addition to tuning variables $X = \partial^2 z/\partial x^2$ and $Y = \partial^2 z/\partial y^2$ we introduced new tuning variables $\frac{1}{2}\left(3(X)^2 - 1\right)$ and $\frac{1}{2}\left(3(Y)^2 - 1\right)$. The rationale for their inclusion was that these are the non-linear functions for which linear reconstruction is (with reasonable assumptions) most accurate from populations of LIF neurons tuned to X and Y (Eliasmith and Anderson, 2003). However, the fit to the Katsuyama et al. (2010) data remained poor despite these measures.

We considered whether a linear-nonlinear receptive field model with depth inputs might produce a better fit. Such models are essentially cosine tuning models with multiple input variables on a grid. However, the depth stimuli in this case (see Equation 3) consisted of linear combinations of x^2 and y^2, so any receptive-field model over the depth field has an equivalent cosine tuning model over K_1 and K_2. Therefore, the neuron is not cosine tuned to either depth or the curvature parameters.

Figure 5 shows an example of a more complex non-linear neuron model that fits the data. This model is based on non-linear interactions between nearby inputs on the same dendrite, which suggest that pyramidal cells may function similarly to multilayer perceptrons (Polsky et al., 2004). The input to this model was a 3×3 depth grid. The model contained 50 dendritic branches, each of which was cosine tuned to the depths. The linear kernels (analogous to preferred directions) were random. The output of each branch was a sigmoid function of the point-wise product of the depth stimulus and the linear kernel. The spike rate was a least-squares optimal weighted sum of the branch outputs. This was found using a matrix pseudoinverse that used 14 singular values. We also created another version of this model (not shown) in which the tuning curve was augmented with additional stimuli (completing the outer circle of points in **Figure 5B**) and it was assumed that the neuron would respond to these stimuli at the background spike rate. This version of the model therefore fit 26 points, and we used 20 singular values in the pseudoinverse. The fit was similar in this case.

We also constructed another alternative model of this cell that was based on a more detailed model of V3A activity. Specifically, instead of a 3×3 depth grid, this model received input from seven non-linear functions of depth at each point. Five of these were Gaussian functions based on "tuned near," "tuned zero," and "tuned far" neurons (Poggio et al., 1988). Two were sigmoidal functions based on "near" and "far" tuning (Poggio et al., 1988). This model (not shown) reproduced the tuning curve somewhat less accurately than the non-linear cell model above. This was the case regardless of minor variations in the set of input tuning functions and their parameters.

Figure 6 shows a cosine-tuning fit of data from Tsutsui et al. (2002). This tuning curve is an average over multiple cells that were tuned to depth gradients of visual stimuli. The best fitting cosine-tuning model has a notably different shape than the aggregate data. In particular, the actual spike rates are fairly constant far away from the preferred stimulus, while the model spike rates continue to decrease farther from the preferred stimulus.

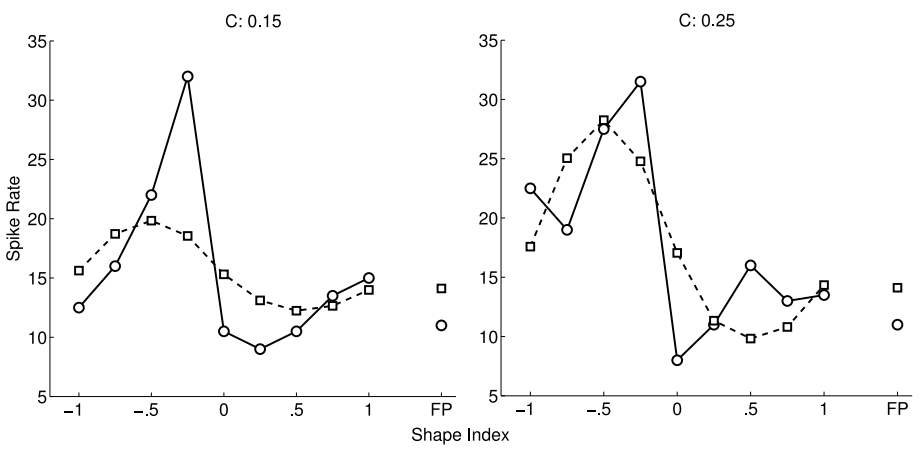


FIGURE 4 | Fit of CIP model (squares) to tuning curve (circles) of an example neuron (0.04 ± 5.22 spikes/s; mean error ± *SD*). The tuning curve is replotted from Katsuyama et al. (2010), with permission from Elsevier. In our model of CIP, neurons are cosine-tuned to five dimensions: depth, horizontal and vertical first spatial derivatives of depth, and horizontal and vertical second spatial derivatives of depth. The stimuli in Katsuyama et al. (2010) varied only in terms of the second derivatives. We also added non-linear tuning functions to improve the fit (see text). The left and right tuning curves are for two different levels of curvedness.

Rosenberg et al. (2013) provide several additional CIP tuning curves over 49 different plane stimuli. Some of these tuning curves are clearly not consistent with cosine tuning for first derivatives of depth or disparity, e.g., with multimodal responses to surface tilt. We fit the non-linear model of **Figure 5** to seven of these tuning curves (their Figures 4, 5B). Using 20 singular values, the correlations between data and our best model fits were $r = 0.98 \pm 0.01$ *SD* for the four tuning curves in their Figure 4, and $r = 0.78 \pm 0.09$ *SD* for the three tuning curves in their Figure 5B. (These fits are somewhat closer than fits reported by Rosenberg et al. to Bingham functions, which is unsurprising as our model has more parameters.) Using 40 singular values, our correlations improved to $r = 0.91 \pm 0.02$ *SD* for the tuning curves in their Figure 5B.

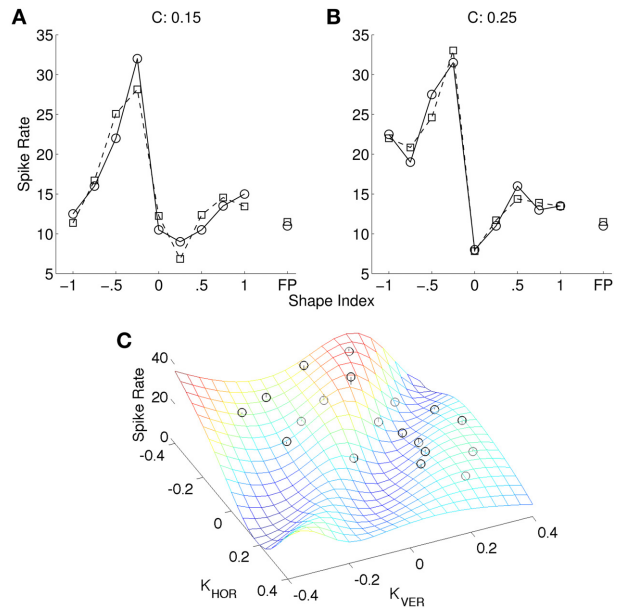


FIGURE 5 | (A) Non-linear model (squares) of same neuron as in **Figure 4** (circles). **(B)** The same spike rates as **(A)** (black circles), re-plotted as a function of $\partial^2 z/\partial x^2$ and $\partial^2 z/\partial y^2$, and the best model fit (mesh) (0.00 ± 1.82 spikes/s; mean error $\pm$ *SD*). The data plots (black circles) are adapted from Katsuyama et al. (2010), with permission from Elsevier.

In summary, the spike rates of these CIP neurons varied with the first and second spatial derivatives of depth, but not in a way that is consistent with cosine tuning to either the depth map, its first and second derivatives, or low-order polynomial functions of these derivatives. Other models, which are physiologically plausible but more complex, fit the data more closely.

3.2. AIP TUNING

Figure 7 shows an example cosine-tuning fit of an augmented tuning curve in superquadric space. This fit is based on a noise-free LIF neuron. For this dataset the shapes were rotated only in one dimension, so we avoided angle discontinuities by using a 2D direction vector in place of the angle. The optimized parameters were the 8-dimensional preferred direction vector $\tilde{\phi}$, the bias b, and the membrane time constant τ_{RC}. Across the 36 points in the augmented tuning curve, the spike rate error (difference between augmented and model spike rates) was 0.70 ± 1.57 (mean $\pm$ *SD*).

Figure 8 shows the means and standard deviations of spike-rate errors for each of the augmented tuning curves. Good fits were obtained for some of the neurons (#1 and #3 in Murata et al., 2000, Figure 10, and the second in Figure 11, which we label #5). This was true for both size-invariant and size-selective augmented tuning curves. Neuron #1 had low spike rates for the stimuli that we studied. Neurons #3 was highly selective for cylinders, and #5 was more broadly tuned but also preferred cylinders. The worst fits were obtained for neuron #6 which responded strongly to plates and cylinders but not to cubes or spheres.

Figure 9A shows the means and standard deviations of spike-rate errors for each of the augmented tuning curves in an 8-dimensional Isomap space. We plot the results for the 8-dimensional Isomap in order to match the number of superquadric parameters. The cosine tuning errors (-0.88 ± 10.68 spikes/s; mean $\pm$ *SD*) were larger than those in the superquadric space (-0.53 ± 6.75 spikes/s). The difference between these variances was significant according to Levene's test [$W_{(1,\,910)} = 41.3; p < 0.001$].

Figure 9B shows how the error declined with higher-dimensional Isomaps. Error variance with the 16-dimensional

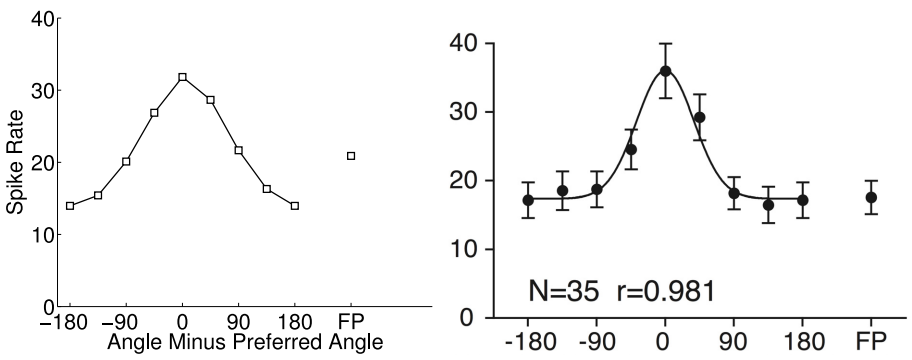


FIGURE 6 | Cosine tuning model (left) of spike rate data aggregated across neurons (right) (0.03 ± 3.00 spikes/s; mean error $\pm$ *SD*). The right panel is from Tsutsui et al. (2002). Reprinted with permission from AAAS. In which *N* is the number of neurons and *r* is the regression coefficient.

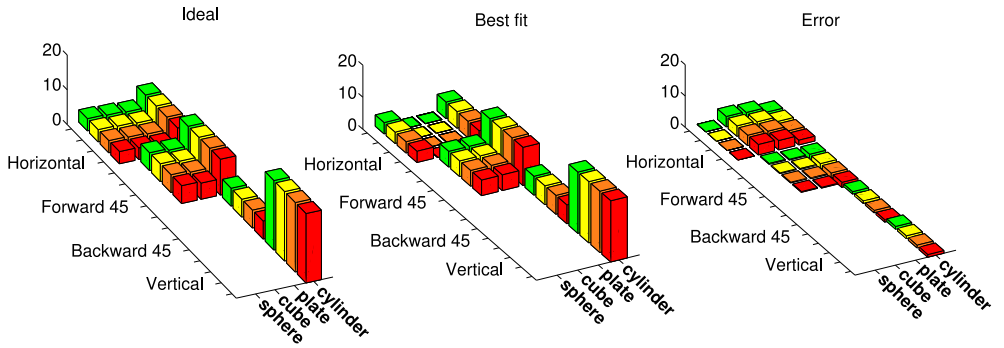


FIGURE 7 | Best fit of a model neuron that is cosine-tuned over superquadric parameters to an augmented tuning curve. This augmented tuning curve is size-invariant. Color corresponds to the size of object (see **Figure 1**) **Left**: Augmented tuning curve. This includes data replotted with permission from Murata et al. (2000). **Center**: Best fit of a cosine-tuned neuron to the augmented tuning curve. **Right**: Error (ideal minus model augmented tuning curve).

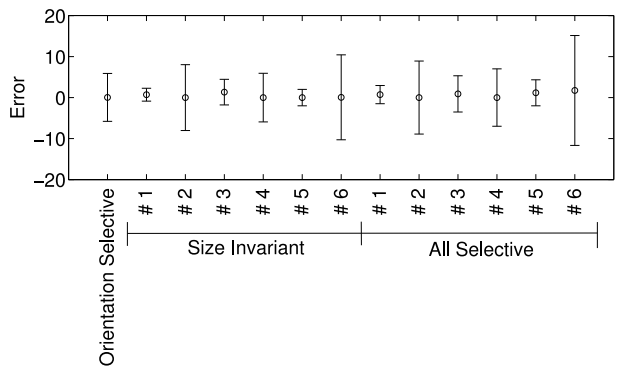


FIGURE 8 | Quality of fit of cosine tuning model over superquadric parameters with various augmented tuning curves. The plots shows, for each tuning curve, the mean ± *SD* of the errors over the grid of shapes, sizes, and orientations shown in **Figure 7**. Note that in this model we can trivially achieve invariance to any superquadric parameter by setting the corresponding component of the preferred direction to zero.

Isomap (-1.77 ± 6.35) was not significantly different from that of the 8-parameter superquadric [Levene's Test; $W_{(1,\,910)} = 1.83$; $p = 0.18$]. (Recalculating the variances around 0 instead of -1.77 and -0.89 did not make the difference significant; $p = 0.058$). The cosine-tuning fits were excellent in the 32-dimensional Isomap space, with significantly lower variance [-0.17 ± 1.29 spikes/s; $W_{(1,\,910)} = 316.2$; $p < 0.001$]. This higher-dimensional shape representation is therefore consistent with the data and with the augmented tuning curves.

3.3. MAPPING FROM CIP TO AIP

We trained multi-layer perceptrons in order to understand whether the superquadric or Isomap models of AIP were more consistent with mapping from CIP input. Because CIP neurons are sensitive to depth and to first and second spatial derivatives of depth, we used these as inputs to the networks. Specifically the inputs consisted of 16×16 depth maps, their 16×16 horizontal and vertical derivatives, and their 16×16 horizontal and vertical second derivatives. The derivatives were approximated by convolving with 3×3 kernels (e.g., $[1\ 1\ 1]^T[1\ 0\ -1]$ and $[1\ 1\ 1]^T[0.5\ -1\ 0.5]$). The total number of inputs was therefore $16 \times 16 \times 5 = 1280$. The hidden layers had logistic activation functions. The weights and biases were trained with the backpropagation algorithm in Matlab's Neural Network Toolbox. The output layer had a linear activation function in order to model the input to cosine-tuned neurons, as described in the Methods. A dataset of 40000 rotated superquadric objects was generated, from which depth and curvature images were derived. This dataset was divided into 28000 objects for training the network and 12000 objects to validate the results obtained in the training.

Figure 10 shows results from networks with two hidden layers, the first with 600 units and the second with 300 units. The scatter plots show the network's output vs. the actual values of the validation dataset. In **Figure 10A** is the network's result for the superquadric shape parameter ϵ_1. The other scatterplots in **Figures 10C,E** illustrate the network's approximation of the scale and orientation parameters A_1 and θ_1. Approximation of the other six parameters was similar (e.g., the scatterplots for ϵ_2 and ϵ_3 resemble that for ϵ_1). The scatterplots **Figures 10B,D,F** illustrate the network's approximation of Isomap parameters. The first, fourth, and seventh dimensions are shown as illustrative examples.

Approximation of the Isomap parameters was much more accurate than approximation of the superquadric parameters. This outcome was very consistent across a variety of networks of different sizes, with one or two hidden layers, with pre-training of hidden layers as autoencoders, etc. We also experimented with networks that contained a hidden layer of LIF neurons with random preferred directions over various local kernels, and optimal linear estimates of the shape parameters from the hidden-layer activity (Eliasmith and Anderson, 2003). The results were also similar in this case, although (as expected) more neurons were required to achieve performance like that of the more fully-optimized multilayer perceptrons.

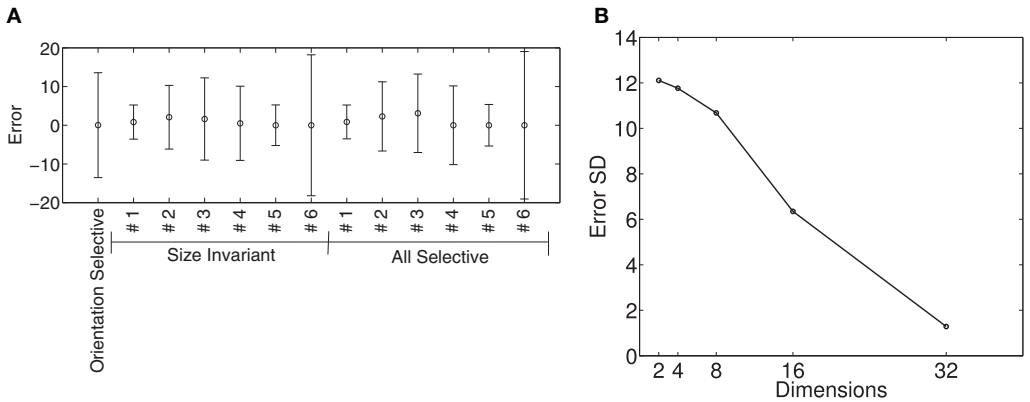


FIGURE 9 | Quality of fit of cosine tuning model over Isomap parameters with the same augmented tuning curves. (A), Mean ± *SD* of errors with 8-dimensional Isomap (the same number of parameters as the superquadric family used in **Figure 8**). Across tuning curves the error is 0.89 ± 10.68. **(B)**, Standard deviation of error over all augmented tuning curves vs. dimension of the Isomap. The error declines sharply with increasing dimension.

Figure 11 compares the distribution of the network's Isomap approximation errors with the distribution of pairwise distances between shape examples in our database. The errors were much smaller than typical distances between examples.

We also experimented with a wide variety of larger networks, including convolutional networks, using the cuda-convnet package (Krizhevsky et al., 2012). These networks did not substantially outperform the multilayer perceptron of **Figures 10**, **11** (lowest mean Euclidean error 0.066 as opposed to 0.081 in **Figure 11**). We also trained some convolutional networks with only the depth map as input, and with a 3 × 3 kernel in the first convolutional layer. Interestingly, some of the resulting kernels resembled the kernels that we created manually to approximate the first and second derivatives.

4. DISCUSSION

This study examined the neural code for three-dimensional shape in visual-dominant AIP neurons. AIP is critical for hand preshaping in grasping, and these neurons encode properties that are relevant to grasping including object shape, size, and orientation.

Our motivation for testing superquadric parameters as a model of AIP tuning was that superquadrics have been used in robotics, in a role that we take to be similar to the role of AIP in the primate brain. Specifically, they have been used as compact approximate representations of point clouds on which to base grasp planning. Such a representation is useful because it allows generalization from training examples to unseen examples, e.g., by interpolating between known solutions for known sets of parameters. An alternative approach in robotics is to cluster point clouds into discrete shape categories (Detry et al., 2013). We see the Isomap as an intermediate approach with some of the advantages of both superquadric fitting and clustering. The Isomap is data-driven and adapts to the statistics of the environment (like clustering), but its parameters make up a low-dimensional and continuous space (like those of superquadrics). Furthermore, unlike the superquadric representation, the Isomap representation does not have large discontinuities between very similar shapes.

We found that cosine tuning on a 32-dimensional Isomap accounted well for the tuning curves of object-selective AIP neurons. We also found that, in contrast with superquadric parameters, the Isomap parameters could be approximated fairly well by various neural networks with CIP-like input.

4.1. AUGMENTED TUNING CURVES

Available AIP data includes the responses of individual neurons to only a few different shapes, in fact fewer shapes than there are parameters in even the simplest superquadric model. To more vigorously test the different shape parameterizations as a basis for plausible neural tuning, and to incorporate additional aggregate information on shape tuning (e.g., the fact that most visual-dominant AIP neurons are orientation selective), we created "augmented" tuning curves that included both data and extrapolations of the data. It is likely that some of these augmented tuning curves were unrealistic. While the general trends in our AIP fitting results are informative (e.g., that Isomap fits improve and outperform superquadrics as dimensions increase), the details depend on our augmentation assumptions. For example, we found that the Isomap error declined more rapidly when we excluded orientation-selective/shape-invariant tuning curves from the analysis. This limitation does not affect interpretation of our other main result, i.e., that superquadrics were poorly approximated by feedforward neural networks while Isomaps were well approximated.

Future modeling would be facilitated by tuning curves with greater numbers of data points. For example, the dataset in Lehky et al. (2011) includes responses of 674 inferotemporal neurons to a common set of 806 images. A relatively extensive AIP dataset was recently collected (Schaffelhofer and Scherberger, 2014), but no tuning curves from this dataset have yet been published.

4.2. COSINE TUNING

We were primarily interested in cosine-tuning models for several reasons, not least because cosine tuning is widespread in the brain (see many examples in Zhang and Sejnowski, 1999). Linear-nonlinear receptive field models of the early visual system

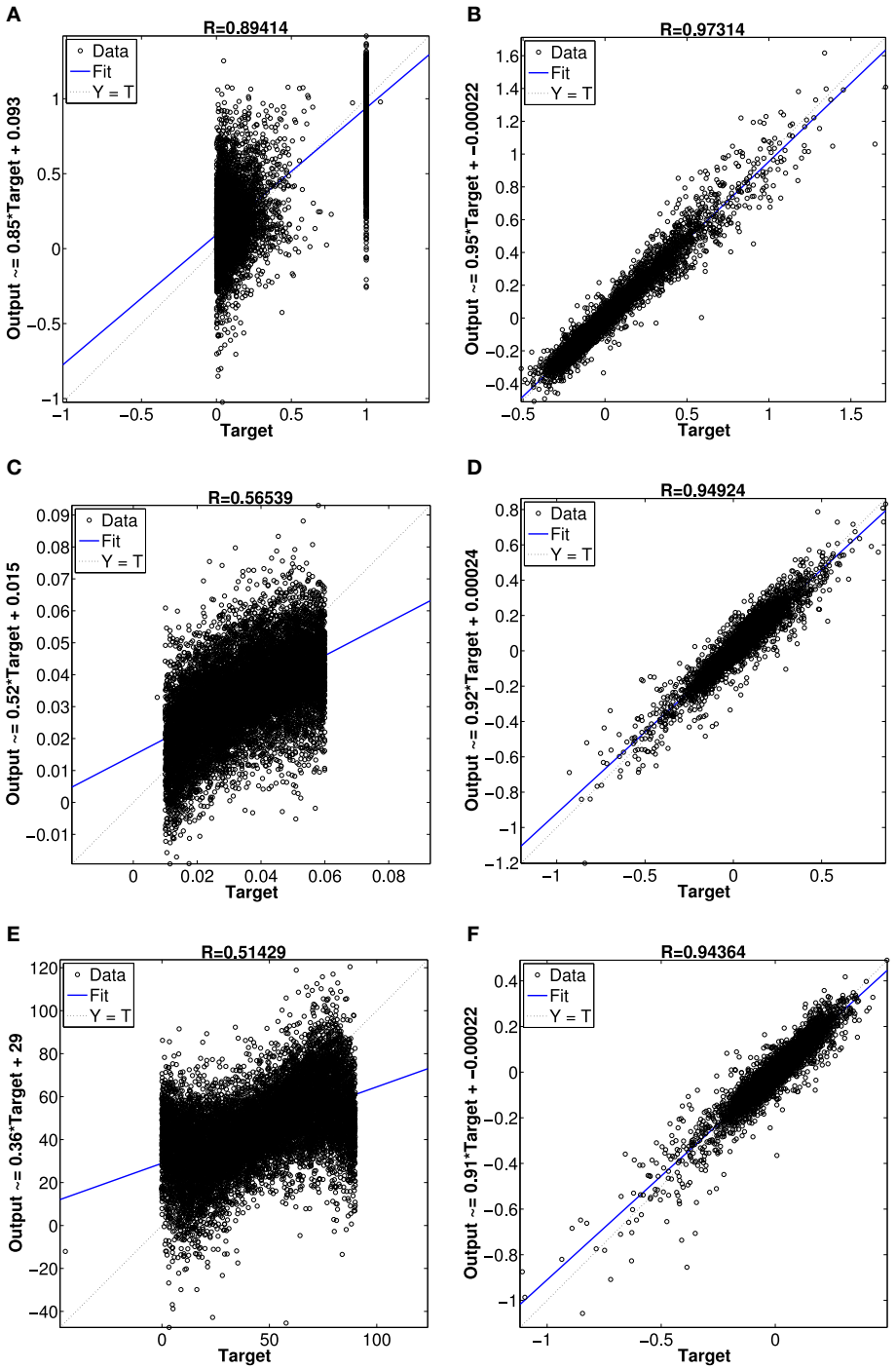


FIGURE 10 | Regression plot comparison between neural network approximations of superquadrics parameters and Isomap parameters. **(A)** superquadric epsilon parameter, **(B)** Isomap dimension 1 parameter, **(C)** superquadric scale parameter, **(D)** Isomap dimension 4 parameter, **(E)** superquadric rotation angle parameter and **(F)** Isomap dimension 7 parameter.

are another kind of cosine tuning, with multiple tuning variables on a 2D grid. Furthermore, a practical advantage of cosine tuning models is that they require only $n + 1$ tuning parameters for n stimulus variables (in contrast a full n-dimensional Gaussian tuning curve has $n + n^2$ parameters). This is important because published tuning curves in CIP and AIP consist of relatively few points, so models with large numbers of parameters may be underconstrained. Cosine tuning is also physiologically realistic in that it can arise from linear synaptic integration. For example, if a matrix W of synaptic weights has n large singular values, then

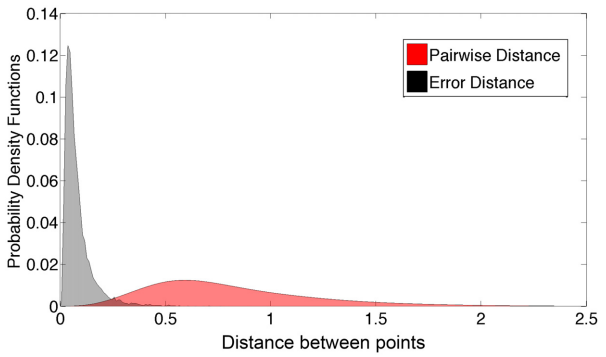


FIGURE 11 | Histogram of Euclidian distances in Isomap space vs. the root sum square error.

the post-synaptic neurons are tuned to a n-dimensional space (if $W = U\Sigma V^T$ then the preferred directions are in the first n columns of U). Cosine tuning curves are also optimal for linear decoding (Salinas and Abbott, 1994). There are also many neurons that do not appear to be cosine tuned, for example speed-tuned neurons in the middle temporal area (Nover et al., 2005). However, where applicable, cosine tuning models provide rich insight into neural activity. We therefore attempted to fit such models to the data where possible. Many AIP tuning curves over similar stimuli with different curvatures vary smoothly and monotonically (Srivastava et al., 2009), consistent with cosine tuning.

Cosine tuning to modest numbers of Isomap parameters (relative to the 256-element depth maps on which they were based) accounted for the AIP data and for our augmented AIP tuning curves.

In contrast, we concluded that the CIP neurons we modeled were not cosine tuned to the stimulus variables with which they have been examined. CIP has been proposed to encode first and second derivatives of depth (Orban et al., 2006). Various neurons in CIP respond to disparity gradient (Shikata et al., 1996; Sakata et al., 1998), texture gradient (Tsutsui et al., 2001), and/or perspective cues for oriented surfaces (Tsutsui et al., 2001). (Accordingly, visual-dominant AIP neurons also respond to monocular visual cues as well as disparity cues, and respond most strongly when disparity and other depth cues are congruent (Romero et al., 2013). Sakata et al. (1998) describe various neurons in CIP as axis-orientation-selective and surface-orientation-selective. The former were sensitive to the orientation of a long cylinder, consistent with two-dimensional tuning for horizontal and vertical curvature. The latter were selective for the orientation of a flat plate, consistent with two-dimensional tuning for depth gradient. Furthermore, Sakata et al. (1998) also recorded a neuron that preferred a cylinder of certain diameter which was tilted back and to the right, but did not respond strongly to a square column of similar dimensions. This suggests selectivity for both first and second derivatives within the same neuron. Katsuyama et al. (2010) recorded CIP responses to curved surfaces that varied in terms of their second derivatives. Tuning to the first and second derivatives of depth is physiologically plausible in that these quantities are linear functions of the depth field, which is available from V3A. We therefore attempted to fit models that were cosine tuned over these variables, but we obtained poor fits.

While CIP neurons are certainly responsive to these variables (and more complex non-linear models of tuning to these variables fit the data closely) it is possible that there are other related variables that provide a more elegant account of these neurons' responses. Notably, some CIP neurons prefer intermediate cylinder diameters (Sakata et al., 1998), whereas cosine tuning for curvature would be constrained to monotonic changes with respect to curvature. Also, some of the neurons in Rosenberg et al. (2013) are clearly non-cosine-tuned for depth slope.

Some CIP tuning curves (see e.g., **Figure 6**) seem to be fairly similar to rectified cosine functions (Salinas and Abbott, 1994) with a negative offset, except that their baseline rates are not zero. In general, spike sorting limitations, which cannot be completely avoided in extracellular recordings (Harris et al., 2000), are a potential source of uncertainty in tuning curves. However, if misclassification rates had been substantial then multi-peaked tuning curves might have been expected, and none were reported in these studies.

4.3. RELATIONSHIP TO SHAPE REPRESENTATION IN IT

Area IT has been shown to represent medial axes and surfaces of objects (Yamane et al., 2008; Hung et al., 2012). AIP has significant connections with IT areas including the lower bank of the superior temporal sulcus (STS), specifically areas TEa and TEm (Borra et al., 2008). These areas partially correspond to functional area TEs, which encodes curvature of depth (Janssen et al., 2000) similarly to CIP. However, AIP responds to depth differences much earlier than TEs (Srivastava et al., 2009). It is possible that a shape representation in IT, with some similarities to that in CIP, provides longer latency reinforcement and/or correction of shape representation in AIP.

4.4. FUTURE WORK

A key direction for future work is to test how well the Isomap shape representation works for robotic grasp planning. This would provide important information about the functional plausibility of this representation. For example, if Isomap-based shape parameters cannot be used to shape a hand for effective grasping, this will strongly suggest that there are critical differences between AIP tuning parameters and Isomap parameters. On the other hand, if the Isomap representation performs well, it may suggest a new biologically-inspired approach for robotic grasping.

An apparent advantage of the Isomap approach is that it is data-driven and makes no prior assumptions about shapes. It would be informative to build Isomaps for less idealized shapes that monkeys might grasp in nature.

Other non-linear dimension-reduction methods (e.g., Yan et al., 2007) could also be compared with the Isomap in terms of fitting AIP data and providing an effective basis for grasp planning. We would expect differences relative to Isomap tuning to be subtle relative to available AIP data, but perhaps distinct advantages would appear in a grasp control system. One interesting possibility would be to emphasize features that are related to reward or performance (Bar-Gad et al., 2003).

Another important direction for future work is to extend the model to include motor-dominant AIP neurons and to F5 neurons as in e.g., Theys et al. (2012, 2013) and Raos et al. (2006).

Finally, our models produced constant spike rates in response to static inputs. A more sophisticated future model would account for response timing and dynamics (Sakaguchi et al., 2010). The Neural Engineering Framework (Eliasmith and Anderson, 2003) provides a principled approach to modeling dynamics in systems of spiking neurons.

ACKNOWLEDGMENTS

Financial support was provided by CrossWing Inc.; NSERC and Mitacs (Canada); DAAD-NRF (South Africa); and the Spanish Ministry of Education and Consejo Social UPM (Spain). We thank Paul Calamai, Renaud Detry, and André Nel for helpful discussions.

REFERENCES

Adams, D. L., and Zeki, S. (2001). Functional organization of macaque v3 for stereoscopic depth. *J. Neurophysiol.* 86, 2195–2203. Available online at: http://jn.physiology.org/content/86/5/2195

Anzai, A., Chowdhury, S. A., and DeAngelis, G. C. (2011). Coding of stereoscopic depth information in visual areas v3 and v3a. *J. Neurosci.* 31, 10270–10282. doi: 10.1523/JNEUROSCI.5956-10.2011

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog. Neurobiol.* 71, 439–473. doi: 10.1016/j.pneurobio.2003.12.001

Biegelbauer, G., and Vincze, M. (2007). "Efficient 3D object detection by fitting superquadrics to range image data for Robot's object manipulation," in *2007 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, NJ: IEEE), 1086–1091. doi: 10.1109/ROBOT.2007.363129

Borra, E., Belmalih, A., Calzavara, R., Gerbella, M., Murata, A., Rozzi, S., et al. (2008). Cortical connections of the macaque anterior intraparietal (AIP) area. *Cereb. Cortex* 18, 1094–1111. doi: 10.1093/cercor/bhm146

Carandini, M. (2004). Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS Biol.* 2:E264. doi: 10.1371/journal.pbio.0020264

Clower, D. M., Dum, R. P., and Strick, P. L. (2005). Basal ganglia and cerebellar inputs to 'AIP'. *Cereb. Cortex* 15, 913–920. doi: 10.1093/cercor/bhh190

Coleman, T. F., and Li, Y. (1994). On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Prog.* 67, 189–224. doi: 10.1007/BF01582221

de Vries, S. C., Kappers, A. M., and Koenderink, J. J. (1993). Shape from stereo: a systematic approach using quadratic surfaces. *Percept. Psychophys.* 53, 71–80. doi: 10.3758/BF03211716

Detry, R., Ek, C. H., Madry, M., and Kragic, D. (2013). "Learning a dictionary of prototypical grasp-predicting parts from grasping experience," in *2013 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, NJ: IEEE), 601–608. doi: 10.1109/ICRA.2013.6630635

Duncan, K., Sarkar, S., Alqasemi, R., and Dubey, R. (2013). "Multi-scale superquadric fitting for efficient shape and pose recovery of unknown objects," in *2013 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, NJ: IEEE), 4238–4243. doi: 10.1109/ICRA.2013.6631176

Eliasmith, C., and Anderson, C. H. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205. doi: 10.1126/science.1225266

Fagg, A. H., and Arbib, M. A. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw.* 11, 1277–1303. doi: 10.1016/S0893-6080(98)00047-1

Felleman, D. J., Burkhalter, A., and Van Essen, D. C. (1997). Cortical connections of areas V3 and VP of macaque monkey extrastriate visual cortex. *J. Comp. Neurol.* 379, 21–47. doi: 10.1002/(SICI)1096-9861(19970303)379:1<21::AID-CNE3>3.0.CO;2-K

Fitzgerald, P. J., Lane, J. W., Thakur, P. H., and Hsiao, S. S. (2004). Receptive field properties of the macaque second somatosensory cortex: evidence for multiple functional representations. *J. Neurosci.* 24, 11193–11204. doi: 10.1523/JNEUROSCI.3481-04.2004

Fogassi, L., Gallese, V., Buccino, G., Craighero, L., Fadiga, L., and Rizzolatti, G. (2001). Cortical mechanism for the visual guidance of hand grasping movements in the monkey: a reversible inactivation study. *Brain* 124(Pt 3), 571–586. doi: 10.1093/brain/124.3.571

Gallese, V., Murata, A., Kaseda, M., Niki, N., and Sakata, H. (1994). Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport* 5, 1525–1529. doi: 10.1097/00001756-199407000-00029

Goldfeder, C., Allen, P. K., Lackner, C., and Pelossof, R. (2007). "Grasp planning via decomposition trees," in *2007 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, NJ: IEEE), 4679–4684. doi: 10.1109/ROBOT.2007.364200

Gregoriou, G. G., Borra, E., Matelli, M., and Luppino, G. (2006). Architectonic organization of the inferior parietal convexity of the macaque monkey. *J. Comp. Neurol.* 496, 422–451. doi: 10.1002/cne.20933

Harris, K. D., Henze, D. A., Csicsvari, J., Hirase, H., and Buzsáki, G. (2000). Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* 84, 401–414.

Haykin, S. (1999). *Neural Networks and Learning Machines, 3rd Edn*. Upper Saddle River, NJ: Prentice Hall PTR.

Huebner, K., Ruthotto, S., and Kragic, D. (2008). "Minimum volume bounding box decomposition for shape approximation in robot grasping," in *2008 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, NJ: IEEE), 1628–1633. doi: 10.1109/ROBOT.2008.4543434

Hung, C.-C., Carlson, E., and Connor, C. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron* 74, 1099–1113. doi: 10.1016/j.neuron.2012.04.029

Ikeuchi, K., and Hebert, M. (1996). "Task-oriented vision," in *Exploratory Vision, Springer Series in Perception Engineering*, eds M. Landy, L. Maloney, and M. Pavel (New York, NY: Springer), 257–277. doi: 10.1007/978-1-4612-3984-0_11

Janssen, P., Vogels, R., and Orban, G. A. (2000). Three-dimensional shape coding in inferior temporal cortex. *Neuron* 27, 385–397. doi: 10.1016/S0896-6273(00)00045-3

Jeannerod, M., Arbib, M., Rizzolatti, G., and Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends Neurosci.* 18, 314–320. doi: 10.1016/0166-2236(95)93921-J

Katsuyama, N., Yamashita, A., Sawada, K., Naganuma, T., Sakata, H., and Taira, M. (2010). Functional and histological properties of caudal intraparietal area of macaque monkey. *Neuroscience* 167, 1–10. doi: 10.1016/j.neuroscience.2010.01.028

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc.), 1097–1105.

Krubitzer, L., Clarey, J., Tweedale, R., Elston, G., and Calford, M. (1995). A redefinition of somatosensory areas in the lateral sulcus of macaque monkeys. *J. Neurosci.* 15, 3821–3839.

Kumar, S., Goldgof, D., and Bowyer, K. (1995). On recovering hyperquadrics from range data. *IEEE Trans. Patt. Anal. Mach. Intell.* 17, 1079–1083. doi: 10.1109/34.473234

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE* (Piscataway, NJ: IEEE), 2278–2324. doi: 10.1109/5.726791

Lehky, S. R., Kiani, R., Esteky, H., and Tanaka, K. (2011). Statistics of visual responses in primate inferotemporal cortex to object stimuli. *J. Neurophysiol.* 106, 1097–1117. doi: 10.1152/jn.00990.2010

Luppino, G., Murata, A., Govoni, P., and Matelli, M. (1999). Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). *Exp. Brain Res.* 128, 181–187. doi: 10.1007/s002210050833

Murata, A., Gallese, V., Luppino, G., Kaseda, M., and Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *J. Neurophysiol.* 83, 2580–2601. Available online at: http://jn.physiology.org/content/83/5/2580

Nakamura, H., Kuroda, T., Wakita, M., Kusunoki, M., Kato, A., Mikami, A., et al. (2001). From three-dimensional space vision to prehensile hand movements:

the lateral intraparietal area links the area V3A and the anterior intraparietal area in macaques. *J. Neurosci.* 21, 8174–8187.

Nover, H., Anderson, C. H., and DeAngelis, G. C. (2005). A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *J. Neurosci.* 25, 10049–10060. doi: 10.1523/JNEUROSCI.1661-05.2005

Orban, G. A., Janssen, P., and Vogels, R. (2006). Extracting 3D structure from disparity. *Trends Neurosci.* 29, 466–473. doi: 10.1016/j.tins.2006.06.012

Oztop, E., Imamizu, H., Cheng, G., and Kawato, M. (2006). A computational model of anterior intraparietal (AIP) neurons. *Neurocomputing* 69, 1354–1361. doi: 10.1016/j.neucom.2005.12.106

Poggio, G. F., Gonzalez, F., and Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *J. Neurosci.* 8, 4531–4550.

Polsky, A., Mel, B. W., and Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. *Nat. Neurosci.* 7, 621–627. doi: 10.1038/nn1253

Prevete, R., Tessitore, G., Catanzariti, E., and Tamburrini, G. (2011). Perceiving affordances: a computational investigation of grasping affordances. *Cogn. Syst. Res.* 12, 122–133. doi: 10.1016/j.cogsys.2010.07.005

Raos, V., Umiltá, M.-A., Murata, A., Fogassi, L., and Gallese, V. (2006). Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey. *J. Neurophysiol.* 95, 709–729. doi: 10.1152/jn.00463.2005

Rizzolatti, G., Gentilucci, M., Camarda, R., Gallese, V., Luppino, G., Matelli, M., et al. (1990). Neurons related to reaching-grasping arm movements in the rostral part of area 6 (area 6aβ). *Exp. Brain Res.* 82, 337–350. doi: 10.1007/BF00231253

Romero, M. C., Van Dromme, I. C. L., and Janssen, P. (2013). The role of binocular disparity in stereoscopic images of objects in the macaque anterior intraparietal area. *PLoS ONE* 8:e55340. doi: 10.1371/journal.pone.0055340

Rosenberg, A., Cowan, N. J., and Angelaki, D. E. (2013). The visual representation of 3d object orientation in parietal cortex. *J. Neurosci.* 33, 19352–19361. doi: 10.1523/JNEUROSCI.3174-13.2013

Sakaguchi, Y., Ishida, F., Shimizu, T., and Murata, A. (2010). Time course of information representation of macaque AIP neurons in hand manipulation task revealed by information analysis. *J. Neurophysiol.* 104, 3625–3643. doi: 10.1152/jn.00125.2010

Sakata, H., Taira, M., Kusunoki, M., Murata, A., and Tanaka, Y. (1997). The TINS lecture. the parietal association cortex in depth perception and visual control of hand action. *Trends Neurosci.* 20, 350–357. doi: 10.1016/S0166-2236(97)01067-9

Sakata, H., Taira, M., Kusunoki, M., Murata, A., Tanaka, Y., and Tsutsui, K. (1998). Neural coding of 3D features of objects for hand action in the parietal cortex of the monkey. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1363–1373. doi: 10.1098/rstb.1998.0290

Salinas, E., and Abbott, L. F. (1994). Vector reconstruction from firing rates. *J. Comput. Neurosci.* 1, 89–107. doi: 10.1007/BF00962720

Schaffelhofer, S., and Scherberger, H. (2014). "From vision to action: a comparative population study of hand grasping areas AIP, F5, and M1," in *Bernstein Conference 2014* (Göttingen). doi: 10.12751/nncn.bc2014.0253

Shikata, E., Tanaka, Y., Nakamura, H., Taira, M., and Sakata, H. (1996). Selectivity of the parietal visual neurones in 3D orientation of surface of stereoscopic stimuli. *Neuroreport* 7, 2389–2394. doi: 10.1097/00001756-199610020-00022

Solina, F., and Bajcsy, R. (1990). Recovery of parametric models from range images: the case for superquadrics with global deformations. *IEEE Trans. Patt. Anal. Mach. Intell.* 12, 131–147. doi: 10.1109/34.44401

Srivastava, S., Orban, G. A., Mazire, P. A. D., and Janssen, P. (2009). A distinct representation of three-dimensional shape in macaque anterior intraparietal area: fast, metric, and coarse. *J. Neurosci.* 29, 10613–10626. doi: 10.1523/JNEUROSCI.6016-08.2009

Taira, M., Tsutsui, K.-I., Jiang, M., Yara, K., and Sakata, H. (2000). Parietal neurons represent surface orientation from the gradient of binocular disparity. *J. Neurophysiol.* 83, 3140–3146. Available online at: http://jn.physiology.org/content/83/5/3140

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Theys, T., Pani, P., van Loon, J., Goffin, J., and Janssen, P. (2012). Selectivity for three-dimensional shape and grasping-related activity in the macaque ventral premotor cortex. *J. Neurosci.* 32, 12038–12050. doi: 10.1523/JNEUROSCI.1790-12.2012

Theys, T., Pani, P., van Loon, J., Goffin, J., and Janssen, P. (2013). Three-dimensional shape coding in grasping circuits: a comparison between the anterior intraparietal area and ventral premotor area F5a. *J. Cogn. Neurosci.* 25, 352–364. doi: 10.1162/jocn-a-00332

Tsao, D. Y., Vanduffel, W., Sasaki, Y., Fize, D., Knutsen, T. A., Mandeville, J. B., et al. (2003). Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. *Neuron* 39, 555–568. doi: 10.1016/S0896-6273(03)00459-8

Tsutsui, K., Jiang, M., Yara, K., Sakata, H., and Taira, M. (2001). Integration of perspective and disparity cues in surface-orientation-selective neurons of area CIP. *J. Neurophysiol.* 86, 2856–2867.

Tsutsui, K.-I., Sakata, H., Naganuma, T., and Taira, M. (2002). Neural correlates for perception of 3D surface orientation from texture gradient. *Science* 298, 409–412. doi: 10.1126/science.1074128

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., and Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360. doi: 10.1038/nn.2202

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Patt. Anal. Mach. Intell.* 29, 40–51. doi: 10.1109/TPAMI.2007.250598

Zhang, K., and Sejnowski, T. J. (1999). A theory of geometric constraints on neural activity for natural three-dimensional movement. *J. Neurosci.* 19, 3122–3145.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 May 2014; accepted: 26 September 2014; published online: 27 October 2014.

Citation: Rezai O, Kleinhans A, Matallanas E, Selby B and Tripp BP (2014) Modeling the shape hierarchy for visually guided grasping. Front. Comput. Neurosci. ***8****:132. doi: 10.3389/fncom.2014.00132*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Rezai, Kleinhans, Matallanas, Selby and Tripp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

REVIEW ARTICLE
published: 05 August 2014
doi: 10.3389/fncom.2014.00084

Feature integration and object representations along the dorsal stream visual hierarchy

***Carolyn Jeane Perry*[1,2]* and *Mazyar Fallah*[1,2,3,4]**

[1] *Visual Perception and Attention Laboratory, School of Kinesiology and Health Science, York University, Toronto, ON, Canada*
[2] *Centre for Vision Research, York University, Toronto, ON, Canada*
[3] *Departments of Biology and Psychology, York University, Toronto, ON, Canada*
[4] *Canadian Action and Perception Network, York University, Toronto, ON, Canada*

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Da-Hui Wang, Beijing Normal University, China
Christopher Pack, McGill University, Canada

****Correspondence:***
Carolyn Jeane Perry, Visual Perception and Attention Laboratory, School of Kinesiology and Health Science, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada
e-mail: carolynperry08@gmail.com

The visual system is split into two processing streams: a ventral stream that receives color and form information and a dorsal stream that receives motion information. Each stream processes that information hierarchically, with each stage building upon the previous. In the ventral stream this leads to the formation of object representations that ultimately allow for object recognition regardless of changes in the surrounding environment. In the dorsal stream, this hierarchical processing has classically been thought to lead to the computation of complex motion in three dimensions. However, there is evidence to suggest that there is integration of both dorsal and ventral stream information into motion computation processes, giving rise to intermediate object representations, which facilitate object selection and decision making mechanisms in the dorsal stream. First we review the hierarchical processing of motion along the dorsal stream and the building up of object representations along the ventral stream. Then we discuss recent work on the integration of ventral and dorsal stream features that lead to intermediate object representations in the dorsal stream. Finally we propose a framework describing how and at what stage different features are integrated into dorsal visual stream object representations. Determining the integration of features along the dorsal stream is necessary to understand not only how the dorsal stream builds up an object representation but also which computations are performed on object representations instead of local features.

Keywords: feature integration, dorsal pathway, object representation, motion processing, decision making

INTRODUCTION

Classically, visual processing from the retina onwards is described as following two general principles. One, the processing of different types of visual information is anatomically segregated into two visual streams, and two, each stream is comprised of hierarchical processing where each stage builds upon the previous stage, becoming increasingly more complex. In the ventral pathway this ultimately results in an ability to recognize objects in spite of changes in the surrounding environment or changes in certain object features (i.e., position, orientation, viewing angle, size, etc). In the dorsal pathway this hierarchical processing produces computations of complex motion of objects within the environment around us, either as we are stationary or moving through that environment. Because of this functional separation, there are many models of object representation in the ventral stream (see Peissig and Tarr, 2007 for a review) and many models of motion processing in the dorsal stream (for reviews see Burr and Thompson, 2011; Nishida, 2011), but motion processing research has been mostly devoid of investigations as to the nature or existence of object representations in the dorsal stream. In fact, the vision for action theory of dorsal stream function (Goodale and Milner, 1992; Goodale, 2008, 2013) would suggest that even though there might not be an internal representation of the object as a whole (see Farivar, 2009 for an alternative view), there are representations of features of an object that are relevant for action in real time. Evidence for this comes from spared functions in visual agnosia wherein damage to the ventral pathway eliminates the ability to recognize objects but spares scaling and orientation of the hand when grasping objects (Goodale et al., 1991, 1994; Milner et al., 2012). In addition, parietal regions of the dorsal pathway involved in reaching and grasping show selectivities for the orientation, shape and size of objects (Taira et al., 1990; Gallese et al., 1994; Murata et al., 2000; Fattori et al., 2005).

More recently, investigations into cross-talk between the two visual streams suggest that there are object representations in the dorsal stream (Schiller, 1993; Sereno and Maunsell, 1998; Tsutsui et al., 2001; Sereno et al., 2002; Peuskens et al., 2004; Durand et al., 2007; Lehky and Sereno, 2007; Wannig et al., 2007; Konen and Kastner, 2008; Tchernikov and Fallah, 2010; Perry and Fallah, 2012). It is important to note however, that this object representation would not necessarily be one that gives rise to object recognition, as in the ventral stream. For example, it has been shown that recognition of objects constructed

from coherently moving dots (structure-from-motion) is severely impaired in visual agnosiacs (Huberle et al., 2012). These cross-talk studies suggest however, that the motion computations that occur within the dorsal stream can benefit from an intermediate object representation that includes different features of the object. This intermediate object representation would allow for selection of one moving object over others contained within the visual field as seen with flankers and crowding (Livne and Sagi, 2007; Malania et al., 2007; Sayim et al., 2008; Manassi et al., 2012; Chicherov et al., 2014), and superimposed surfaces (Valdes-Sosa et al., 1998; Rodríguez et al., 2002; Mitchell et al., 2003; Reynolds et al., 2003; Stoner et al., 2005; Fallah et al., 2007; Wannig et al., 2007).

In this review we will first give a brief overview of the hierarchical nature of feature processing in both the ventral and dorsal pathways. Various models of the ventral stream have been proposed wherein each integrates features to build up an object representation (scale invariant feature transform (SIFT): Lowe, 1987; Neocognitron: Fukushima, 1975; hierarchical model and X (HMAX): Riesenhuber and Poggio, 1999, and others. For review see Poggio and Ullman, 2013), often based on behavioral and neurophysiological studies (Cowey and Weiskrantz, 1967; Gross et al., 1971, 1972; Dean, 1976; Marr and Nishihara, 1978; Biederman, 1987; Biederman and Cooper, 1991). However, the dorsal stream has generally been relegated to models and algorithms that build up more complex motion representations, from the prior stage's processing (Marr and Ullman, 1981; Adelson and Bergen, 1985; Cavanagh and Mather, 1989; Taub et al., 1997; Krekelberg and Albright, 2005; Pack et al., 2006; Tsui and Pack, 2011; Mineault et al., 2012; Krekelberg and van Wezel, 2013; Patterson et al., 2014; for review see Burr and Thompson, 2011). This may be due to the fact that many behavioral and neurophysiological studies of the dorsal stream have used paradigms that are focused on individual motion features instead of object representations. While feature integration and object representations that lead to object based selection are fairly well understood concepts within the context of the ventral pathway, less is known about how and where these processes occur in the dorsal pathway. We will systematically review the studies that do shed light into which stages of the dorsal stream use object representations vs. motion features. Our aims are to provide a framework for object representations within the dorsal stream and propose where the anatomical locations of these representations may be. We find that motion features but not object representations are used up to global motion processing, as is found in area middle temporal (MT). The next stage of processing, area medial superior temporal (MST), relies on intermediate object representations based on smooth pursuit and glass pattern studies. Finally, intermediate object representations can be used by the decision making circuitry further down the dorsal stream (e.g., area lateral intraparietal (LIP)), which results in faster decisions. It should be noted that the review of literature presented here is strictly limited to those processes that are pertinent to the current discussion and thus is not by any means exhaustive.

HIERARCHICAL VISUAL PROCESSING

DORSAL PATHWAY

The dorsal visual pathway is specialized for motion processing. Much research has determined the hierarchical nature of motion processing wherein each stage builds upon the previous stage's output leading to understanding of the algorithms and connectivity to produce models of the different stages of motion processing (Marr and Ullman, 1981; Adelson and Bergen, 1985; Cavanagh and Mather, 1989; Taub et al., 1997; Krekelberg and Albright, 2005; Pack et al., 2006; Tsui and Pack, 2011; Mineault et al., 2012;

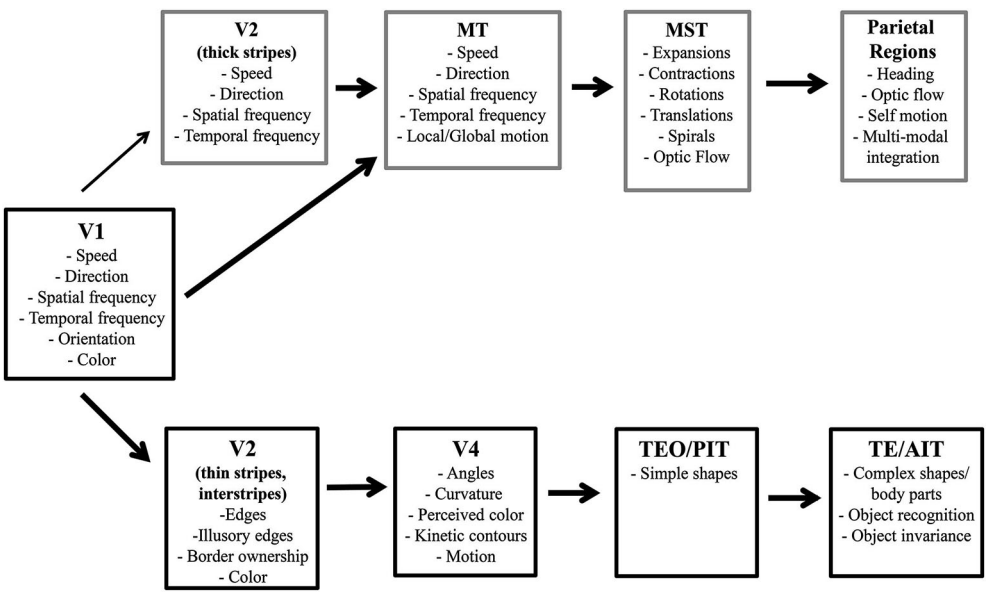


FIGURE 1 | Hierarchy of visual processing in ventral and dorsal streams. Gray boxes, from V2 on, depict select features processed at each region along the dorsal pathway. Black boxes, from V2 on, represent features processed along the ventral pathway.

Krekelberg and van Wezel, 2013; Patterson et al., 2014; for review see Burr and Thompson, 2011). It is important to note that these models focus on the transformation of motion information and not its integration into object representations. Although motion can produce form cues to be used in representing objects in the ventral stream, e.g., structure-from-motion (Johansson, 1973, 1976; Siegel and Andersen, 1988; Bradley et al., 1998; Grunewald et al., 2002; Jordan et al., 2006), object representation in the dorsal stream has not been historically focussed upon. This section briefly reviews the anatomical and functional hierarchy for motion processing (see **Figure 1** for an overview).

V1

Magnocellular cells in the retina and lateral geniculate nucleus (LGN) provide the input to motion processing in the dorsal pathway. These cells are sensitive to low luminance and also to lower spatial and higher temporal frequencies, but are not sensitive to color. They project to layer 4Cα in the primary visual cortex (V1). In V1complex cells are sensitive to the motion of oriented moving edges, bars or gratings (Hubel and Wiesel, 1968; Hubel et al., 1978; Adelson and Bergen, 1985) and show direction selectivity (Orban et al., 1986; Movshon and Newsome, 1996). Complex cells also show the combined spatiotemporal frequency tuning necessary for early speed selectivity (Orban et al., 1986; Priebe et al., 2006). In addition, it has been shown that V1 cells respond only to the local (or component) motion contained in complex patterns (Movshon and Newsome, 1996).

V2

Motion information, from layer 4B in V1, projects to the thick stripes in V2 (Hubel and Livingstone, 1987; Levitt et al., 1994). Although not traditionally thought to play a central role in motion processing, the thick stripes in V2 provide the second largest input to area MT (DeYoe and Van Essen, 1985; Shipp and Zeki, 1985; Born and Bradley, 2005) and it has recently been suggested that directional maps could first emerge in V2 (Lu et al., 2010; however, see Gegenfurtner et al., 1997 for an alternative view).

MT

While MT is the next stage of motion processing after V2, it also receives significant input directly from V1 (Felleman and Van Essen, 1991; Born and Bradley, 2005). MT cells are sensitive to many features associated with 2D motion such as direction (Maunsell and Van Essen, 1983; Albright, 1984; Lagae et al., 1993), speed (Maunsell and Van Essen, 1983; Lagae et al., 1993; Perrone and Thiele, 2001; Priebe et al., 2003; Brooks et al., 2011), and spatial frequency (Priebe et al., 2003; Brooks et al., 2011). The increase in receptive field size and the unique characteristics of MT cells allow for the processing of both local (component) and global (pattern/random dot kinetograms) motion (Pack and Born, 2001; gratings: Adelson and Movshon, 1982; Rodman and Albright, 1989; random dot kinetograms (RDKs): Britten et al., 1992; Snowden et al., 1992). This allows MT to both integrate the motion of multiple dots or incongruent motions created by edges within the same object, and also to separate multiple moving objects from each other. It is important to note that neurons in area MT have been shown to not be color selective (Maunsell and Van Essen, 1983; Shipp and Zeki, 1985; Zeki et al., 1991; Dobkins and Albright, 1994; Gegenfurtner et al., 1994).

MST

With the local and global 2D motion information from area MT, area MST has been implicated in processing complex, 3D motion and in the start of computations of optic flow and self-motion which are dependent on the analysis of 3D motion. Area MST has been anatomically divided into lateral (MSTl) and dorsal (MSTd) regions, where MSTl is thought to be intricately involved in computing the velocity signals of object trajectories used in the maintenance of pursuit eye movements (Tanaka et al., 1993; Ilg, 2008). In comparison, neurons in MSTd are selective for rotations and expansion/contraction motion (Saito et al., 1986), or their combination, aka spiral motion (Graziano et al., 1994; Mineault et al., 2012). MSTd neurons are also selective for optic flow (Duffy and Wurtz, 1991a,b). In fact MSTd neurons can take optic flow and compute the heading or direction of self-motion (Duffy and Wurtz, 1995; Gu et al., 2006).

Beyond MST

After MST, the dorsal pathway continues into the posterior parietal cortex. Motion processing therein involves more complicated optic flow and self-motion patterns, including the motion of objects while the viewer is also moving (Phinney and Siegel, 2000; Raffi and Siegel, 2007; Raffi et al., 2010; Chen et al., 2013; Raffi et al., 2014;). For example, cells in area 7a are tuned to distinguish between *types* of optic flow (Siegel and Read, 1997), and neurons in caudal pole of the superior parietal lobule (Brodmann area 5) (PEc) can combine optic flow information with signals regarding the position of the head and eye (Raffi et al., 2014).

VENTRAL PATHWAY

The ventral visual pathway processes form and color information in a hierarchical stream that builds up separately and then integrates into intermediate and full object representations (Marr and Nishihara, 1978; Biederman, 1987; Biederman and Cooper, 1991) ending with object recognition (Cowey and Weiskrantz, 1967; Gross et al., 1971, 1972; Dean, 1976). Thus, hierarchical models of the object representation and recognition focus on feature integration in the ventral stream (SIFT: Lowe, 1987; Neocognitron: Fukushima, 1975; HMAX: Riesenhuber and Poggio, 1999, and others. For review see Poggio and Ullman, 2013). This section briefly reviews the anatomical and functional hierarchy for building up an object in the ventral pathway (see **Figure 1** for an overview).

V1

Input to V1 in the ventral pathway comes mainly from the parvocellular layers of the LGN with additional magnocellular input (Ferrera et al., 1992, 1994). Parvocellular cells, sensitive to color, high contrasts, and high spatial and low temporal frequencies, project to layer 4Cβ of V1 which is subsequently divided into color blobs and form interblobs. Blobs are color selective but contrast and size invariant (Solomon et al., 2004; Solomon and Lennie, 2005), and untuned for orientation (Livingstone and Hubel, 1987;

Ts'o and Gilbert, 1988; Roe and Ts'o, 1999; Landisman and Ts'o, 2002; Shipp and Zeki, 2002). Interblobs are orientation selective for multiple stimulus types, i.e., edges, bars, gratings (Hubel and Wiesel, 1968; Hubel et al., 1978). Both blobs and interblobs process features without regard to objects, although feedback can produce object-based modulation (Roelfsema et al., 1998) and may be involved in representing objects (Fallah and Reynolds, 2001; Roelfsema and Spekreijse, 2001).

V2

While color processing (interstripes) changes little from that seen in V1, there is notable progression in form processing (thin stripes). V2 neurons are sensitive to the orientation of edges that are defined either by illusory contours or texture (von der Heydt et al., 1984; Peterhans and von der Heydt, 1989; von der Heydt and Peterhans, 1989). V2 cells also encode border ownership (Zhou et al., 2000) which is the first stage of assigning an oriented edge to an object representation. Thus contour-based object representation starts in V2.

V4

Neurons in V4 are tuned for hue that is unaffected by luminance and not limited to a set of colors along the cardinal color axes (red-green, blue-yellow) as seen in V1 (Conway and Livingstone, 2006; Conway et al., 2007). Center-surround interactions produce encoding of perceived color instead of physical color (Schein and Desimone, 1990). Thus, V4 is the first representation of perceived color which is the earliest stage at which color should be incorporated into an ecologically valid object representation.

Form processing in V4 combines multiple, spatially-adjacent, orientation responses seen in V1 and V2 to encode angles and curvatures (Pasupathy and Connor, 1999). These responses advance the nascent object representation from border ownership (Orban, 2008) to responses that are dependent on the placement of the curvature with respect to the center of the shape (Pasupathy and Connor, 2001).

Selection for the orientation of contours created between moving objects (kinetic contours) emerges in V4 (Mysore et al., 2006). Accordingly, a subset of V4 neurons are directionally selective (Ferrera et al., 1992, 1994; Li et al., 2013). Therefore, it should be noted that the intermediate object representations in area V4 can include motion features as well as color and shape.

IT cortex

Inferior temporal (IT) cortex has a range of object property complexity starting with simpler features posteriorly (PIT or TEO: Tanaka et al., 1991; Kobatake and Tanaka, 1994) that increase in complexity as processing moves anteriorly (AIT or TE) to represent objects and perform object recognition (Cowey and Weiskrantz, 1967; Gross et al., 1971, 1972; Dean, 1976). This includes complex shapes, combinations of color or texture with shape (Gross et al., 1972; Desimone et al., 1984; Tanaka et al., 1991), and body parts (faces or hands: see Gross, 2008 for a review). In addition, responses in IT cortex are position and size invariant (Sato et al., 1980; Schwartz et al., 1983; Rolls and Baylis, 1986; Ito et al., 1995; Logothetis and Pauls, 1995) and also invariant to changes in luminance, texture, and relative motion (Sáry et al., 1993). Combined, these characteristics make IT ideal for representing objects despite changes in the surrounding environment and retinal image.

FEATURE INTEGRATION IN THE DORSAL STREAM

Classically, as presented above, it is thought that the ventral pathway is involved in the creation of object representations and categorizations that allow for recognition, object-based selection and decision making processes. Comparatively, the early dorsal stream is most often thought to be specialized for motion processing. Growing evidence suggests however, that processing in the dorsal stream may also allow for object based selection and decision making, which is consistent with later dorsal stream involvement in visumotor guidance, e.g., vision for action (Goodale and Milner, 1992; Goodale, 2008, 2013). In the ventral stream, the object-file theory (Kahneman et al., 1992) has been supported by growing empirical evidence (Mitroff et al., 2005, 2007, 2009; Noles et al., 2005). Object-files collect, store and update information regarding specific objects over time. They are considered to be mid-level representations of objects that do not rely on higher-level object categorizations.

While motion processing studies have focused on individual motion features like direction or speed discriminations of a single moving stimulus, these motion computations could instead be working on intermediate object representations. We hypothesize that later dorsal stream processing occurs on intermediate object representations formed by feature integration instead of on independent motion features. Further we propose that the intermediate object representations also integrate ventral stream information such as color or form. Here we present evidence that support the presence of intermediate (or mid-level) object representations in the dorsal stream, resulting from both ventral and dorsal stream features being integrated into an object-file.

There are multiple ways to investigate the mechanism and timing of feature integration (Cavanagh et al., 1984; Kahneman et al., 1992; Croner and Albright, 1997; Mitroff et al., 2005; Bodelón et al., 2007; Perry and Fallah, 2012 among others). To study feature integration in the dorsal pathway, it is practical to utilize stimuli that activate motion processing regions. Area MT is well known to be involved in direction computations of moving stimuli including the global motion of RDKs (Britten et al., 1992; Snowden et al., 1992). The use of coherently moving, superimposed RDK's that produce the perception of two superimposed objects moving in different directions controls for spatial location, allowing for investigation of object properties (Valdes-Sosa et al., 1998; Rodríguez et al., 2002; Mitchell et al., 2003; Reynolds et al., 2003; Stoner et al., 2005; Fallah et al., 2007; Wannig et al., 2007). In addition, direction discrimination of two superimposed surfaces becomes more difficult as the presentation time decreases (Valdes-Sosa et al., 1998), suggesting that there is a limitation in speed of processing.

Using two superimposed RDKs does, however, create a perceptual illusion known as direction repulsion. Instead of the

directions of the two superimposed surfaces being integrated, the directions are perceived as being repulsed away from the real directions of motion (Marshak and Sekuler, 1979; Mather and Moulden, 1980; Hiris and Blake, 1996; Braddick et al., 2002; Curran and Benton, 2003). This phenomenon can also be observed with superimposed gratings under conditions that produce motion transparency (Kim and Wilson, 1996). Direction repulsion is the result of inhibitory, repulsive interactions (Marshak and Sekuler, 1979; Mather and Moulden, 1980; Wilson and Kim, 1994; Kim and Wilson, 1996; Perry et al., 2014) between the directions of motion at the level of global motion processing in area MT (Wilson and Kim, 1994; Kim and Wilson, 1996; Benton and Curran, 2003). We will present studies on the integration of features into the dorsal stream wherein the direction repulsion paradigm is used to distinguish between perceptual alterations in the magnitude of direction repulsion and processing speeds needed to make the perceptual decisions (Perry and Fallah, 2012; Perry et al., 2014). The results provide insight into where features are integrated and when an intermediate object representation is likely to occur.

INTEGRATION OF COLOR

Color is a feature that is processed in the ventral stream through input from parvocellular cells.

Many neuronal studies have found that neurons in the dorsal pathway are not sensitive to color (Maunsell and Van Essen, 1983; Shipp and Zeki, 1985; Zeki et al., 1991; Dobkins and Albright, 1994; Gegenfurtner et al., 1994). In fact, ecologically speaking, color is an irrelevant feature when it comes to processing motion, as in the color of a ball should not matter when attempting to catch it. In spite of this, a number of studies have found that color does in fact alter different aspects of motion processing (Croner and Albright, 1997, 1999; Tchernikov and Fallah, 2010). This would suggest that there is integration of color with motion information in the dorsal stream.

We investigated the effects of color on direction repulsion (**Figure 2**) to determine whether cross-stream feature integration affects direction discrimination, which would support the use of intermediate object representations in motion processing. Two superimposed, coherently moving RDK's were presented, initially for 2000 ms. Each surface could move in one of 12 directions relative to either the vertical or horizontal axes, and both directions created angle differences between the two surfaces ranging between 70° and 110°. If participants correctly determined the directions of both surfaces $\geq$7/8 times, the presentation time decreased, if participants failed to meet this criterion, the time increased. This process continued until participants completed a double reversal. The time needed to process both surfaces correctly (Presentation Time) was estimated to within $\pm$50 ms. Direction repulsion was calculated as being the angle difference between the perceived directions of motion and the actual directions of the surfaces.

If segmenting the two superimposed surfaces by color (**Figure 2B**) reduced direction repulsion, compared to when both surfaces were the same color (**Figure 2A**), this would suggest that color information from the ventral stream is integrated into motion processing in the dorsal stream prior to or at the time that global motion processing is computed, e.g., the stage where mutual inhibition gives rise to repulsion.

Previous work found that when segmenting coherently moving dots of one color from distractor dots of a different color in the same RDK, color acts as a filter that allows for improvements in direction discriminations, behaviorally in humans and animals (Croner and Albright, 1997) and in the responses of area MT neurons (Croner and Albright, 1999). In this case, color would be gated earlier (in V2) allowing for the suppression of distractor colored input to MT. This effectively allows MT to process the coherently moving dots as if they were appearing alone and in turn improves direction computation. Thus when the distractor color is known, color filters can suppress input to motion processing, a finding that has been replicated in superimposed surfaces (Wannig et al., 2007). Based on these findings, we hypothesized that integrating the color with the motion of the two superimposed surfaces might also allow for the surfaces to be individually filtered by color and in turn reduce direction repulsion.

Surprisingly, when selecting between multiple moving surfaces that are different colors, direction discrimination is unchanged from that seen when both surfaces are the same color (**Figure 3A**). Therefore, the global motion processing of a moving RDK is not performed on intermediate object representations, but instead relies on processing the individual motion features. There is however, a large decrease (43% reduction) in the processing time needed to correctly determine both directions of motion. When both surfaces are the same color, processing both directions took almost 1500 ms, but when the surfaces were different colors, processing time was reduced to $\sim$840 ms (**Figure 3B**; Perry and Fallah, 2012). We have suggested previously (Perry and Fallah, 2012; Perry et al., 2014) that it is most likely processing time is reduced through increasing the speed of the decision making process. **Figure 4A** depicts the steps necessary to perform the task of judging the directions of two superimposed surfaces, and the time needed for each step (Perry and Fallah, 2012). The superimposed dot fields are first segmented (SG) into two surfaces, and then the direction of one surface is processed (D1). This would include (**Figure 4B**) sequential recruitment (Nakayama and Silverman, 1984; McKee and Welch, 1985; Mikami et al., 1986), global motion processing, mutual inhibition (repulsion), and information accumulation for decision making (Shadlen and Newsome, 1996; Huk and Shadlen, 2005; Palmer et al., 2005; Zaksas and Pasternak, 2006; Hussar and Pasternak, 2013). Attention is switched (SW) to the second surface, and then the direction of the other surface is processed (D2). When both surfaces are the same color, correctly processing the direction of both surfaces takes more than 1000 ms (**Figure 4A**), but when the surfaces are segmented by color, the direction of both surfaces is correctly processed in under 1000 ms (**Figure 4C**; Perry and Fallah, 2012), a $\sim$650 ms decrease in processing time. It could be that the time needed to segment (SG) the two surfaces is reduced when each surface is a different color. However, as segmentation is speeded by not more than 25 ms in texture-defined objects (Caputo and Casco, 1999) this is unlikely the sole mechanism underlying such a large decrease in processing time. Alternatively, switching

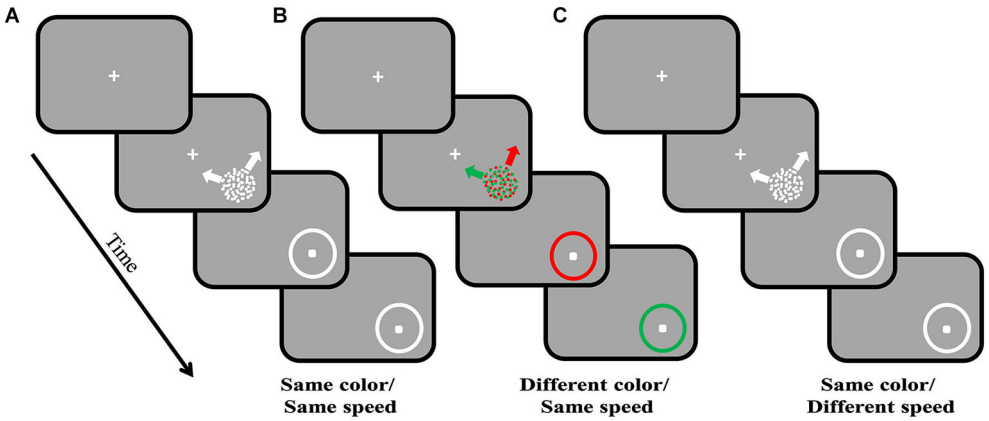


FIGURE 2 | Direction repulsion staircase paradigm. Each trial commences with the appearance of a centrally located fixation point. Once fixation is maintained for 200 ms, the visual stimulus, two superimposed, coherently moving in different directions, random dot kinetograms (RDK's), are presented in the lower right quadrant. In **(A)** the two surfaces are the same color and move at the same speed. In **(B)** the surfaces are the same speed but are different colors, and in **(C)** the surfaces are the same colors but different speeds. The two surfaces are presented for a variable amount of time (staircase procedure). Once they are removed, participants use a mouse to indicate the two directions of motion by clicking on the response circle, once for each direction. In **(B)** participants are required to give the direction for the indicated colored surface in order; the order is randomly assigned between trials. Initially the visual stimulus is presented for 2000 ms, and based on participant's ability to correctly determine both directions of motion, this time will either increase or decrease in successive blocks of trials. Once participants reach a double reversal of presentation times, the time needed to process both directions of motion can be estimated to within ±50 ms. Direction repulsion is calculated as the difference between the angle created by the two clicks on the response circle and the angle created by the two real directions of motion.

attention (SW) between the two surfaces may be speeded when each surface is a different color. Switching attention between serially presented objects in the same location (as in attentional blink) requires only a few hundred milliseconds (Raymond et al., 1992)—but can be attenuated by around 100 ms when targets and probes are less similar (Raymond et al., 1995). Again this mechanism is not sufficient by itself to produce the decrease in processing time. Therefore, there must be a reduction in the time needed to process each direction for such a large decrease in processing time to occur.

In order for color to reduce direction processing time (**Figure 4C**), color input would likely have to affect either the sequential recruitment or decision-making mechanisms including information accumulation (**Figure 4B**) since it does not affect global motion processing (the mutual inhibition circuit). First, MT needs to associate individual dots across two frames (sequential recruitment: Mikami et al., 1986) and pool that information across enough dots (Britten et al., 1992; Snowden et al., 1992) to determine the global motions of the two surfaces. If color worked on sequential recruitment processes, each dot would only need to be compared to dots of the same color across frames, reducing the possibilities by half, speeding up the process immensely. However, by acting as a color filter on sequential recruitment, this color filtering would also be expected to reduce the direction repulsion illusion as each set of colored dots would be processed individually as described earlier (Croner and Albright, 1997, 1999). Instead, there was no change in direction discrimination when two moving surfaces were superimposed (Perry and Fallah, 2012) which indicates that color could not be used to filter out the second surface and reduce the possibilities during sequential recruitment. Alternatively, the integration of color with motion could affect decision-making. Direction discriminations take the information from motion processing in area MT (Albright, 1984; Mikami et al., 1986; Newsome and Paré, 1988; Salzman et al., 1992), and pass it downstream, to areas like LIP, where it is accumulated and a decision threshold reached (Shadlen and Newsome, 1996; Huk and Shadlen, 2005; Zaksas and Pasternak, 2006; Hussar and Pasternak, 2013). If the two surfaces are identical except for their direction of motion, the direction of each surface interferes with the accumulation of direction information for the other surface (**Figure 5A**—Palmer et al., 2005). This interference results in a noisy walk to the decision threshold (accumulator model—Palmer et al., 2005). That is, a decision-making neuron accumulating information to make a decision of rightward motion, would treat input from directional cells preferring rightward motion as positive evidence towards reaching threshold, but input from cells preferring downward motion interferes reducing the accumulated evidence. This produces a noisy walk to threshold. More positive evidence would need to be accumulated before threshold is reached, which means more processing time is needed. With a second feature (color) added to each surface, the two sources of input can be distinguished and selected between. This selection can reduce or eliminate the input from the interfering surface, which reduces the noise in the walk towards the decision threshold, increasing the slope and thus reducing processing time (**Figure 5B**). Therefore, this requires that the accumulation of information for direction discrimination works on intermediate object representations in which color is integrated with motion. This intermediate object representation gives the advantage of allowing for competitive selection of objects (e.g., biased competition: Desimone and Duncan, 1995; Desimone, 1998; Reynolds et al., 2003; Fallah et al., 2007)

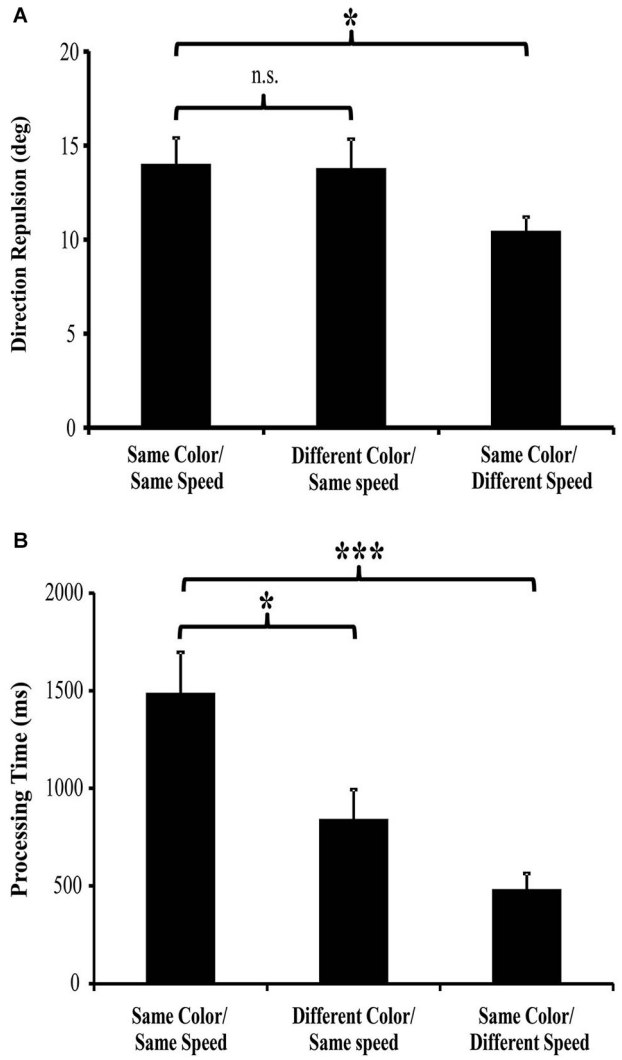


FIGURE 3 | Direction repulsion and processing time results (from Perry and Fallah, 2012; Perry et al., 2014). (A) There was no significant modulation of direction repulsion with the addition of color (Different Color/Same Speed: 13.79° ± 1.54 SEM) when compared to the Same color/Same speed (14.02° ± 1.39 SEM) condition. However, direction repulsion in the Same color/Different speed condition (10.47° ± 0.74 SEM) was significantly less than in the Same color/Same speed condition. **(B)** Processing time in both the Different color/Same speed (842 ms ± 209 SEM) and Same color/Different speed (483 ms ± 81 SEM) conditions was significantly less than in the Same color/Same speed (1488 ms ± 208 SEM) condition.

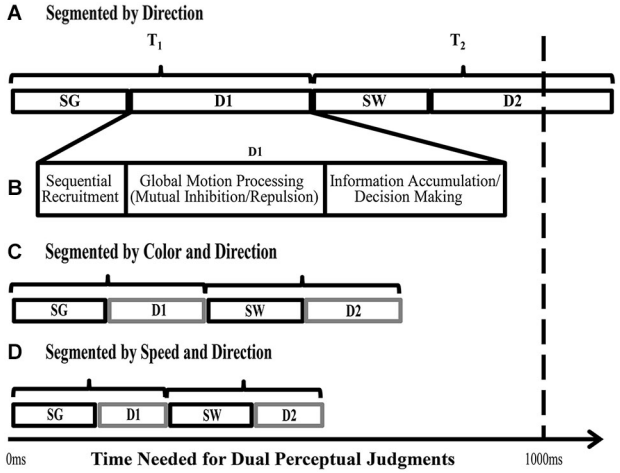


FIGURE 4 | Stages required for direction judgments of two superimposed objects. Based on the task described in **Figure 3**. SG = time needed for *Segmentation* of the two fields of dots into two surfaces, based on different directions of motion, SW = time needed to *Switch* processing from one surface to the other, D1 and D2 = the time needed to process the *Directions* of each superimposed surface (includes sequential recruitment, global motion computation, information accumulation and decision making; shown in detail in **(B)**. **(A)** When the two surfaces differ only in direction, the time needed to complete all the stages involved in the task takes more than 1000 ms on average (Perry and Fallah, 2012). **(B)** Depicts the processes needed to determine the direction of motion of one surface (D1). **(C)** When the surfaces differ in color as well as direction, processing time significantly decreases to less than 1000 ms (Perry and Fallah, 2012). **(D)** When the surfaces differ in speed as well as direction, the time needed to process both directions is reduced further. As the initial segmentation (SG) and attentional switch time (SW) do not appreciably decrease with additional distinguishing features, we propose that the time needed to complete the task decreases as a result of speeded decision making processes (D1 and D2—see text for details) and correspondingly, in **(B)** and **(C)** D1 and D2 are depicted as requiring less time than in **(A)** (adapted from Perry and Fallah, 2012).

at later stages of dorsal stream computations such as decision making.

In summary, changes in processing time, due to speeded decision making processes (as proposed above), with no alteration in direction discrimination, suggest that color is integrated into dorsal stream intermediate object representations after global motion processing. This allows for decision-making processes to use those object representations to reach decision thresholds faster.

INTEGRATION OF SPEED

Unlike with color, previous investigations of direction repulsion have shown that when two superimposed surfaces are of different speeds (Marshak and Sekuler, 1979; Curran and Benton, 2003; Perry et al., 2014) or different spatial frequencies (Kim and Wilson, 1996), direction discrimination improves; direction repulsion is attenuated. Given that spatial frequency, speed and direction are all co-processed within MT (Maunsell and Van Essen, 1983; Albright, 1984; Lagae et al., 1993; Perrone and Thiele, 2001), this is perhaps not surprising. Comparison of movement between two frames give us all three of these features. The spatial location of an object from one frame to the next can be used to calculate direction and speed, while spatial frequency can be extracted from the number of times an object appeared over a given distance. So this information comes in together as a single input and does not require integration; it is inherent based on the movement of the stimulus. Consistent with this, neurons in MT are simultaneously selective for multiple motion features, such as speed and direction. Consequently, a neurons response to one feature (direction for example) can be altered by the response of that same neuron to a different motion feature (such as speed), and as a

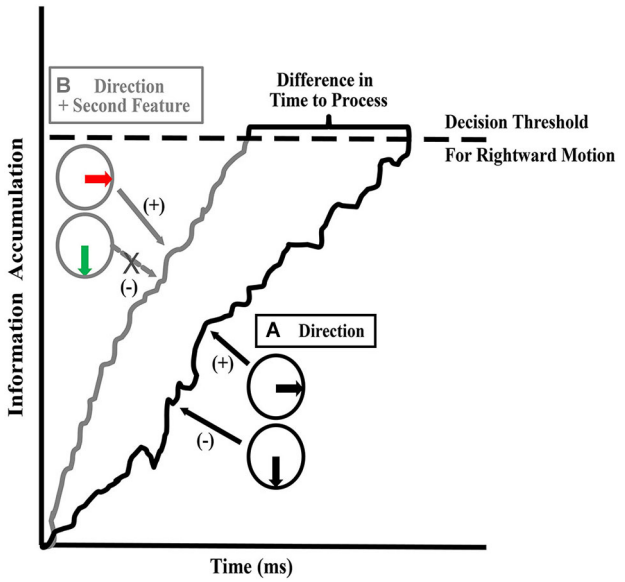


FIGURE 5 | Information accumulation and decision threshold. Hypothesized stage at which processing time is reduced. Areas downstream of MT accumulate motion information in order to arrive at a decision. The figure depicts information accumulation for the rightward direction. When accumulating evidence in support of the rightward direction (+), the evidence is reduced by noise created by the presence of the other surface (–). **(A)** When only direction (one feature) differs between the surfaces, interference between the directions of each surface creates a noisy walk: i.e., incongruent input that reduces the accumulated evidence for the rightward direction. This extends the time needed to reach the decision threshold. **(B)** When direction and a second feature such as color or speed differs between the surfaces, the second feature can be used to reduce the interference caused by the other surface (by allowing competitive selection to override the influence of the second surface) in the walk to threshold, thus reducing the time needed to reach a decision threshold.

result can be considered to be conjoined, i.e., the processing of one feature affects processing of a different feature (Maunsell and Van Essen, 1983; Albright, 1984; Lagae et al., 1993; Perrone and Thiele, 2001). Based on co-processing, motion processing is reflective then of the presented combination of conjoined features. This occurs without the need for a bound object representation. For example, perception of speed can be distorted under a number of different viewing conditions (Krekelberg et al., 2006a,b). A reduction in contrast reduces perceived speed in slow moving stimuli (Thompson, 1982) and increases perceived speed of fast moving stimuli (Thompson et al., 2006). Perceived speed is also dependent upon spatial frequency (Priebe et al., 2003). And finally the perception of direction is sensitive to motion processing conjunctions: direction discrimination becomes more accurate when superimposed surfaces are different speeds (Marshak and Sekuler, 1979; Curran and Benton, 2003; Perry et al., 2014) or different spatial frequencies (Kim and Wilson, 1996).

These examples suggest that direction computation occurs on conjoined dorsal stream features such as direction and speed or direction and spatial frequency information. Using the same paradigm as described in section Integration of color, but with surfaces that are segmented by differences in speed (**Figure 2C**), we tested whether speed, while conjoined with direction for discrimination, could also be used as a distinguishing feature in intermediate object representations like color is (Section Integration of color) and similarly speed up decision making circuitry (Perry et al., 2014). As with color (Perry and Fallah, 2012), we found that differences in the speeds of two superimposed surfaces decreased processing time (**Figure 3B**). In fact, processing time was lower than that seen when the surfaces were segmented by color (Speed-segmented: 483 ms vs. Color-segmented: 841 ms). It could be that velocity, conjoined speed and direction, is the signal that becomes a part of the object representation. If that were the case however, processing time would not be altered as velocity would comprise a single object feature and there would be no other independent feature for use by selection mechanisms to reduce the noise in the walk to threshold (**Figure 5**) and reach a decision threshold more quickly. Instead these results suggest that speed information is treated as an independent feature in an intermediate object representation that is used by decision making circuitry to speed processing times (**Figure 4D**; Perry et al., 2014) similar to the effect of color (Perry and Fallah, 2012). Independent in this case simply means that in spite of the fact that speed is co-processed with direction, and their conjunction attenuates direction repulsion during direction computations, speed *alone* can be utilized as a distinguishing feature to select between the object representations when accumulating information for the perceptual decision.

Unlike the effects of color integration, speed differences reduced direction repulsion which further supports that direction discrimination is modulated by other motion features that are conjoined (processed together) in the dorsal pathway. However, ventral stream features, such as color, do not affect motion until after global motion processing occurs. It has been suggested (Marshak and Sekuler, 1979; Mather and Moulden, 1980) that direction repulsion arises due to inhibitory interactions between populations of neurons, a theory recently formalized (**Figure 6**—adapted from Perry et al., 2014). In mutual inhibition, the responses of neurons to one direction are inhibited by the responses of neurons to a second direction (**Figures 6A,B**) and the amount of inhibition determines the magnitude of direction repulsion. As the angle between the two directions increases, direction repulsion diminishes (Marshak and Sekuler, 1979; Mather and Moulden, 1980) which suggests that mutual inhibition is dependent upon the overlap in tuning between the neurons responding to the two directions (**Figures 6A,B**). When the surfaces are identical except for direction (**Figure 7A**) mutual inhibition and direction repulsion is based solely on the overlap between the tuning curves. Since color is not integrated into motion until after this computation, differences in color do not change the overlap between the two populations and direction repulsion is unaffected (**Figure 7B**). However, when the surfaces are segmented by dorsal stream features such as speed (**Figure 7C**) or spatial frequency (**Figure 7D**) the overlap is reduced due to tuning in multi-dimensional feature space and direction repulsion is decreased. Dorsal stream features are conjoined to produce

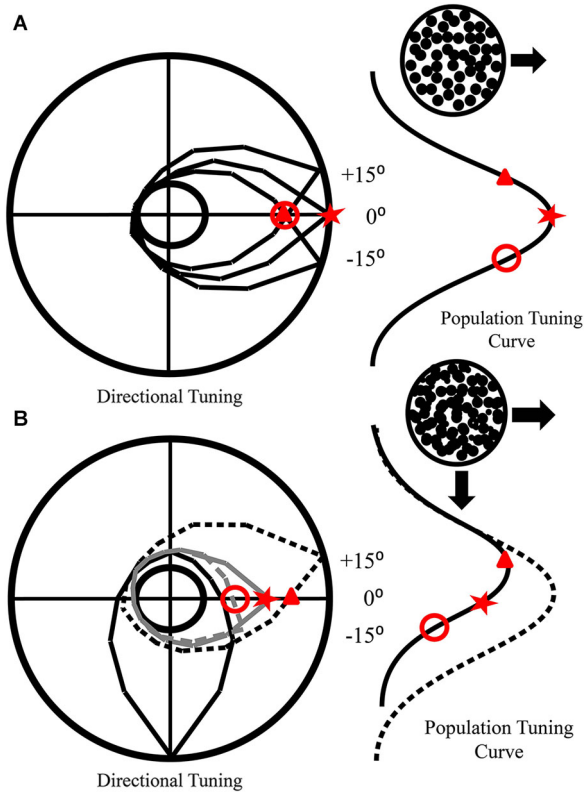


FIGURE 6 | How mutual inhibition produces direction repulsion. **(A)** Individual tuning curves of neurons preferring the direction of motion (0°) and nearby directions (±15°) are presented in the polar plot. The population tuning curve that arises from their responses to rightward motion is also depicted. **(B)** The addition of a second surface moving downwards produces inhibition of other directional neurons. This inhibition drops off as the difference in preferred directions increases. Hence, the −15° neuron is more strongly inhibited (gray dashes) than the 0° neuron (solid gray), while the +15° neuron (black dashes) is not inhibited. This produces a population tuning curve that is shifted away (solid black) from the real direction of motion (dotted line). As the inhibition is mutual, a similar shift would occur for neurons responding to downward motion.

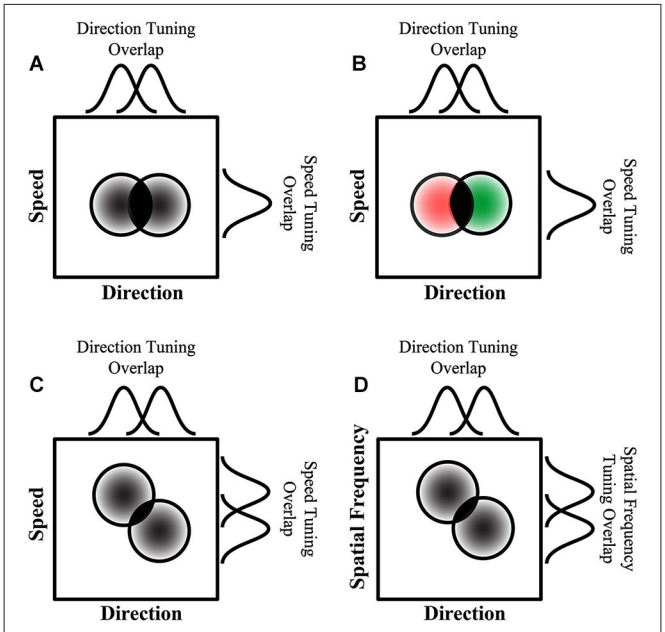


FIGURE 7 | How additional features affect mutual inhibition and direction repulsion. **(A)** Two surfaces that only differ in direction produce direction repulsion whose magnitude is dependent on the area of overlap between their tuning curves (directional tuning curves—top, and two dimensional tuning curves—circles, overlap depicted in solid black). **(B)** When the surfaces are different colors, there is no change in the direction tuning curve overlap which is consistent with color not affecting direction repulsion. However, when a second motion feature that is co-processed with direction, such as speed **(C)** or spatial frequency **(D)**, the population of neurons responding to each direction is segregated based on both features and as a result there is a reduction in the two-dimensional tuning curve overlap (solid black overlap in circles) which results in attenuated direction repulsion (overlaps in **(C)** and **(D)** are smaller than in **(A)** and **(B)**. **(A–D)** The circular plots represent multi-dimensional tuning, while the curves above and to the right of each plot represent the tuning in each dimension respectively (adapted from Perry et al., 2014).

multi-dimensional tuning and thus do not require integration into an object representation. This is supported by the fact that color, which is part of the object, does not affect this circuitry (**Figure 7B**). Overall, as direction repulsion is thought to arise from a local circuit in area MT governing global motion processing, the formation of an intermediate object representation that includes speed and color information likely occurs after that stage.

INTEGRATION OF FORM

Artists have long known how to depict motion in still images using features such as speed-lines (the "wake" of a moving object). These non-moving streaks have been shown to affect human perception of motion (Geisler, 1999; Burr and Ross, 2002) by providing a direction input along the orientation of the streak which can either enhance discrimination of a congruently moving stimulus or interfere with incongruent or orthogonal direction discrimination. This motion streak effect is thought to occur as early as V1, supported by computational (Geisler, 1999) and neurophysiological (Geisler et al., 2001) studies. Thus, speed-lines affect the perception of direction by, in effect, producing motion input for use along the dorsal stream. Similarly, glass patterns, paired dots that appear and disappear randomly on a display, give rise to the perception of bistable directions of motion along the contour of the pattern in the absence of underlying motion signals (Glass, 1969; Ross et al., 2000). These spatial patterns produce motion signals that are represented along with magnocellular motion signals in area MT and ST (Krekelberg et al., 2003), and integrate with real motion signals in perceiving direction (Burr and Ross, 2002).

In essence, these form inputs to the dorsal stream provide the equivalent of motion input to mid-level areas in the dorsal stream starting in area MT (Krekelberg et al., 2003). It is likely that the motion produced by these form inputs is then integrated into the object file as motion features (speed, direction) instead of form features. These effects differ from color which is integrated as its own feature into an intermediate object representation later in the dorsal stream hierarchy. That still leaves an open question as to whether other ventral stream features that do not give rise

to the perception of motion could also be integrated into dorsal stream object files. Other features could be tested with the same direction repulsion paradigm as described earlier. For example, direction repulsion and processing time could be determined for surfaces distinguished by different contrast levels. As the dorsal stream saturates at much lower contrast than the ventral stream (Heuer and Britten, 2002), if decision-making processing time is affected by contrast differences that are above the saturation point for the dorsal stream, then the dorsal stream object file integrates ventral stream contrast information. Additionally, would a size difference between the dots of the two surfaces result in speeded perceptual decision-making similar to the effects of color? The effects of shape (varying the form of the RDK elements, i.e., dots vs. squares vs. triangles) also needs to be tested.

INTERMEDIATE OBJECT REPRESENTATIONS IN THE DORSAL STREAM

Thus far, the evidence presented suggests two main concepts. First, global direction computations are based on the co-processing of dorsal stream motion information. Surfaces segmented by speed or spatial frequency (but not color) result in an improvement in direction computations and thus an attenuation of direction repulsion. Secondly, both speed and color are integrated into a dorsal stream intermediate object representation (or object file) which in turn is used by decision making processes to speed processing times. Speed and direction would need to be independent features in a dorsal stream object file, because this allows for awareness of changes in one dimension independent of the other velocity feature. For example, a moving ball provides velocity information (conjoined speed and direction). If it changes speed but continues to move in the same direction, the population of MT cells that would respond to the conjoined speed/direction selectivity changes. Without independence of these motion features in the object representation, switching underlying MT populations would mark a change in all of the conjoined features. Instead, with independence observers are aware of the speed changing while the direction does not. Thus a dorsal stream object file can denote changes in speed or changes in direction independently. We propose that the dorsal stream object file would also include ventral stream information such as color. Decision-making then works on object files instead of direction information alone, and therefore distinguishing features in the object files can be used to selectively focus decision-making on the relevant direction information.

The features that are placed in the object file are dependent upon which features are important to completing the specified task (Harel et al., 2014). Theoretically then, using the direction repulsion paradigm as an example, task relevant would mean that any feature that distinguished the two superimposed surfaces from each other would be a feature added to the object file. This is what occurred with both speed and color, and therefore it would be logical to extrapolate that other task relevant features would also be included in an object file. We have previously suggested (Section Integration of form) how other form features, such as size, shape and contrast, could be tested for integration into a dorsal stream object file.

We propose that global motion processing occurs on conjoined motion features such as speed and direction, whereas the accumulation of perceptual information to reach a decision is performed on intermediate object representations. While these hypotheses are yet to be directly tested at the neurophysiological level (e.g., in animal models), in the next section we propose the likely neural substrates and dorsal stream areas subserving each of these processes, based on known properties of these areas.

POSSIBLE LOCATION OF OBJECT REPRESENTATIONS IN THE DORSAL STREAM

Figure 8 provides an overview of processing along both the ventral and dorsal pathways with known object representations in the ventral stream and hypothesized object representations in the dorsal stream. Given that object files are considered to be mid-level representations, and are found at intermediate stages of ventral stream processing, they should similarly be found in and around area MT in the dorsal stream. Perceived color is processed in area V4, and thus color processing would need to reach this stage before being incorporated into an object representation in either the ventral or dorsal stream. Color is not integrated with direction prior to direction computation circuits in MT as the addition of color did not reduce direction repulsion. However, color and speed *did* reduce the time needed to fully process both directions of motion. Therefore while global motion direction computations which are computed in area MT are not performed on object files, color and speed are integrated into an object file after direction computation in MT.

Evidence of motion computations relying on object representations comes from smooth pursuit. Color is known to affect smooth pursuit eye movements to moving surfaces (Tchernikov and Fallah, 2010) which are dependent upon the processing of velocity signals for both the surface and the background in area MST (Dürsteler and Wurtz, 1988; Komatsu and Wurtz, 1988, 1989; Thier and Erickson, 1992; Ilg, 2008). Intuitively, eye movements should be color blind. Instead color biases selection of one superimposed surface over the other based on a color hierarchy, and the competition between the two colored surfaces modulates the speed of pursuit (Tchernikov and Fallah, 2010). This suggests that it is not only the reaching and grasping systems later in the dorsal stream that work on object features, but as part of the vision for action pathway, smooth pursuit computations are based on object files. Thus the integration of color into the dorsal stream object file may occur as early as area MST, or at least before the frontal eye fields (FEF) generate the motor plan.

After MST in the dorsal stream, area LIP in the parietal lobe has been shown to be involved in the accumulation of motion information for perceptual decision-making (Shadlen and Newsome, 1996; Huk and Shadlen, 2005; Palmer et al., 2005). This stage of processing works on object files as color and speed differences reduce the time needed to reach the decision threshold. Beyond this stage, a number of areas in the posterior parietal cortex are selective for objects, a function necessary for visuomotor guidance of grasping. Such object selectivity has been found in areas anterior intraparietal (AIP) and 7a (Taira et al., 1990; Murata et al., 2000; Phinney and Siegel, 2000).

This hypothetical framework for object representations in the dorsal stream (**Figure 8**) can be tested in future neurophysiological studies. Specifically, global motion processing in area MT neurons and the concomitant direction repulsion of the population tuning should not be affected by the addition of color differences. Whereas responses of neurons in area MST that give rise to pursuit motion should be modulated by the color differences in superimposed surfaces (Tchernikov and Fallah, 2010). Finally, decision-making neurons in area LIP should show steeper slopes and reach decision thresholds faster when a second distinguishing feature such as color or speed is present.

OTHER EVIDENCE FOR DORSAL STREAM OBJECT REPRESENTATIONS

Other studies have shown selection of objects in the dorsal stream that upon reflection would support intermediate object representations. For example, judging the direction of a brief translation of one of two counter-rotating superimposed surfaces is improved when that surface is selected by color (Valdes-Sosa et al., 2000), an effect the authors attributed to the use of object files by the dorsal stream. The different motions between the two surfaces provides noise in accumulating direction information, but reducing noise through selection of that object file would speed processing such that the decision threshold could be reached during the brief translation period. Similarly, if the object file is selected by a transient motion feature capturing attention, selection of that object file is maintained and again improves the discrimination of a subsequent brief translation (Reynolds et al., 2003) along with modulating the visually evoked N1 component, a marker of selective attention (Pinilla et al., 2001; Khoe et al., 2005). In fact, when one of two superimposed surfaces is selected by a color segmentation cue, the selective advantage for processing brief translations of that surface survives the removal of color differences (Mitchell et al., 2003), once again showing that selection is maintained via an object file. In fact, concurrent judgments of simple form (square or circle) and motion are impaired when made across two superimposed surfaces compared to when they are made for the same surface (Rodríguez et al., 2002). This is similar to Duncan (1984), which showed that attending to an object representation allows judgments of multiple ventral stream form features "for free" but that there was a cost associated with having to make judgments across two superimposed objects. Together, these studies suggest that there are also object representations in the later stages of the dorsal stream. Furthermore, competitive selection processes work not only on objects in the ventral stream (Desimone, 1998; Reynolds et al., 2003; Fallah et al., 2007), but also on objects in the dorsal stream.

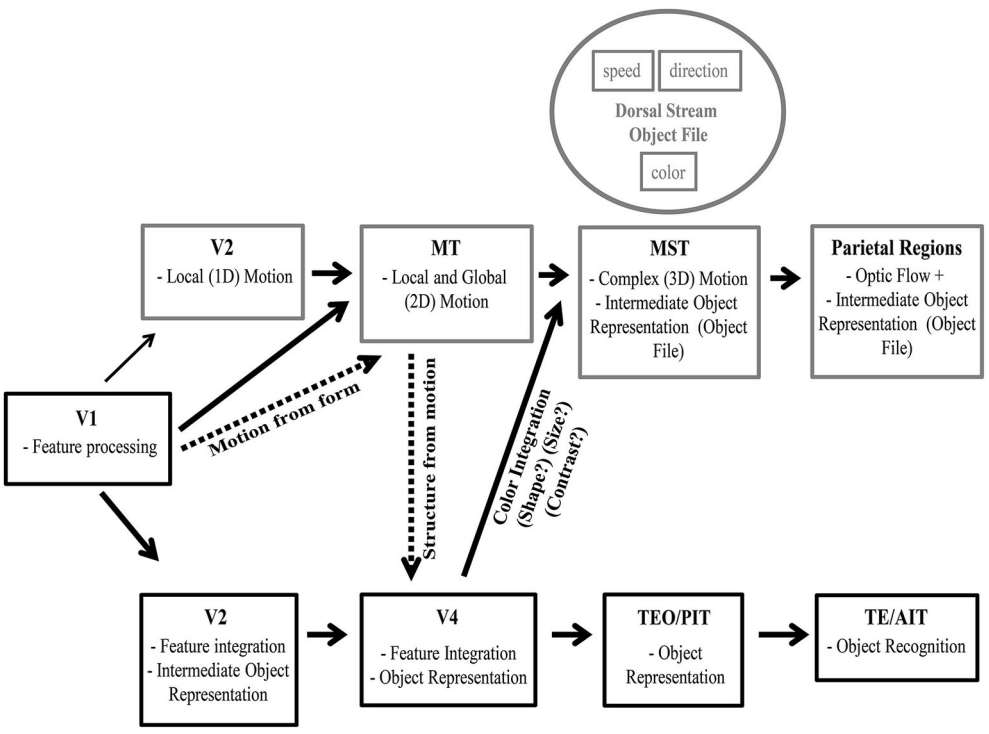


FIGURE 8 | Intermediate object representation model. Visual processing along the ventral stream is depicted along with known object representations starting in area V2. We also depict visual processing along the dorsal stream with the hypothetical stages which process dorsal stream object files. As visual processing progresses along the dorsal pathway stimulus parameters are calculated and this information is provided to area MT. In MT, information regarding speed, direction and spatial frequency are co-processed forming multidimensional selectivity. After local and global motion processing circuits in MT, an intermediate object representation is formed that incorporates independent motion features (such as speed and direction) and ventral stream features (such as color, with other features such as shape and size to be determined). This intermediate object representation is in place prior to decision making circuitry that represents motion or guides action.

VISION FOR ACTION

The dorsal stream object representation would not need to progress to the level of object recognition however. As already discussed, the vision for action theory states that the dorsal pathway's reaching and grasping system uses object features as a means of guiding action in real time. With damage to the ventral stream, patients can still orient their hand and scale their grip according to the orientation and shape of the item to be grasped. This does not require that the object is fully processed through to recognition, just that a list of features associated with a specific object be available for selection (Freiwald, 2007). An object file would provide such a list from which different features could be used to select the correct object among multiple, even superimposed, objects (Valdes-Sosa et al., 1998, 2000; Pinilla et al., 2001; Wannig et al., 2007; Perry and Fallah, 2012; Perry et al., 2014).

DORSAL TO VENTRAL INTEGRATION

Our proposal is that the dorsal stream integrates features, from both the dorsal and ventral pathways, into an object representation that can be used by decision making circuitry (contained within the dorsal stream) for selection purposes. A similar process occurs in the ventral stream, and it is not only features processed within the ventral stream that are integrated to form object representations used in object recognition and decision making. As early as V4, motion information from the dorsal pathway is used to define stationary edges that occur between moving stimuli (kinetic boundaries—Mysore et al., 2006). However, MT also plays a role in segmentation mechanisms (Born and Bradley, 2005) as a necessary component of surface reconstruction (Andersen and Bradley, 1998). This is what allows MT to separate the motion of multiple moving stimuli from each other (Snowden et al., 1991; Stoner and Albright, 1996), even under conditions of occlusion (Nowlan and Sejnowski, 1995), and to separate moving objects from background (Bradley and Andersen, 1998; Born and Bradley, 2005). Similarly, superimposed dots patterns, moving in opposite directions and moving at variable speeds can be integrated to create a percept of a rotating cylinder. This indicates that processing along the dorsal pathway also allows for perception of 3D structures (Bradley et al., 1998; Dodd et al., 2001). Moving dots are also known to give rise to human shape percepts. Moreover, this perception of biological motion goes beyond shape and form processing. Higher order features, such as gender, are also derived from biological motion (Barclay et al., 1978; Mather and Murdoch, 1994; Jordan et al., 2006). As gender is derived from the global, not local motion, and gender adapts with prolonged exposure to biological motion (Jordan et al., 2006), this occurs at a stage beyond area MT. Biological motion is represented in the superior temporal polysensory area (STP: Perrett et al., 1989) and as such is an object representation later along the dorsal stream, which gives rise to gender representation.

ALTERNATIVE LOCATION FOR THE OBJECT REPRESENTATION

While evidence supports the dorsal stream decision-making processes working on object representations, the site for these representations are unknown. We have suggested that intermediate object representations are built up at later stages in the dorsal stream (**Figure 8**). However, these decision making circuits in the dorsal stream could instead be modulated by object representations contained in the ventral pathway.

For this to occur, motion information would have to be a tag (e.g., Finger of INSTantiation (FINST): Pylyshyn, 1989, 1994) associated with object processing in the ventral stream, which would then have to be passed back to the dorsal stream in time for direction decisions to be made. While this is possible, Occam's razor suggests the more parsimonious explanation of dorsal stream object files is likely the correct one. There is a means of testing whether intermediate object representations occur in the dorsal stream. As visual agnosiacs have damage to the ventral stream but retain certain form information used to guide grasps, they could be tested to see whether motion decision-making could be sped up without ventral stream object representations. If so, then there must be dorsal stream intermediate object representations separate from those in the ventral stream. Such intermediate object representations would not give rise to recognition but would incorporate the form features maintained in the dorsal stream to provide real-time visual guidance for actions such as hand orientation, grip scaling, and pincer grip locations (Goodale et al., 1991, 1994; Milner et al., 2012). Note that even if the intermediate object representation was to be created in the ventral stream, it would still be used by decision-making areas in the dorsal stream. The areas that give rise to the object representation would change, but the later stages of dorsal stream processing would still be dependent on object representations, not just motion information.

CONCLUSIONS

We have provided a framework for not only how the dorsal stream extracts motion information but also builds up an object representation that is used in decision making processes. The hierarchical nature of visual processing, in both the ventral and dorsal pathways, provides the basis for where an object representation in the dorsal pathway would exist. Both color and speed information, as independent object features, are integrated into motion processing circuits beyond direction computations (such as in area MT) and prior to decision making and attentional selection (such as in area LIP). In fact, color-dependent smooth pursuit may indicate an intermediate object representation occurs as early as area MST. It is also likely that later parietal areas that guide grasping, such as AIP, may also contain the requisite circuitry for intermediate object representations in the dorsal stream. We have suggested that this object representation would not give rise to object recognition as in the ventral stream but instead would contain a list of object features upon which decisions could be made and actions performed. Object files are a possible mechanism through which information necessary for dorsal stream decision making and selection could be collected and updated as needed. The use of dorsal stream information for the creation of objects in the ventral pathway supports our proposal of parallel mechanisms existing in the dorsal stream. Testing visual agnosiacs on dorsal stream decision making, requiring the

use of object representations, would be a way to determine if the dorsal pathway alone can support these intermediate object representations.

ACKNOWLEDGMENTS

The authors are supported by funds provided by an NSERC Discovery Grant to Mazyar Fallah and an NSERC Alexander Graham Bell Canadian Doctoral Scholarship to Carolyn J. Perry.

REFERENCES

Adelson, E. H., and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* 2, 284–299. doi: 10.1364/josaa.2.000284

Adelson, E. H., and Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature* 300, 523–525. doi: 10.1038/300523a0

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.* 52, 1106–1130.

Andersen, R. A., and Bradley, D. C. (1998). Perception of three-dimensional structure from motion. *Trends Cogn. Sci.* 2, 222–228. doi: 10.1016/s1364-6613(98)01181-4

Barclay, C. D., Cutting, J. E., and Kozlowski, L. T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Percept. Psychophys.* 23, 145–152. doi: 10.3758/bf03208295

Benton, C. P., and Curran, W. (2003). Direction repulsion goes global. *Curr. Biol.* 13, 767–771. doi: 10.1016/s0960-9822(03)00285-9

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295x.94.2.115

Biederman, I., and Cooper, E. E. (1991). Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cogn. Psychol.* 23, 393–419. doi: 10.1016/0010-0285(91)90014-f

Bodelón, C., Fallah, M., and Reynolds, J. H. (2007). Temporal resolution for the perception of features and conjunctions. *J. Neurosci.* 27, 725–730. doi: 10.1523/jneurosci.3860-06.2007

Born, R. T., and Bradley, D. C. (2005). Structure and function of visual area MT. *Annu. Rev. Neurosci.* 28, 157–189. doi: 10.1146/annurev.neuro.26.041002.131052

Braddick, O. J., Wishart, K. A., and Curran, W. (2002). Directional performance in motion transparency. *Vision Res.* 42, 1237–1248. doi: 10.1016/s0042-6989(02)00018-4

Bradley, D. C., and Andersen, R. A. (1998). Center-surround antagonism based on disparity in primate area MT. *J. Neurosci.* 18, 7552–7565.

Bradley, D. C., Chang, G. C., and Andersen, R. A. (1998). Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* 392, 714–717. doi: 10.1038/33688

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* 12, 4745–4765.

Brooks, K. R., Morris, T., and Thompson, P. (2011). Contrast and stimulus complexity moderate the relationship between spatial frequency and perceived speed: implications for MT models of speed perception. *J. Vis.* 11:19. doi: 10.1167/11.14.19

Burr, D., and Thompson, P. (2011). Motion psychophysics: 1985–2010. *Vision Res.* 51, 1431–1456. doi: 10.1016/j.visres.2011.02.008

Burr, D. C., and Ross, J. (2002). Direct evidence that "speedlines" influence motion mechanisms. *J. Neurosci.* 22, 8661–8664.

Caputo, G., and Casco, C. (1999). A visual evoked potential correlate of global figure-ground segmentation. *Vision Res.* 39, 1597–1610. doi: 10.1016/s0042-6989(98)00270-3

Cavanagh, P., and Mather, G. (1989). Motion: the long and short of it. *Spat. Vis.* 4, 103–129. doi: 10.1163/156856889x00077

Cavanagh, P., Tyler, C. W., and Favreau, O. E. (1984). Perceived velocity of moving chromatic gratings. *J. Opt. Soc. Am. A* 1, 893–899. doi: 10.1364/josaa.1.000893

Chen, X., DeAngelis, G. C., and Angelaki, D. E. (2013). Eye-centered representation of optic flow tuning in the ventral intraparietal area. *J. Neurosci.* 33, 18574–18582. doi: 10.1523/jneurosci.2837-13.2013

Chicherov, V., Plomp, G., and Herzog, M. H. (2014). Neural correlates of visual crowding. *Neuroimage* 93(Pt. 1), 23–31. doi: 10.1016/j.neuroimage.2014.02.021

Conway, B. R., and Livingstone, M. S. (2006). Spatial and temporal properties of cone signals in alert macaque primary visual cortex. *J. Neurosci.* 26, 10826–10846. doi: 10.1523/jneurosci.2091-06.2006

Conway, B. R., Moeller, S., and Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron* 56, 560–573. doi: 10.1016/j.neuron.2007.10.008

Cowey, A., and Weiskrantz, L. (1967). A comparison of the effects of inferotemporal and striate cortex lesions on the visual behaviour of rhesus monkeys. *Q. J. Exp. Psychol.* 19, 246–253. doi: 10.1080/14640746708400099

Croner, L. J., and Albright, T. D. (1997). Image segmentation enhances discrimination of motion in visual noise. *Vision Res.* 37, 1415–1427. doi: 10.1016/s0042-6989(96)00299-4

Croner, L. J., and Albright, T. D. (1999). Segmentation by color influences responses of motion-sensitive neurons in the cortical middle temporal visual area. *J. Neurosci.* 19, 3935–3951.

Curran, W., and Benton, C. P. (2003). Speed tuning of direction repulsion describes an inverted U-function. *Vision Res.* 43, 1847–1853. doi: 10.1016/s0042-6989(03)00302-x

Dean, P. (1976). Effects of inferotemporal lesions on the behavior of monkeys. *Psychol. Bull.* 83, 41–71. doi: 10.1037//0033-2909.83.1.41

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1245–1255. doi: 10.1098/rstb.1998.0280

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4, 2051–2062.

Desimone, R., and Duncan, J. (1995). Neural mechanism of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

DeYoe, E. A., and Van Essen, D. C. (1985). Segregation of efferent connections and receptive field properties in visual area V2 of the macaque. *Nature* 317, 58–61. doi: 10.1038/317058a0

Dobkins, K. R., and Albright, T. D. (1994). What happens if it changes color when it moves?: the nature of chromatic input to macaque visual area MT. *J. Neurosci.* 14, 4854–4870.

Dodd, J. V., Krug, K., Cumming, B. G., and Parker, A. J. (2001). Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *J. Neurosci.* 21, 4809–4821.

Duffy, C. J., and Wurtz, R. H. (1991a). Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *J. Neurophysiol.* 65, 1329–1345.

Duffy, C. J., and Wurtz, R. H. (1991b). Sensitivity of MST neurons to optic flow stimuli. II. Mechanisms of response selectivity revealed by small-field stimuli. *J. Neurophysiol.* 65, 1346–1359.

Duffy, C. J., and Wurtz, R. H. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *J. Neurosci.* 15(7 Pt. 2), 5192–5208.

Durand, J. B., Nelissen, K., Joly, O., Wardak, C., Todd, J. T., Norman, J. F., et al. (2007). Anterior regions of monkey parietal cortex process visual 3D shape. *Neuron* 55, 493–505. doi: 10.1016/j.neuron.2007.06.040

Duncan, J. (1984). Selective attention and the organization of visual information. *J. Exp. Psychol. Gen.* 113, 501–517. doi: 10.1037/0096-3445.113.4.501

Dürsteler, M. R., and Wurtz, R. H. (1988). Pursuit and optokinetic deficits following chemical lesions of cortical areas MT and MST. *J. Neurophysiol.* 60, 940–965.

Fallah, M., and Reynolds, J. H. (2001). Attention! V1 neurons lining up for inspection. *Neuron* 31, 674–675. doi: 10.1016/s0896-6273(01)00425-1

Fallah, M., Stoner, G. R., and Reynolds, J. H. (2007). Stimulus-specific competitive selection in macaque extrastriate visual area V4. *Proc. Natl. Acad. Sci. U S A* 104, 4165–4169. doi: 10.1073/pnas.0611722104

Farivar, R. (2009). Dorsal-ventral integration in object recognition. *Brain Res. Rev.* 61, 144–153. doi: 10.1016/j.brainresrev.2009.05.006

Fattori, P., Kutz, D. F., Breveglieri, R., Marzocchi, N., and Galletti, C. (2005). Spatial tuning of reaching activity in the medial parieto-occipital cortex (area V6A) of macaque monkey. *Eur. J. Neurosci.* 22, 956–972. doi: 10.1111/j.1460-9568.2005.04288.x

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Ferrera, P., Nealey, T. A., and Maunsell, J. H. R. (1992). Mixed parvocellular and magnocellular geniculate signals in visual area V4. *Nature* 358, 756–758. doi: 10.1038/358756a0

Ferrera, P., Nealey, T. A., and Maunsell, J. H. R. (1994). Responses in macaque visual area V4 following inactivation of the parvocellular and magnocellular LGN pathways. *J. Neurosci.* 14, 2080–2088.

Freiwald, W. A. (2007). Attention to objects made of features. *Trends Cogn. Sci.* 11, 453–454. doi: 10.1016/j.tics.2007.08.010

Fukushima, K. (1975). Cognitron: a self-organizing multilayered neural network. *Biol. Cybern.* 20, 121–136. doi: 10.1007/bf00342633

Gallese, V., Murata, A., Kaseda, M., Niki, N., and Sakata, H. (1994). Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport* 5, 1525–1529. doi: 10.1097/00001756-199407000-00029

Gegenfurtner, K. R., Kiper, D. C., and Levitt, J. B. (1997). Functional properties of neurons in macaque area V3. *J. Neurophysiol.* 77, 1906–1923.

Gegenfurtner, K. R., Kiper, D. C., Beusmans, J. M., Carandini, M., Zaidi, Q., and Movshon, J. A. (1994). Chromatic properties of neurons in macaque MT. *Vis. Neurosci.* 11, 455–466. doi: 10.1017/s095252380000239x

Geisler, W. S. (1999). Motion streaks provide a spatial code for motion direction. *Nature* 400, 65–69. doi: 10.1038/21886

Geisler, W. S., Albrecht, D. G., Crane, A. M., and Stern, L. (2001). Motion direction signals in the primary visual cortex of cat and monkey. *Vis. Neurosci.* 18, 501–516. doi: 10.1017/s0952523801184014

Glass, L. (1969). Moiré effect from random dots. *Nature* 223, 578–580. doi: 10.1038/223578a0

Goodale, M. A. (2008). Action without perception in human vision. *Cogn. Neuropsychol.* 25, 891–919. doi: 10.1080/02643290801961984

Goodale, M. A. (2013). Separate visual systems for perception and action: a framework for understanding cortical visual impairment. *Dev. Med. Child Neurol.* 55(Suppl. 4), 9–12. doi: 10.1111/dmcn.12299

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Goodale, M. A., Meenan, J. P., Bulthoff, H. H., Nicolle, D. A., Murphy, K. J., and Racicot, C. I. (1994). Separate neural pathways for the visual analysis of object shape in perception and prehension. *Curr. Biol.* 4, 604–610. doi: 10.1016/s0960-9822(00)00132-9

Goodale, M. A., Milner, A. D., Jakobson, L. S., and Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature* 349, 154–156. doi: 10.1038/34914a0

Graziano, M. S., Andersen, R. A., and Snowden, R. J. (1994). Tuning of MST neurons to spiral motions. *J. Neurosci.* 14, 54–67.

Gross, C. G. (2008). Single neuron studies of inferior temporal cortex. *Neuropsychologia* 46, 841–852. doi: 10.1016/j.neuropsychologia.2007.11.009

Gross, C. G., Cowey, A., and Manning, F. J. (1971). Further analysis of visual discrimination deficits following foveal prestriate and inferotemporal lesions in rhesus monkeys. *J. Comp. Physiol. Psychol.* 76, 1–7. doi: 10.1037/h0031039

Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35, 96–111.

Grunewald, A., Bradley, D. C., and Andersen, R. A. (2002). Neural correlates of structure-from-motion perception in macaque V1 and MT. *J. Neurosci.* 22, 6195–6207.

Gu, Y., Watkins, P. V., Angelaki, D. E., and DeAngelis, G. C. (2006). Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *J. Neurosci.* 26, 73–85. doi: 10.1523/jneurosci.2356-05.2006

Harel, A., Kravitz, D. J., and Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proc. Natl. Acad. Sci. U S A* 111, E962–E971. doi: 10.1073/pnas.1312567111

Heuer, H. W., and Britten, K. H. (2002). Contrast dependence of response normalization in area MT of the rhesus macaque. *J. Neurophysiol.* 88, 3398–3408. doi: 10.1152/jn.00255.2002

Hiris, E., and Blake, R. (1996). Direction repulsion in motion transparency. *Vis. Neurosci.* 13, 187–197. doi: 10.1017/s0952523800007227

Hubel, D. H., and Livingstone, M. S. (1987). Segregation of form, color and stereopsis in primate area 18. *J. Neurosci.* 7, 3378–3415.

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hubel, D. H., Wiesel, T. N., and Stryker, M. P. (1978). Anatomical demonstration of orientation columns in macaque monkey. *J. Comp. Neurol.* 177, 361–380. doi: 10.1002/cne.901770302

Huberle, E., Rupek, P., Lappe, M., and Karnath, H. O. (2012). Perception of biological motion in visual agnosia. *Front. Behav. Neurosci.* 6:56. doi: 10.3389/fnbeh.2012.00056

Huk, A. C., and Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.* 25, 10420–10436. doi: 10.1523/jneurosci.4684-04.2005

Hussar, C. R., and Pasternak, T. (2013). Common rules guide comparisons of speed and direction of motion in the dorsolateral prefrontal cortex. *J. Neurosci.* 33, 972–986. doi: 10.1523/jneurosci.4075-12.2013

Ilg, U. J. (2008). The role of areas MT and MST in coding of visual motion underlying the execution of smooth pursuit. *Vision Res.* 48, 2062–2069. doi: 10.1016/j.visres.2008.04.015

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percep. Psychophys.* 14, 201–211. doi: 10.3758/bf03212378

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception: an experimental and theoretical analysis of calculus-like functions in visual data processing. *Pysch. Res.* 38, 379–393. doi: 10.1007/bf00309043

Jordan, H., Fallah, M., and Stoner, G. R. (2006). Adaptation of gender derived from biological motion. *Nat. Neurosci.* 9, 738–739. doi: 10.1038/nn1710

Kahneman, D., Treisman, A., and Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cogn. Psychol.* 24, 175–219. doi: 10.1016/0010-0285(92)90007-o

Khoe, W., Mitchell, J. F., Reynolds, J. H., and Hillyard, S. A. (2005). Exogenous attentional selection of transparent superimposed surfaces modulates early event-related potentials. *Vision Res.* 45, 3004–3014. doi: 10.1016/j.visres.2005.04.021

Kim, J., and Wilson, H. R. (1996). Direction repulsion between components in motion transparency. *Vision Res.* 36, 1177–1187. doi: 10.1016/0042-6989(95)00153-0

Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* 71, 856–867.

Komatsu, H., and Wurtz, R. H. (1988). Relation of cortical areas MT and MST to pursuit eye movements. III. Interaction with full-field visual stimulation. *J. Neurophysiol.* 60, 621–644.

Komatsu, H., and Wurtz, R. H. (1989). Modulation of pursuit eye movements by stimulation of cortical areas MT and MST. *J. Neurophysiol.* 62, 31–47.

Konen, C. S., and Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nat. Neurosci.* 11, 224–231. doi: 10.1038/nn2036

Krekelberg, B., and Albright, T. D. (2005). Motion mechanisms in macaque MT. *J. Neurophysiol.* 93, 2908–2921. doi: 10.1152/jn.00473.2004

Krekelberg, B., Dannenberg, S., Hoffmann, K. P., Bremmer, F., and Ross, J. (2003). Neural correlates of implied motion. *Nature* 424, 674–677. doi: 10.1038/nature01852

Krekelberg, B., and van Wezel, R. J. (2013). Neural mechanisms of speed perception: transparent motion. *J. Neurophysiol.* 110, 2007–2018. doi: 10.1152/jn.00333.2013

Krekelberg, B., van Wezel, R. J., and Albright, T. D. (2006a). Adaptation in macaque MT reduces perceived speed and improves speed discrimination. *J. Neurophysiol.* 95, 255–270. doi: 10.1152/jn.00750.2005

Krekelberg, B., van Wezel, R. J., and Albright, T. D. (2006b). Interactions between speed and contrast tuning in the middle temporal area: implications for the neural code for speed. *J. Neurosci.* 26, 8988–8998. doi: 10.1523/jneurosci.1983-06.2006

Lagae, L., Raiguel, S., and Orban, G. A. (1993). Speed and direction selectivity of macaque middle temporal neurons. *J. Neurophysiol.* 69, 19–39.

Landisman, C. E., and Ts'o, D. Y. (2002). Color processing in macaque striate cortex: electrophysiological properties. *J. Neurophysiol.* 87, 3138–3151. doi: 10.1152/jn.00957.1999

Lehky, S. R., and Sereno, A. B. (2007). Comparison of shape encoding in primate dorsal and ventral visual pathways. *J. Neurophysiol.* 97, 307–319. doi: 10.1152/jn.00168.2006

Levitt, J. B., Kiper, D. C., and Movshon, J. A. (1994). Receptive fields and functional architecture of macaque V2. *J. Neurophysiol.* 71, 2517–2542.

Li, P., Zhu, S., Chen, M., Han, C., Xu, H., Hu, J., et al. (2013). A motion direction preference map in monkey V4. *Neuron* 78, 376–388. doi: 10.1016/j.neuron.2013.02.024

Livingstone, M. S., and Hubel, D. H. (1987). Connections between layer 4B of area 17 and the thick cytochrome oxidase stripes of area 18 in the squirrel monkey. *J. Neurosci.* 7, 3371–3377.

Livne, T., and Sagi, D. (2007). Configuration influence on crowding. *J. Vis.* 7, 1–12. doi: 10.1167/7.2.4

Logothetis, N. K., and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex* 5, 270–288. doi: 10.1093/cercor/5.3.270

Lowe, D. G. (1987). The viewpoint consistency constraint. *Int. J. Comput. Vis.* 1, 57–72. doi: 10.1007/bf00128526

Lu, H. D., Chen, G., Tanigawa, H., and Roe, A. W. (2010). A motion direction map in macaque V2. *Neuron* 68, 1002–1013. doi: 10.1016/j.neuron.2010.11.020

Malania, M., Herzog, M. H., and Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *J. Vis.* 7, 1–7. doi: 10.1167/7.2.1

Manassi, M., Sayim, B., and Herzog, M. H. (2012). Grouping, pooling and when bigger is better in visual crowding. *J. Vis.* 12:13. doi: 10.1167/12.10.13

Marr, D., and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B Biol. Sci.* 200, 269–294. doi: 10.1098/rspb.1978.0020

Marr, D., and Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proc. R. Soc. Lond. B Biol. Sci.* 211, 151–180. doi: 10.1098/rspb.1981.0001

Marshak, W., and Sekuler, R. (1979). Mutual repulsion between moving visual targets. *Science* 205, 1399–1401. doi: 10.1126/science.472756

Mather, G., and Moulden, B. (1980). A simultaneous shift in apparent direction: further evidence for a "distribution-shift" model of direction coding. *Q. J. Exp. Psychol.* 32, 325–333. doi: 10.1080/14640748008401168

Mather, G., and Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. B Biol. Sci.* 258, 273–279. doi: 10.1098/rspb.1994.0173

Maunsell, J. H., and Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed and orientation. *J. Neurophysiol.* 49, 1127–1147.

McKee, S. P., and Welch, L. (1985). Sequential recruitment in the discrimination of velocity. *J. Opt. Soc. Am. A* 2, 243–251. doi: 10.1364/josaa.2.000243

Mikami, A., Newsome, W. T., and Wurtz, R. H. (1986). Motion selectivity in macaque visual cortex. I. Mechanisms of direction and speed selectivity in extrastriate area MT. *J. Neurophysiol.* 55, 1308–1327.

Milner, A. D., Ganel, T., and Goodale, M. A. (2012). Does grasping in patient D.F. depend on vision? *Trends Cogn. Sci.* 16, 256–257. doi: 10.1016/j.tics.2012.03.004

Mineault, P. J., Khawaja, F. A., Butts, D. A., and Pack, C. C. (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proc. Natl. Acad. Sci. U S A* 109, E972–E980. doi: 10.1073/pnas.1115685109

Mitchell, J. F., Stoner, G. R., Fallah, M., and Reynolds, J. H. (2003). Attentional selection of superimposed surfaces cannot be explained by modulation of the gain of color channels. *Vision Res.* 43, 1323–1328. doi: 10.1016/s0042-6989(03)00123-8

Mitroff, S. R., Arita, J. T., and Fleck, M. S. (2009). Staying in bounds: contextual constraints on object-file coherence. *Vis. Cogn.* 17, 195–211. doi: 10.1037/e527342012-017

Mitroff, S. R., Scholl, B. J., and Noles, N. S. (2007). Object files can be purely episodic. *Perception* 36, 1730–1735. doi: 10.1068/p5804

Mitroff, S. R., Scholl, B. J., and Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition* 96, 67–92. doi: 10.1016/j.cognition.2004.03.008

Movshon, J. A., and Newsome, W. T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *J. Neurosci.* 16, 7733–7741.

Murata, A., Gallese, V., Luppino, G., Kaseda, M., and Sakata, H. (2000). Selectivity for the shape, size and orientation of objects for grasping in neurons of monkey parietal area AIP. *J. Neurophysiol.* 83, 2580–2601.

Mysore, S. G., Vogels, R., Raiguel, S. E., and Orban, G. A. (2006). Processing of kinetic boundaries in macaque V4. *J. Neurophysiol.* 95, 1864–1880. doi: 10.1152/jn.00627.2005

Nakayama, K., and Silverman, G. H. (1984). Temporal and spatial characteristics of the upper displacement for motion in random dots. *Vision Res.* 24, 293–299. doi: 10.1016/0042-6989(84)90054-3

Newsome, W. T., and Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* 8, 2201–2211.

Nishida, S. (2011). Advancement of motion psychophysics: review 2001–2010. *J. Vis.* 11:11. doi: 10.1167/11.5.11

Noles, N. S., Scholl, B. J., and Mitroff, S. R. (2005). The persistence of object file representations. *Percep. Psychophys.* 67, 324–334. doi: 10.3758/BF03206495

Nowlan, S. J., and Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *J. Neurosci.* 15, 1195–1214.

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiol. Rev.* 88, 59–89. doi: 10.1152/physrev.00008.2007

Orban, G. A., Kennedy, H., and Bullier, J. (1986). Velocity sensitivity and direction selectivity of neurons in areas V1 and V2 of the monkey: influence of eccentricity. *J. Neurophysiol.* 56, 462–480.

Pack, C. C., and Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* 409, 1040–1042. doi: 10.1038/35059085

Pack, C. C., Conway, B. R., Born, R. T., and Livingstone, M. S. (2006). Spatiotemporal structure of nonlinear subunits in macaque visual cortex. *J. Neurosci.* 26, 893–907. doi: 10.1523/jneurosci.3226-05.2006

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis.* 5, 376–404. doi: 10.1167/5.5.1

Pasupathy, A., and Connor, C. E. (1999). Responses to contour features in macaque area V4. *J. Neurophysiol.* 82, 2490–2502.

Pasupathy, A., and Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* 86, 2505–2519.

Patterson, C. A., Wissig, S. C., and Kohn, A. (2014). Adaptation disrupts motion integration in the primate dorsal stream. *Neuron* 81, 674–686. doi: 10.1016/j.neuron.2013.11.022

Peissig, J. J., and Tarr, M. J. (2007). Visual object recognition: do we know more now than we did 20 years ago? *Annu. Rev. Psychol.* 58, 75–96. doi: 10.1146/annurev.psych.58.102904.190114

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., et al. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *J. Exp. Biol.* 146, 87–113.

Perrone, J. A., and Thiele, A. (2001). Speed skills: measuring the visual speed analyzing properties of primate MT neurons. *Nat. Neurosci.* 4, 526–532. doi: 10.1038/87480

Perry, C. J., and Fallah, M. (2012). Color improves speed of processing but not perception in a motion illusion. *Front. Psychol.* 3:92. doi: 10.3389/fpsyg.2012.00092

Perry, C. J., Tahiri, A., and Fallah, M. (2014). Feature integration within and across visual streams occurs at different visual processing stages. *J. Vis.* 14:10. doi: 10.1167/14.2.10

Peterhans, E., and von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *J. Neurosci.* 9, 1749–1763.

Peuskens, H., Claeys, K. G., Todd, J. T., Norman, J. F., Van Hecke, P., and Orban, G. A. (2004). Attention to 3-D shape, 3-D motion and texture in 3-D structure from motion displays. *J. Cogn. Neurosci.* 16, 665–682. doi: 10.1162/089892904323057371

Phinney, R. E., and Siegel, R. M. (2000). Speed selectivity for optic flow in area 7a of the behaving macaque. *Cereb. Cortex* 10, 413–421. doi: 10.1093/cercor/10.4.413

Pinilla, T., Cobo, A., Torres, K., and Valdes-Sosa, M. (2001). Attentional shifts between surfaces: effects on detection and early brain potentials. *Vision Res.* 41, 1619–1630. doi: 10.1016/s0042-6989(01)00039-6

Poggio, T., and Ullman, S. (2013). Vision: are models of object recognition catching up with the brain? *Ann. N Y Acad. Sci.* 1305, 72–82. doi: 10.1111/nyas.12148

Priebe, N. J., Cassanello, C. R., and Lisberger, S. G. (2003). The neural representation of speed in macaque area MT/V5. *J. Neurosci.* 23, 5650–5661. doi: 10.1038/14819

Priebe, N. J., Lisberger, S. G., and Movshon, J. A. (2006). Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. *J. Neurosci.* 26, 2941–2950. doi: 10.1523/jneurosci.3936-05.2006

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. *Cognition* 32, 65–97. doi: 10.1016/0010-0277(89)90014-0

Pylyshyn, Z. (1994). Some primitive mechanisms of spatial attention. *Cognition* 50, 363–384. doi: 10.1016/0010-0277(94)90036-1

Raffi, M., and Siegel, R. M. (2007). A functional architecture of optic flow in the inferior parietal lobule of the behaving monkey. *PLoS One* 2:e200. doi: 10.1371/journal.pone.0000200

Raffi, M., Carrozzini, C., Maioli, M. G., and Squatrito, S. (2010). Multimodal representation of optic flow in area PEc of macaque monkey. *Neuroscience* 171, 1241–1255. doi: 10.1016/j.neuroscience.2010.09.026

Raffi, M., Persiani, M., Piras, A., and Squatrito, S. (2014). Optic flow neurons in area PEc integrate eye and head position signals. *Neurosci. Lett.* 568, 23–28. doi: 10.1016/j.neulet.2014.03.042

Raymond, J. E., Shapiro, K. L., and Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* 18, 849–860. doi: 10.1037/0096-1523.18.3.849

Raymond, J. E., Shapiro, K. L., and Arnell, K. M. (1995). Similarity determines the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 653–662. doi: 10.1037//0096-1523.21.3.653

Reynolds, J. H., Alborzian, S., and Stoner, G. R. (2003). Exogenously cued attention triggers competitive selection of surfaces. *Vision Res.* 43, 59–66. doi: 10.1016/s0042-6989(02)00403-0

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rodman, H. R., and Albright, T. D. (1989). Single-unit analysis of pattern-motion selective properties in the middle temporal visual area (MT). *Exp. Brain Res.* 75, 53–64. doi: 10.1007/bf00248530

Rodríguez, V., Valdes-Sosa, M., and Freiwald, W. (2002). Dividing attention between form and motion during transparent surface perception. *Brain Res. Cogn. Brain Res.* 13, 187–193. doi: 10.1016/s0926-6410(01)00111-2

Roe, A. W., and Ts'o, D. Y. (1999). Specificity of color connectivity between primate V1 and V2. *J. Neurophysiol.* 82, 2719–2730.

Roelfsema, P. R., and Spekreijse, H. (2001). The representation of erroneously perceived stimuli in the primary visual cortex. *Neuron* 31, 853–863. doi: 10.1016/s0896-6273(01)00408-1

Roelfsema, P. R., Lamme, V. A., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395, 376–381. doi: 10.1038/26475

Rolls, E. T., and Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp. Brain Res.* 65, 38–48. doi: 10.1007/bf00243828

Ross, J., Badcock, D. R., and Hayes, A. (2000). Coherent global motion in the absence of coherent velocity signals. *Curr. Biol.* 10, 679–682. doi: 10.1016/s0960-9822(00)00524-8

Saito, H., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., and Iwai, E. (1986). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *J. Neurosci.* 6, 145–157.

Salzman, C. D., Murasugi, C. M., Britten, K. H., and Newsome, W. T. (1992). Microstimulation in visual area MT: effects on direction discrimination performance. *J. Neurosci.* 12, 2331–2355.

Sáry, G., Vogels, R., and Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science* 260, 995–997. doi: 10.1126/science.8493538

Sato, T., Kawamura, T., and Iwai, E. (1980). Responsiveness of inferotemporal single units to visual pattern stimuli in monkeys performing discrimination. *Exp. Brain Res.* 38, 313–319. doi: 10.1007/bf00236651

Sayim, B., Westheimer, G., and Herzog, M. H. (2008). Contrast polarity, chromaticity and stereoscopic depth modulate contextual interactions in vernier acuity. *J. Vis.* 8:12. doi: 10.1167/8.8.12

Schein, S. J., and Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. *J. Neurosci.* 10, 3369–3389.

Schiller, P. H. (1993). The effects of V4 and middle temporal (MT) area lesions on visual performance in the rhesus monkey. *Vis. Neurosci.* 10, 717–746. doi: 10.1017/s0952523800005423

Schwartz, E. L., Desimone, R., Albright, T. D., and Gross, C. G. (1983). Shape recognition and inferior temporal neurons. *Proc. Natl. Acad. Sci. U S A* 80, 5776–5778. doi: 10.1073/pnas.80.18.5776

Sereno, A. B., and Maunsell, J. H. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature* 395, 500–503. doi: 10.1038/26752

Sereno, M. E., Trinath, T., Augath, M., and Logothetis, N. K. (2002). Three-dimensional shape representation in monkey cortex. *Neuron* 33, 635–652. doi: 10.1016/s0896-6273(02)00598-6

Shadlen, M. N., and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proc. Natl. Acad. Sci. U S A* 93, 628–633. doi: 10.1073/pnas.93.2.628

Shipp, S., and Zeki, S. (1985). Segregation of pathways leading from area V2 to areas V4 and V5 of macaque monkey visual cortex. *Nature* 315, 322–325. doi: 10.1038/315322a0

Shipp, S., and Zeki, S. (2002). The functional organization of area V2, I: specialization across stripes and layers. *Vis. Neurosci.* 19, 187–210. doi: 10.1017/s0952523802191164

Siegel, R. M., and Andersen, R. A. (1988). Perception of three-dimensional structure from motion in monkey and man. *Nature* 331, 259–261. doi: 10.1038/331259a0

Siegel, R. M., and Read, H. L. (1997). Analysis of optic flow in the monkey parietal area 7a. *Cereb. Cortex* 7, 327–346. doi: 10.1093/cercor/7.4.327

Snowden, R. J., Treue, S., and Andersen, R. A. (1992). The response of neurons in areas V1 and MT of the alert rhesus monkey to moving random dot patterns. *Exp. Brain Res.* 88, 389–400. doi: 10.1007/bf02259114

Snowden, R. J., Treue, S., Erickson, R. G., and Andersen, R. A. (1991). The response of area MT and V1 neurons to transparent motion. *J. Neurosci.* 11, 2768–2785.

Solomon, S. G., and Lennie, P. (2005). Chromatic gain controls in visual cortical neurons. *J. Neurosci.* 25, 4779–4792. doi: 10.1523/jneurosci.5316-04.2005

Solomon, S. G., Peirce, J. W., and Lennie, P. (2004). The impact of suppressive surrounds on chromatic properties of cortical neurons. *J. Neurosci.* 24, 148–160. doi: 10.1523/jneurosci.3036-03.2004

Stoner, G. R., and Albright, T. D. (1996). The interpretation of visual motion: evidence for surface segmentation mechanisms. *Vision Res.* 36, 1291–1310. doi: 10.1016/0042-6989(95)00195-6

Stoner, G. R., Mitchell, J. F., Fallah, M., and Reynolds, J. H. (2005). Interacting competitive selection in attention and binocular rivalry. *Prog. Brain Res.* 149, 227–234. doi: 10.1016/s0079-6123(05)49016-0

Taira, M., Mine, S., Georgopoulos, A. P., Murata, A., and Sakata, H. (1990). Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Exp. Brain Res.* 83, 29–36. doi: 10.1007/bf00232190

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.* 66, 170–189.

Tanaka, K., Sugita, Y., Moriya, M., and Saito, H. (1993). Analysis of object motion in the ventral part of the medial superior temporal area of the macaque visual cortex. *J. Neurophysiol.* 69, 128–142.

Taub, E., Victor, J. D., and Conte, M. M. (1997). Nonlinear preprin short-range motion. *Vision Res.* 37, 1459–1477. doi: 10.1016/s0042-6989(96)-7

Tchernikov, I., and Fallah, M. (2010). A color hierarchy for automatic target selection. *PLoS One* 5:e9338. doi: 10.1371/journal.pone.0009338

Thier, P., and Erickson, R. G. (1992). Responses of visual-tracking neurons from cortical area MST-I to visual, eye and head motion. *Eur. J. Neurosci.* 4, 539–553. doi: 10.1111/j.1460-9568.1992.tb00904.x

Thompson, P. (1982). Perceived rate of movement depends on contrast. *Vision Res.* 22, 377–380. doi: 10.1016/0042-6989(82)90153-5

Thompson, P., Brooks, K., and Hammett, S. T. (2006). Speed can go up as well as down at low contrast: implications for models of motion perception. *Vision Res.* 46, 782–786. doi: 10.1016/j.visres.2005.08.005

Ts'o, D. Y., and Gilbert, C. D. (1988). The organization of chromatic and spatial interactions in the primate striate cortex. *J. Neurosci.* 8, 1712–1727.

Tsui, J. M., and Pack, C. C. (2011). Contrast sensitivity of MT receptive field centers and surrounds. *J. Neurophysiol.* 106, 1888–1900. doi: 10.1152/jn.00165.2011

Tsutsui, K., Jiang, M., Yara, K., Sakata, H., and Taira, M. (2001). Integration of perspective and disparity cues in surface-orientation-selective neurons of area CIP. *J. Neurophysiol.* 86, 2856–2867.

Valdes-Sosa, M., Cobo, A., and Pinilla, T. (1998). Transparent motion and object-based attention. *Cognition* 66, B13–B23. doi: 10.1016/s0010-0277(98)00012-2

Valdes-Sosa, M., Cobo, A., and Pinilla, T. (2000). Attention to object files defined by transparent motion. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 488–505. doi: 10.1037/0096-1523.26.2.488

von der Heydt, R., and Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J. Neurosci.* 9, 1731–1748.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science* 224, 1260–1262. doi: 10.1126/science.6539501

Wannig, A., Rodriguez, V., and Freiwald, W. A. (2007). Attention to surfaces modulates motion processing in extrastriate area MT. *Neuron* 54, 639–651. doi: 10.1016/j.neuron.2007.05.001

Wilson, H. R., and Kim, J. (1994). A model for motion coherence and transparency. *Vis. Neurosci.* 11, 1205–1220. doi: 10.1017/s0952523800007008

Zaksas, D., and Pasternak, T. (2006). Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J. Neurosci.* 26, 11726–11742. doi: 10.1523/jneurosci.3420-06.2006

Zeki, S., Watson, J. D., Lueck, C. J., Friston, K. J., Kennard, C., and Frackowiak, R. S. (1991). A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* 11, 641–649.

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 April 2014; accepted: 16 July 2014; published online: 05 August 2014.
*Citation: Perry CJ and Fallah M (2014) Feature integration and object representations along the dorsal stream visual hierarchy. Front. Comput. Neurosci. **8**:84. doi: 10.3389/fncom.2014.00084*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Perry and Fallah. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 15 December 2014
doi: 10.3389/fncom.2014.00160

Sparsey™: event recognition via deep hierarchical sparse distributed codes

Gerard J. Rinkus*†

Neurithmic Systems LLC, Newton, MA, USA

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Alessandro Treves, Scuola Internazionale Superiore di Studi Avanzati, Italy
Marc Pomplun, University of Massachusetts Boston, USA

***Correspondence:**
Gerard J. Rinkus, Neurithmic Systems LLC, 275 Grove St., Suite 2-4069, Newton, MA, USA
e-mail: grinkus@brandeis.edu

†Present address:
Gerard J. Rinkus, Visiting Scientist, Lisman Lab, Biology, Brandeis University, Waltham, MA, USA

The visual cortex's hierarchical, multi-level organization is captured in many biologically inspired computational vision models, the general idea being that progressively larger scale (spatially/temporally) and more complex visual features are represented in progressively higher areas. However, most earlier models use localist representations (codes) in each representational field (which we equate with the cortical macrocolumn, "mac"), at each level. In localism, each represented feature/concept/event (hereinafter "item") is coded by a single unit. The model we describe, Sparsey, is hierarchical as well but crucially, it uses sparse distributed coding (SDC) in every mac in all levels. In SDC, each represented item is coded by a small subset of the mac's units. The SDCs of different items can overlap and the size of overlap between items can be used to represent their similarity. The difference between localism and SDC is crucial because SDC allows the two essential operations of associative memory, storing a new item and retrieving the best-matching stored item, to be done in fixed time for the life of the model. Since the model's core algorithm, which does both storage and retrieval (inference), makes a single pass over all macs on each time step, the overall model's storage/retrieval operation is also fixed-time, a criterion we consider essential for scalability to the huge ("Big Data") problems. A 2010 paper described a nonhierarchical version of this model in the context of purely spatial pattern processing. Here, we elaborate a fully hierarchical model (arbitrary numbers of levels and macs per level), describing novel model principles like progressive critical periods, dynamic modulation of principal cells' activation functions based on a mac-level familiarity measure, representation of multiple simultaneously active hypotheses, a novel method of time warp invariant recognition, and we report results showing learning/recognition of spatiotemporal patterns.

Keywords: sparse distributed codes, cortical hierarchy, sequence recognition, event recognition, deep learning, critical periods, time warp invariance

INTRODUCTION

In this paper, we provide the hierarchical elaboration of the macro/mini-column model of cortical computation described in Rinkus (1996, 2010) which is now named Sparsey. We report results of initial experiments involving multi-level models with multiple macrocolumns ("macs") per level, processing spatiotemporal patterns, i.e., "events." In particular, we show: (a) single-trial unsupervised learning of sequences where this learning results in the formation of hierarchical spatiotemporal memory traces; and (b) recognition of training sequences, i.e., exact or nearly exact reactivation of complete hierarchical traces over all frames of a sequence. The canonical macrocolumnar algorithm—which probabilistically chooses a sparse distributed code (SDC) as a function of a mac's entire input, i.e., its bottom-up (U), horizontal (H), and top-down (D) input vectors, at a given moment—operates similarly, modulo parameters, in both learning and recognition, in all macs at all levels. Computationally, Sparsey's most important property is that a mac both stores (learns) new input items—which in general are temporal-context-dependent inputs, i.e., particular spatiotemporal *moments*—and retrieves the spatiotemporally closest-matching stored item in time that remains fixed as the number of items stored in the mac increases. This property depends critically on the use of SDCs, is essential for scalability to "Big Data" problems, and *has not been shown for any other computational model, biologically inspired or not!*

The model has a number of other interesting neurally plausible properties, including the following. (1) A "critical period" concept wherein learning is frozen in a mac's afferent synaptic projections when those projections reach a threshold saturation. In a hierarchical setting, freezing will occur beginning with the lowest level macs (analogous to primary sensory cortex) and progress upward over the course of experience. (2) A "progressive persistence" property wherein the activation duration (persistence) of the "neurons" (and thus of the SDCs which are sets of co-active neurons) increases with level; there is some evidence for increasing persistence along the ventral visual path (Rolls and Tovee, 1994; Uusitalo et al., 1997; Gauthier et al., 2012). This allows an SDC in a mac at level J to associate with sequences of SDCs in Level J-1 macs with which it is connected, i.e., a chunking (compression) mechanism. In particular, this provides a means to

learn in unsupervised fashion perceptual invariances produced by continuous transforms occurring in the environment (e.g., rotation, translation, etc.). Rolls' VisNet model, introduced in Rolls (1992) and reviewed in Rolls (2012), uses a similar concept to explain learning of naturally-experienced transforms, although his trace-learning-rule-based implementation differs markedly from ours. (3) During learning, an SDC is chosen on the basis of signals arriving from all active afferent neurons in the mac's total (U, H, and D) receptive field (RF). However, during retrieval, if the highest-order match, i.e., involving all three (U, H, and D) input sources, falls below a threshold, the mac considers a progression of lower-order matches, e.g., involving only its U and D inputs, but ignoring its H inputs, and if that also falls below a threshold, a match involving only its U inputs. This "back-off" protocol, in conjunction with progressive persistence, allows a protocol by which the model can rapidly—*crucially, the protocol does not increase the time complexity of closest-match retrieval*—compare a test sequence (e.g., video snippet) not only to the set of all sequences *actually* experienced and stored, but to a much larger space of nonlinearly time-warped variants of the actually-experienced sequences. (4) During retrieval, multiple competing hypotheses can momentarily (i.e., for one or several frames) be co-active in any given mac and resolve to a single hypothesis as subsequent disambiguating information enters.

While the results reported herein are specifically for the unsupervised learning case, Sparsey also implements supervised learning in the form of cross-modal unsupervised learning, where one of the input modalities is treated as a label modality. That is, if the same label is co-presented with multiple (arbitrarily different) inputs in another (raw sensory) modality, then a single internal representation of that label can be associated with the multiple (arbitrarily different) internal representations of the sensory inputs. That internal representation of the label then *de facto* constitutes a representation of the class that includes all those sensory inputs regardless of how different they are, providing the model a means to learn essentially arbitrarily nonlinear categories (invariances), i.e., instances of what Bengio terms "AI Set" problems (Bengio, 2007). Although we describe this principle in this paper, its full elaboration and demonstration in the context of supervised learning will be treated in a future paper.

Regarding the model's possible neural realization, our primary concern is that all of the model's formal structural and dynamic properties/mechanisms be *plausibly realizable* by known neural principles. For example, we do not give a detailed neural model of the winner-take-all (WTA) competition that we hypothesize to take place in the model's minicolumns, but rather rely on the plausibility of any of the many detailed models of WTA competition in the literature, (e.g., Grossberg, 1973; Yu et al., 2002; Knoblich et al., 2007; Oster et al., 2009; Jitsev, 2010). Nor do we give a detailed neural model for the mac's computation of the overall spatiotemporal familiarity of its input (the "*G*" measure), or for the G-contingent modulation of neurons' activation functions. Furthermore, the model relies only upon binary neurons and a simple synaptic learning model. This paper is really most centrally an explanation of why and how the use of SDC in conjunction with hierarchy provides a computationally efficient, scalable, and neurally plausible solution to event (i.e., single- or multimodal spatiotemporal pattern) learning and recognition.

OVERALL MODEL CONCEPT

The remarkable structural homogeneity across the neocortical sheet suggests a canonical circuit/algorithm, i.e., a core computational module, operating similarly in all regions (Douglas et al., 1989; Douglas and Martin, 2004). In addition, DiCarlo et al. (2012) present compelling first-principles arguments based on computational efficiency and evolution for a macrocolumn-sized canonical functional module whose goal they describe as "cortically local subspace untangling." We also identify the canonical functional module with the cortical "macrocolumn" (a.k.a. "hypercolumn" in V1, or "barrel"-related volumes in rat/mouse primary somatosensory cortex), i.e., a volume of cortex, ∼200–500 um in diameter, and will refer to it as a "mac." In our view, the mac's essential function, or "meta job description," in the terms of DiCarlo et al. (2012), is to operate as a semi-autonomous content-addressable memory. That is, the mac:

(a) assigns (stores, learns) neural codes, specifically *sparse distributed codes* (SDCs), representing its global (i.e., combined U, H, and D) input patterns; and
(b) retrieves (reactivates) stored codes, i.e., *memories*, on subsequent occasions when the global input pattern matches a stored code sufficiently closely.

If the mac's learning process ensures that *similar inputs map to similar code*s (SISC), as Sparsey's does, then operating as a content addressable memory is functionally equivalent to local subspace untangling.

Although the majority of neurophysiological studies through the decades have formalized the responses of cortical neurons in terms of purely spatial receptive fields (RFs), evidence revealing the truly spatiotemporal nature of neuronal RFs is accumulating (DeAngelis et al., 1993, 1999; Rust et al., 2005; Gavornik and Bear, 2014; Ramirez et al., 2014). In our mac model, time is discrete: U signals arrive from neurons active on the current time step while H and D signals arrive from neurons active on the previous time step. We can view the combined U, H, and D inputs as a "context-dependent U input" (where the H and D signals are considered the "context") or more holistically, as an overall particular spatiotemporal *moment* (as suggested earlier).

As will be described in detail, the first step of the mac's canonical algorithm, during both learning and retrieval, is to combine its U, H, and D inputs to yield a (scalar) judgment, *G*, as to the spatiotemporal *familiarity* of the current *moment*. Provided the number of codes stored in the mac is small enough, *G* measures the *spatiotemporal similarity* of the best matching stored moment, *x*, to the current moment, *I*.

$$G = \arg\max_{x} (sim(I, x))$$

Figure I-1 shows the envisioned correspondence of Sparsey to the cortical macrocolumn. In particular, we view the mac's subpopulation of L2/3 pyramidals as the actual repository of SDCs. And even more specifically, we postulate that the ∼20 L2/3

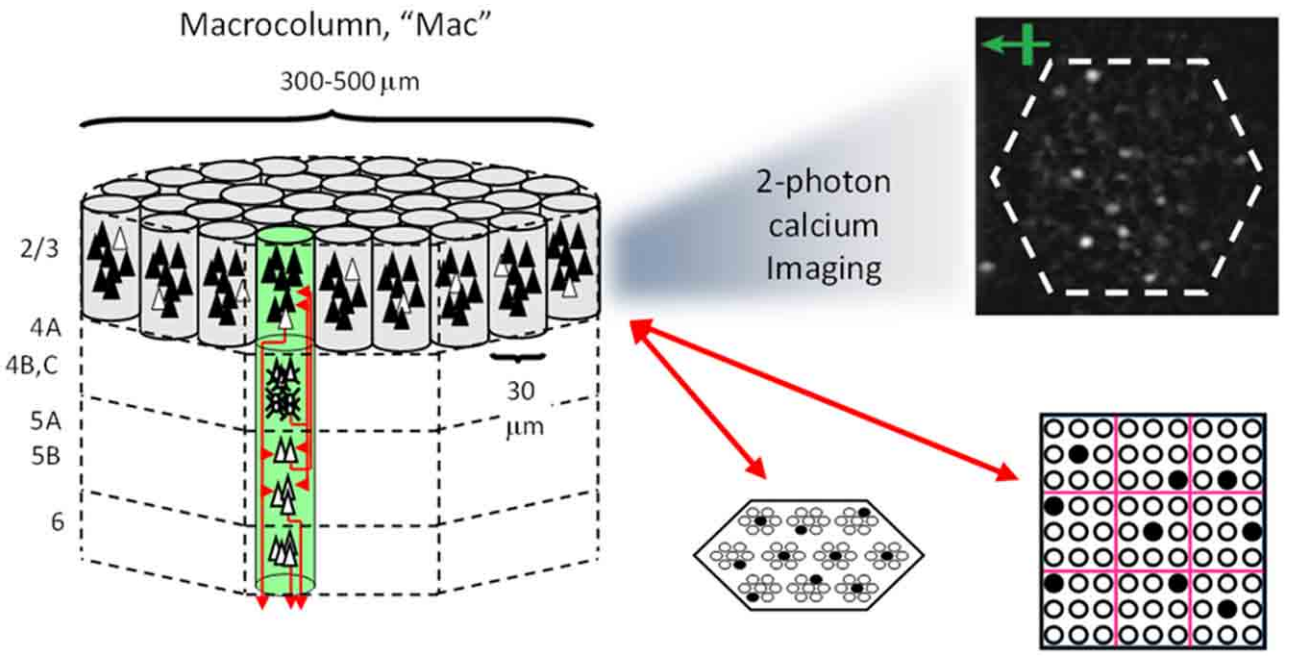


FIGURE I-1 | Proposed correspondence between the cortical macrocolumn and Sparsey's mac. Left: schematic of a cortical macrocolumn composed of ∼70 minicolumns (green cylinder). SDCs representing context-dependent inputs reside in mac's L2/3 population. An SDC is a set composed of one active L2/3 pyramidal cell per minicolumn. **Upper Right**: 2-photon calcium image of activity in a mac-sized area of cat V1 given a left-moving vertical bar in the mac's RF; we have added dashed hexagonal boundary to suggest the boundary of macrocolumn/hypercolumn module (adapted from Ohki et al., 2005). **Lower Right**: two formats that we use to depict macs; they show only the L2/3 cells. The hexagonal format mac has 10 minicolumns each with seven cells. The rectangular format mac has nine minicolumns each with nine cells. Note that in these formats, active cells are black (or red as in many subsequent figures); inactive cells are white.

pyramidals in each of the mac's ∼70 minicolumns function in WTA fashion. Thus, a single SDC code will consist of 70 L2/3 pyramidals, one per minicolumn. Note: we also refer to minicolumns as *competitive modules* (CMs). Two-photon calcium imaging movies, e.g., Ohki et al. (2005), Sadovsky and MacLean (2014), provide some support for the existence of such macrocolumnar SDCs as they show numerous instances of ensembles, consisting of from several to hundreds of neurons, often spanning several 100 um, turning on and off as tightly synchronized wholes. We anticipate that the recently developed super-fast voltage sensor ASAP1 (St-Pierre et al., 2014) may allow much higher fidelity testing of SDCs and Sparsey in general.

Figure I-2 (left) illustrates the three afferent projections to a particular mac at level L1 (analog of cortical V1), M_i^1 (i.e., the *ith* mac at level L1). The red hexagon at L0 indicates M_i^1's aperture onto the thalamic representation of the visual space, i.e., its classical receptive field (RF), which we can refer to more specifically as M_i^1's U-RF. This aperture consists of about 40 binary pixels connected all-to-all with M_i^1's cells; black arrows show representative U-weights (U-wts) from two active pixels. Note that we assume that visual inputs to the model are filtered to single-pixel-wide edges and binarized. The blue semi-transparent prism represents the full bundle of U-wts comprising M_i^1's U-RF.

The all-to-all *U-connectivity* within the blue prism is essential because the concept of the RF of a mac *as a whole, not of an individual cell*, is central to our theory. This is because the "atomic coding unit," or equivalently, the "atomic unit of meaning" in this theory is the SDC, i.e., a *set* of cells. The activation of a mac, during both learning and recognition, consists in the activation of an entire SDC, i.e., simultaneous activation of one cell in every minicolumn. Similarly, deactivation of a mac consists in the simultaneous deactivation of *all* cells comprising the SDC (though in general, some of the cells contained in a mac's currently active SDC might also be contained in the next SDC to become active in that mac). Thus, in order to be able to view an SDC as *collectively* (or *atomically*) representing the input to a mac as a whole, all cells in a mac must have the same RF (the same set of afferent cells). This scenario is assumed throughout this report.

In **Figure I-2**, magenta lines represent the D-wts comprising M_i^1's afferent D projection, or D-RF. In this case, M_i^1's D-RF consists of only one L2 (analog of V2) mac, M_j^2, which is all-to-all connected to M_i^1 (representative D-wts from just two of M_j^2's cells are shown). Any given mac also receives complete H-projections from all nearby macs in its own level (including itself) whose centers fall within a parameter-specifiable radius of its own center. Signals propagating via H-wts are defined to take one time step (one sequence item) to propagate. Green arrows show a small representative sample of H-wts mediating signals arriving form cells active on the prior time step (gray). Red indicates cells active on current time step. At right of **Figure I-2**, we zoom in on one of M_i^1's minicolumns (CMs) to emphasize that every cell in a CM has the same H-, U-, and D-RFs. **Figure I-3** further illustrates

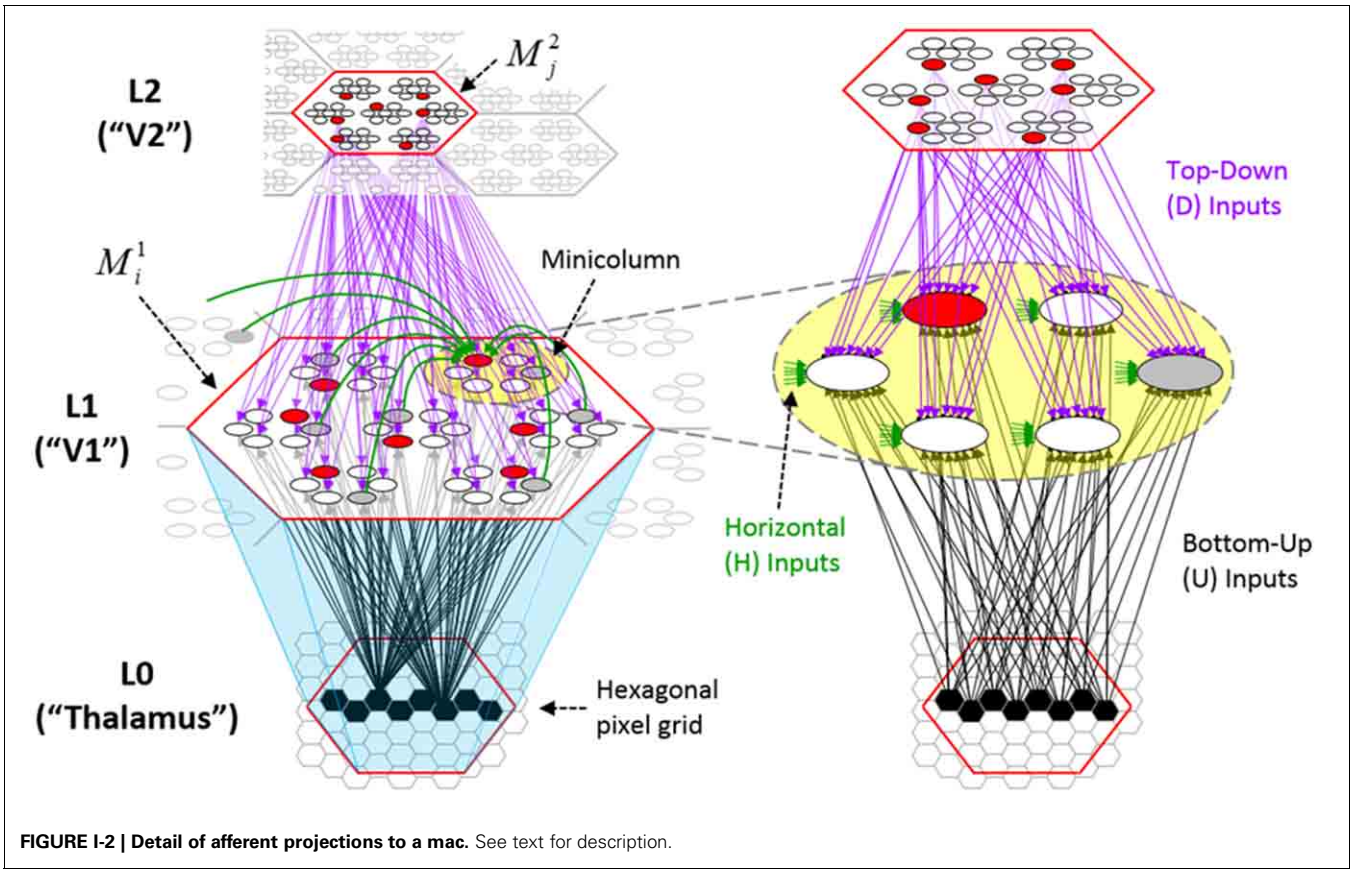


FIGURE I-2 | Detail of afferent projections to a mac. See text for description.

(using the rectangular format for depicting macs) the concept that all cells in a given mac have the same U-, H-, and D-RFs and that those RFs respect the borders of the source macs. Each cell in the L1 mac, $M^1_{(2,2)}$ (here we use an alternate (x,y) coordinate indexing convention for the macs), receives a D-wt from all cells in all five L2 macs indicated, an H-wt from all cells in $M^1_{(2,2)}$ and its N, S, E, and W neighboring macs (green shading), and a U-wt from all 36 cells in the indicated aperture.

The hierarchical organization of visual cortex is captured in many biologically inspired computational vision models with the general idea being that progressively larger scale (both spatially and temporally) and more complex visual features are represented in progressively higher areas (Riesenhuber and Poggio, 1999; Serre et al., 2005). Our cortical model, Sparsey, is hierarchical as well, but as noted above, a crucial, in fact, the most crucial difference between Sparsey and most other biologically inspired vision models is that Sparsey encodes information at all levels of the hierarchy, and in every mac at every level, with SDCs. This stands in contrast to models that use *localist representations*, e.g., all published versions of the HMAX family of models, (e.g., Murray and Kreutz-Delgado, 2007; Serre et al., 2007) and other cortically-inspired hierarchical models (Kouh and Poggio, 2008; Litvak and Ullman, 2009; Jitsev, 2010) and the majority of graphical probability-based models (e.g., hidden Markov models, Bayesian nets, dynamic Bayesian nets). There are several other models for which SDC is central, e.g., SDM (Kanerva, 1988, 1994, 2009; Jockel, 2009), Convergence-Zone Memory (Moll and Miikkulainen, 1997), Associative-Projective Neural Networks (Rachkovskij, 2001; Rachkovskij and Kussul, 2001), Cogent Confabulation (Hecht-Nielsen, 2005), Valiant's "positive shared" representations (Valiant, 2006; Feldman and Valiant, 2009), and Numenta's Grok (described in Numenta white papers). However, none of these models has been substantially elaborated or demonstrated in an explicitly hierarchical architecture and most have not been substantially elaborated for the spatiotemporal case.

Figure I-4 illustrates the difference between a localist, e.g., an HMAX-like, model and the SDC-based Sparsey model. The input level (analogous to thalamus) is the same in both cases: each small gray/red hexagon in the input level represents the aperture (U-RF) of a single V1 mac (gray/red hexagon). In **Figure I-4A**, the representation used in each mac (at all levels) is *localist*, i.e., each feature is represented by a single cell and at any one time, only one cell (feature) is active (red) in any given mac (here the cell is depicted with an icon representing the feature it represents). In contrast, in **Figure I-4B**, any particular feature is represented by a *set* of co-active cells (red), one in each of a mac's minicolumns: compare the two macs at lower left of **Figure I-4A** with the corresponding macs in **Figure I-4B** (blue and brown arrows). Any given cell will generally participate in the codes of many different features. A yellow call-out shows codes for other features stored in the mac, besides the feature that is currently active. If you look closely, you can see that for some macs, some cells are active in more than one of the codes.

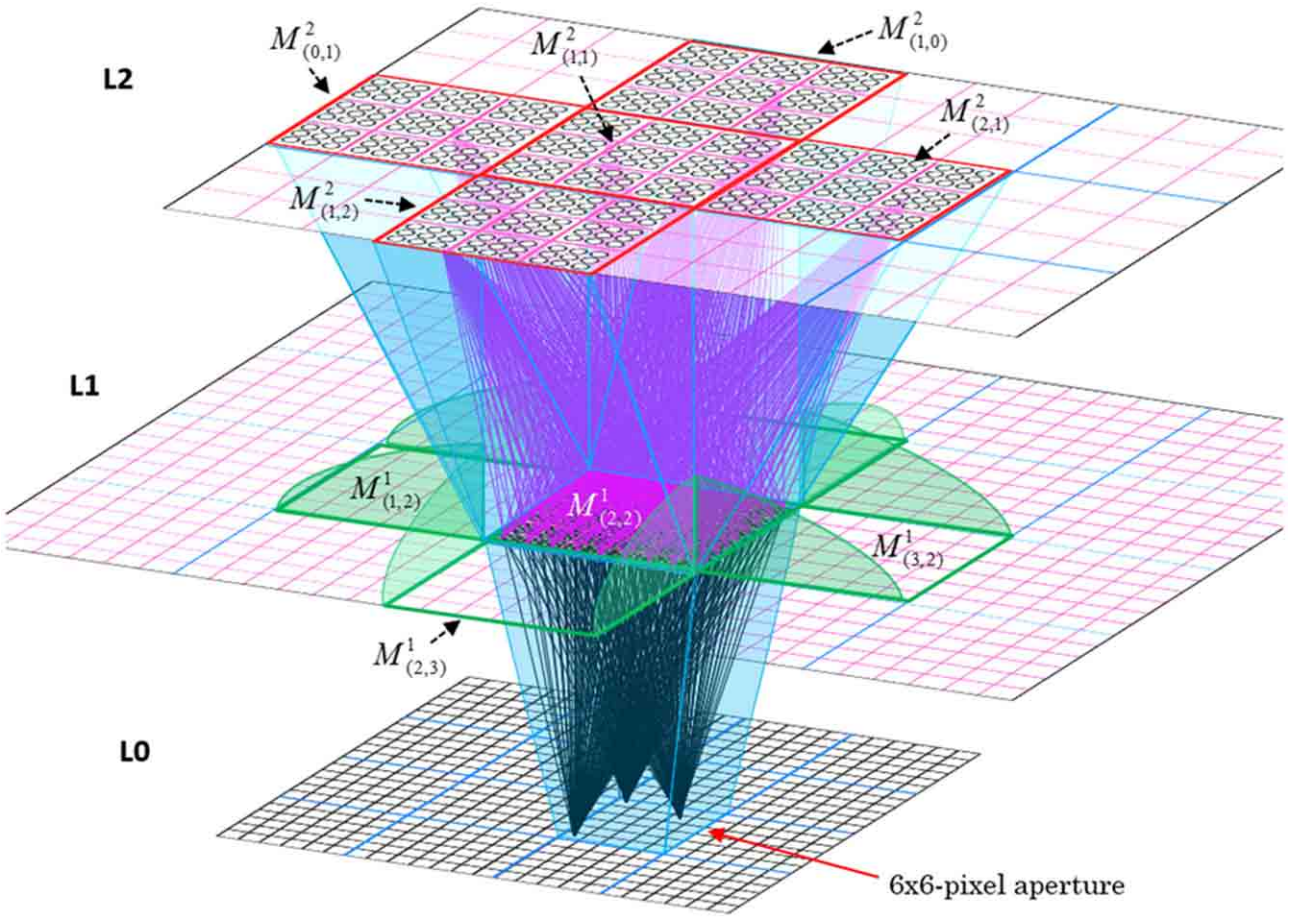


FIGURE I-3 | Connectivity scheme. Within each of the three afferent projections, H, U, and D, to a mac, $M^1_{(2,2)}$ (where the mac index is now in terms of (x,y) coordinates in the level, and we have switched to the rectangular mac topology), the connectivity is full and respects mac borders. L1 is a 5 × 4 sheet of macs (blue borders), each consisting of 36 minicolumns (pink borders), but the scale is too small to see the individual cells within minicolumns. L2 is a 4 × 3 sheet of macs, each consisting of nine CMs, each consisting of nine cells.

Looking at **Figure I-4A**, adapted from Serre et al. (2005), one can see the basic principle of hierarchical compositionality in action. The two neighboring apertures (pink) over the dog's nose lead to activation of cells representing a vertical and a horizontal feature in neighboring V1 macs. Due to the convergence/divergence of U-projections to V2, both of these cells project to the cells in the left-hand V2 mac. Each of these cells projects to multiple cells in that V2 mac, however, only the red (active) cell representing an "upper left corner" feature, is maximally activated by the conjunction of these two V1 features. Similarly, the U-signals from the cell representing the "diagonal" feature active in the right-hand V1 mac will combine with signals representing features in nearby apertures to activate the appropriate higher-level feature in the V2 mac whose U-RF includes these apertures (small dashed circles in the input level). Note that some notion of competition (e.g., the "max" operation in HMAX models) operates amongst the cells of a mac such that at any one time, only one cell (one feature) can be active.

We underscore that in **Figure I-4**, we depict simple (solid border) and complex (dashed border) features within individual macs, implying that complex and simple features can compete with each other. We believe that the distinction between simple and complex features may be largely due to coarseness of older experimental methods (e.g., using synthetic low-dimensional stimuli): newer studies are revealing far more precise tuning functions (Nandy et al., 2013), including temporal context specificity, even as early as V1 (DeAngelis et al., 1993, 1999), and in other modalities, somatosensory (Ramirez et al., 2014) and auditory (Theunissen and Elie, 2014).

The same hierarchical compositional scheme as between V1 and V2 continues up the hierarchy (some levels not shown), causing activation of progressively higher-level features. At higher levels, we typically call them concepts, e.g., the visual concept of "Jennifer Aniston," the visual concept of the class of dogs, the visual concept of a particular dog, etc. We show most of the features at higher levels with dashed outlines to indicate that they are *complex* features, i.e., features with particular, perhaps many, dimensions of invariance, most of which are learned through experience. In Sparsey, the particular invariances are learned from scratch and will generally vary from one feature/concept to

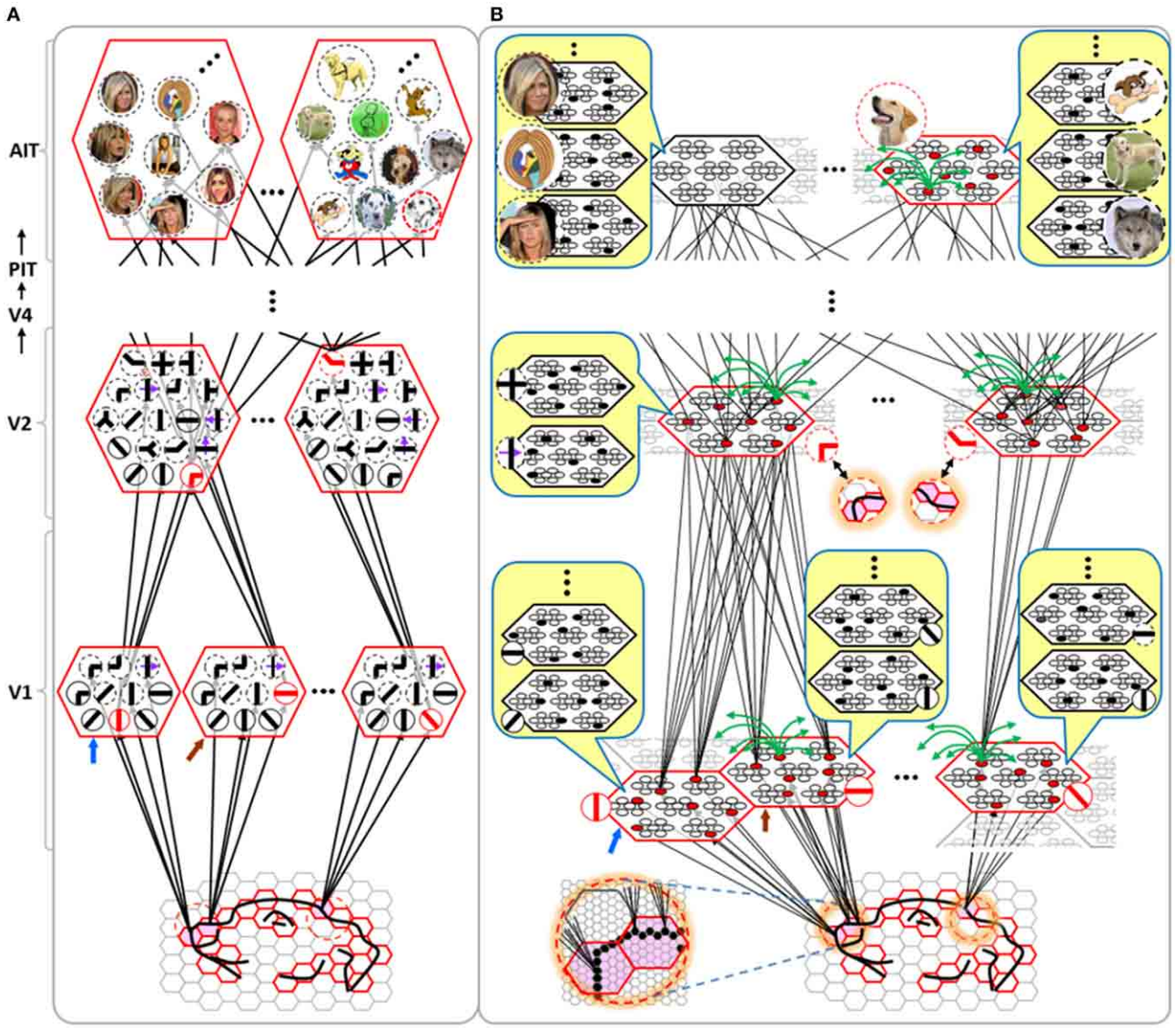


FIGURE I-4 | Comparison of a localist (A) and an SDC-based (B) hierarchical vision model. See text.

another, including within the same mac. The particular features shown in the different macs in this example are purely notional: it is the overall hierarchical compositionality principle that is important, not the particular features shown, nor the particular cortical regions in which they are shown.

The hierarchical compositional process described above in the context of the localist model of **Figure I-4A** applies to the SDC-based model in **Figure I-4B** as well. However, features/concepts are now represented by sets of cells rather than single cells. Thus, the vertical and horizontal features forming part of the dog's nose are represented with SDCs in their respective V1 macs (blue and brown arrows, respectively), rather than with single cells. The U-signals propagating from these two V1 macs converge on the cells of the left-hand V2 mac and combine, via Sparsey's *code selection algorithm* (CSA) (described in Section Sparsey's Core Algorithm), to activate the SDC representing the "corner" feature, and similarly on up the hierarchy. Each of the orange outlined insets at V2 shows the input level aperture of the corresponding mac, emphasizing the idea that the precise input pattern is mapped into the closest-matching stored feature, in this example, a "upper left 90° corner" at left and a "NNE-pointing 135° angle" at right. The inset at bottom of **Figure I-4B** zooms in to show that the U-signals to V1 arise from individual pixels of the apertures (which would correspond to individual LGN projection cells).

In the past, IT cells have generally been depicted as being *narrowly selective* to particular objects (Desimone et al., 1984; Kreiman et al., 2006; Kiani et al., 2007; Rust and DiCarlo, 2010). However, as DiCarlo et al. (2012) point out, the data overwhelmingly support the view of individual IT cells as having a "diversity of selectivity"; that is, individual IT cells generally respond to

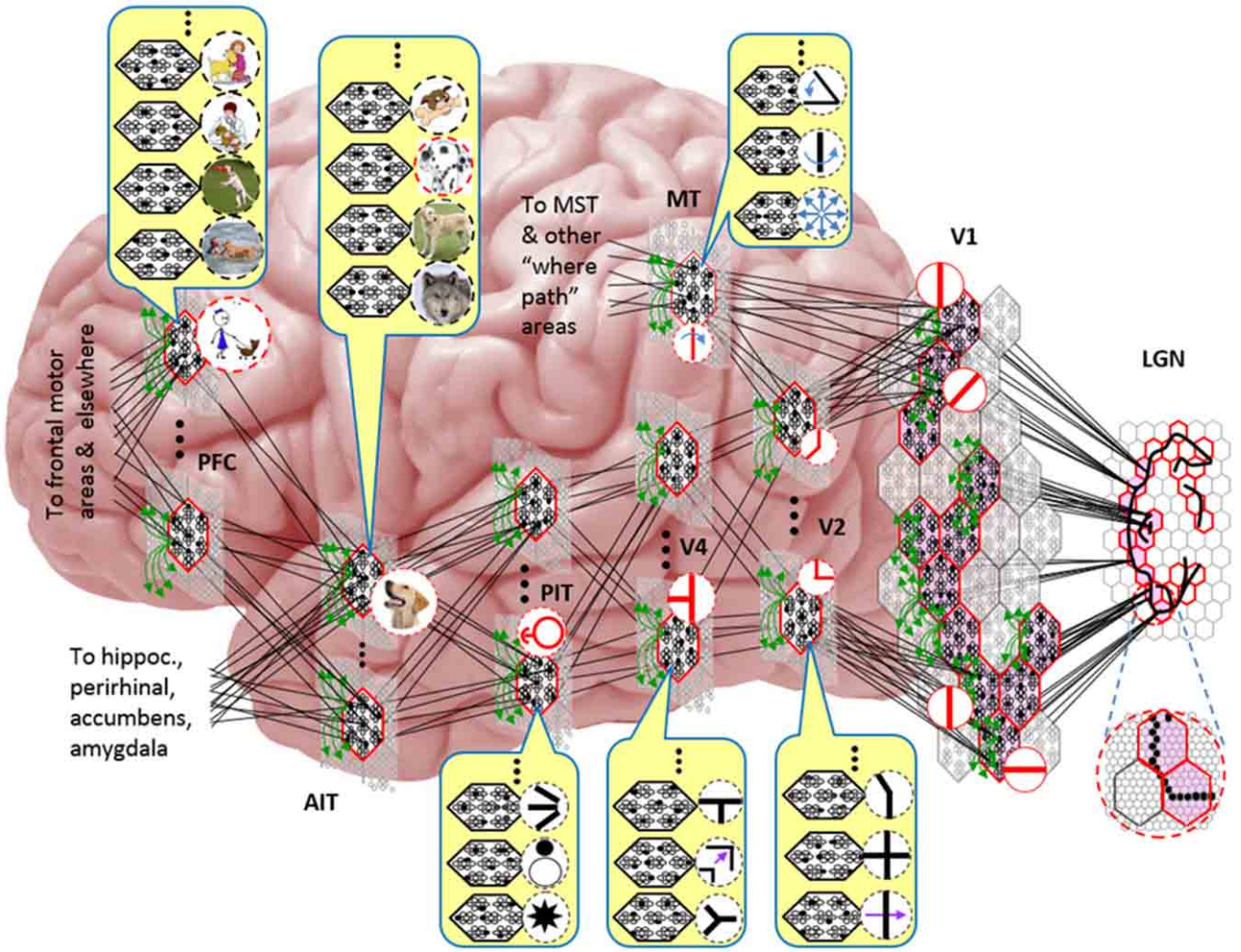


FIGURE I-5 | Notional mapping of Sparsey to brain.

many different objects and in that sense are much more broadly tuned. This diversity is notionally suggested in **Figures I-4B**, **I-5** in that individual cells are seen to participate in multiple SDCs representing different images/concepts. However, the particular input (stimulus) dimensions for which any given cell ultimately demonstrates some degree of invariance is not prescribed a priori. Rather they emerge essentially idiosyncratically over the history of a cell's inclusions in SDCs of particular experienced moments. Thus, the dimensions of invariance in the tuning functions of even immediately neighboring cells may generally end up quite different.

Figure I-5 embellishes the scheme shown in **Figure I-4B** and (turning it sideways) casts it onto the physical brain. We add paths from V1 and V2 to an MT representation as well. We add a notional PFC representation in which a higher-level concept involving the dog, i.e., the fact that it is being walked, is active. We show a more complete tiling of macs at V1 than in **Figure I-4B** to emphasize that only V1 macs that have a sufficient fraction of active pixels, e.g., an edge contour, in their aperture become active (pink). In general, we expect the fraction of active macs to decrease with level. As this and prior figures suggest, we currently model the macs as having no overlap with each other (i.e., they tile the local region), though their RFs [as well as their *projective fields* (PFs)] can overlap. However, we expect that in the real brain, macs can physically overlap. That is, any given minicolumn could be contained in multiple overlapping macs, where only one of those macs can be active at any given moment. The degree of overlap could vary by region, possibly generally increasing anteriorly. If so, then this would partially explain (in conjunction with the extremely limited view of population activity that single/few-unit electrophysiology has provided through most of the history of neuroscience) why there has been little evidence thus far for macs in more frontal regions.

SPARSE DISTRIBUTED CODES vs. LOCALIST CODES

One important difference between SDC and localist representation is that the space of representations (codes) for a mac using SDC is exponentially larger than for a mac using a localist representation. Specifically, if Q is the number of CMs in a mac and K is the number of cells per CM, then there are K^Q unique SDC

codes for that mac. A localist mac of the same size only has $Q \times K$ unique codes. Note that it is not the case that an SDC-based mac can use that entire code space, i.e., store K^Q features. Rather, the limiting factor on the number of codes storable in an SDC-based mac is the fraction of the mac's afferent synaptic weights that are set high (our model uses effectively binary weights), i.e., degree of *saturation*. In fact, the number of codes storable such that all stored codes can be retrieved with some prescribed average retrieval accuracy (error), is probably a vanishingly small fraction of the entire code space. However, real macrocolumns have $Q \approx 70$ minicolumns, each with $K \approx 20$ L2/3 principal cells: a "vanishingly small fraction" of 20^{70} can of course still be a large absolute number of codes.

While the difference in code space size between localist and SDC models is important, it is the distributed nature of the SDC codes *per se* that is most important. Many have pointed out a key property of SDC which is that since codes overlap, the number of cells in common between two codes can be used to represent their similarity. For example, if a given mac has $Q = 100$ CMs, then there are 101 possible degrees of intersection between codes, and thus 101 degrees of similarity, which can be represented between concepts stored in that mac. The details of the process/algorithm that assigns codes to inputs determines the specific definition of similarity implemented. We will discuss the similarity metric(s) implemented and implementable in Sparsey throughout the sequel.

However, as stated earlier, the most important distinction between *localism* and SDC is that SDC allows the two essential operations of associative (content-addressable) memory, storing new inputs and retrieving the *best-matching* stored input, to be done *in fixed time for the life of the model.* That is, given a model of a fixed size (dominated by the number of weights), and which therefore has a particular limit on the amount, C, of information that it can store and retrieve subject to a prescribed average retrieval accuracy (error), the time it takes to either store (learn) a new input or retrieve the best-matching stored input (memory) remains constant regardless of how much information has been stored, so long as that amount remains less than C. *There is no other extant model, including all HMAX models, all convolutional network (CN) models, all Deep Learning (DL) models, all other models in the class of graphical probability models (GPMs), and the locality-sensitive hashing models, for which this capability—constant storage and best-match retrieval time over the life of the system—has been demonstrated.* All these other classes of models realize the benefits of hierarchy *per se*, i.e., the principle of hierarchical compositionality which is critical for rapidly learning highly nonlinear category boundaries, as described in Bengio et al. (2012), but only Sparsey also realizes the speed benefit, and therefore ultimately, the scalability benefit, of SDC. We state the algorithm in Section Sparsey's Core Algorithm. The reader can see by inspection of the CSA (**Table I-1**) that it has a fixed number of steps; in particular, it does *not* iterate over stored items.

Another way of understanding the computational power of SDC compared to localism is as follows. We stated above that in a localist representation such as in **Figure I-4A**, only one cell, representing one hypothesis can be active at a time. The other cells in the mac might, at some point prior to the choice of a final winner, have a distribution of sub-threshold voltages that reflects the likelihood distribution over all represented hypotheses. But ultimately, only one cell will win, i.e., go supra-threshold and spike. Consequently, only that one cell, and thus that one hypothesis, will materially influence the next time step's decision process in the same mac (via the recurrent H matrix) and in any other downstream macs.

In contrast, because SDCs physically overlap, if one particular SDC (and thus, the hypothesis that it represents) is *fully* active in a mac, i.e., if all Q of that code's cells are active, *then all other codes (and thus, their associated hypotheses) stored in that mac are also simultaneously physically partially active in proportion to the size of their intersections with the single fully active code.* Furthermore, if the process/algorithm that assigns the codes to inputs has enforced the *similar-inputs-to-similar-codes* (SISC) property, then all stored inputs (hypotheses) are active with strength in descending order of similarity to the fully active hypothesis. We assume that more similar inputs generally reflect more similar world states and that world state similarity correlates with likelihood. In this case, the single fully active code also physically functions as the *full likelihood distribution over all SDCs (hypotheses) stored in a mac.* **Figure I-6** illustrates this concept. We show five hypothetical SDCs, denoted with $\phi()$, for five input items, A-E (the actual input items are not shown here), which have been stored in the mac shown. At right, we show the decreasing intersections of the codes with $\phi(A)$. Thus, when code $\phi(A)$ is (fully) active, $\phi(B)$ is 4/7 active, $\phi(C)$ is 3/7 active, etc. Since cells representing *all* of these hypotheses, not just the most likely hypothesis, A, actually spike, it follows that *all of these hypotheses physically influence the next time step's decision processes*, i.e., the resulting likelihood distributions, active on the next time step in the same and all downstream macs.

We believe this difference to be fundamentally important. In particular, it means that performing a single execution of the *fixed-time* CSA transmits the influence of *every* represented hypothesis, regardless of how strongly active a hypothesis is, to every hypothesis represented in downstream macs. We emphasize that the representation of a hypothesis's probability (or likelihood) in our model—i.e., as the fraction of a given hypothesis's full code (of Q cells) that is active—differs fundamentally from existing representations in which single neurons encode such probabilities in their strengths of activation (e.g., firing rates) as described in the recent review of Pouget et al. (2013).

SPARSEY'S CORE ALGORITHM

During learning, Sparsey's core algorithm, the code selection algorithm (CSA), operates on every time step (frame) in every mac of every level, resulting in activation of a set of cells (an SDC) in the mac. The CSA can also be used, with one major variation, during retrieval (recognition). However, there is a much simpler retrieval algorithm, essentially just the first few steps of the CSA, which is preferable if the system "knows" that it is in retrieval mode. Note that this is not the natural condition for autonomous systems: in general, the system must be able to decide for itself, on a frame-by-frame basis, whether it needs to be in learning mode (if, and to what extent, the input is novel) or retrieval mode (if the input is completely familiar). We first describe the CSA's

Table I-1 | The CSA during learning.

	Equation	Short description
1	$Active(m) = \begin{cases} true & \Upsilon(m) < \delta(m) \\ true & \pi_U^- \le \pi_U(m) \le \pi_U^+ \\ false & otherwise \end{cases}$	Determine if mac m will become active
2	$u(i) = \sum_{j \in \mathrm{RF_U}} x(j,t) \times F(\zeta(j,t)) \times w(j,i)$ $h(i) = \sum_{j \in \mathrm{RF_H}} x(j,t-1) \times F(\zeta(j,t-1)) \times w(j,i)$ $d(i) = \sum_{j \in \mathrm{RF_D}} x(j,t-1) \times F(\zeta(j,t-1)) \times w(j,i)$	Compute the raw U, H, and D input summations
3	$U(i) = \begin{cases} \max\left(1, u(i)/\pi_U^- \times w_{\max}\right) & L = 1 \\ \max\left(1, u(i) \min\left(\pi_U^-, \pi_U^*\right) \times Q \times w_{\max}\right) & L > 1 \end{cases}$. $H(i) = \max\left(1, h(i)/\min\left(\pi_H^-, \pi_H^*\right) \times Q \times w_{\max}\right)$ $D(i) = \max\left(1, d(i) \min\left(\pi_D^-, \pi_D^*\right) \times Q \times w_{\max}\right)$	Compute normalized, filtered input summations
4	$V(i) = \begin{cases} H(i)^{\lambda_H} \times U(i)^{\lambda_U(t)} \times D(i)^{\lambda_D} & t \ge 1 \\ U(i)^{\lambda_U(0)} & t = 0 \end{cases}$	Compute local evidential support for each cell
5a	$\zeta_q = \sum_{i=0}^{K} \left[V(i) > V_\zeta\right]$	(a) Compute #cells representing a maximally competing hypothesis in each CM. (b) Compute # of maximally active hypotheses, ζ, in the mac
5b	$\zeta = \sum_{j=0}^{Q-1} \zeta_q / Q$	
6	$F(\zeta) = \begin{cases} \zeta^A & 1 \le \zeta \le B \\ 0 & \zeta > B \end{cases}$	Compute the *multiple competing hypotheses* (MCH) correction factor, $F(\zeta)$, for the mac
7	$\hat{V}_j = \max_{i \in C_j} \{V(i)\}$	Find the max V, $\hat{V}_j$, in each CM, C_j
8	$G = \sum_{q=1}^{Q} \hat{V}_k / Q$	Compute G as the average $\hat{V}$-value over the Q CMs
9	$\eta = 1 + \left(\left[\frac{G - G^-}{1 - G^-}\right]^+\right)^\gamma \times \chi \times K$	Determine the expansivity of the sigmoid activation function
10	$\psi(i) = \frac{(\eta - 1)}{\left(1 + \sigma_1 e^{-\sigma_2 (V(i) - \sigma_3)}\right)^{\sigma_4}} + 1$	Apply sigmoid activation function (which collapses to the constant function when $G < G^-$) to each cell
11	$\rho(i) = \frac{\psi(i)}{\sum_{k \in CM} \psi(k)}$	In each CM, normalize the relative probabilities of winning (ψ) to final probabilities (ρ) of winning
12	Select a final winner in each CM according to the ρ distribution in that CM, i.e., soft max	

learning mode, then its variation for retrieval, then its much simpler retrieval mode. See **Table I-2** for definitions of symbols used in equations and throughout the paper.

CSA: LEARNING MODE

The overall goal of the CSA when in learning mode is to assign codes to a mac's inputs in adherence with the SISC property, i.e., more similar overall inputs to a mac are mapped to more highly intersecting SDCs. With respect to each of a mac's individual afferent RFs, U, H, and D, the similarity metric is extremely primitive: the similarity of two patterns in an afferent RF is simply an increasing function of *the number of features* in common between the two patterns, thus embodying only what Bengio et al. (2012) refer to as the weakest of priors, the *smoothness* prior. However, the CSA multiplicatively combines these component similarity measures and, because the H and D signals carry temporal information reflecting the history of the sequence being processed, the CSA implements a spatiotemporal similarity metric. Nevertheless, the ability to learn arbitrarily complex *nonlinear* similarity metrics (i.e., category boundaries, or invariances), requires a hierarchical network of macs and the ability for an individual SDC, e.g., active in one mac, to associate with multiple (perhaps arbitrarily different) SDCs in one or more other macs. We elaborate more on Sparsey's implementation of this capability in Section Learning arbitrarily complex nonlinear similarity metrics.

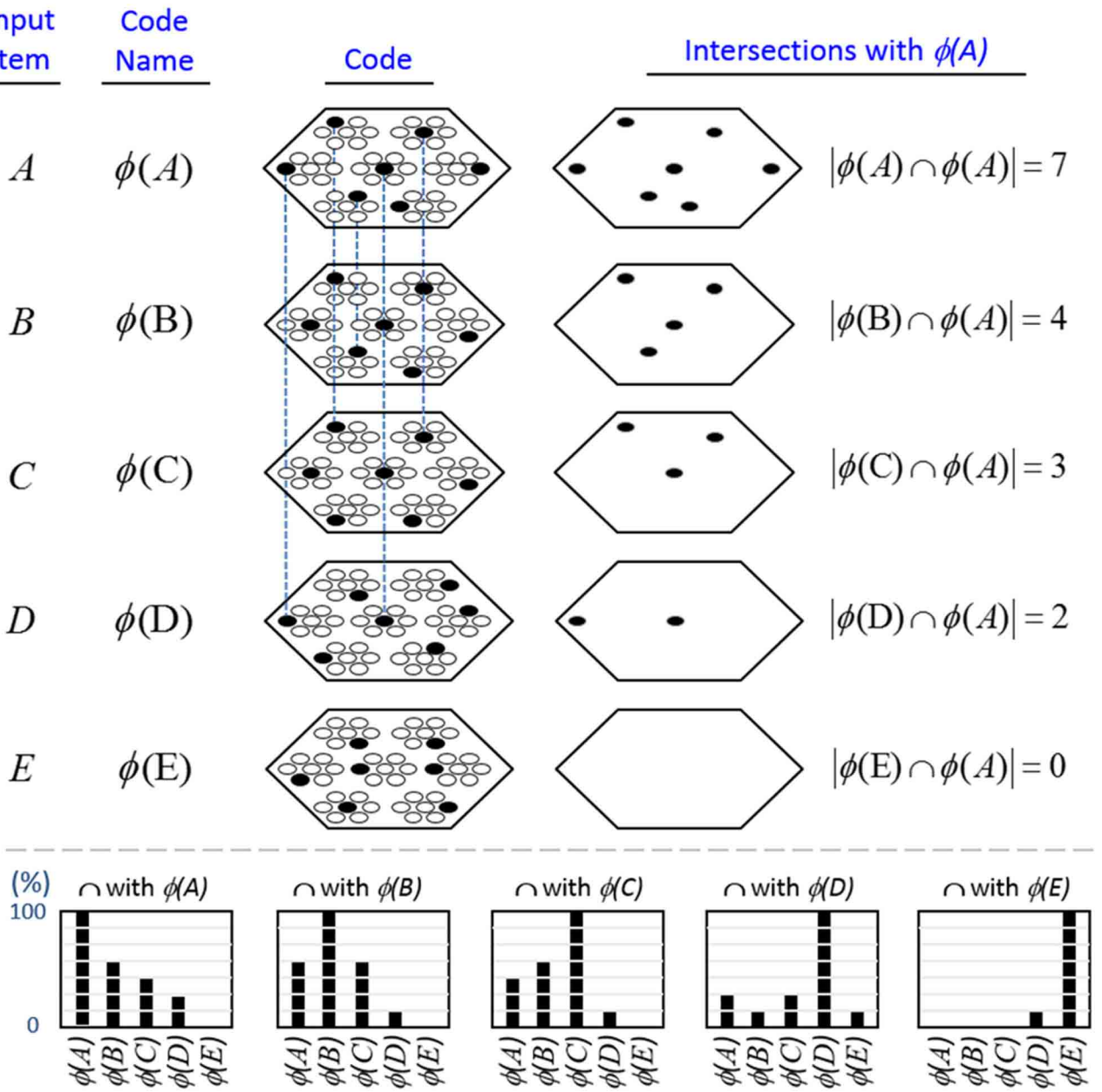


FIGURE I-6 | If the process that assigns SDCs to inputs enforces the *similar-input-to-similar-codes* (SISC) property, then the currently active code in a mac simultaneously physically functions as the entire likelihood distribution over all hypotheses stored in the mac. At bottom, we show the activation strength distribution over all five codes (stored hypotheses), when each of the five codes is fully active. If SISC was enforced when these codes were assigned (learned), then these distributions are interpretable as likelihood distributions. See text for further discussion.

The CSA has 12 steps which can be broken into two phases. Phase 1 (Steps 1–7) culminates in computation of the *familiarity*, G (normalized to [0,1]), of the overall (H, U, and D) input to the mac as a whole, i.e., G is a function of the *global* state of the mac. To first approximation, G is the similarity of the current overall input to the closest-matching previously stored (learned) overall input. As we will see, computing G involves a round of deterministic (hard max) competition resulting in one winning cell in each of the Q CMs. In Phase 2 (Steps 8–12), the activation function of the cells is modified based on G and a second round of competition occurs, resulting in the final set of Q winners, i.e., the activated code in the mac on the current time step. The second round of competition is *probabilistic* (*soft* max), i.e., the winner in each CM is chosen as a draw from a probability distribution over the CM's K cells.

In neural terms, each of the CSA's two competitive rounds entail the principal cells in each CM integrating their inputs,

engaging the local inhibitory circuitry, resulting in a single spiking winner. The difference is that the cell activation functions (F/I-curves) used during the second round of integration will generally be very different from those used during the first round. Broadly, the goal is as follows: as G approaches 1, make cells with larger inputs compared to others in the CM increasingly likely to win in the second round, whereas as G approaches 0, make all cells in a CM equally likely to win in the second round. We discuss this further in Section Neural implementation of CSA.

We now describe the steps of the CSA in learning mode. We will refer to the generic "circuit model" in **Figure II-1** in describing some of the steps. The figure has two internal levels with one small mac at each level, but the focus, in describing the algorithm, will be on the L1 mac, M_j^1, highlighted in yellow. M_j^1 consists of $Q = 4$ CMs, each with $K = 3$ cells. Gray arrows represent the U-wts from the input level, L0, consisting of 12 binary pixels. Magenta arrows represent the D-wts from the L2 mac. Green lines depict a subset of the H-wts. The representation of where the different afferents arrive on the cells is not intended to be veridical. The depicted "Max" operations are the hard max operations of CSA Step 7. The blue arrows portray the mac-global G-based modulation of the cellular V-to-ψ map (essentially, the F/I curve). The probabilistic draw operation is not explicitly depicted in this circuit model.

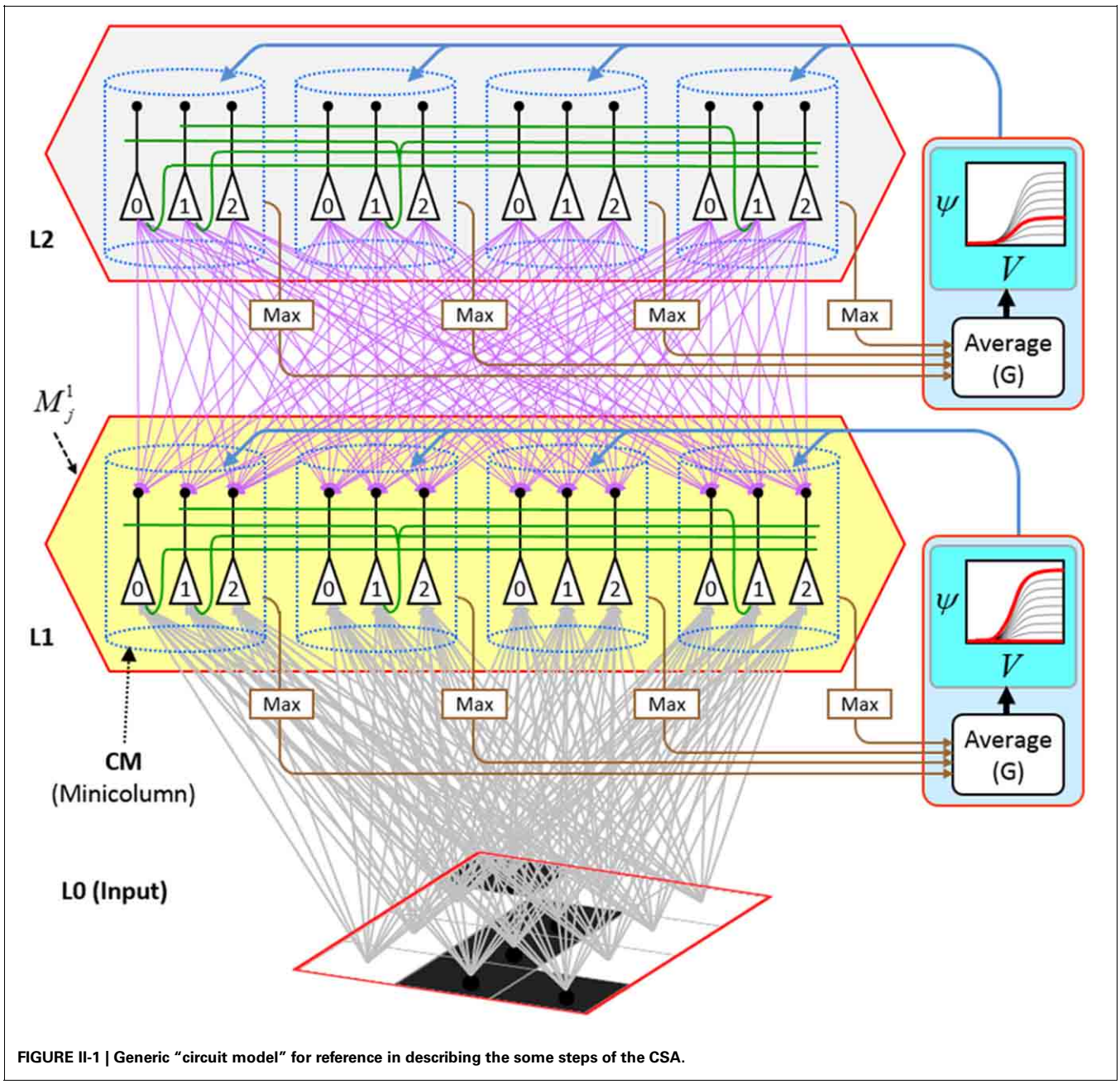


FIGURE II-1 | Generic "circuit model" for reference in describing the some steps of the CSA.

Step 1: Determine if the mac will become active

As shown in Equation (1), during learning, a mac, m, becomes active if either of two conditions hold: (a) if the number of active features in its U-RF, $\pi_U(m)$, is between π_U^- and π_U^+; or (b) if it is already active but the number of frames that it has been on for, i.e., its *code age*, $\Upsilon(m)$, is less than its persistence, $\delta(m)$. That is, during learning, we want to ensure that codes remain on for their entire prescribed persistence durations. We currently have no conditions on the number of active features in the H and D RFs.

$$Active(m) = \begin{cases} true & \Upsilon(m) < \delta(m) \\ true & \pi_U^- \leq \pi_U(m) \leq \pi_U^+ \\ false & otherwise \end{cases} \quad (1)$$

Step 2: Compute raw U, H, and D-summations for each cell, i, in the mac

Every cell, i, in the mac computes its three weighted input summations, $u(i)$, as in Equation (2a). RF_U is a synonym for U-RF. $a(j, t)$ is pre-synaptic cell j's activation, which is binary, on the current frame. Note that the synapses are *effectively binary*. Although the weight range is [0,127], pre-post correlation causes a weight to increase immediately to $w_{\max} = 127$ and the asymptotic weight distribution will have a tight cluster around 0 (for weights that are effectively "0") and around 127 (for weights that are effectively "1"). The learning policy and mechanics are described in Section Learning policy and mechanics. $F(\zeta(j, t))$ is a term needed to adjust the weights of afferent signals from cells in macs in which multiple competing hypotheses (MCHs) are active. If the number of MCHs (ζ) is small then we want to boost the weights of those signals, but if it gets too high, in which case we refer to the source mac as being *muddled*, those signals will generally only serve to decrease SNR in target macs and so we disregard them. Computing and dealing with MCHs is described in Steps 5 and 6. $h(i)$ and $d(i)$ are computed in analogous fashion Equations (2b) and (2c), with the slight change that H and D signals are modeled as originating from codes active on the previous time step $(t - 1)$.

$$u(i) = \sum_{j \in \mathrm{RF_U}} a(j, t) \times F(\zeta(j, t)) \times w(j, i) \quad (2a)$$

$$h(i) = \sum_{j \in \mathrm{RF_H}} a(j, t-1) \times F(\zeta(j, t-1)) \times w(j, i) \quad (2b)$$

$$d(i) = \sum_{j \in \mathrm{RF_D}} a(j, t-1) \times F(\zeta(j, t-1)) \times w(j, i) \quad (2c)$$

Step 3: Normalize and filter the raw summations

The summations, $u(i)$, $h(i)$, and $d(i)$, are normalized to [0,1] interval, yielding $U(i)$, $H(i)$, and $D(i)$. We explained above that a mac m only becomes active if the number of active features in its U-RF, $\pi_U(m)$, is between π_U^- and π_U^+, referred to as the lower and upper *mac activation bounds*. Given our assumption that visual inputs to the model are filtered to single-pixel-wide edges and binarized, we expect relatively straight or low-curvature edges roughly spanning the diameter of an L0 aperture to occur rather frequently in natural imagery. **Figure II-2** shows two examples of such inputs, as frames of sequences, involving either only a single L0 aperture (panel A) or a region consisting of three L0 apertures, i.e., as might comprise the U-RFs of an L2 mac (e.g., as in **Figure I-4B**). The general problem, treated in this figure, is that the number of features present in a mac's U-RF, $\pi_U(m)$, may vary from one frame to the next. Note that for macs at L2 and higher, the number of features present in an RF is the *number of active macs* in that RF, not the total number of active cells in that RF. The policy implemented in Sparsey is that inputs with different numbers of active features compete with each other on an equal footing. Thus, normalizers (denominators) in Equations (3a–c) use the lower mac activation bound, π_U^-, π_H^-, and π_D^-. This necessitates hard limiting the maximum possible normalized value to 1, so that inputs with between π_U^- and π_U^+ active features yield normalized values confined to [0,1]. There is one additional nuance. As noted above, if a mac in m's U-RF is muddled, then we disregard all signals from it, i.e., they are not included in the u-summations of m's cells. However, since that mac is active, it will be included in the number of active features, $\pi_U(m)$. Thus, we should normalize by the number of active, *nonmuddled* macs in m's U-RF (not simply the number of active macs): we denote this value as π_U^*. Finally, note that when the afferent feature is represented by a mac, that feature is actually being represented by the simultaneous activation of, and thus, inputs from, Q cells; thus the denominator must be adjusted accordingly, i.e., multiplied by Q and by the maximum weight of a synapse, $w_{\max}$.

$$U(i) = \begin{cases} \max(1,\, u(i)/\pi_U^- \times w_{\max}) & L = 1 \\ \max(1,\, u(i)/\min(\pi_U^-, \pi_U^*) \times Q \times w_{\max}) & L > 1 \end{cases} \quad (3a)$$

$$H(i) = \max\left(1,\, h(i)/\min\left(\pi_H^-, \pi_H^*\right) \times Q \times w_{\max}\right) \quad (3b)$$

$$D(i) = \max\left(1,\, d(i)/\min\left(\pi_D^-, \pi_D^*\right) \times Q \times w_{\max}\right) \quad (3c)$$

Step 4: Compute overall local support for each cell in the mac

The overall *local* (to the individual cell) measure, $V(i)$, of evidence/support that cell i should be activated is computed by multiplying filtered versions of the normalized inputs as in Equation (4). $V(i)$ can also be viewed as the normalized degree of match of cell i's total afferent (including U, H, and D) synaptic weight vector to its total input pattern. We emphasize that the V measure is *not* a measure of support for a *single* hypothesis, since an individual cell does not represent a single hypothesis. Rather, in terms of hypotheses, $V(i)$ can be viewed as the local support for the *set* of hypotheses whose representations (codes) include cell i. The individual normalized summations are raised to powers (λ), which allows control of the relative sensitivities of V to the different input sources (U, H, and D). Currently, the U-sensitivity parameter, λ_U, varies with time (index of frame with respect to beginning of sequence). We will add time-dependence to the H and D sensitivity parameters as well and explore the space of policies regarding these schedules in the future. In general terms, these parameters (along with many others) influence the shapes of the boundaries of the categories learned by a mac.

$$V(i) = \begin{cases} H(i)^{\lambda_H} \times U(i)^{\lambda_U(t)} \times D(i)^{\lambda_D} & t \geq 1 \\ U(i)^{\lambda_U(0)} & t = 0 \end{cases} \quad (4)$$

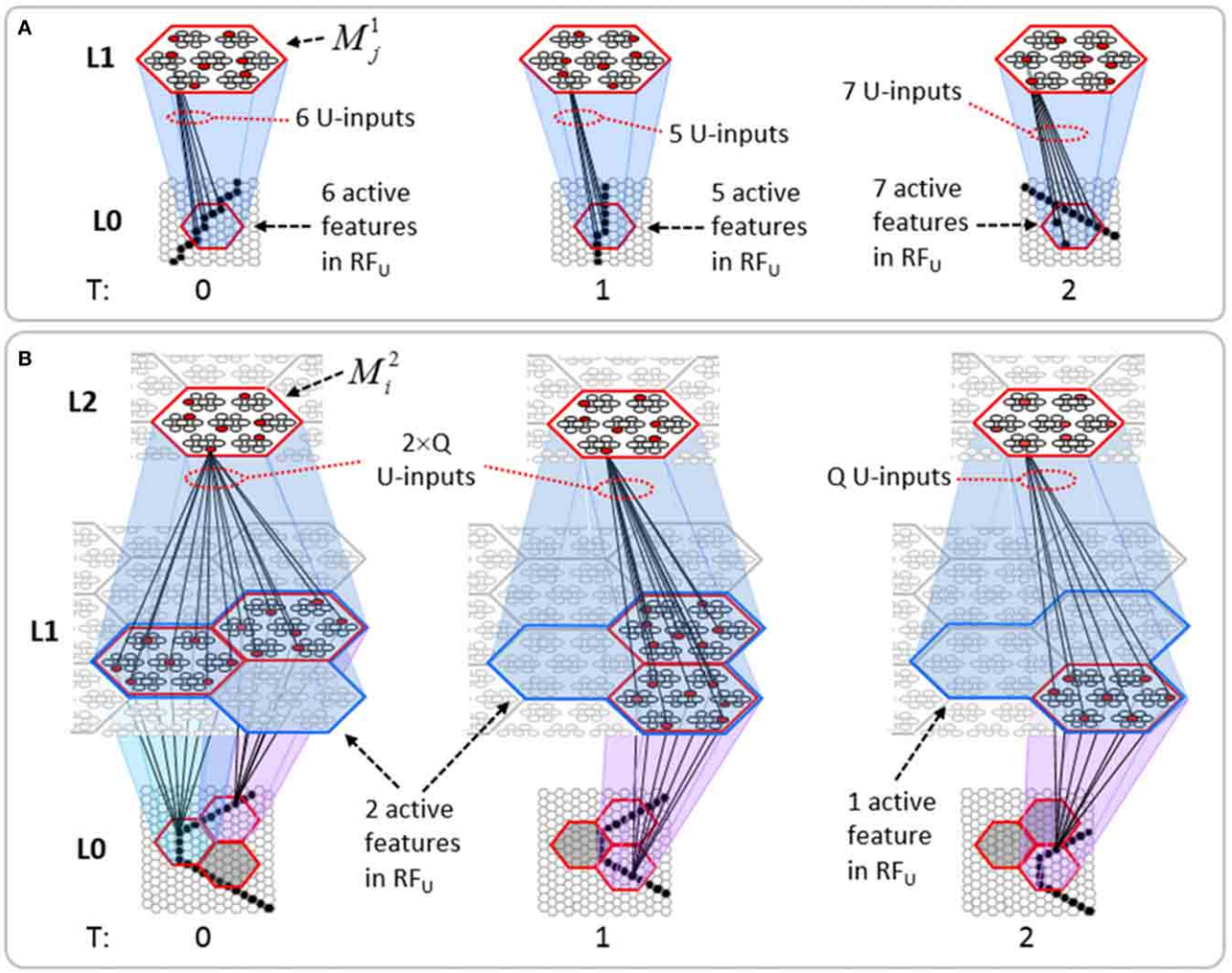


FIGURE II-2 | The mac's normalization policy must be able to deal with inputs of different sizes, i.e., inputs having different numbers of active features. (A) An edge rotates through the aperture over three time steps, but the number of active features (in this case, pixels) varies from one time step (moment) to the next. In order for the mac to be able to recognize the 5-pixel input ($T = 1$) just as strongly as the 6 or 7-pixel inputs, the u-summations must be divided by 5. **(B)** The U-RFs of macs at L2 and higher consist of an integer number of subjacent level macs, e.g., here, M_i^2's U-RF consists of three L1 macs (blue border). Each active mac in M_i^2's U-RF represents one feature. As for panel a, the number of active features varies across moments, but in this case, the variation is in increments/decrements of Q synaptic inputs. Grayed-out apertures have too few active pixels for their associated L1 macs to become active.

As described in Section CSA: Retrieval Mode, during retrieval, this step is significantly generalized to provide an extremely powerful, general, and efficient mechanism for dealing with arbitrary, nonlinear invariances, most notably, nonlinear time-warping of sequences.

Step 5: Compute the number of competing hypotheses that will be active in the mac once the final code for this frame is activated

To motivate the need for keeping track of the number of competing hypotheses active in a mac, we consider the case of *complex* sequences, in which the same input item occurs multiple times and in multiple contexts. **Figure II-3** portrays a minimal example in which item B occurs as the middle state of sequences [ABC] and [DBE]. Here, the model's single internal level, L1, consists of just one mac, with $Q = 4$ CMs, each with $K = 4$ cell. **Figure II-3A** shows notional codes (SDCs) chosen on the three time steps of [ABC]. The code name convention here is that ϕ denotes a code, the superscript "1" indicates the model level at which code resides. The subscript indicates the specific moment of the sequence that the code represents; thus, it is necessary for the subscript to specify the full temporal context, from start of sequence, leading up to the current input item. Successively active codes are chained together, resulting in spatiotemporal memory traces that represent sequences. Green lines indicate the H-wts that are increased from one code to the next. Black lines indicate the U-wts that are increased from currently active pixels to currently active L1 cells (red). Thus, as described earlier, e.g., in **Figure I-2**, individual cells learn spatiotemporal

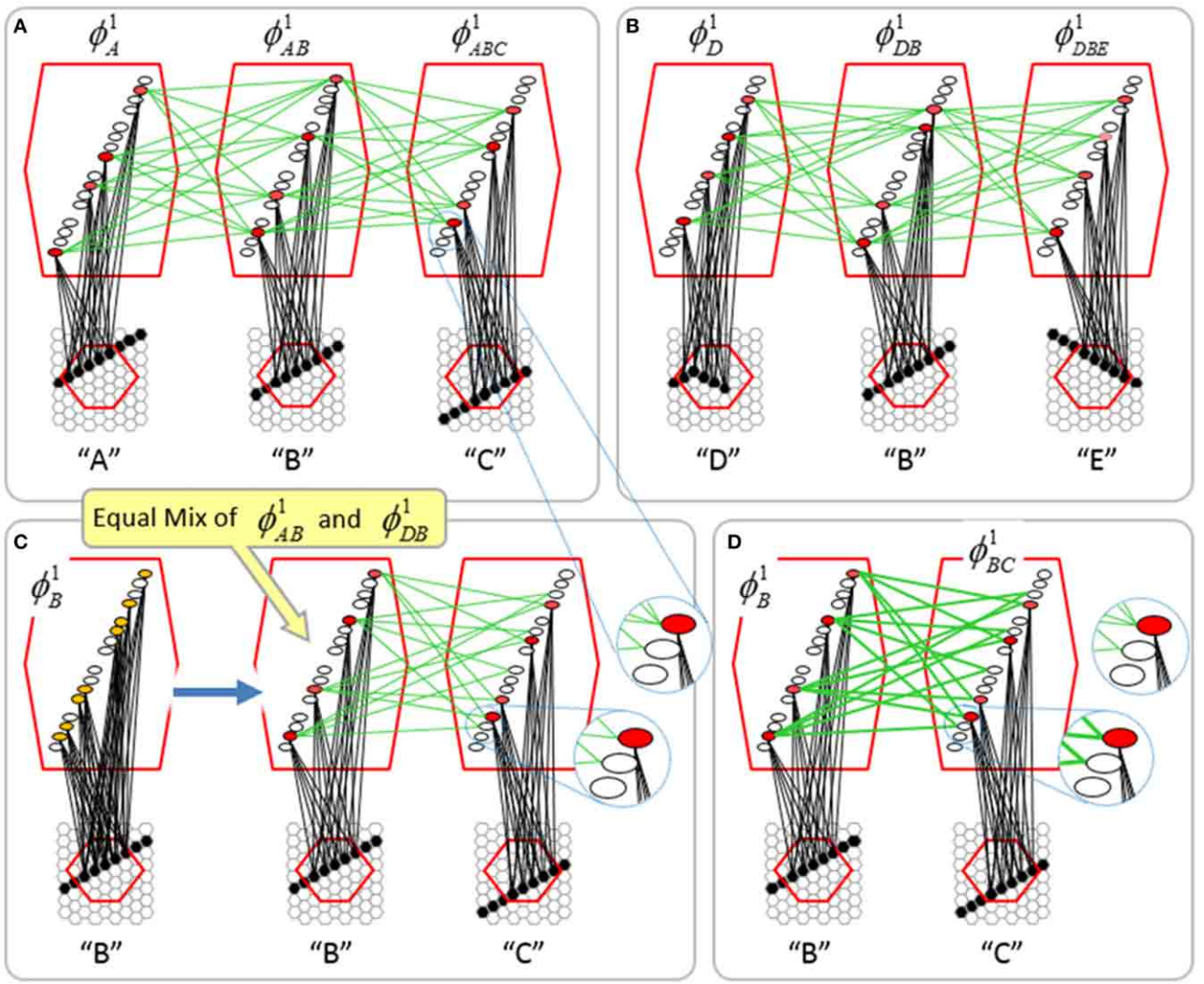


FIGURE II-3 | Portrayal of reason why macs need to know how many multiple competing hypotheses (MCHs) are/were active in their afferent macs. (A) Memory trace of 3-item sequence, [ABC]. This model has a single internal level with one mac consisting of Q = 4 CMs, each with K = 4 cells. We show notional SDCs (sets of red cells) for each of the three items. The green lines represent increased H-wts in the recurrent H-matrix: the trace is shown unrolled in time in time. **(B)** A notional memory trace of sequence [DBE]. The SDC chosen for item B differs from that in [ABC] because of the different temporal context signals, i.e., from the code for item D rather than the code for item A. **(C)** We prompt with item B, the model enters a state that has equal measures of both of B's previously assigned SDCs. Thus multiple (here, two) hypotheses are equally active. **(D)** If the model can detect that multiple hypotheses are active in this mac, then it can boost its efferent H-signals (multiplying them by the number of MCHs), in which case the combined H and U signals when the next item, here "C", is presents, causing the SDC for the moment [ABC] to become fully active. See text for more details.

inputs in correlated fashion, as whole SDCs. Learning is described more thoroughly in Section Learning Policy and Mechanics.

As portrayed in **Figure II-3B**, if [ABC] has been previously learned, then when item B of another sequence, [DBC], is encountered, the CSA will generally cause a different SDC, here, ϕ^1_{DB}, to be chosen. ϕ^1_{DB} will be H-associated with whatever code is activated for the next item, in this case ϕ^1_{DBE} for item E. This choosing of codes in a context-dependent way (where the dependency has no fixed Markov order and in practice can be extremely long), enables subsequent recognition of complex sequences without confusion.

However, what if in some future recognition test instance, we prompt the network with item B, i.e., as the first item of the sequence, as shown in **Figure II-3C**? In this case, there are no active H-wts and so the computation of local support Equation (4) depends only on the U-wts. But, the pixels comprising item B have been fully associated with the two codes, ϕ^1_{AB} and ϕ^1_{DB}, which have been assigned to the two moments when item B was presented, [A**B**] and [D**B**]. We show the two maximally implicated (more specifically, maximally U-implicated) cells in each CM as orange to indicate that a choice between them in each CM has not yet been made. However, by the time the CSA completes for the frame when item B is presented, one winner must

be chosen in each CM (as will become clear as we continue to explain the CSA throughout the remainder of section Sparsey's Core Algorithm). And, because it is the case in each CM, that both orange cells are equally implicated, we choose winners randomly between them, resulting in a code that is an equal mix of the winners from ϕ^1_{AB} and ϕ^1_{DB}. In this case, we refer to the mac as having multiple competing hypotheses active (MCHs), where we specifically mean that all the active hypotheses (in this case, just two) are approximately equally strongly active.

The problem can now be seen at the right of **Figure II-3C** when C is presented. Clearly, once C is presented, the model has enough information to know which of the two learned sequences, or more specifically, which particular moment is intended, [AB**C**] rather than [DB**E**]. However, the cells comprising the code representing that learned moment, ϕ^1_{ABC}, will, at the current test moment (lower inset in **Figure II-3C**), have only half the active H-inputs that they had during the original learning instance (i.e., upper inset in **Figure II-3C**). This leads, once processed through steps 2b, 3b, and 4, to V-values that will be far below $V = 1$, for simplicity, let's say $V = 0.5$, for the cells comprising ϕ^1_{ABC}. As will be explained in the remaining CSA steps, this ultimately leads to the model *not* recognizing the current test trial moment [B**C**] as equivalent to the learning trial moment [AB**C**], and consequently, to activation of a new code that could in general be arbitrarily different from ϕ^1_{ABC}.

However, there is a fairly general solution to this problem where multiple competing hypotheses are present in an active mac code, e.g., in the code for B indicated by the yellow callout. The mac can easily detect when an MCH condition exists. Specifically, it can tally the number cells with $V = 1$—or, allowing some slight tolerance for considering a cell to be *maximally* implicated, cells with $V(i) > V_\zeta$, where V_ζ is close to 1, e.g., $V_\zeta = 0.95$—in each of its Q CMs, as in Equation (5a). It can then sum ζ_q over all Q CMs and divide by Q (and round to the nearest integer, "**rni**"), resulting in the number of MCHs active in the mac, ζ, as in Equation (5b). In this example, $\zeta = 2$, and the principle by which the H-input conditions, specifically the h-summations, for the cells in ϕ^1_{ABC} on this test trial moment [B**C**] can be made the same as they were during the learning trial moment [AB**C**], is simply to multiply all outgoing H-signals from ϕ^1_B by $\zeta = 2$. We indicate the *inflated* H-signals by the thicker green lines in the lower inset at right of **Figure II-3D**. This ultimately leads to $V = 1$ for all four cells comprising ϕ^1_{ABC} and, via the remaining steps of the CSA, reinstatement of ${}^1\phi_{ABC}$ with very high probability (or with certainty, in the simple retrieval mode described in Section CSA: Simple Retrieval Mode), i.e., with recognition of test trial moment [BC] as equivalent to learning trial moment [ABC]. The model has successfully gotten through an ambiguous moment based on presentation of further, disambiguating inputs.

We note here that uniformly boosting the efferent H-signals from ϕ^1_B also causes the h-summations for the four cells comprising the code ϕ^1_{DBE} to be the same as they were in the learning trial moment [DBE]. However, by Equation (4), the V-values depend on the U-inputs as well. In this case, the four cells of ϕ^1_{DBE} have u-summations of zero, which leads to $V = 0$, and ultimately to essentially zero probability of any of these cells winning the competitions in their respective CMs. Though we don't show the example here, if on the test trial, we present E instead of C after B, the situation is reversed; the u-summations of cells comprising the code ϕ^1_{DBE} are the same as they were in the learning trial moment [DB**E**] whereas those of the cells comprising the code ϕ^1_{ABC} are zero, resulting with high probability (or certainty) in reinstatement of ϕ^1_{DBE}.

$$\zeta_q = \sum_{i=0}^{K} \left[V(i) > V_\zeta \right] \tag{5a}$$

$$\zeta = \mathbf{rni}\left(\sum_{j=0}^{Q-1} \zeta_q / Q \right) \tag{5b}$$

Step 6: Compute correction factor for multiple competing hypotheses to be applied to efferent signals from this mac

The example in **Figure II-3** was rather clean in that it involved only two sequences having been learned, containing a total of six moments, [**A**], [A**B**], [AB**C**], [**D**], [D**B**], and [DB**E**], and very little pixel-wise overlap between the items. Thus, cross-talk between the stored codes was minimized. However, in general, macs will store far more codes. If for example, the mac of **Figure II-3** was asked to store 10 moments where B was presented, then, if we prompted the network with B as the first sequence item, we would expect almost all cells in all CMs to have $V = 1$. As discussed in Step 2, when the number of MCHs (ζ) in a mac gets too high, i.e., when the mac is *muddled*, its efferent signals will generally only serve to decrease SNR in target macs (including itself on the next time step via the recurrent H-wts) and so we disregard them. Specifically, when ζ is small, e.g., two or three, we want to boost the value of the signals coming from all active cells in that mac by multiplying by ζ (as in **Figure II-3D**). However, as ζ grows beyond that range, the expected overlap between the competing codes increases and to approximately account for that, we begin to diminish the boost factor as in Equation (6), where A is an exponent less than 1, e.g., 0.7. Further, once ζ reaches a threshold, B, typically set to 3 or 4, we multiply the outgoing weights by 0, thus effectively disregarding the mac completely in downstream computations. We denote the correction factor for MCHs as $F(\zeta)$, defined as in Equation (6). We also use the notation,$F(\zeta(j, t))$, as in Equation (2), where $\zeta(j, t)$ is the number of hypotheses tied for maximal activation strength in the owning mac of a pre-synaptic cell, j, at time (frame) t.

$$F(\zeta) = \begin{cases} \zeta^A & 1 \le \zeta \le B \\ 0 & \zeta > B \end{cases} \tag{6}$$

Step 7: Determine the maximum local support in each of the mac's CMs

Operationally, this step is quite simple: simply find the cell with the highest V-value, $\hat{V}_j$, in each CM, C_j, as in Equation (7).

$$\hat{V}_j = \max_{i \in C_j} \{V(i)\} \tag{7}$$

Conceptually, the cell with $\hat{V}_j$ in a CM is the cell most implicated by the mac's total input (multiple cells can be tied for $\hat{V}_j$), or in other words, the *most likely* winner in the CM. In fact, in the simple retrieval mode (Section CSA: Simple Retrieval Mode), the cell with $\hat{V}_j$ in each CM is chosen winner.

Step 8: Compute the familiarity of the mac's overall input

The average, G, of the maximum V's across the mac's Q CMs is computed as in Equation (8): G is a measure of the *familiarity* of the macs overall input. This is done on every time step (frame), so we sometimes denote G as a function of time, $G(t)$. And, G is computed independently for each activated mac, so we may also use more general notation that indicates mac as well.

$$G = \sum_{q=1}^{Q} \hat{V}_k / Q \tag{8}$$

The main intuition motivating the definition and use of G is as follows. If the mac's current input moment has been experienced in the past, then all active afferent weights (U, H, and D) to the code activated in that instance would have been increased. Thus, in the current moment, all Q cells comprising that code will have $V = 1$. Thus, $G = 1$. Thus, a *familiar* moment must always result in $G = 1$ (assuming that MCHs are accounted for as described above). On the other hand, suppose that the current *overall* input moment is novel, even if sub-components of the current overall input have been experienced exactly before. In this case, provided that few enough codes have been stored in the mac (so that crosstalk remains sufficiently small), there will be at least some CMs, C_j, for which $\hat{V}_j$ is significantly less than 1. Thus, $G < 1$. Moreover, as the examples in the Results section will show, G correlates with the familiarity of the overall mac input. Thus, G measures the familiarity, or inverse novelty, of the global input to the mac.

Note that in the brain, this step requires that the Q cells with $V = \hat{V}_j$ become active (i.e., spike) so that their outputs can be summed and averaged. This constitutes the first of two rounds of competition that occurs within the mac's CMs on each execution of the CSA. However, as explained herein, this set of Q cells will, in general, *not* be identical to (and can often be substantially different from, especially when $G \approx 0$) the finally chosen code for this execution of the CSA (i.e., the code chosen in Step 12).

Step 9: Determine the expansivity/compressivity of the I/O function to be used for the second and final round of competition within the mac's CMs

Determine the range, η, of the sigmoid activation function, which transforms a cell's V-value into its relative (within its own CM) probability of winning, ψ. We refer to that transform as the V-to-ψ map. We refer to χ as the *sigmoid expansion factor* and γ as the *sigmoid expansion exponent*.

$$\eta = 1 + \left(\left[\frac{G - G^{-}}{1 - G^{-}} \right]^{+} \right)^{\gamma} \times \chi \times K \tag{9}$$

As noted several times earlier, the overall goal of the CSA when in learning mode is to assign codes to a mac's inputs in adherence with the SISC property, i.e., more similar overall inputs to a mac are mapped to more highly intersecting SDCs. Given that G represents, to first approximation, the similarity of the closest-matching stored input to the current input, we can restate the goal as follows.

1. as G goes to 1, meaning the input X is completely familiar, we want the probability of reinstating the code ϕ_X that was originally assigned to represent X, to go to 1. It is the cells comprising ϕ_X, which are causing the high G-value. But these are the cells with the maximal V's ($V = \hat{V}_j = 1$) in their respective CMs. Thus, within each CM, C_j, we want to increase the probability of picking the cell with $V = \hat{V}_j$ relative to cells with $V < \hat{V}_j$, i.e., we want to transform the V's via an *expansive* nonlinearity
2. as G goes to 0 (completely novel input), we want the set of winners chosen to have the minimum average intersection with all stored codes. We can achieve that by choosing the winner in each CM from the uniform distribution, i.e., by making all cells in a CM equally likely to win, i.e., transform the V's via a maximally *compressive* nonlinearity.

The first goal is met by making the activation function a very expansive nonlinearity. **Figure II-4** shows how the expansivity of the V-to-ψ map affects cell win probability, and indirectly, whole-code reinstatement probability. All nine panels concern a small example mac with $Q = 6$ CMs each comprised of $K = 7$ cells. Each panel shows hypothetical V and ρ vectors over the cells of the CMs, across two parametrically varying conditions: model "age" (across columns), which we can take as a correlate of the number of stored codes and thus, of the amount of interference (crosstalk) between codes during retrieval, and expansivity (η) (across rows) of the V-to-ψ map. As described shortly, the V-values are first transformed to relative probabilities (ψ) (Step 10), which are then normalized to absolute probabilities (ρ) (Step 11). In all panels, the example V vector in each CM has one cell with $V = 1$ (pink bars). Thus, by Step 8, all panels correspond to a $G = 1$ condition. The other six cells (black bars) in each CM are assigned uniformly randomly chosen values in defined intervals that depend on the age of the model. The intervals for "Early," "Middle," and "Late," are [0.0, 0.1], [0.1, 0.5], and [0.2, 0.8], respectively, simulating the increasing crosstalk with age.

For each age condition, we show the effects of using a V-to-ψ map with three different η-values. Note that in actual operation (specifically, Step 9), all panels would be processed with a V-to-ψ map with the maximal η-value (again, because $G = 1$ in all panels). But our purpose here is just to show the consequences on the final ρ distribution for a given V distribution (the V distribution is the same for all three rows in any given column) as a function of η. And, note that the minimum ψ-value in all cases is 1. Thus, for the "Early" column, the highly expansive V-to-ψ map ($\eta = 300$) (top row) results in a 300/306 $\approx$ 98% probability of selecting the cell with $V = 1$ (pink) in each CM. This results in a $(300/306)^6 \approx 89\%$ probability of choosing the pink cell in *all* $Q = 6$ CMs, i.e., of reinstating the *entire* correct code. In the

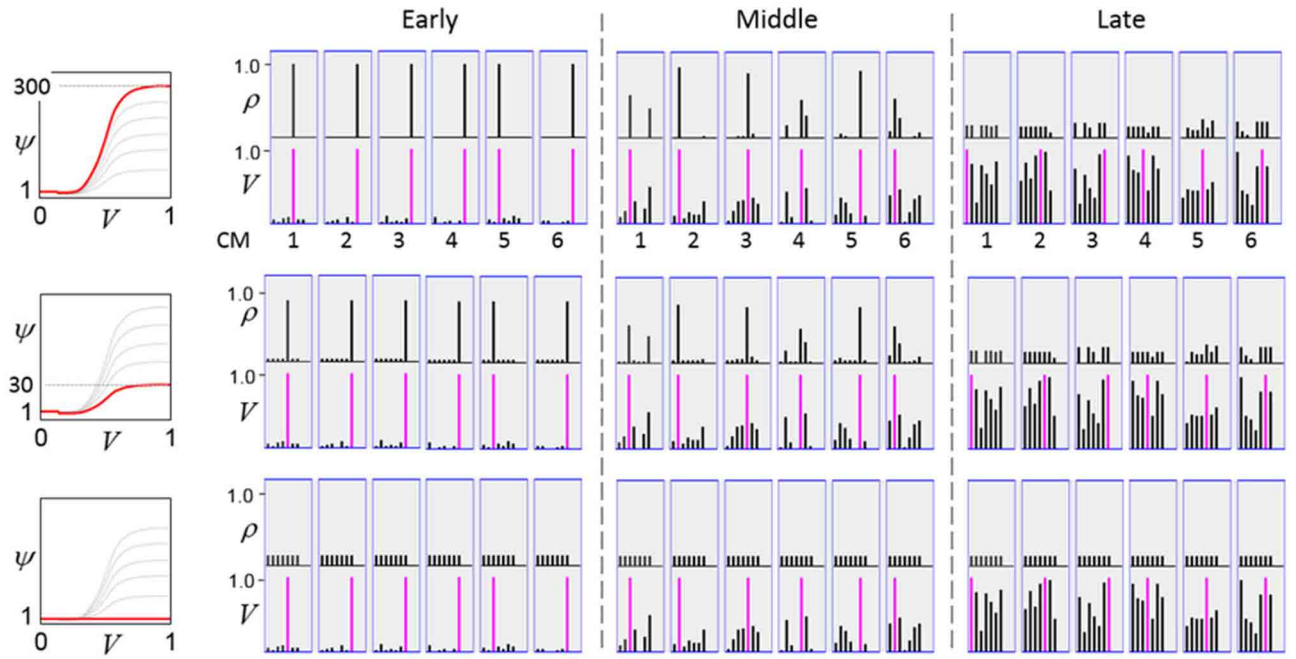


FIGURE II-4 | *G*-based sigmoid transform characteristics. All panels show hypothetical V and ρ vectors over the $K = 7$ cells in each of the $Q = 6$ CMs comprising the mac. In all nine panels, the V vector in each CM has one cell prescribed to have $V = 1$ (pink bars). The V's of the other six cells (black bars) in each CM are drawn randomly from defined intervals that depend on the age (amount of inputs experienced) of the model. For each age condition, we show the effects of using a V-to-ψ map with three different η-values. But our purpose here is just to show the consequences on the final ρ distribution for a given V distribution (the V distribution is the same for all three rows in any given column) as a function of the expansivity/compressivity (η) of the V-to-ψ map. See text for details.

second row, η is reduced to 30. Each of the six black cells ultimately ends up with a 1/36 probability of winning and the pink cell, with a $30/36 = 5/6$ win probability. In this case the likelihood of reinstating the *entire* correct code, is $(5/6)^6 \approx 33\%$. In the bottom row, $\eta = 1$, i.e., the V-to-ψ map has been collapsed to the constant function, $\psi = 1$. As can be seen, all cells, including the cell with $V = 1$ become equally likely to be chosen winner in their respective CMs.

Greater crosstalk can clearly be seen in the "Middle" condition. Consequently, even for $\eta = 300$, several of the cells with nonmaximal V end up with significant final probability ρ of being chosen winner in their respective CMs. The ρ-distributions are slightly further compressed (flatter) when $\eta = 30$, and completely compressed when $\eta = 1$ (bottom row). The "Late" condition is intended to model a later period of the life of the model, after many memories (codes) have been stored in this mac. Thus, when the input pattern associated with any of those stored codes is presented again, many of the cells in each CM will have an appreciable V-value (again, here they are drawn uniformly from [0.2, 0.8]). In this condition, even if $\eta = 300$, the probability of selecting the correct cell (pink) in each CMs is close to chance, as is the chance of reinstating the entire correct code. And the situation only gets worse for lower η-values.

Note that for any particular V distribution in a CM, the relative increase to the final probability of being chosen winner is a smoothly and faster-than-linearly increasing (typically, $\gamma \geq 2$) function of G. Thus, in each CM, the probability that the most highly implicated (by the mac's total input) cell (those corresponding to the pink bars in **Figure II-4**) wins increases smoothly as G goes to 1. (Strictly, this is true only for the portion of the sigmoid nonlinearity with slope > 1). The initial (left) and final (right) portions of the sigmoid are compressive ranges.) And since the overall code is just the result of the Q independent draws, it follows that the expected intersection of the code consisting of the Q most highly implicated cells, i.e., the code of the closest-matching stored input, with the finally chosen code is also an increasing function of G, i.e., thus realizing the "SISC" property.

Step 10: Apply the modulated activation function to all the mac's cells, resulting in a relative probability distribution of winning over the cells of each CM

Apply sigmoid activation function to each cell. Note: the sigmoid collapses to a constant function, $\psi(i) = 1$, when $\eta = 1$ (i.e., when $G < G^-$).

$$\psi(i) = \frac{(\eta - 1)}{(1 + \sigma_1 e^{-\sigma_2 (V(i) - \sigma_3)})^{\sigma_4}} + 1 \qquad (10)$$

In a more general development, the CSA could include additional prior steps for setting any of the other sigmoid parameters, σ_1, σ_2, σ_3, and σ_4, all of which interact to control overall sigmoid expansivity and shape. In particular, in the current implementation,

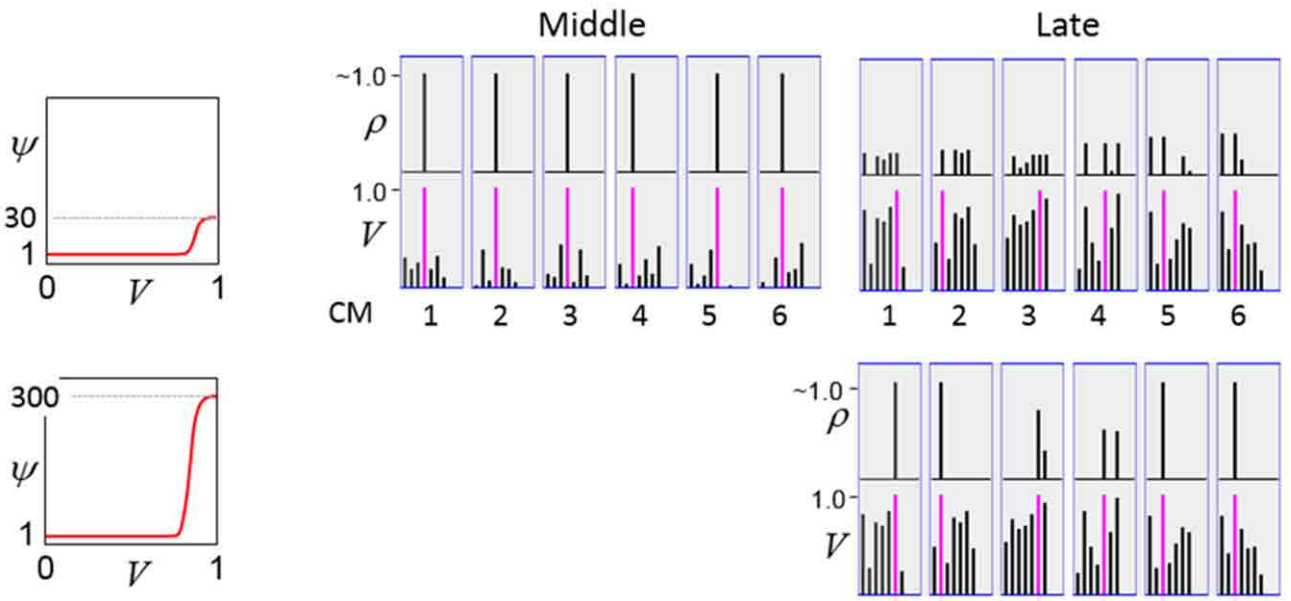


FIGURE II-5 | Moving the inflection point of the sigmoidal *V*-to-ψ map to the right greatly increases the probability of selecting the correct cell despite mounting crosstalk due to a growing number of codes stored in superposition.

the horizontal position of the sigmoid's inflection point is moved rightward as additional codes are stored in a mac. **Figure II-5** shows that doing so greatly increases the probability of choosing the correct cell in each CM and thus, of reinstating the entire correct code, even when many codes have been stored in the mac. In the "Middle" condition, even if $\eta = 30$, the probability of choosing the pink cell in each CM is very close to 1. For the "Late" condition, setting $\eta = 30$ significantly improves the situation relative to the top right panel of **Figure II-4** and setting $\eta = 300$ makes the probability of choosing the correct cell close to 1 in four of the six CMs. Thus, we have a mechanism for keeping memories accessible for longer lifetimes.

Step 11: Convert relative win probability distributions to absolute distributions

In each of the mac's CMs, the ψ-values of the cells are converted to true probabilities of winning (ρ) and the winner is selected by drawing from the ρ distribution, resulting in a final SDC, ϕ, for the mac, as in Equation (11).

$$\rho(i) = \frac{\psi(i)}{\sum_{k \in CM} \psi(k)} \tag{11}$$

Step 12: Pick winners in the mac's CMs, i.e., activate the SDC

The last step of the CSA is just selecting a final winner in each CM according to the ρ distribution in that CM, i.e., soft max. This is the second round of competition. Our hypothesis that the canonical cortical computation involves two rounds of competition is a strong and falsifiable prediction of the model with respect to actual neural dynamics, which we would like to explore further.

The CSA is given in **Table I-1**.

Learning policy and mechanics

Broadly, Sparsey's learning policy can be described as Hebbian with passive weight decay. As noted earlier, the model's synapses are *effectively binary*. By this we mean that although the weight range is [0,127], the several learning related properties conspire to cause the asymptotic weight distribution to tend toward having two spikes, one at 0 and the other at $w_{\max} = 127$, thus effectively being binary.

In actuality, a synapse's weight, $w(j, i)$, where j and i index the pre- and postsynaptic cells, respectively, is determined by two primary variables, its *age*, $\sigma(j, i)$, which is the number of time steps (e.g., video frames) since it was last increased, and its *permanence*, $\theta(j, i)$, which measures how resistant to decrease the weight is (i.e., the passive decay rate). The learning law is implemented as follows. Whenever a synapse's pre- and postsynaptic cells are coactive [i.e., a "pre-post correlation," $a(j) = 1 \wedge a(i) = 1$], its age is set to zero, as in Equation (12a), which has the effect of setting its weight to $w_{\max}$ (as can be seen in the "weight table" of **Figure II-6**, an age of zero always maps to $w_{\max}$). Otherwise, $\sigma(j, i)$ increases by one on each successive time step (across all frames of all sequences presented) on which there is no pre-post correlation Equation (12c), stopping when it gets to the maximum age, $\sigma_{\max}$ Equation (12d). Also note that once a synapse has reached maximum permanence, $\theta_{\max}$, its age stays at zero, i.e., its weight stays at $w_{\max}$ Equation (12b). At any point, the synapse's weight, $w(j,i)$, is gotten by dereferencing $\sigma(j, i)$ and $\theta(j, i)$ from the weight table shown in **Figure II-6**.

The intent of the decay schedule (for any permanence value) is to keep the weight at or near $w_{\max}$ for some initial window of time (number of time steps), $T_\sigma(\theta)$, and then allow it to decay increasingly rapidly toward zero. Thus, the model "assumes" that a pre-post correlation reflects an important / meaningful event

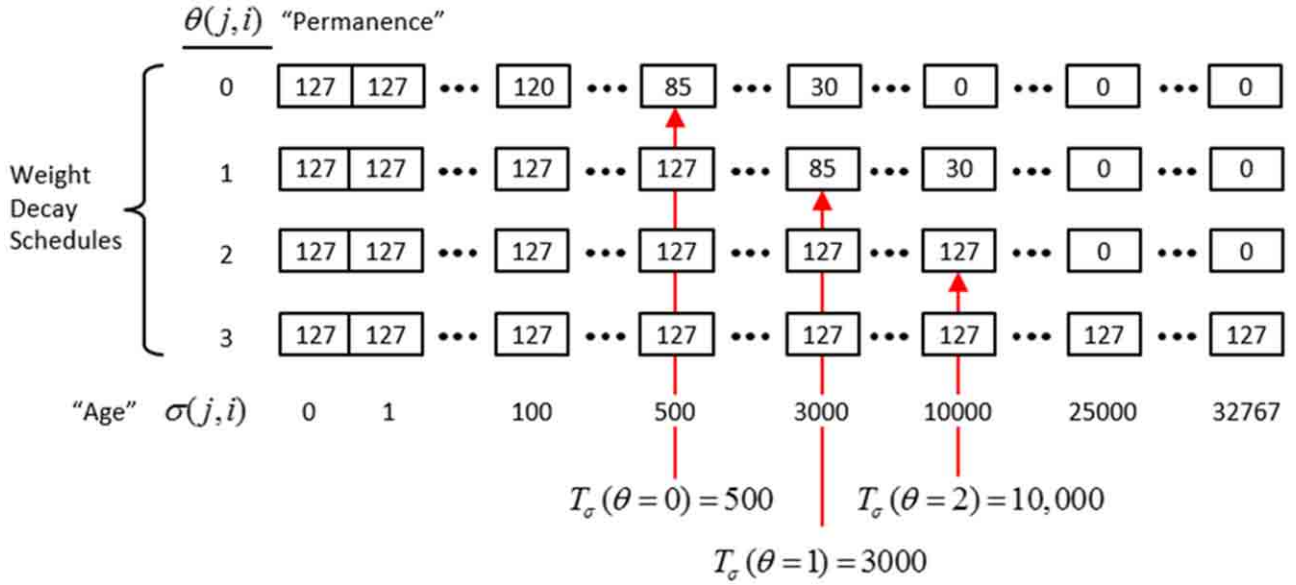


FIGURE II-6 | The "weight table": Indexed by age (columns) and permanence (rows). A synapse's weight is gotten by dereferencing its age, $\sigma(j, i)$, and its permanence, $\theta(j, i)$. See text for details.

in the input space and therefore strongly embeds it in memory (consistent with the notion of episodic memory). If the synapse experiences a second pre-post correlations within the window $T_\sigma(\theta)$, its permanence is incremented as in Equation (13) and $\sigma(j, i)$ is set back to 0 (i.e., its weight is set back to $w_{\max}$); otherwise the age, $\sigma(j, i)$, increases by one with each time step and the weight decreases according to the decay schedule in effect. Thus, pre-post correlations due to noise or spurious events, which will have a much longer expected time to recurrence, will tend to fade from memory. Sparsey's permanence property is closely related to the notion of synaptic tagging (Frey and Morris, 1997; Morris and Frey, 1999; Sajikumar and Frey, 2004; Moncada and Viola, 2007; Barrett et al., 2009).

$$\sigma(j,i) = \begin{cases} 0 & , \quad a(j) = 1 \wedge a(i) = 1 \quad (12a) \\ 0 & , \quad \theta(j,i) = \theta_{\max} \quad (12b) \\ \sigma(j,i) + 1 & , \quad a(j) = 0 \vee a(i) = 0 \quad (12c) \\ \sigma(j,i) & , \quad \sigma(j,i) = \sigma_{\max} \quad (12d) \end{cases}$$

$$\theta(j,i) = \begin{cases} \theta(j,i) + 1 & , \; a(j) = 1 \wedge a(i) = 1 \wedge \sigma(j,i) \leq T_\sigma(\theta(j,i)) \\ \theta(j,i) & , \text{ otherwise} \end{cases} \quad (13)$$

The exact parametric details are less important, but as can be seen in the weight table, the decay rate decreases with $\theta(j, i)$ and the window, $T_\sigma(\theta)$, within which a second pre-post correlation will cause an increase in permanence, increases with $\theta(j, i)$ (three example values shown). Permanence can only increase and in our investigations thus far, we typically make a synaptic weight completely permanent on the second or third within-window pre-post correlation [$\theta_{\max} = 1$ or $\theta_{\max} = 2$, respectively]. The justification of this policy derives from two facts: (a) a mac's input is a sizable set of co-active cells; and (b) due to the SISC property, the probability that a weight will be increased correlates with the strength of the statistical regularity of the input (i.e., the structural permanence of the input feature) causing that increase. These two facts conspire to make the expected time of recurrence of a pre-post correlation exponentially longer for spurious/noisy events than for meaningful (i.e., due to structural regularities of the environment) events.

If we run the model indefinitely, then eventually every synapse will experience two successive pre-post correlations occurring within any predefined window, T_σ. Thus, without some additional mechanism in place, eventually *all* afferent synapses into a mac will be permanently increased to $w_{\max} = 127$ at which point (total saturation of the afferent weight matrices) all information will be lost from the afferent matrices. Therefore, Sparsey implements a "critical period" concept, in which *all* weights leading to a mac are "frozen" (no further learning) once the fraction of weights that have been increased in any *one* of its afferent matrices crosses a threshold. This may seem a rather drastic solution to the classic trade-off that Grossberg termed the "stability-plasticity dilemma" (Grossberg, 1980). However, note that: (a) "critical periods" have been demonstrated in the real brain in vision and other modalities (Wiesel and Hubel, 1963; Barkat et al., 2011; Pandipati and Schoppa, 2012); (b) model parameter settings can readily be found such that in general, all synaptic matrices afferent to a mac approach their respective saturation thresholds roughly at the same time (so that the above rule for freezing a mac does not result in significantly underutilized synaptic matrices); and (c) in Sparsey, freezing of learning is applied on a mac-by-mac basis. We anticipate that in actual operation, the statistics of natural visual input domains (filtered as described earlier, i.e., to binary 1-pixel wide edges) in conjunction with model principles/parameters will result in the tendency for the lowest

level macs to freeze earliest, and progressively higher macs to freeze progressively later, i.e., a "progressive critical periods" concept. Though clearly, if the model as a whole is to be able to learn new inputs throughout its entire "life," parameters must be set so that some macs, logically those at the highest levels, never freeze. We are still in the earliest stages of exploring the vast space of model parameters that influence the pattern of freezing across levels.

The ultimate test of whether the use of critical periods as described above is too drastic or not is how well a model can continue to perform recognition/retrieval (or perform the specific recognition/retrieval-contingent tasks with which it is charged) over its operational lifetime (which will in general entail large numbers of novel inputs), in particular, after many of its lower levels have been frozen.

Learning arbitrarily complex nonlinear similarity metrics

The essential property needed to allow learning of arbitrarily complex *nonlinear* similarity metrics (i.e., category boundaries, or invariances) is the ability for an individual SDC in one mac to associate with multiple, perhaps arbitrarily different, SDCs in one or more other macs. This ability is present *a priori* in Sparsey in the form of the *progressive persistence* property wherein code duration, or *persistence* (δ), (measured in frames) increases with hierarchical level (in most experiments so far, δ doubles with level). For example, the V2 code $\phi_X^{2,j}$ in **Figure II-7A** becomes associated with the V1 code $\phi_Y^{1,i}$ at time t, and because it persists for two time steps, it also becomes associated with $\phi_Z^{1,i}$ at $t+1$. By construction of this example, $\phi_X^{2,j}$ represents (a particular instance of) the spatiotemporal concept, "rightward-moving vertical edge." However, if for the moment, we ignore the fact that these two associations occurred on successive time steps, then we can view $\phi_X^{2,j}$ as representing $XOR\left(\phi_Y^{1,i}, \phi_Z^{1,i}\right)$, i.e., just two different (in fact, pixel-wise disjoint) instances of a vertical edge falling within the U-RF of M_j^2. That is, the U-signals from either of these two input patterns alone (but not together[1]) can cause reinstatement of $\phi_X^{2,j}$. This provides an unsupervised means by which *arbitrarily different*, but temporally contiguous, input images, which may in principle portray any transformation that can be carried out over a two-time-step period and over the spatial extent of the RF in question, can be associated with the *same* object or class (the identity of which is carried by the persisting code, $\phi_X^{2,j}$).

Figure II-7B shows two more instances in which $\phi_X^{2,j}$ is active, denoted t and t' to suggest that they may occur at arbitrary times. If there is a supervisory signal by which $\phi_X^{2,j}$ can be activated whenever desired, then $\phi_X^{2,j}$ will associate with whatever codes are active in its RF (in this example, specifically, its U-RF) at such times. In this case, the two inputs associated with $\phi_X^{2,j}$ are just two different instances of a vertical edge falling within $\phi_X^{2,j}$'s RF. Furthermore, note that the number of active codes (features) in the RF can vary across association events. Thus, $\phi_X^{2,j}$ can serve as a code representing *any invariances* present in the set of codes with which it has been associated.

[1]Indeed, the two codes, $\phi_Y^{1,i}$ and $\phi_Z^{1,i}$, cannot occur together since they occur in the same L1 mac, M_i^1.

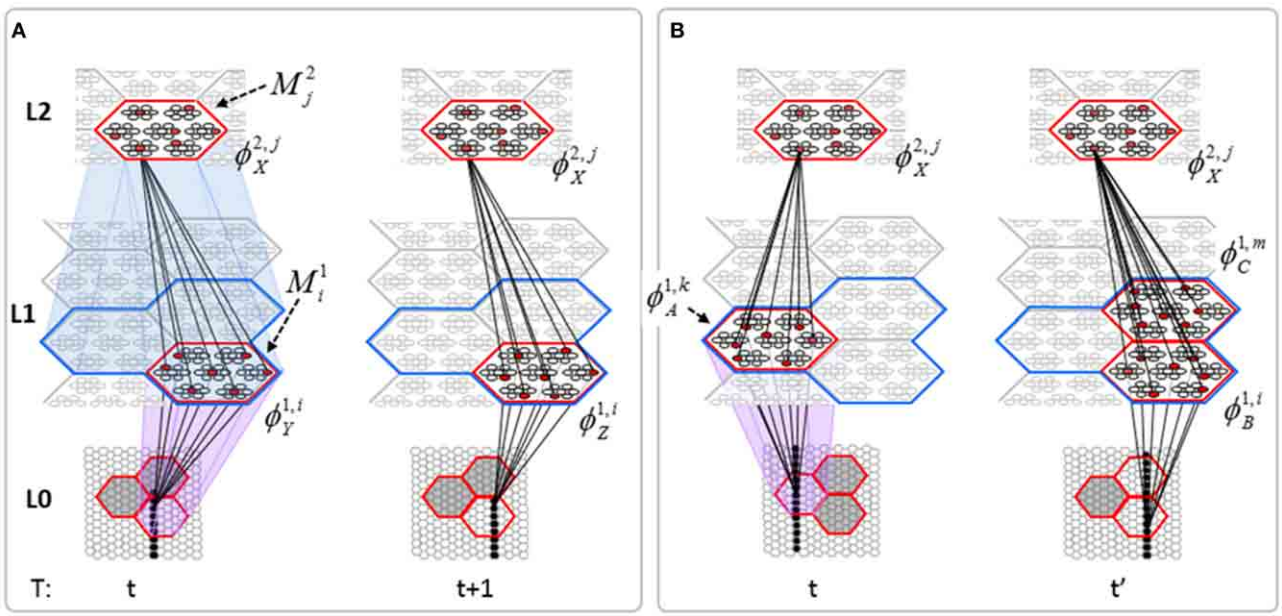


FIGURE II-7 | (A) The basic model property, progressive persistence, allows SDCs in higher level macs to associate with sequences of temporally contiguous SDCs active in macs in their U-RFs. **(B)** More generally, any mechanism which allows a particular code, e.g., $\phi_X^{2,j}$, to be activated under the control of a supervisory signal can cause $\phi_X^{2,j}$ to associate with two or more arbitrarily different codes presented at arbitrarily different times, thus allowing $\phi_X^{2,j}$ to represent arbitrary invariances (classes, similarity metrics).

This is in fact how supervised learning is implemented in Sparsey. That is, the supervised learning signal (label) is essentially just another input modality and supervised learning is therefore treated as a special case of cross-modal unsupervised learning. We have conducted preliminary supervised learning studies involving the MNIST digit recognition database (LeCun et al., 1998) using a model architecture like that in **Figure II-8**. However, to adequately describe the supervised learning architecture, protocol, and theory, would add too much length to this paper and so we save that work for a separate paper. Nevertheless, we are confident that the general framework described here will allow arbitrarily complex *nonlinear* similarity metrics, e.g., functions described as comprising the "AI Set," by Bengio et al. (2012), to be efficiently learned as *unions*, where each element of the union is a *hierarchical spatiotemporal composition* of the locally primitive (i.e., smoothness prior only) similarity metrics embedded in individual macs.

Neural implementation of CSA

Though we identify the broad correspondence of model structures and principles to biological counterparts throughout the paper, we have thus far been less concerned with determining precise neural realizations. Our goal has been to elucidate computationally efficient and biologically plausible mechanisms for generic functions, e.g., the ability to form large numbers of permanent memory traces of arbitrary spatiotemporal events on-the-fly and based on single trials, the ability to subsequently directly (i.e., without any serial search) retrieve the best-matching or most relevant memories, invariance to nonlinear time warping, coherent handling of simultaneous activation of multiple hypotheses, etc. We believe that Sparsey meets these criterion so far. For one thing, it does not require computing any gradients or sampling of distributions, as do the Deep Learning models (Hinton et al., 2006; Salakhutdinov and Hinton, 2012). Nevertheless, we do want to make a few points concerning Sparsey's relation to the brain.

First, we believe it is quite important, both for distinguishing Sparsey from other canonical cortical microcircuit models and for falsifiability, that the CSA really does entail two rounds, in quick succession, in which the mac's principal cells integrate their inputs, resulting in at least one of the cells in each CM reaching threshold and sending action potentials to the local inhibitory circuitry, which then fires, thus keeping all other cells in the CM from spiking (according to any number of detailed biophysical

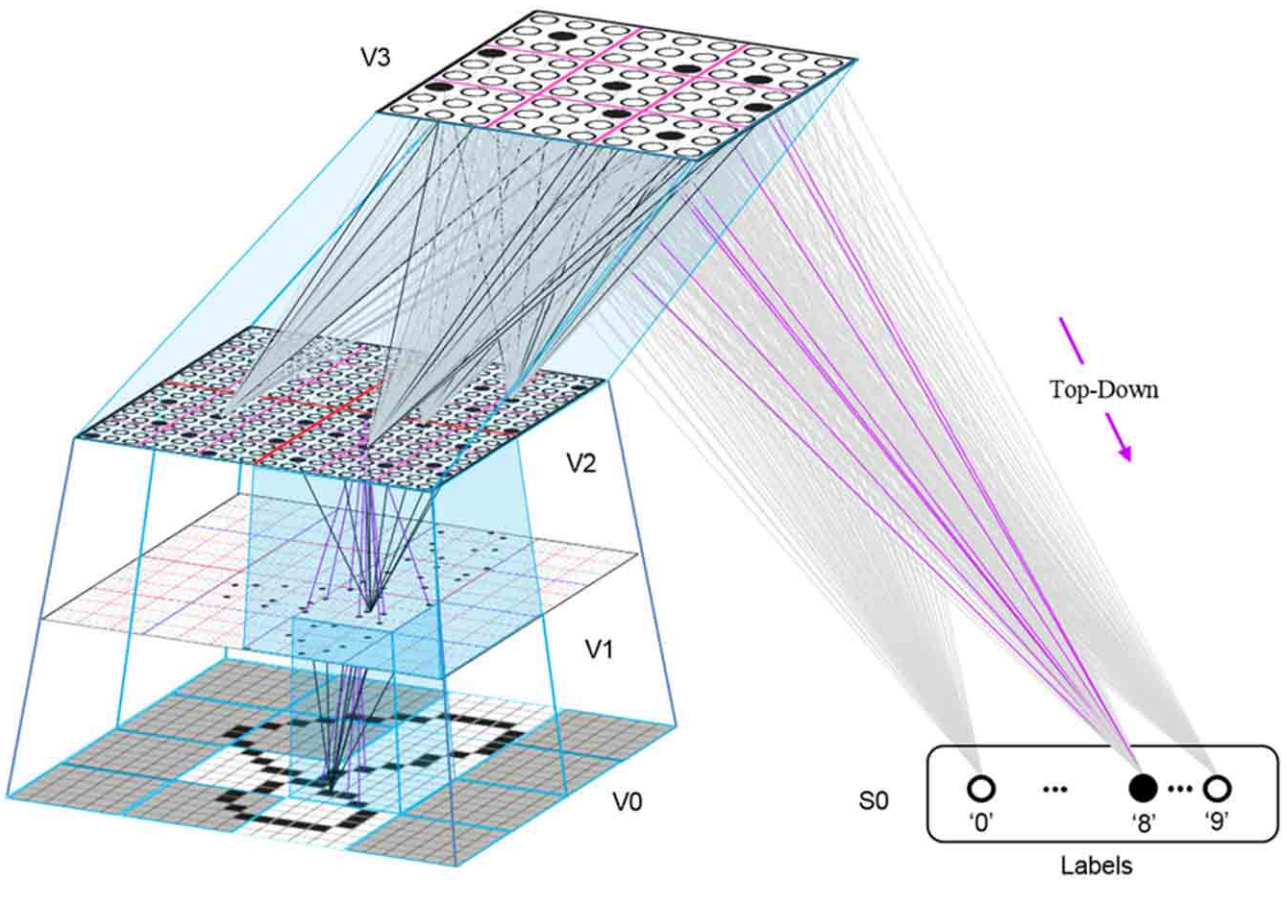


FIGURE II-8 | The 4-level model used in preliminary supervised learning studies involving the MNIST digit recognition task. This shows the recognition test trial in which the "8" was presented, giving rise to a flow of U-signals activating codes in macs throughout the hierarchy, and finally a top-down flow from the activated V3 code to the Label field, where the unit with maximal D-summation, the "8" unit, wins.

mechanisms, e.g., Jitsev, 2010). The first round winners' outputs (in addition to engaging the local inhibitory circuitry to suppress the other cells in their respective CMs) are averaged to yield *G*. And *G* then drives a modulation of the cell activation functions (as described in Sections Step 9: Determine the expansivity/compressivity of the I/O function to be used for the second and final round of competition within the mac's CMs and Step 10: Apply the modulated activation function to all the mac's cells, resulting in a relative probability distribution of winning over the cells of each CM) in preparation for the second round of competition. Due to the modulated activation functions, the second (and final) round winners will generally differ from the first round winners. Specifically, the intersection of the set of second round winners with the first round winners increases with *G*. In Rinkus (2010), we speculated that some combination of neuromodulators could underlie this behavior, but we have not yet refined that hypothesis.

Second, we note that Sparsey is a highly simplified/reduced model of cortical processing. It lacks analogs of layers 4, 5, or 6, and does not explicitly model inhibitory cells. In addition, it uses binary (nonspiking) neurons, effectively binary weights with variable permanence, and a simple Hebbian learning scheme with passive decay. The general consensus is that L4 is the main recipient of feedforward signals (from thalamus or from earlier cortical stages), whereas L2/3 receives horizontal (intrinsic) and top-down inputs. And, L5 and L6 project to earlier cortical stages and to subcortical structures and are involved in local feedback loops with L2/3. While numerous studies provide more detailed specifications fitting the above supra/infra-granular canonical circuit motif (Douglas and Martin, 2004), numerous details are yet to be understood and various new studies force significant modification/clarification of the canonical view, e.g., that L5/6 cells are also activated directly by U (specifically, thalamic) input Constantinople and Bruno (2013) and that thalamic input to L1 is much more substantial than previously thought (Rubio-Garrido et al., 2009).

In any case, while realizing the generic functionalities noted above has thus far required only a single population (layer) of principal cells, which best matches the L2/3 pyramidals, we anticipate incorporating modeling of other layers as needed. In particular, in its current "1-layer" form, Sparsey can be viewed as carrying out spatiotemporal processing underlying perception/recognition of spatiotemporal patterns and thinking, but without the accompanying motor output. Incorporating a "motor side" to the model will surely minimally force a move to a "2-layer" concept, i.e., supragranular (L2/3 and L4) and infragranular (L5 and L6).

CSA: RETRIEVAL MODE

In this section, we will first motivate the need for introducing some complexity to the computation of *G* when in retrieval mode and then describe the modification. We begin by thinking about how the model should respond to test trials involving previously learned sequences corrupted in particular ways. For example, if the model has learned the sequence S1 = [BOUNDARY] in the past and is now presented with S2 = [BOUNDRY], should it decide that S2 is functionally equivalent to S1? That is, should it respond equivalently to S2 and S1? More precisely, should its internal state at the end of processing S2 be the same as it was at the end of processing S1? The reader will probably agree that it should. We all encounter spelling errors like this all the time and read right through them. Similarly, if one encountered S3 = [BBOUNDARY], S4 = [BBOOUUNNDDAARRYY], S5 = [BOUNNNNNNDARY], or any of numerous other variations, he/she would likely decide it was an instance of S1. We could think of all these variations (corruptions) simply as omissions/repetitions. However, we prefer to think of this class of corruptions as instances of the class of *nonlinearly time-warped* instances of (discrete) sequences. Thus, S2 can be thought of as an instance of S1 that is presented at the same speed as during learning up until item "D" is reached, at which time the process presenting the items momentarily speeds up (e.g., doubles its speed) so that "A" is presented but then replaced by "R" before the model's next sampling period. Then the process slows back down to its original speed and item "Y" is sampled. Thus, S2 is a nonlinearly time-warped instance of S1. We can construct similar explanations, involving the underlying process producing the sequences undergoing a schedule of speedups and slowdowns relative to the original learning speed, for S3, S4, etc. In fact, S4 is even simpler; it's just a uniform slowing down, to half speed, of the whole process.

Of course, there are limits to how much we want a system to generalize regarding these warpings. And the final equivalence classes, in particular for processing language, must be experience-dependent and idiosyncratic. For example, should a model think that S6 = [COD] is just an instance of S7 = [CLOUDS], produced twice as fast as during the learning instance? In general, probably not. Furthermore, we have not even considered in these examples the fact that the individual sequence items are actually pixel patterns which can themselves be noisy, partially occluded, etc., which would of course influence the normative category decisions. Nevertheless, the ubiquity of instances such as described above, not just in the realm of language, but in lower-level raw sensory inputs, suggests that a model have some mechanism for dealing with them, i.e., some mechanism for treating moments produced by nonlinearly time-warping as equivalent.

Our explanation of the modified *G* computation in retrieval mode uses an example involving a 3-level model that has only one mac at each level. **Figure II-9** shows representative samples of the U, H, and D learning that occurs as the model is presented with the sequence, [BOTH]. Note that the model is unrolled in time here, i.e., the model is pictured at four successive time steps and in particular, the origin and destination cell populations of the increased H synapses (green) are the same. This figure illustrates several key concepts. First, learning a sequence involves increasing the H-wts from the previously active code to the currently active code. The D-wts (magenta) are also increased from the previously active code (in this case, in the L2 mac) to the currently active destination code in the L1 mac. Note however that the U-wts (blue) are increased from the currently active input (L0 code) to the currently active L1 code. We show the full set of afferent U, H, and D wts that are increased for one cell—the winner in the upper left CM of the L1 mac—at each time step. Thus, this figure emphasizes that, on each moment, individual cells become

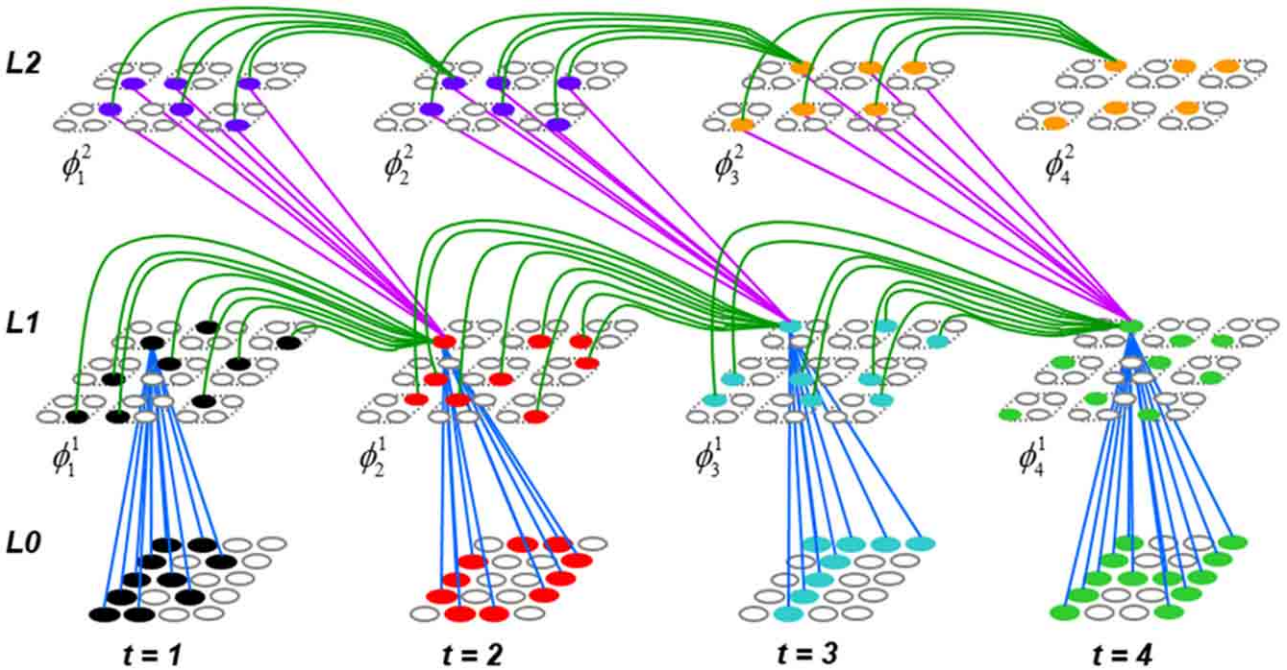


FIGURE II-9 | The formation of a hierarchical spatiotemporal memory trace, unrolled in time, of the input sequence, [BOTH]. We only show representative samples of the increased weights on each frame. The model has one L1 mac with $Q_1 = 9$ CMs, each with $K = 4$ cells and one L2 mac with $Q_2 = 6$ CMs, each with $K = 4$ cells. The resulting trace can be said to have been produced using both *chaining* (increasing H-wts between successively active codes at the same level) and *chunking* increasing U and D wts between single higher-level (L2) codes and multiple lower-level (L1) codes. See text for detailed explanation.

associated with their entire afferent input (spatiotemporal context) in one fell swoop. Though we only show this occurring for one cell on each frame, all winners in a mac code will receive the same weight increases simultaneously. Thus, we can say not only that individual cells become associated with the mac's entire spatiotemporal contexts but that whole mac codes become associated with the mac's entire spatiotemporal contexts.

The second key concept illustrated is progressive persistence, in this case, that L2 codes persist for twice as long as L1 codes. Cell color in this figure is used to make persistence clear. Thus, the first L2 code that becomes active D-associates with two L1 codes. And, because of the modeling decision that D-wts are increased from previously active to currently active codes, the two L1 codes are those at $t = 2$ and $t = 3$. The second L2 code to become active (orange) D-associates with the L2 code at $t = 3$ and would associate with a $t = 4$ L1 code if one occurred.

Having illustrated (in **Figure II-9**) the nature of the hierarchical spatiotemporal memory trace that the model forms for [BOTH], **Figure II-10** compares model conditions when processing one particular moment—the second moment—of a test trial that is identical to the learning trial (**Figure II-10A**) to conditions when processing the second moment of a time-warped instance of the learning trial—specifically, a moment at which the item that originally appeared as the third item of the learning trial, "T," now appears as the second item immediately after "B," i.e., "O" has been omitted (**Figure II-10B**). We can represent the two test trial moments as [B**O**] and [B**T**], respectively, where bolding indicates the frame currently being processed and the nonbolded letters indicate the context leading up to the current moment. The first thing to say is that the second moment of the time-warped instance is simply a novel moment. Thus, the caveat we mentioned above applies. That is, deciding whether a particular novel input moment should be considered a time-warped instance of a known moment or as a new moment altogether cannot be done absolutely.

Figure II-10A shows the case where the test trial moment [B**O**] is identical to the learning trial moment [B**O**]. The main point to see here is that, given the weight increases that will have occurred on the learning trial, all three input vectors, U, H, and D, will be maximal (equal to 1) for the red cell (which is in ϕ_2^1) in each L1 CM. At right (yellow), we zoom in on the conditions only for the upper left L1 CM, but the conditions are statistically similar for all L1 CMs. We show that for the red cell, $U = 1$, $H = 1$, and $D = 1$. The blue cell (which is in ϕ_3^1) also has maximal D-support and the blue, green, and black cells have nonzero U inputs (their U-inputs are not shown in the main figure to minimize clutter), due to the pixel overlap amongst the four input patterns, but they all have $H = 0$. Thus, according to Equation (4) of the CSA (**Table II-1**), the red cell has $V = U \times H \times D = 1$, whereas the others have $V = 0$. We refer to red cell as having a "3-way match" in that all three evidence vectors are maximal and agree. Also, we refer to the G version computed using all three input vectors as G_{HUD}. Thus, in this case, where the test moment is identical to a learned moment, CSA Equation (4) is sufficient as is.

However, as shown in **Figure II-10B**, when an item ("O") has been omitted with respect to the learning trial, the H and D vectors to the red cell will no longer agree with its U vector. Various policies could be imagined for handling this situation.

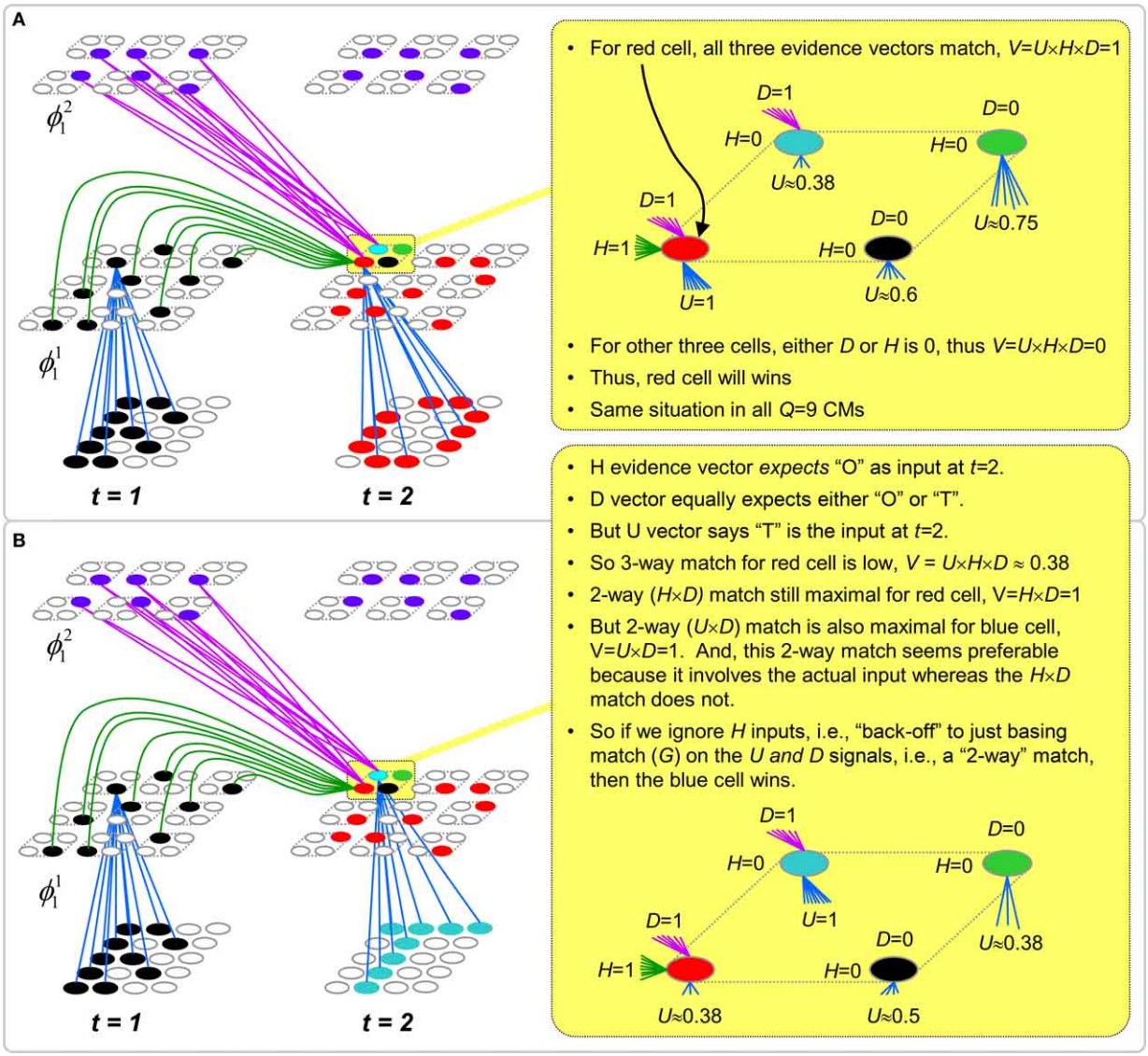


FIGURE II-10 | Motivation for the Back-off Strategy for computing *G* in retrieval mode. (A) Detail of conditions that exist at L1 when processing the second moment [B**O**] of a test trial that is identical to the learning trial (in **Figure II-8**). **(B)** Detail of conditions that exist at L1 when processing the second moment [BT] of a test trial that is a time-warped version of the learning trial, specifically, a sequence that is sped up by 2x at $t = 1$, causing the "O" to be missed and the "T" to occur immediately after the "B." See text for detailed discussion.

The model could simply consider such a case as being a novel moment, [B**T**]. This would require no modification to the CSA. Or, as discussed earlier, the model could *check* to see whether the current moment could have resulted from a nonlinear time-warping process, and should therefore be judged identical to some previously learned moment. In this case, the current moment [B**T**] is identical to the learning trial moment [BO**T**] if we assume that the process presenting the sequence to the model sped up by 2x at $t = 1$, causing the "O" to be missed.

So, how does the model *check* this possibility? It is quite simple. All it needs to do is disregard the H signals when computing the V's (CSA Step 4). In other words, it "backs off" from the more stringent 3-way G_{HUD} match criterion to the more permissive 2-way G_{UD} criterion. Note that the model begins by computing the highest-order G available at the current moment, in this case, using all three input vectors. Only if that highest-order G falls below a threshold, which we typically set rather high, e.g., $G_{HUD} = 0.9$, does it bother to compute the next lower order version(s) of G, i.e., G_{UD}, G_{HU}, and G_{HD}. Similarly, only if whichever 2-way version has been considered falls below another threshold, which is typically set even higher than the first, e.g., $G^{+}_{UD} = 0.95$, does the model *back-off* to the next lower order match criterion.

In this example, $G_{UD} = 1$, meaning that there is a code stored in the L1 mac—specifically, the set of blue cells assigned as the L1 code at $t = 3$ of the learning trial (**Figure II-9**)—which yields a perfect 2-way match. Thus, there is no need to back-off to the "1-way" match criterion, G_U. However, there are many naturally

occurring instances in which backing all the way off to the lowest-order criterion (i.e., basing the V-values and thus, the G, on only the U signals, ignoring the H and D signals) is appropriate. There are myriad policy considerations regarding possible precedence orders of the different G versions and whether or not and under what conditions the various versions should be considered. We are actively exploring these issues, but cannot delve into this topic in this paper.

Figure II-11 completes this example by showing that the back-off policy allows the model to keep pace with nonlinearly time-warped instances of previously learned sequences. That is, the model's internal state (i.e., the codes active in the macs) can either advance more quickly (as in this example) or slow down (not demonstrated herein) to stay in sync with the sequence being presented. **Figure II-11A** is given for comparison, showing the full memory trace that becomes active during a retrieval trial for an exact duplicate of the training trial, [BOTH]. In this case, no back-off would be required because all signals at all times would be the same during retrieval as they were during learning. **Figure II-11B** shows the trace that obtains, using the back-off protocol, throughout presentation of the nonlinearly time-warped instance of the training trial, [BTH].

The back-off from G_{HUD} to G_{UD} occurs in the L1 mac at $t = 2$ (as was described in **Figure II-10B**). Since $G_{UD} = 1$, the V-to-ψ map is made very expansive, resulting in activation, at $t = 2$ of the test trial, of the code, ϕ_3^1 (blue cells), which was

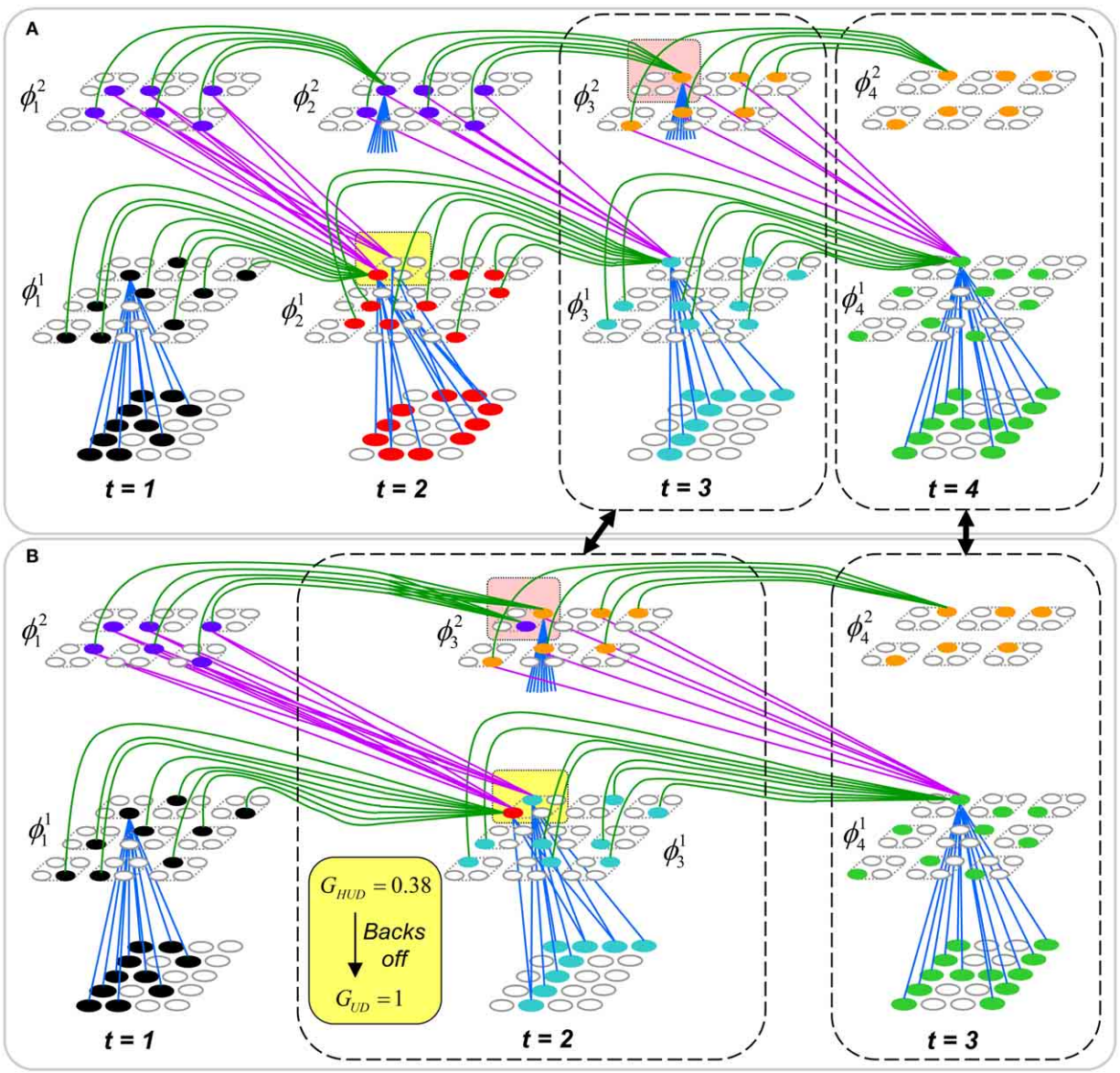


FIGURE II-11 | This figure shows the complete test trial traces for: (A) an exact duplicate of the training trial, [BOTH]; and (B) the nonlinearly time-warped instance of the training trial, [BTH]. In **(B)**, back-off from G_{HUD} to G_{UD} occurs in the L1 mac at $t = 2$ (as was described in **Figure II-10B**), which allows entire internal state of the model (i.e., at L1 and L2) to "catch up" with the momentary speed up of the sequence. The remainder of the sequence and the associated internal trace then obtains the same as during learning. See text for detailed description.

originally activated at $t = 3$ in the learning trial. Thus, the back-off has allowed the model's internal state (in L1) to "catch up" to the momentarily sped up process that is producing the input sequence. Once ϕ_3^1 is activated, it sends U-signals to L2 (blue signals converging on orange cell in rose highlight box). This results in the L2 code, ϕ_3^2 (orange cells), being activated without requiring any back-off. That's because the L2 code from which H signals arrive at $t = 2$, ϕ_1^2 (purple cells) increased its weights not only onto itself (at $t = 2$ of the learning trial) but also onto ϕ_3^2 at $t = 3$ of the learning trial. Thus, the six cells comprising ϕ_3^2 (orange) yield $G_{HU} = 1$ (note that G_{HU} is the highest order G version available at L2 since there is no higher level). Consequently, a maximally expansive V-to-ψ map is used in the L2 mac, resulting in reinstatement of ϕ_3^2. At this point—$t = 2$ of the test trial—the entire internal state of the model (i.e., at L1 and L2) is identical to its state at $t = 3$ of the learning trial (two central dashed boxes connected by double-headed black arrow): the model, as a whole, has "caught up" with the momentary speed up of the sequence. The remainder of the sequence proceeds the same as it did during learning, i.e., state at $t = 3$ of retrieval trial equals state at $t = 4$ of learning trial.

The final, and really the most important, point of this section is that Sparsey's back-off policy *does not change* the time complexity of the CSA: it still runs with *fixed time complexity*, which is essential in terms of scalability to real-world problems. True, expanding the logic to compute multiple versions of G increases the absolute number of computer operations required by a single execution of the CSA. However, the number of possible G versions is small and more to the point, fixed. Thus, adding the back-off logic adds only a fixed number of operations to the CSA and so does not change the CSA's time complexity.

During each execution of the CSA, *all stored codes* compete with each other. In general, the set of stored codes will correspond to moments spanning a large range of *Markov orders*. For example, in **Figure II-9**, the four moments, [**B**], [B**O**], [BO**T**], and [BOT**H**], are stored, which are of progressively greater Markov order. *During each moment of retrieval, they all compete*. More specifically, they all compete first using the highest-order G, and then if necessary, using progressively lower-order G's. However, it is crucial to see that with back-off, not only are the explicitly stored (i.e., actually experienced) moments compared, but so are a far larger number of time-warped versions of the actually-experienced moments. For example in **Figures II-10B**, **II-11B**, the moment [B**T**], which never actually occurred competes and wins (by virtue of back-off) over the moment [B**O**], which did occur. And crucially, as noted above, all these comparisons take place with fixed time complexity! Space does not permit here, but the above mechanism and reasoning generalizes to arbitrarily deep hierarchies. As the number of levels increases, with persistence doubling at each level, the space of hypothetical nonlinearly time-warped versions of actually experienced moments, which will materially compete with the actual moments (on every frame and in every mac) grows exponentially. And, we emphasize that these exponentially increasing spaces of never-actually-experienced hypotheses are *envelopes* around the actually-experienced moments: thus, the invariances implicitly represented by these envelopes are (a) learned and (b) idiosyncratic to the specific experience of the model.

CSA: SIMPLE RETRIEVAL MODE

Both the learning mode CSA and the retrieval mode CSA described above, which is just the learning mode CSA augmented by the back-off protocol, involve the G-based modification of the cell activation functions and the second, probabilistic round of competition for choosing the final code (CSA Steps 8–12, **Table I-1**). If the model is operating as a truly autonomous agent, then it, or rather any of its constituent macs, may be presented with a truly novel input pattern at every moment experienced. Thus, a mac must be prepared to learn, i.e., assign a new SDC, at every moment[2]. As described in earlier sections, the CSA's two competitive stages, with the second, probabilistic stage using the G-modulated cell activation functions, satisfies the requirements for autonomous operation. That is, as G decreases, the expected intersection of the final code (for the current frame) chosen with the closest matching stored code decreases to chance, which results in the occurrence of novel pre-post correlations, and thus new learning. On the other hand, as G increases toward 1, the expected intersection of the finally chosen code with the closest matching stored code increases to complete, which results in no (or at least, statistically, very few) novel pre-post correlations and thus no new learning.

However, if the model "knows" that is operating in pure retrieval mode, i.e., that at each moment each mac should simply activate the code of the learned moment that most closely matches its current input moment, then there is no advantage to having the second G-dependent probabilistic stage of competition. In fact, the optimal strategy in this case is simply to choose the cell with the highest V-value in each CM. The transfer of global information (G) back into the local (within each CM) winner selection processes, which occurs in steps 8–12, does not help and in fact, can only hurt (i.e., it can only reduce the probability of the maximally likely cell in a given CM winning). Thus, in this "simple retrieval mode," in which the model knows that it will not be asked to learn anything new, the optimal algorithm is just the first seven steps of the CSA given in **Table I-1**, but augmented with the back-off protocol described in the previous section. Thus, we do not state the simple retrieval mode of the CSA separately. We will clearly indicate which of the two retrieval modes is used in the studies reported in the next section.

We emphasize that the *deterministic* "simple retrieval mode" algorithm cannot be used during learning because it would result in essentially mapping all of the mac's input patterns to one or a very small number of codes, vastly over-utilizing only a tiny fraction of the mac's cells and vastly decreasing the number of codes (amount of information) that can be stored in the mac.

However, based on first principles, it seems plausible that for the vast majority of Sparsey's envisioned operational regime, i.e., the regime in which the number of codes stored in the macs (or more specifically, the fraction of synapses that have

[2]Actually, in a hierarchical model faced with the prospect of possibly having to learn something new on every moment of its operational lifetime, it's sufficient only that at least one mac (which would typically be at the highest level) be prepared to learn at every moment (cf. earlier discussion of critical periods).

been increased) remains below a threshold, the simple retrieval mode should always do better (on average) than the probabilistic retrieval mode Specifically, recall that in probabilistic retrieval mode, the winner in a CM is chosen *as a draw* from the V distribution. Depending on the particular shape/statistics of the V distribution, the cell with the maximum V might therefore be chosen winner only a small fraction of the time. Yet, that max-V cell is the most likely cell given the total evidence (from the U, H, and D signals) arriving at the mac. In simple retrieval mode, the max-V cell always wins. Again, provided that the fraction of the mac's afferent synapses that have been increased remains low enough, simply choosing the max-V cell as winner yields higher expected accuracy.

DEFINITIONS OF SYMBOLS USED HEREIN

Table I-2 | Major symbols in CSA equations.

Symbol	Definition	Symbol	Definition
Active(m)	Whether mac m is active or not	$\lambda_{U(t)}$	Power to which U is raised prior to being multiplied with H and D signals. It can vary as a function of time from beginning of the sequence (snippet) being processed
$\Upsilon(m)$	Age, in number of time steps (frames), of the currently active code in mac m	λ_H, λ_D	Analogous to $\lambda_{U(t)}$ except that for now they are not functions of time
Q,Q_i	Number of CMs per mac; same but for a specific level, i	$\delta(m)$	Persistence, in number of time steps, of mac m. Currently, all macs of a given level have the same persistence
K,K_i	Number of cells per CM;; same but for a specific level, i	$M_{2,3}$ M_4^3	The mac at coordinates (2,3) (when the level is unambiguous). Alternate notation: Mac with index "4" at level "3."
$u(i)$, $h(i)$,$d(i)$	Raw sum of weighted signals from cells comprising cell i's U-RF. $h(i)$,$d(i)$ are analogous	$U(i)$, $H(i)$, $D(i)$	$U(i)$ is the normalized $u(i)$, to [0,1] range. $H(i)$,$D(i)$ are analogous
U^-	Lower threshold below which a cell's U-value is considered 0.H^-,D^-analogous	U^+	Upper threshold above which a cell's U-value is considered 1. H^+,D^+ analogous
π_U	The # of active features in a mac's U-RF	π_U^*	Number of active features in a mac's U-RF, which are active in macs with $\zeta \leq B$
π_U^-, π_U^+	Lower and upper bounds on the number of active features that must be present in a mac's U-RF for that mac to activate	$V(i)$	Overall local evidence that cell i should become active. Product of functions of $U(i)$, $H(i)$, and $D(i)$
$\hat{V}_j$	Maximum $V(i)$ in CM, C_j.	V_ζ	Threshold for a cell to be considered as part of an active hypothesis
G $G(t)$	Average $\hat{V}$-value over a mac's Q CMs. It is a measure of the familiarity of a mac's total input, normalized to [0,1]	G^-	Threshold below which the mac's G-value is considered effectively zero
χ	The sigmoid expansion factor	γ	The sigmoid expansion exponent
η	Range of the V-to-ψ map, which transforms a cell's V-value into its relative (within its own CM) probability of winning, ψ	σ_1, σ_2 σ_3, σ_4	Parameters that interact to control overall sigmoid expansivity and shape, e.g., horizontal position of inflection pt., etc
$a(j, t)$	Activation (0,1) of cell j at time t	M_j	Number of macs in Level j
ζ_q	# of cells in CM q with $V(i) > V_\zeta$. Typically, V_ζ is set close to 1, e.g., 0.95	ζ	The number of maximally active hypotheses, ζ, in a mac
$F(\zeta)$	The correction factor for increasing the weights of outgoing signals from cells in macs that have *multiple competing hypotheses* (MCHs), i.e., $\zeta > 1$	$F(\zeta(j, t))$	The MCH correction factor $F(\zeta)$ at time t for mac that contains cell j
A	Exponent (<1.0) for discounting MCH correction factor when $\zeta > 1$	B	Threshold on ζ above which we ignore completely signals from the source mac
$\psi(i)$	The relative probability of activating cell i in a mac	$\rho(i)$	The absolute probability of activating cell i in a mac
G_U	G computed based only on the U signals to a mac. Similarly, G_{HUD} is G computed based on all three input vectors, U, H, and D. Similarly, for G_{HU}, G_{UD}, G_{HD}, G_H		
$G_{HU}^+(t)$	Threshold below which we back off to the next lower-order (or more generally, the next-considered) version of G. Here, we suggest that this threshold can be a function of time (frame)		
U-RF, H-RF, D-RF	U-RF is a bottom-up receptive field. Can be applied to single cells or to whole macs. For cells/macs at L1 the U-RF is a set (or array) of individual binary cells (e.g., pixels). For cells/macs at higher levels, the U-RF is a set (array) of macs. H-RF and D-RF are analogous, but they always consist of a set (array) of macs		

RESULTS

STUDY 1: SPATIOTEMPORAL SISC PROPERTY

Study 1 is an *unsupervised learning* study that demonstrates that Sparsey maps spatiotemporally more similar inputs to more highly intersecting SDCs, i.e., the *similar-inputs-to-similar-codes* (SISC) property. This is an instance of what others have referred to as the "smoothness prior" (Bengio et al., 2012). The model instance used here has a 12 × 12-pixel input level (L0) and one internal level (L1) consisting of one mac with $Q = 25$ CMs, each with $K = 9$ cells, as in **Figure III-1B**. The set of six 2-frame sequences (S0–S5) used in this study are shown in **Figure III-1A**. All sequences have the same second item, X, while the pixel-wise overlap of the sequence-initial item with S0's first item, A, decreases across sequences, S1 = [BX], S2 = [CX], etc. Thus, the spatiotemporal similarity of the second frame of each sequence with the second frame of S0 drops across sequences (even though the *purely spatial* similarity of the second frame remains the same at 100%). We will show that the codes assigned to the second frame of the progressively spatiotemporally less similar sequences have progressively smaller intersection with the code assigned to the second frame of S0.

During learning, on each frame of an input sequence, an L1 code is chosen using the learning mode CSA (**Table I-1**). Then, associative learning occurs from active L0 units (active pixels) to active L1 units: these U-wts are set high, i.e., they are effectively binary. Also, on the second frame ($T = 1$), H wts from L1 units active at $T = 0$ to currently active L1 units are set high. **Figure III-1C** shows the memory trace assigned to S0. The trace

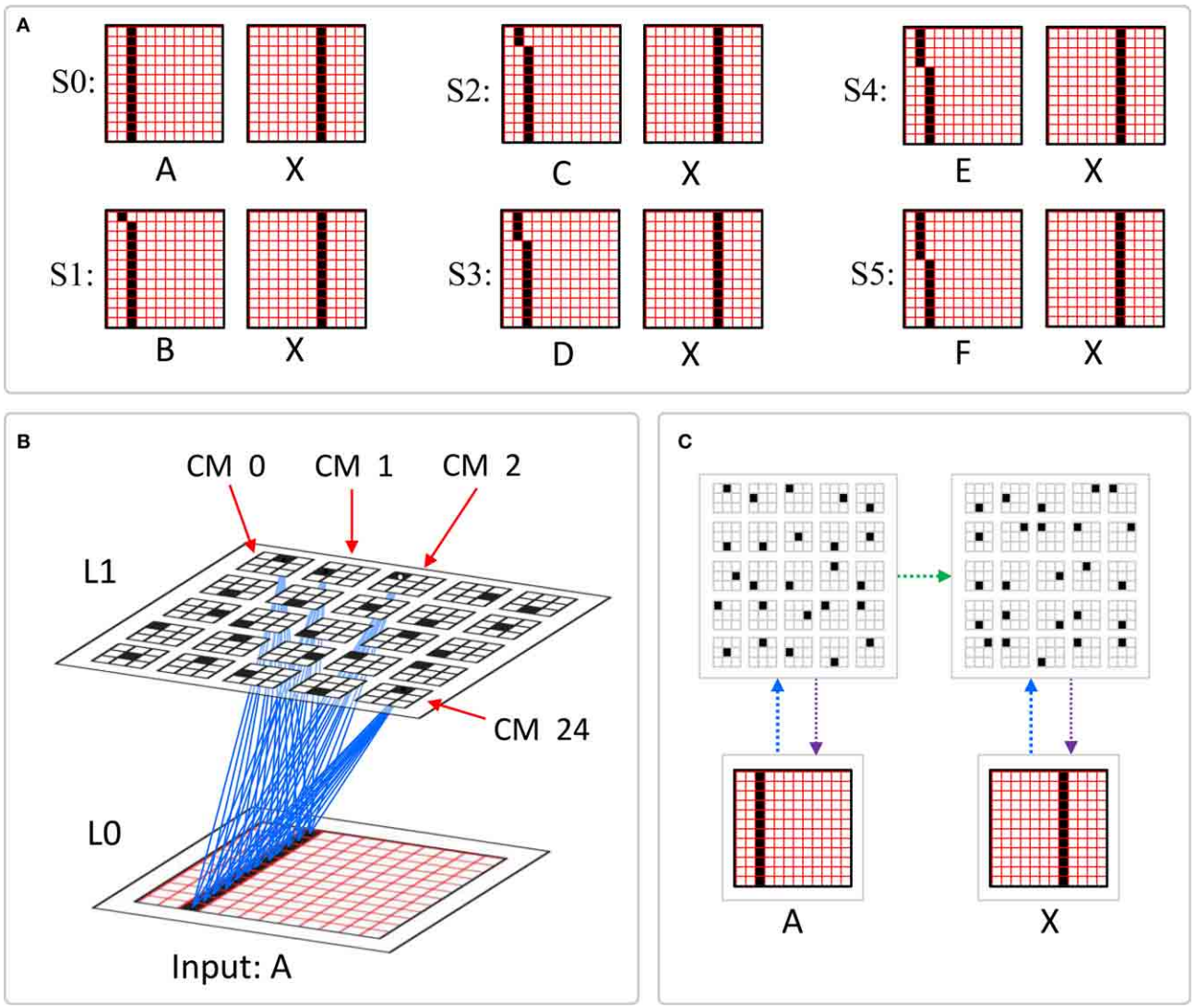


FIGURE III-1 | (A) The six 2-frame sequences used in Study 1. **(B)** The model whose internal level consists of one mac comprised of $Q = 25$ CMs, each with $K = 9$ cells. A subset of the U-wts (blue) increased at $T = 0$ of sequence S0 = [AX] from the active pixels to the cells comprising the winning SDC. **(C)** The memory trace assigned to S0 to which we will compare (in **Figure III-2**) the memory traces assigned to the other five sequences in this figure. The green arrow represents the learning that occurs in the recurrent H-matrix from the 25 winners at $T = 0$, when A is presented, to the 25 winners at $T = 1$, when X is presented. The blue (magenta) arrows represent the learning in the U (D) matrix on each of the two time steps.

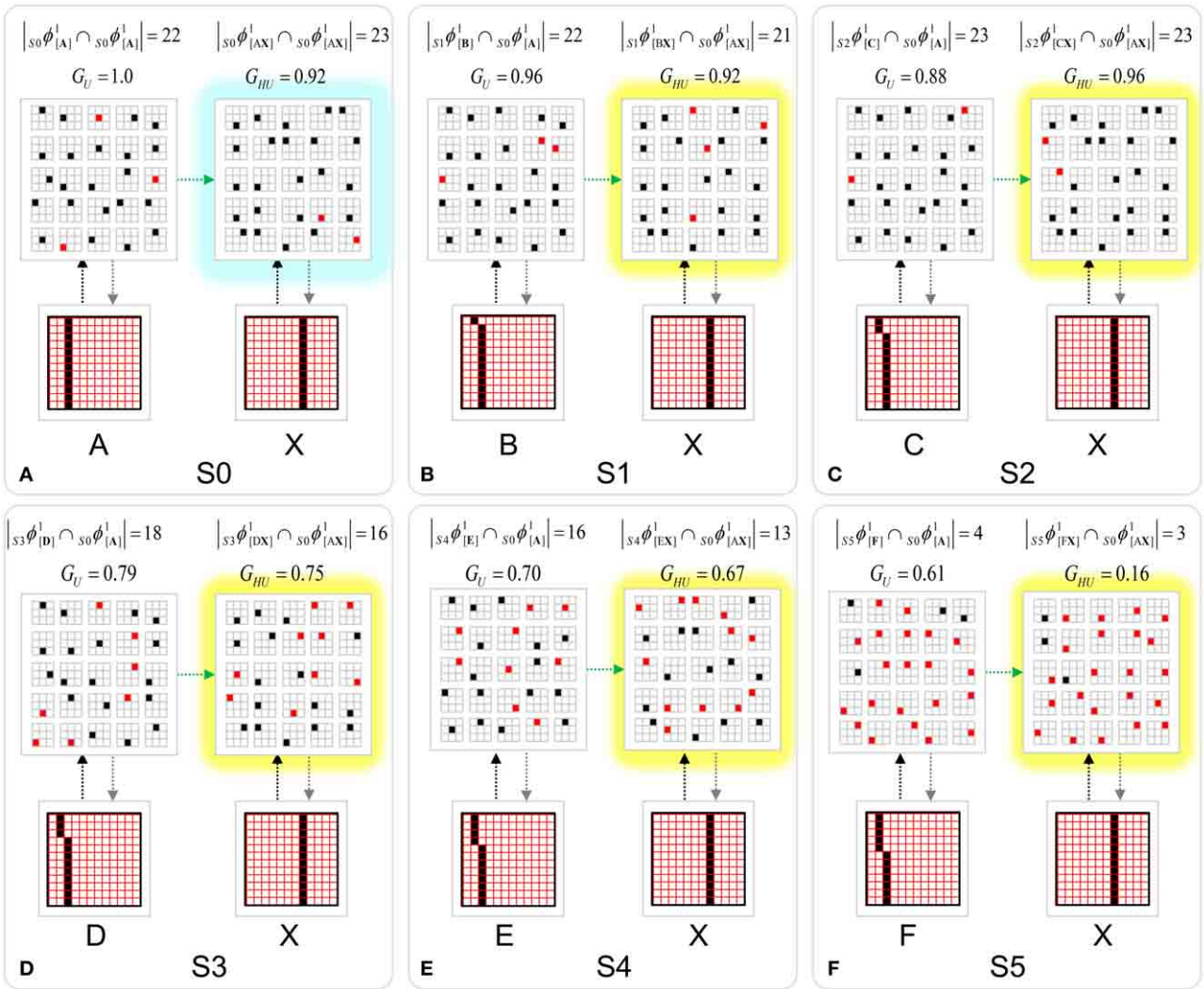


FIGURE III-2 | Memory traces assigned to specific instances of the six sequences of Study 1. The basic SISC property can be seen (across panels **A–F**) in the decreasing intersection size of the L1 codes assigned to the second moment of each sequence (highlighted in yellow) to the L1 code assigned to the second moment of **(S0)** (in **Figure III-1C**), [A**X**] (black units are those that do intersect, red are those that do not). The *G*-values are the model's estimates of spatiotemporal similarity of the current moment. Note that the same trend of intersection size decreasing with similarity can be seen in comparing the first moments of each sequence, **(S1–S5)**, with the first moment of **(S0)**. However, strictly that is a purely spatial similarity measure since no temporal context signals present on the first moment of a sequence.

consists of two SDCs. One might also refer to the set of weight increases made during presentation of S0 as the "memory trace," however, it is the sequence of SDCs across time steps which, unless otherwise stated, we refer to as the memory trace of a sequence. Note that because [AX] is the first sequence presented to the model, the particular units chosen on both frames of S0 are chosen at random.

Figure III-2 shows, in panels B–F, the memory traces assigned to five sequences, [BX], [CX], [DX], [EX], and [FX], which are progressively less spatiotemporally similar to [AX]. In addition, **Figure III-2A** shows the memory trace reactivated in response to a second presentation of [AX]. For each of the experiments represented by the six panels of **Figure III-2**, the sequence shown is presented as the second sequence experienced by the model. For example, when S4 = [EX] is presented, it is presented to the model after the model has only learned [AX], *not* the rest of the intervening sequences, S1–S3.

The main result visible in **Figure III-2** is that in comparing the L1 codes assigned to frame 2 of each sequence, S1–S5, to the L1 code assigned to frame 2 of S0 (in **Figure III-1C**), we see progressively smaller intersection. These five L1 codes are highlighted in yellow. Black units are units which are the same as for frame 2 of sequence [AX] (**Figure III-1C**); red units are different[3]. Thus, on the second moment, [B**X**], of sequence S1,

[3]If we viewed the presentations of S1–S5 as recognition trials in which we were presenting progressively more perturbed variants of [AX], then these red units would be considered errors. However, in this case, we are viewing these

Table II-1 | Code similarity decreases with spatiotemporal similarity of moments.

Decreasing similarity of 1st moment	Decreasing similarity of 2nd moment
$\left\lvert {}_{S0}\phi^1_{[A]} \cap_{S0} \phi^1_{[A]} \right\rvert = 22$ (88%)	$\left\lvert {}_{S0}\phi^1_{[AX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 23$ (92%)
$\left\lvert {}_{S1}\phi^1_{[B]} \cap_{S0} \phi^1_{[A]} \right\rvert = 22$ (88%)	$\left\lvert {}_{S1}\phi^1_{[BX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 21$ (84%)
$\left\lvert {}_{S2}\phi^1_{[C]} \cap_{S0} \phi^1_{[A]} \right\rvert = 23$ (92%)	$\left\lvert {}_{S2}\phi^1_{[CX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 23$ (92%)
$\left\lvert {}_{S3}\phi^1_{[D]} \cap_{S0} \phi^1_{[A]} \right\rvert = 18$ (72%)	$\left\lvert {}_{S3}\phi^1_{[DX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 16$ (64%)
$\left\lvert {}_{S4}\phi^1_{[E]} \cap_{S0} \phi^1_{[A]} \right\rvert = 16$ (64%)	$\left\lvert {}_{S4}\phi^1_{[EX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 13$ (52%)
$\left\lvert {}_{S5}\phi^1_{[F]} \cap_{S0} \phi^1_{[A]} \right\rvert = 4$ (16%)	$\left\lvert {}_{S5}\phi^1_{[FX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 3$ (12%) (~chance)

the code assigned, ${}_{S1}\phi^1_{[\mathbf{BX}]}$, has 21 out of the maximum possible 25 units in common with the code, ${}_{S0}\phi^1_{[\mathbf{AX}]}$, assigned to the second moment, [A**X**], of S0, i.e., $\left\lvert {}_{S1}\phi^1_{[\mathbf{BX}]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 21$. Note that we have slightly generalized the code name convention: the lead subscript indicates the sequence in which the code occurs. As the spatiotemporal similarity of the second sequence moment with [A**X**] decreases further across panels c-f, the intersection of the assigned code with ${}_{S0}\phi^1_{[\mathbf{AX}]}$ trends downward, despite the fact that in this particular instance, $\left\lvert {}_{S2}\phi^1_{[CX]} \cap_{S0} \phi^1_{[AX]} \right\rvert = 23$ even though [C**X**] must clearly be considered less similar to [A**X**] than [B**X**] is to [A***X***]. Despite this statistical blip, the codes assigned for the remaining progressively less spatiotemporally similar moments, [D**X**], [E**X**], and [F**X**], have monotonically decreasing intersection with ${}_{S0}\phi^1_{[\mathbf{AX}]}$ as summarized in the right column of **Table II-1**. In fact, the same trend obtains with respect to the first sequence moment as well (left column). However, note that in the latter case, it is purely spatial similarity in the input space that is relevant (since no temporal context information is present on the first moment of a sequence).

We emphasize that each of the memory traces shown in **Figure III-2** is a particular instance. The winner in a CM is chosen as a draw from a likelihood distribution over the CM's units, i.e., "softmax" (CSA Step 12), *not* by simply choosing the max likelihood unit, i.e., plain ("hard") max. Thus, we will generally see some variation in the chosen codes across instances of the same experiment and the amount of variation will increase as the similarity of the test sequence to the learned sequence, [AX], decreases. This statistical variation, for example, is why the memory trace in **Figure III-2A** is not perfect. Due to the statistical nature of Sparsey's CSA, demonstration of the SISC property requires running many instances of each of the experiments shown in **Figure III-2** and reporting average results. Such a protocol was followed in Study 2.

as presentations of similar but not identical sequences to S0, in which case it is appropriate for the model to assign unique codes. In this case, the red units are not errors, but simply just different from the unit chosen in the corresponding CM in frame 2 of S0.

STUDY 2: SINGLE-TRIAL LEARNING OF SETS OF LONGER SEQUENCES

Study 2 demonstrates single-trial learning of longer and more complex sequences, derived from natural video, by a model with multiple internal levels. We presented eight 20-frame 24 × 24-pixel, natural-derived, snippets (movies), produced from the KTH Video data set (Schuldt et al., 2004). All 160 frames of the eight snippets are shown in **Figure III-3**. These are taken from instances of people waving their arms. See example video. The snippets were presented once each.

The model in Study 2 had 4 levels, a total of 21 macs, 3285 cells ("neurons"), and 1,880,568 synapses[4]. As shown in **Figure III-4**, the first internal level (L1) had 16 macs, each consisting of $Q_1 = 9$ CMs, each having $K_1 = 16$ cells. L2 had 4 macs, each having of $Q_2 = 9$ CMs, each having of $K_2 = 9$ cells. The top level (L3) consisted of one mac, consisting of $Q_3 = 9$ CMs, each with $K_3 = 9$ cells. The semi-transparent blue prisms indicate the bottom-up (U) wiring scheme. Each 6 × 6-pixel *aperture* of the input level, L0, *U-connects* to all 9 × 16 = 144 cells in the corresponding L1 mac. Each L1 mac U-connects to all 9 × 9 = 81 cells in the overlying L2 mac. All four L1 macs forming one quadrant of level L1 U-connect to the same overlying L2 mac (i.e., convergence). All four L2 macs U-connect to all 9 × 9 = 81 L3 cells (more convergence). The figure is a snapshot of the model while processing frame 15 of Snippet 1. L1 mac activation criteria were set in this study so that an L1 mac would only become active if between 5 and 7 (of the 36) pixels in its aperture were active: apertures with too few or too many active pixels are grayed out. Criteria were set to allow an L2 mac to become active if between 1 and 4 of its four afferent L1 macs were active, and to allow the L3 mac to become active if between 1 and 4 of its four afferent L2 macs were active.

Before discussing the features learned by several of the model's macs, we first report the recognition accuracy. The core accuracy measure, $\Gamma(x, x')$, is the similarity (normalized intersection) of the codes (SDCs) active in a given mac M^j_i on a given frame t during the learning and test presentations of a snippet x as in Equation (14), where we normalize by the fixed size Q_j of codes in macs at level j. Note that the test presentation of x is denoted as x'. We can then average over all macs at all levels to get the recognition accuracy for the whole network on frame t of the test trial for x', $R_t(x')$, as in Equation (15). (Note that since these studies involve exact-match recognition, where the test and training snippets are identical, we can drop x from the notation.) We can then average over all T frames of the test trial to get the recognition accuracy of the entire hierarchical spatiotemporal memory trace for snippet x', $R^*(x')$, as in Equation (16). We also report the full network accuracy on just the last frame of the test snippet, $R^{\Omega}(x')$, which is just Equation (15) with t equal to the final frame of the snippet.

$$\Gamma^{j,i}_t(x, x') = \left\lvert {}_x\phi^{j,i}_t \cap_{x'} \phi^{j,i}_t \right\rvert / Q_j \tag{14}$$

[4]The model included an additional 82,944 top-down (D) synapses from cells at the first internal level (L1) to cells at the input level (L0). However, these synapses are neither required for nor used during recognition and thus, are not counted in computation of information storage capacity in bits/synapse.

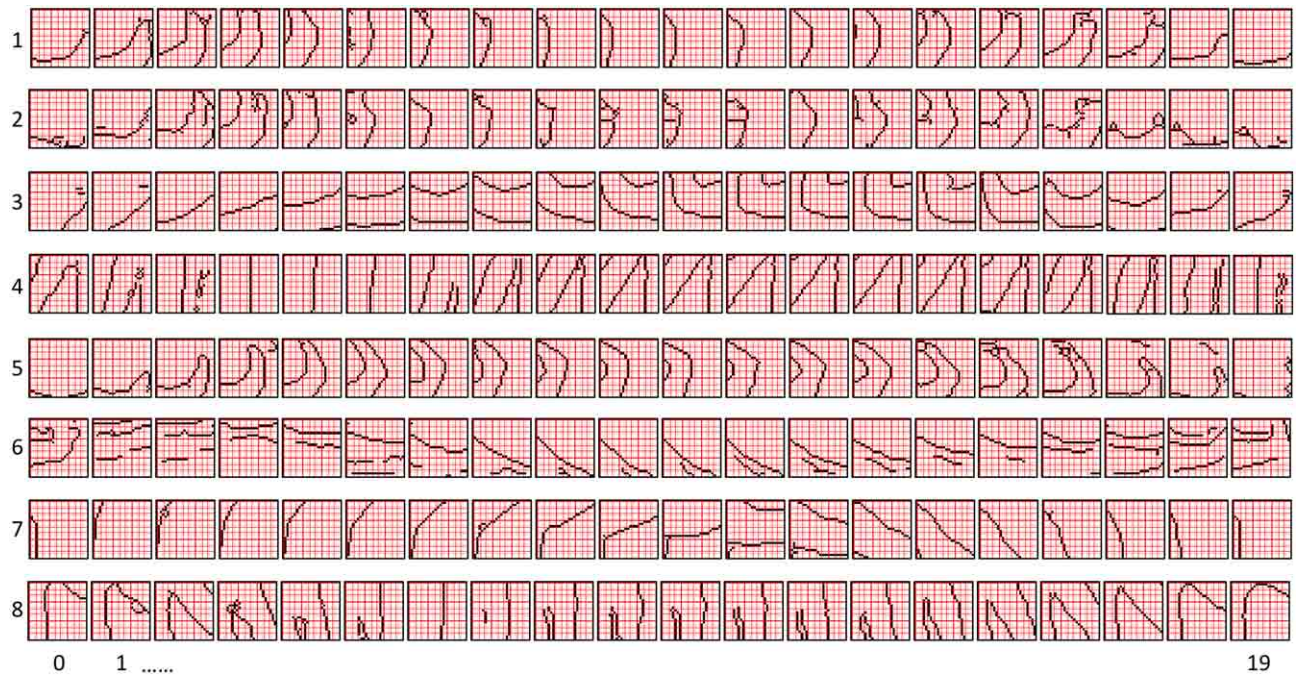


FIGURE III-3 | The frames of the eight snippets used in Study 2.

$$R_t(x') = \sum_{j=1}^{J} \sum_{i=1}^{M_j} \Gamma_t^{j,i}\left(x, x'\right) \tag{15}$$

$$R^*(x') = \sum_{t} R_t\left(x'\right)/T \tag{16}$$

Table II-2 reports $R^*(x')$ and $R^{\Omega}(x')$ for all snippets (and broken down by level as well) and averaged across all snippets (bottom row). It provides these results using the two CSA retrieval modes described in Section Sparsey's Core Algorithm, the *probabilistic* mode (columns 5 and 6), which is identical to the learning mode except that it uses the back-off protocol, and the *simple* mode (columns 3 and 4), which simply chooses the cell with the maximum V in each CM as winner (i.e., without using the mac-global information, G). The first point to make regarding Sparsey's performance on this set is that using the simple retrieval mode, it achieves an overall accuracy across all frames of all episodes of 85% and across all final frames of 91%. One can readily see that the simple retrieval mode does far better than the probabilistic mode. But again, the simple mode presumes that the model "knows" that it is operating purely in retrieval mode.

As noted above, these are *exact-match* recognition tests: the test sequences are identical to the training sequences. One might therefore be underwhelmed by anything less than 100% recognition. After all, in classification experiments, perfect classification of all training inputs is typically considered a basic sanity test. However, our R measures are not reporting the *class* of the test sequences: they are reporting the detailed comparison of the hierarchical, spatiotemporal patterns of activation that occur during the test and training trials. (Note: we refer to the activation pattern that transpires on the test trial as a *memory trace* and to the one that transpires on the training trial as the *learning trace*.) In this study, these traces span four levels, 20 time steps, involve precisely ordered activation of 1–2 thousand neurons, and are formed with one trial. **Figure III-5** gives some idea of this complexity: it shows the full 4-level learning trace for the first four frames of Sequence (Snippet) 1.

Thus, despite being less than perfect on this exact-match recognition experiment, we consider this performance (in the simple retrieval mode) to be good. Bear in mind that these experiments reflect very little in the way of parameter optimization: the model parameter space is very large and its exploration will be ongoing for quite some time. Moreover, we anticipate that there are many possible straightforward model modifications that would likely boost performance without increasing the model's time complexity for either learning or retrieval. For example, many of the static parameters in the CSA equations could be made dynamic, e.g., to depend on temporal offset from start of sequence, or on degrees of saturation of weight matrices, etc. There is a very large landscape to explore here. Furthermore, as noted, this study involved only unsupervised learning. As discussed in Section Learning arbitrarily complex nonlinear similarity metrics, the addition of supervised learning to the model greatly increases its capabilities, i.e., to learning arbitrarily nonlinear (spatiotemporal) categories. However, we do not report supervised learning studies in this paper.

While the simple mode performance is good, we note that the model in this case has almost 1.9 million weights. Thus, the information storage capacity here is rather low, ∼0.018

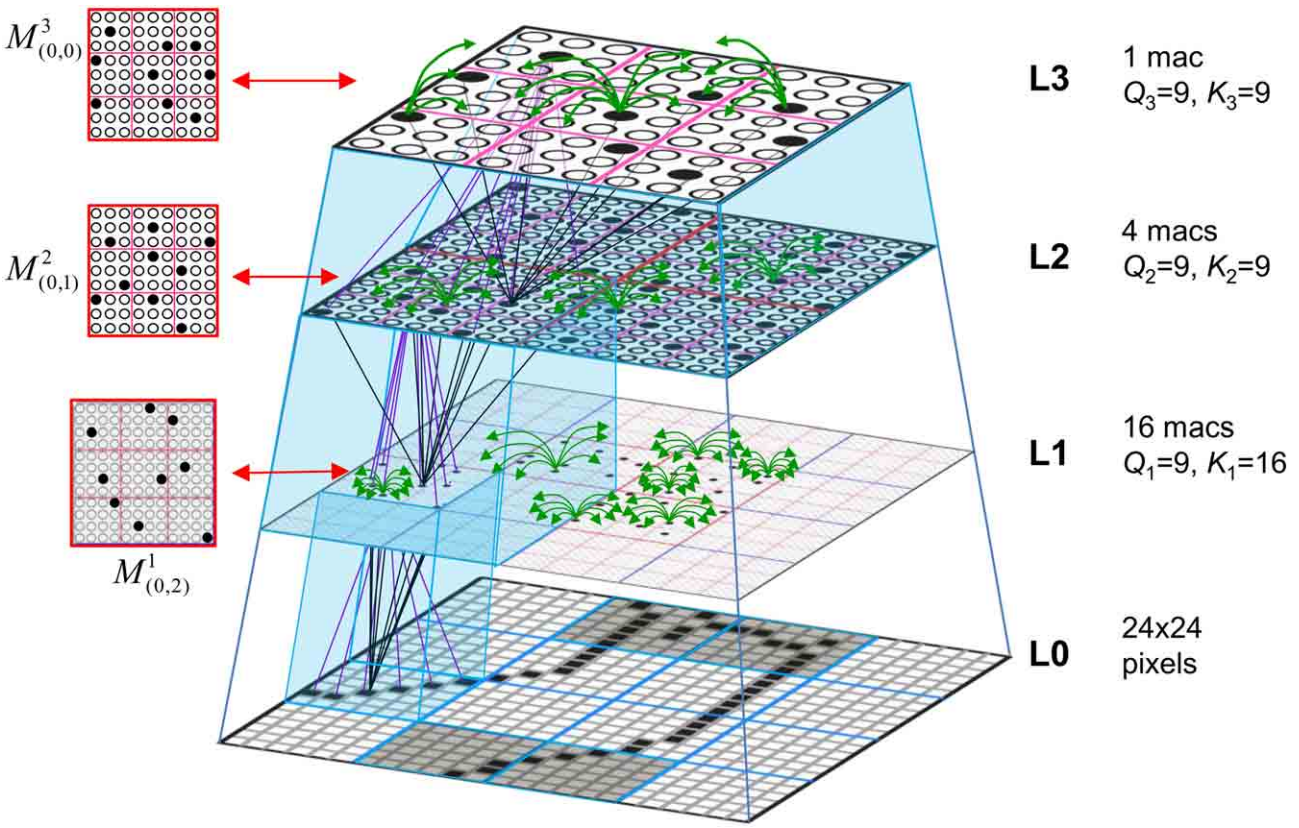


FIGURE III-4 | The 4-level model used in Study 2. Blue semi-transparent prisms show the U-RFs of the macs at the prism tops, i.e., the L3 mac $M^3_{(0,0)}$'s U-RF contains all four L2 macs, each L2 mac's U-RF contains the four underlying L1 macs, and each L1 mac's U-RF is the underlying 6 × 6-pixel L0 aperture. At left, we show plan views of individual macs showing their active codes at the particular spatiotemporal moment depicted ($T = 15$ of Snippet 1). The four gray L0 apertures have too few/many active pixels for their associated L1 macs to activate. Subsets of the U (black), H (green), and D (magenta) wts are shown.

bits/synapse. However, in the course of our investigations, we were routinely able to achieve the same or better performance on this data set with much smaller networks, e.g., 4-level networks with a total of 331,000 weights[5], yielding a storage capacity of ∼0.1 bits/synapse, which is within an order of magnitude of the theoretical maximum for associative memory, ∼0.69 bits/synapse (Willshaw et al., 1969). Those results are given in the last two columns **Table II-2**.

While this unsupervised learning study involves only the exact-match condition (the test inputs are identical to the training inputs), the more typical goal of an unsupervised learning study is to show that the model learns the *higher-order statistical structure* of the input space, or in terms we used earlier, that the model maps similar inputs to similar codes (SISC). Study 3 involves the nonexact-match condition (the test inputs differ from the training inputs) and directly demonstrates that the model retrieves the spatiotemporally best matching stored input given a novel input.

The effect of the lower and upper mac activation bounds on the number of active features needed for a mac to activate (see Section Step 3: Normalize and filter the raw summations) can also be seen in **Figure III-5**. For L1, $\pi^{1,-}_U = 5$ and $\pi^{1,+}_U = 7$ (we've added the level index to the superscript since these parameters can vary by level): thus only a few of the 16 L1 macs become active on each frame, e.g., five on Frame 0, six on Frame 1, etc. One such criterion-meeting L1 mac, M^1_{12}, and its L0 aperture (with six active pixels) are highlighted in yellow in Frame 0. As noted in Section Normalize and filter the raw summations, for L2 and higher, the number of active features equals the number of active macs in a mac's U-RF. In this simulation, the bounds for L2 macs were $\pi^{2,-}_U = 1$ and $\pi^{2,+}_U = 4$ and the bounds for the L3 mac were $\pi^{3,-}_U = 1$ and $\pi^{3,+}_U = 3$. Thus, on Frame 1, we can see that L2 mac M^2_2 (yellow) is active because the number of active features in its U-RF, $\pi^{2,2}_U(1) = 1$, meets the criteria:

$$\pi^{2,-}_U = 1 \leq \pi^{2,2}_U(1) = 1 \leq \pi^{2,+}_U = 4$$

[5]The smaller model had 4 levels, a total of 21 macrocolumns (macs), 1692 cells ("neurons"), and 343,116 synapses. It had an additional 32,256 D synapses from L1 cells to L0 cells. However, these synapses are neither required for nor used during recognition and thus, are not counted in the computation of information storage capacity in bits/synapse. L1 consisted of 16 macs, each with of $Q1 = 4$ CMs, and each CM consisting of $K1 = 14$ cells. L2 had 4 macs, each having of $Q2 = 4$ CMs, each having of $K2 = 12$ cells. The top level (L3) consisted of one mac, consisting of $Q3 = 4$ CMs, each with $K3 = 7$ cells.

Table II-2 | Recognition accuracy for Study 2.

Snippet	Level	Larger model				Smaller model	
		Simple mode		Probabilistic mode		Simple mode	
		$R^{d*}(x')$	$R^{\Omega}(x')$	$R^{*}(x')$	$R^{\Omega}(x')$	$R^{*}(x')$	$R^{\Omega}(x')$
1	L3	0.56	0.78	0.46	0.67	0.53	1.00
	L2	0.90	1.00	0.73	0.50	0.77	1.00
	L1	0.95	0.97	0.71	0.53	0.82	0.88
2	L3	0.53	0.33	0.29	0	0.39	0.50
	L2	0.84	1.00	0.39	0	0.55	0.50
	L1	0.86	1.00	0.55	0.37	0.72	0.75
3	L3	0.57	0.56	0.26	0.67	0.69	0.75
	L2	0.92	1.00	0.65	1.00	0.92	1.00
	L1	0.98	1.00	0.74	0.89	0.98	1.00
4	L3	0.85	0.78	0.86	0.78	0.90	1.00
	L2	0.94	1.00	0.86	1.00	0.91	1.00
	L1	0.94	1.00	0.83	0.95	0.93	1.00
5	L3	0.63	0.89	0.46	0.56	0.84	1.00
	L2	0.91	1.00	0.86	1.00	0.90	1.00
	L1	0.94	1.00	0.85	0.93	0.93	1.00
6	L3	0.86	1.00	0.69	1.00	0.81	1.00
	L2	0.97	1.00	0.82	0.83	0.92	1.00
	L1	0.96	1.00	0.82	0.93	0.95	1.00
7	L3	0.64	0.78	0.55	0.44	0.84	0.75
	L2	0.95	1.00	0.91	0.33	0.99	0.88
	L1	0.98	1.00	0.84	0.41	1.00	1.00
8	L3	0.79	0.78	0.55	0.78	0.93	1.00
	L2	0.95	1.00	0.85	0.89	0.94	1.00
	L1	0.94	1.00	0.84	0.81	0.93	1.00
All Snippets	All Levels	85.0	91.0	0.68	0.68	0.84	0.92

All accuracies expressed as decimal (between 0 and 1). $R^{*}(x')$ is averaged over all 20 frames of snippet, x', where in this case (the exact-match test case), x' is identical to the training snippet, x. $R^{\Omega}(x')$ is the accuracy only on the final frame of snippet x'.

M_1^2 (rose) and M_3^2 (no color) also activate because they also meet the criteria:

$$\pi_U^{2,-} = 1 \leq \pi_U^{2,1}(1) = 2 \leq \pi_U^{2,+} = 4$$
$$\pi_U^{2,-} = 1 \leq \pi_U^{2,3}(1) = 3 \leq \pi_U^{2,+} = 4$$

The blue boxes indicate that L3 mac M_0^3's U-RF is the entire L2 level; M_0^3 is active on all four frames because it meets its mac activation bound criteria on all for frames.

The *progressive persistence* property can also be seen in **Figure III-5**. The persistence at L2 is two frames, i.e., $\delta^2 = 2$. Thus, the L2 code (the set of 9 black cells) that becomes active in M_2^2 on Frame 0 remains active on Frame 1. That same L2 code, which (following earlier notation) we can denote, $\phi_0^{2,2}$, becomes D-associated with the L1 codes active in its U-RF on Frames 0 and 1, denoted $\phi_1^{1,12}$ and $\phi_2^{1,9}$, respectively. Magenta lines show the increased D-wts from one of the cells in $\phi_0^{2,2}$ to the L1 codes, $\phi_1^{1,12}$ and $\phi_2^{1,9}$, though the same increases would occur from the other eight cells comprising $\phi_0^{2,2}$ $(= \phi_1^{2,2})$ as well. Similarly, the code that becomes active in M_2^2 on Frame 2 remains active on Frame 3. L3 persistence is $\delta^3 = 4$, thus the code activated in M_0^3 on Frame 0 remains active until Frame 3.

The reader may note a discrepancy at L3 between the progressive persistence policy, which says that (during learning) once active, an L3 code will remain active for 4 frames, and the activation bounds, which in this simulation says that an L3 mac will only become active if it has between 1 and 3 active features in its U-RF, whereas on Frames 3 and 4, there are four active features in M_0^3 's U-RF. The resolution is that persistence trumps the activation criteria: that is, the policy, during learning, is to allow a mac that has *already* become active to remain active for its full persistence regardless of how the number of active features in its U-RF changes throughout its persistence.

We also note that though not shown in **Figure III-5**, large numbers of (U, H, and D) synapses are increased within/between macs on each of these frames. This is especially true early in the system's life, when most input patterns that occur will be novel. In general, as more and more frames are experienced, fewer and fewer synapses are increased with each new frame. However, as described in Section Learning policy and mechanics, the model has a "freezing" policy wherein, once a critical fraction of the weights of any of a mac's three afferent projections (U, H, or D) have been increased, *all* of that mac's afferent projections are frozen, preventing any further codes (i.e., features) from being stored in its basis. Freezing is necessary in order to avoid oversaturating the weight matrices, which would lead to information (memory) loss. Once a mac's learning is frozen, the set of features that has been stored in it, remains its permanent lexicon, or *basis*, for perceiving/recognizing all future inputs to it. Note that even if a mac's afferent matrices are frozen, its efferent matrices are not, meaning that previously stored codes in a frozen mac can still be efferently-associated with other codes following freezing.

Although none of the macs in the model in this study became frozen, the codes that were stored in the various macs across the 160 frames of the input set still constitute their learned feature bases. **Figure III-6** shows the complete set of criteria-meeting inputs, i.e., having between $\pi_U^{1,-} = 5$ and $\pi_U^{1,+} = 7$ active pixels, which present to L0 Aperture 0 across all 160 frames. These 45 inputs constitute the learned feature basis of L1 mac M_0^1. Note the near-canonical nature of many of the patterns, e.g., perfect, or near-perfect vertical, horizontal, diagonal edges.

As another example, **Figure III-7** shows the complete set of unique, criteria-meeting patterns that occurred in Aperture 8, and were stored in M_8^1 over the course of the training set. Here, we manually ordered them so as to emphasize the "canonicalness" of the resulting features. In this case, seven of these features (blue underbars) occurred at least twice during the 160 frames. It is perhaps surprising that given such a small number of frames derived from natural video, the resulting basis can be so canonical. Moreover, several of these features are already beginning to recur in the input stream even within the first 160 frames of this model's experience. These phenomena are due to the conjunction of the preprocessing (1-pixel wide edges and binarization), the small aperture size, and the L1 mac activation criteria. Similar bases were learned in the other 14 L1 macs as well. These findings give us confidence that freezing L1 macs even very early in the

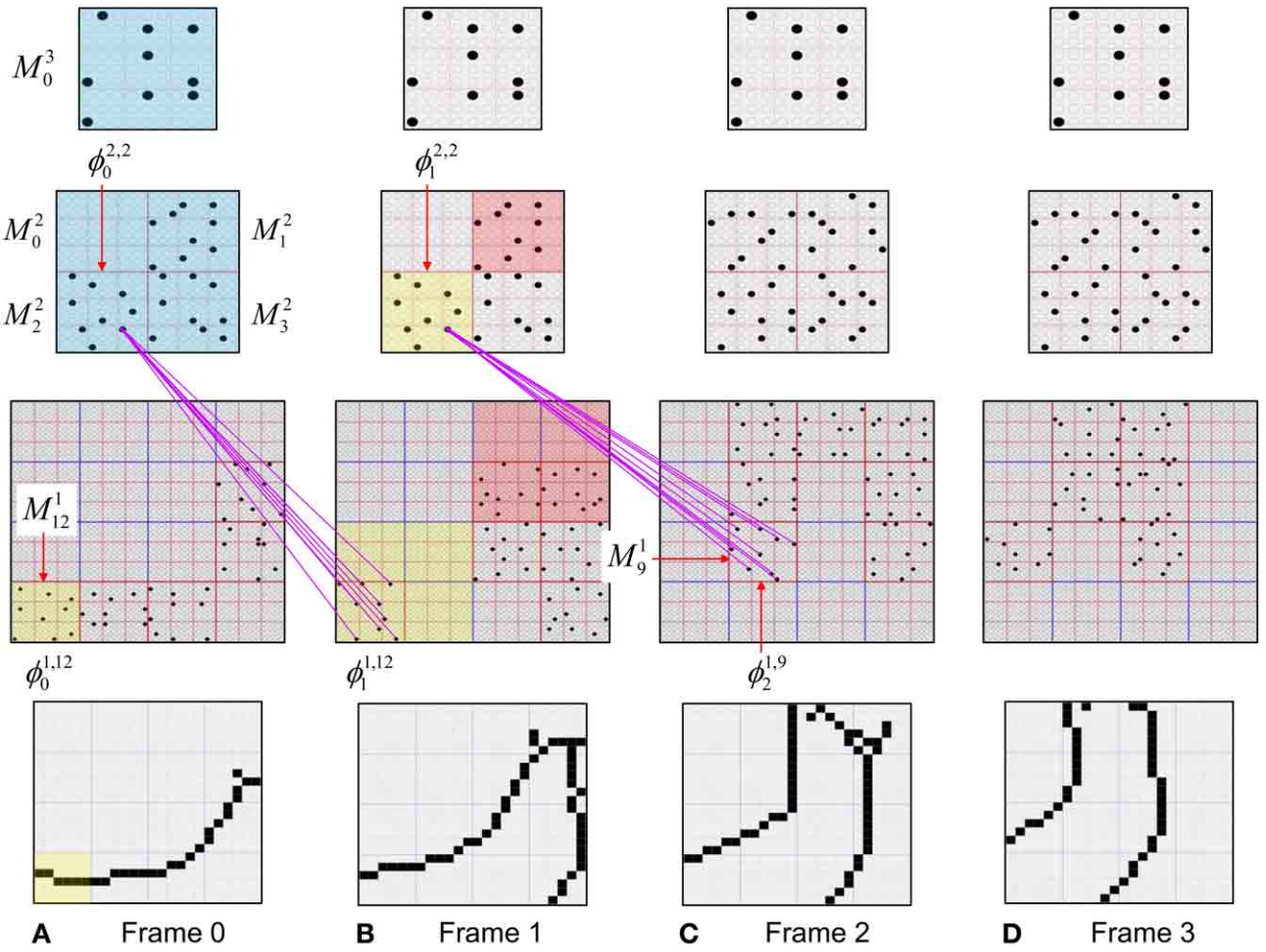


FIGURE III-5 | The inputs (L0) and codes that become active at all three internal levels on the first four frames of Snippet 1 (panels A–D). Note that the mac codes that become active at L2 persist for two time steps. Thus, the code active in mac, M_2^2, on frames 0 and 1 can be referenced by two names, $\phi_0^{2,2}$ and $\phi_1^{2,2}$. The magenta lines show the D-wts from one of the cells comprising $\phi_0^{2,2}$, which are increased onto $\phi_1^{1,12}$ on frame 1 and onto $\phi_2^{1,9}$ on frame 2. See text for further discussion.

"life" of the model, e.g., after a few hundred features have been stored, will allow the macs to parse/recognize all future inputs with quite sufficient fidelity. We feel these results provide an illuminating framework for understanding the various *critical period* phenomena observed in the visual and other modalities of biological brains (Wiesel and Hubel, 1963; Barkat et al., 2011; Pandipati and Schoppa, 2012).

We used the same protocol as above to catalog the input patterns learned by, and stored in, the L2 macs and in the L3 mac. **Figure III-8** shows 78 of the 112 unique, criteria-meeting patterns that occurred in the 12 × 12-pixel region comprising the U-RF of L2 mac M_0^2, throughout the 160 frames of the training set (the thich-outlined green and red pairs are duplicates). This region is the union of the U-RFs of the four L1 macs, M_0^1, M_1^1, M_4^1, and M_5^1. The gray/yellow 6 × 6 quadrants are L0 apertures in which too many ($> \pi_U^{1,+} = 7$)/too few ($< \pi_U^{1,-} = 5$) pixels were active for the L1 mac to activate. Thus, when any of the 12 × 12 patterns in the figure occurs, the actual input passed up to M_0^2 will be from codes active only in the L1 macs whose 6 × 6 RFs are *not* gray or yellow.

As can be seen in **Figure III-8**, the spatial extent of the L2 RF has doubled in width and height compared to L1 RF. Thus, the space of possible inputs in such an RF is exponentially larger. Nevertheless, most of these larger features still have low intrinsic dimensionality, e.g., an essentially straight or low-curvature edge across the whole 12 × 12 RF. Even the more complex features such as the angle features in the bottom row, i.e., two straight/low-curvature segments with a single "elbow" point (thick pink outline), have rather low intrinsic dimensionality (i.e., we can give short verbal descriptions of them). Again, these are canonical-looking features, and they end up in the basis of M_0^2, but they were not hand-engineered. The number of active features (quadrants that are neither gray nor yellow) in each 12 × 12 pattern varies from 0 (in which case, M_0^2 will not become active on that frame, thick blue outline) to 4. Thus, M_0^2 learns input patterns having varying numbers of features (varying complexities).

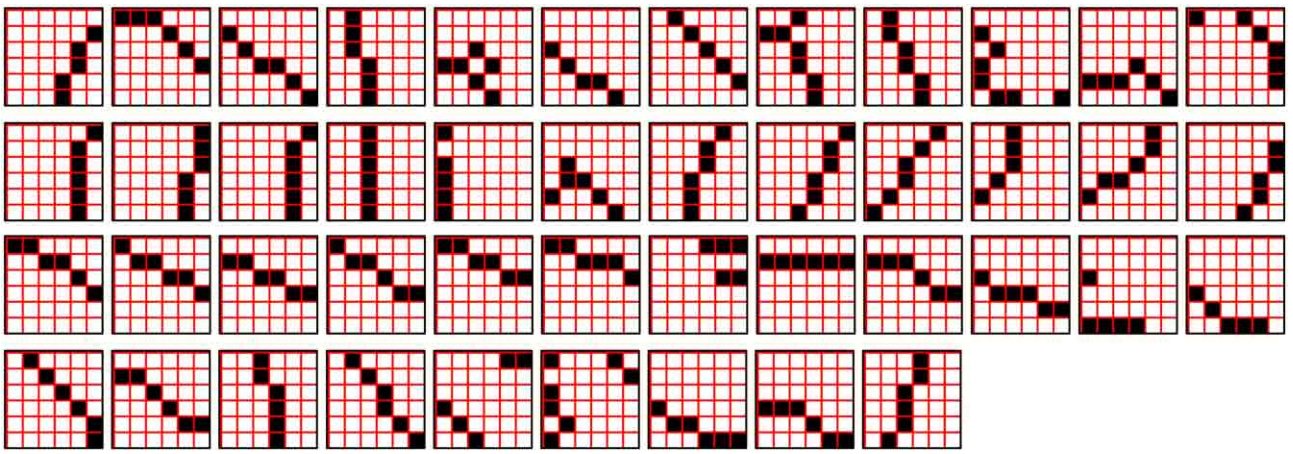


FIGURE III-6 | The set of all unique patterns with between $\pi_U^{1,-} = 5$ and $\pi_U^{1,+} = 7$ active pixels that occurred in L0 Aperture 0 (and are stored in mac M_0^1) throughout the 160 frames of the training set.

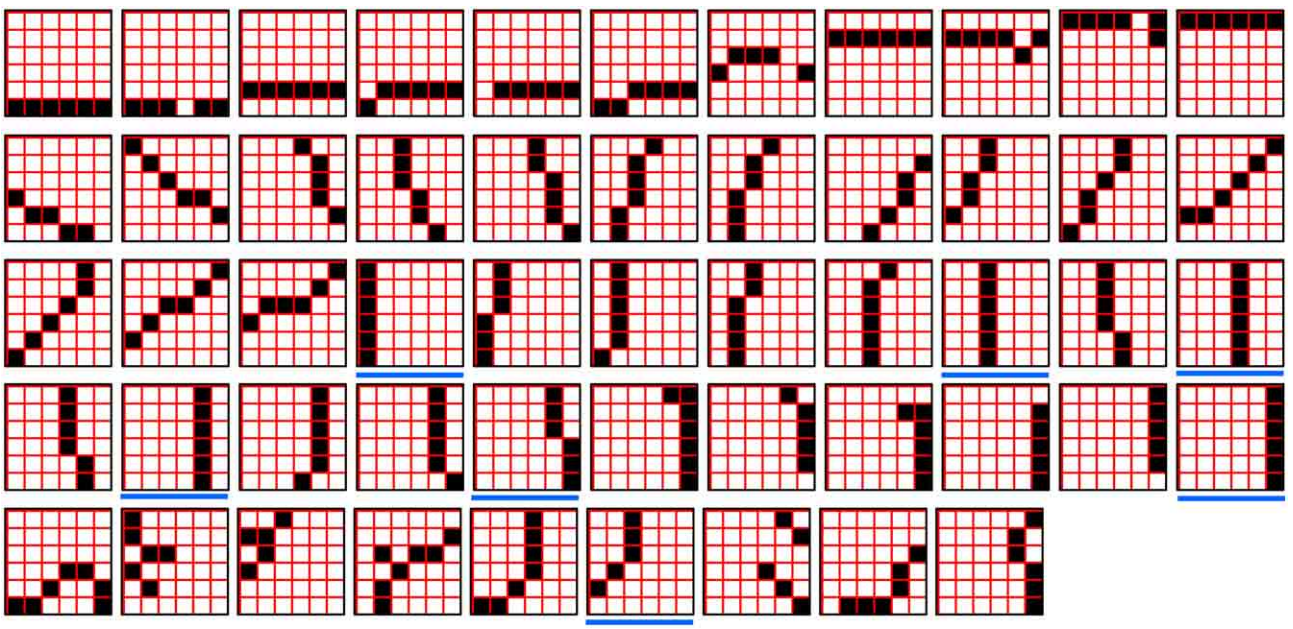


FIGURE III-7 | The set of all unique patterns with between $\pi_U^{1,-} = 5$ and $\pi_U^{1,+} = 7$ active pixels that occur in L0 Aperture 8 (and are stored in mac M_8^1) throughout the 160 frames of the training.

Thus, it is also the case that during retrievals, all these features, of varying complexities, formally compete with each other. In general, this argues for narrower mac activation ranges, $[\pi_U^-, \pi_U^+]$, because narrower ranges make normalization easier. Exploration of the interaction of mac activation ranges across levels and with other parameters is another ongoing effort of our research.

Note that since L2 codes persist for two frames, these input patterns, or to be more precise the SDCs in the corresponding L1 macs, will be associated to only roughly half as many codes in M_0^2. Thus, each consecutive pair of two 12×12 panels (in row-major order) would become associated with the same M_0^2 code. **Figure III-9** illustrates this concept for the first pair of 12×12 panels of **Figure III-8** (thick black outline). Thus, the L2 codes formally represent spatiotemporal patterns. Given the discrete nature of our overall framework, i.e., discrete frames, binary pixels, constrained wiring schemes, these larger-scale (both spatially and temporally) spatiotemporal patterns, i.e., L2 features, can be viewed as spatiotemporal *compositions* of lower-level features. A detailed development and analysis of this spatiotemporal compositional aspect is one major focus of current and future studies.

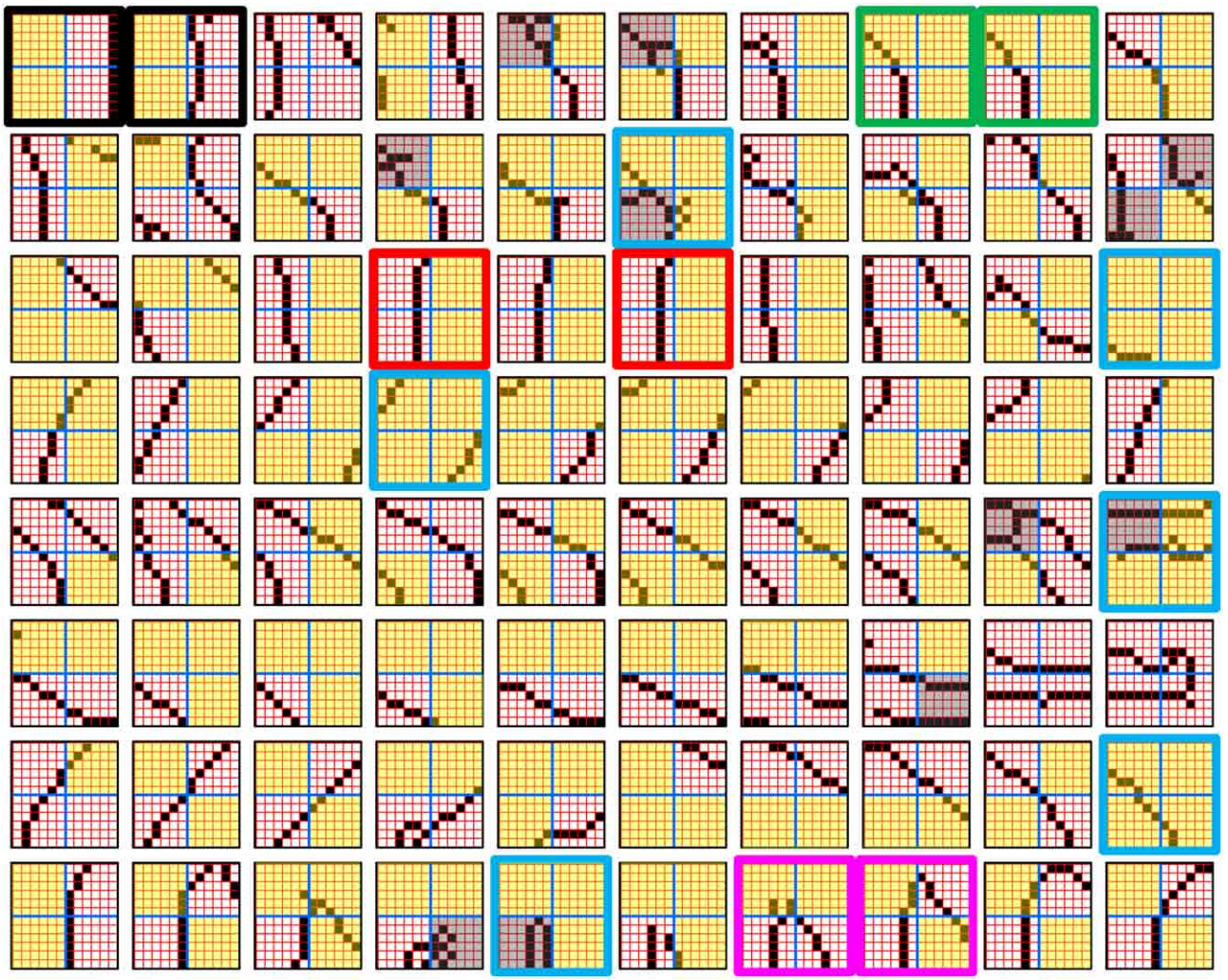


FIGURE III-8 | 78 of the 112 unique patterns with between $\pi_U^{2,-} = 1$ and $\pi_U^{2,+} = 4$ active features that occurred in the U-RF of L2 mac M_0^2 throughout the 160 frames of the training set. Gray/yellow quadrants are ones in which too many/few pixels were active for the corresponding L1 mac to activate.

The concept of operation during learning and also during recognition, is one in which all of the macs across all levels operate, in parallel, on the particular spatiotemporal fragments of the input that they receive, dealing with variation on a fragment-by-fragment basis. Support for this view comes from recent experimental work (Bart and Hegdé, 2012). In subsequent work, we will be quantitatively assessing the similarity of features that occur, over the long time frame of experience, following the initial period in which many of the lower-level macs become frozen, within apertures of the different scales corresponding to the model's different levels, to the (frozen) bases of those macs. The goal will be to assess how well the model is able to represent (and if novel, learn) future inputs using the fixed lexicon of features stored in its lower levels.

Finally, before leaving this section, we want to underscore the very different concept of feature basis present in Sparsey than that present in localist models such as Olshausen and Field (1997). This difference is summarized in terms of four characteristics in **Table II-3**.

STUDY 3: SPATIOTEMPORAL BEST-MATCH RETRIEVAL

In this study, we demonstrate spatiotemporal best-match retrieval as follows. In this case, we are again using a model with one internal level (L1) consisting of one mac with $Q = 9$ CMs; K varies across experiments. In each experimental run, we train the model on a set of random sequences. We then create a noisy version of each training sequence by randomly changing some fraction of the pixels in each of its frames. **Figure III-10** (middle) shows a typical training sequence. **Figure III-10** (top) shows the corresponding randomly produced noisy version of that sequence: one pixel was randomly changed in each frame, which actually yields two pixel-level differences between the original and the noisy frame. Each frame in the training set had between 9 and 12 active pixels, which yields noise levels from $2/9 = 22.2\%$ to $2/12 = 16.7\%$. **Figure III-10** (bottom) shows a sequence produced from

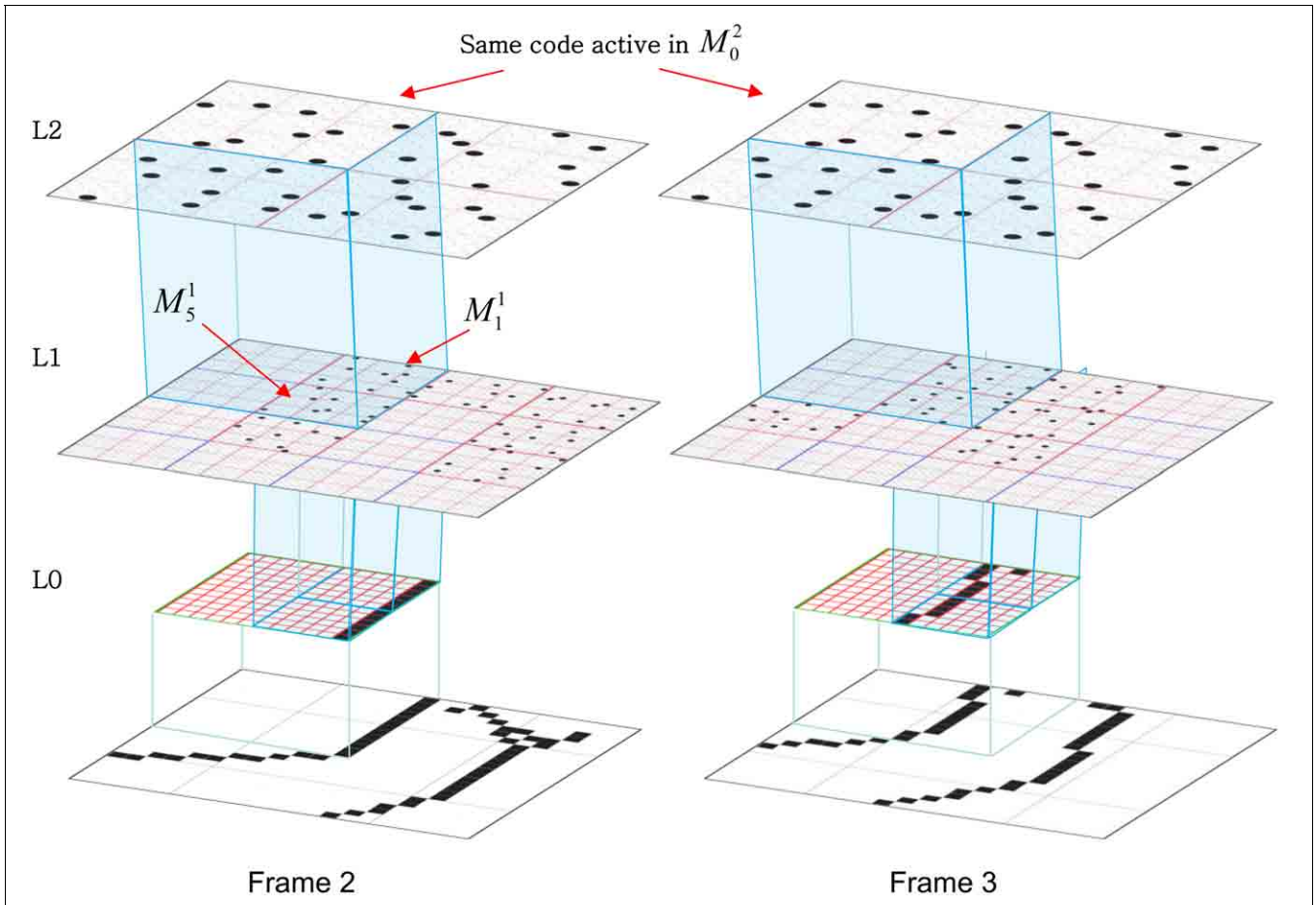


FIGURE III-9 | All L1 codes that become active in M_0^2's U-RF on the two frames depicted will associate with the same M_0^2 code. Notice that that same code is active for both frames. Thus, a total of four unique L1 codes—the two active in M_1^1 on Frames 2 and 3 and the two active in M_5^1 on Frames 2 and 3—will associate with the M_0^2 code shown. There will in addition be H-association from the L1 codes active on Frame 2 to those active on Frame 3 and also from the L2 code recurrently to itself (since it is on for two consecutive frames), and D-associations as well.

the middle one by randomly changing two pixels in each frame, which yields four pixel-level differences and thus noise levels, from 4/9 = 44.4% to 4/12 = 33.3%. In this study, we ran one series of experiments testing with the 1-pixel-changed frames (columns 5–7 of **Table II-4**) and one series testing with the 2-pixels-changed frames (columns 8–10 of **Table II-4**).

Given the random method of creating individual frames of the training set and the high input dimension involved (144), if the fraction of changed pixels is small enough, e.g., <10–20%, then the probability that a changed frame, x' will end up closer to (having higher intersection with) any other frame in the training set than to the frame, x, from which it was created, is extremely small. Moreover, remember that Sparsey actually "sees" each input frame *in the context of the sequence frames leading up to it*, i.e., it computes the spatiotemporal similarity of particular *moments* in time (by virtue of its combining of U and H signals on each time step), not simply the spatial similarity between isolated *snapshots*. Thus, the relevant point is that the probability that a changed moment, e.g., [x′,y′,**z**′], with its exponentially higher dimensionality (144^3), will end up closer to any other moment in the training set than to the moment, [x,y,**z**], from which it was created, is vanishingly small[6]. This condition is required to validate the testing protocol/criterion described above, which compares the L1 code on each test frame to the L1 code on the corresponding training frame. Thus, if we can show that the model activates the exact same sequence of L1 codes in response to the noisy sequence, then we will have shown that the model is doing spatiotemporal *best-match* retrieval.

Columns 5–7 of **Table II-4** show that the model is able to recognize a set of training sequences despite significant noise on every frame. The absolute capacity increases with network size. For example, the network of Row 1 had 6336 weights and showed good recognition of two 10-frame random sequences despite 16.7–22.2% noise on each frame, while the larger network of Row 4 had 39,168 weights and showed very good recognition for 10, similarly noisy, 10-frame sequences, and so on. Columns 8–10 of **Table II-4** show that the model still performs well even for much

[6]The notation [x,y,**z**], with z bolded, indicates the moment on which frame z is being presented as the third frame of a sequence after x and y have been presented as the first and second frames of the sequence.

Table II-3 | Comparison of the concept of "feature basis" present in Sparsey and localist models.

1. ORIGIN OF ANY SINGLE BASIS FEATURE	
Sparsey	A single input pattern experienced, even with a single trial
Localist	An average of many inputs experienced
2. CONTENT OF ANY SINGLE FEATURE	
Sparsey	Multiple spatial phases (i.e., multiple edge segments at different locations in the aperture), as apparent, e.g. in **Figure III-8**, multiple spatial frequencies, multiple orientations (i.e., each of the multiple possible edge segments in the aperture can have a different orientation), and multiple temporal frequencies. Because each Sparsey feature is derived from a single event (and not an average over multiple events), it's not really appropriate to speak of a Sparsey feature as having multiple modes on each of the encoded stimulus dimensions
Localist	A single spatial phase, a single spatial frequency, a single orientation, and a single temporal frequency. In general, a localist feature such as these is unimodal (e.g., Gaussian, Gabor) on each of the encoded stimulus dimensions
3. NUMBER OF UNITS IN THE CODE OF ANY SINGLE FEATURE	
Sparsey	Many. Q, where, in real macrocolumns, Q is order 100
Localist	One
4. NUMBER OF BASIS FEATURES PARTICIPATING IN THE REPRESENTATION OF ANY SINGLE INPUT (MOMENT)	
Sparsey	One. But again, that one active feature is represented by Q active units. Thus, this type of representation is called "sparse" specifically because the number of physical units active in representing any one input is small compared to the total number of physical units. But, these representations can also be sparse in the senses typically used for localist models (below). As noted above, any single active SDC represents the presence of multiple (*but, for most natural inputs, a smallish number of*) spatial phases, spatial frequencies, orientations, and temporal frequencies
Localist	Few, several. These representations are called "sparse" for two reasons • The number of features in a sufficient basis is small compared to the number of all possible features definable on the input space • The number of features active in the representation of any one input is small compared to the number of features in the basis

larger per-frame noise levels. In these tests, which involved randomly changing two pixels on every frame, the frame-wise noise levels varied from 33.3% (on frames which had 12 active pixels) to 44.4% (on frames with 9 active pixels). A key point to note in **Table II-4** is that while the absolute capacities (the number of sequences that are can be stored) are lower for the 2-pixel-changed series compared to the 1-pixel-chaged series, capacity still remains large. The primary reason for lower storage capacity in the 2-pixel-changed case is that because the test input frames are less similar to the training input frames (than in the 1-pixel-changed case), the ρ distributions from which winners in the CMs are chosen [Equation (11) and Step 12 of the CSA] will be flatter, yielding more single-unit errors, thus reducing $R^*(x')$.

Table II-5 gives the detailed (frame-by-frame) accuracies for all sequences for individual runs of Experiments 1 and 4. The top two rows are for Experiment 1 in which the small network could store only two sequences while maintaining reasonably high recognition accuracy. The bottom 10 rows are for a run of Experiment 4 in which the network had $Z = 144$ L1 units and 39,168 weights. The rightmost column, $R^*(x')$, is the average over all 10 frames of a given sequence presentation. It is important to note how the model fails as it is stressed by having to store additional sequences. Specifically, even as accuracy averaged over all sequences falls, a subset of the stored sequences is still recognized perfectly. This can be seen even in the small network example: Seq. 1 is retrieved virtually perfectly. Only a single unit-level error is made on frame 6. Seq. 2 starts out being recalled perfectly for the first few frames but then begins picking up errors in frame 4 and hobbles along for the rest of the sequence. Nevertheless, note that even on the last frame of Seq. 2, the L1 code is still correct in 5 of the 9 CMs. In Experiment 4, we see that 9 of the 10 sequences are recalled virtually perfectly, while one (Seq. 9) begins perfectly but then picks up some errors on frame 5 and then degrades to 0% accuracy by the last frame. It is also important to realize that while the model occasionally makes mistakes, it generally recovers by the next frame. In other examples (not shown here), the model can often recover from more significant errors.

Figure III-11 shows the pair-wise L1 code intersections over the full set of frames experienced over all training and test frames (moments) of the experimental run described in the top two lines of **Table II-5**. Since there were two 10-frame sequences, this is a total of 40 frames. The upper yellow triangle shows the intersections between all codes assigned on the 10 frames of the training presentation of Seq. 1. Similarly for the other triangles down the main diagonal. The top value the green triangle (row 20, col. 1) shows that L1 code "20," i.e., the code activated on the first frame of the *test* presentation of Seq.1 intersects completely (in all $Q = 9$ CMs) with L1 code 0, i.e., the code activated on the first frame of the *training* presentation of Seq. 1. Similarly, for codes, 21 and 1, 22 and 2, etc. Reading down the minor diagonal (between the red lines) tells how well the model does: perfect recognition of all noisy frames of all sequences would yield "9"s all the way down.

Constant-time retrieval

When each frame is presented during a recognition test trial the likelihoods of all codes stored during the learning trial are formally evaluated. They are evaluated in parallel by the constant-time code selection algorithm (CSA). However, at no point does the model produce *explicit* representations of the likelihoods of the individual codes (hypotheses) stored. Such an explicit representation, e.g., a list, of likelihoods would constitute a localist representation of those likelihoods. What the model actually does is make Q draws, one in each CM. However, the *net effect* of making these Q draws (*soft-max* operations) is that a *hard-max* over *all* stored hypotheses is evaluated. This is true whether the model has stored a single 5-frame sequence, or a single 500-frame sequence, or 100 5-frame sequences. And crucially, because the numbers of CMs, and thus units, and weights, are

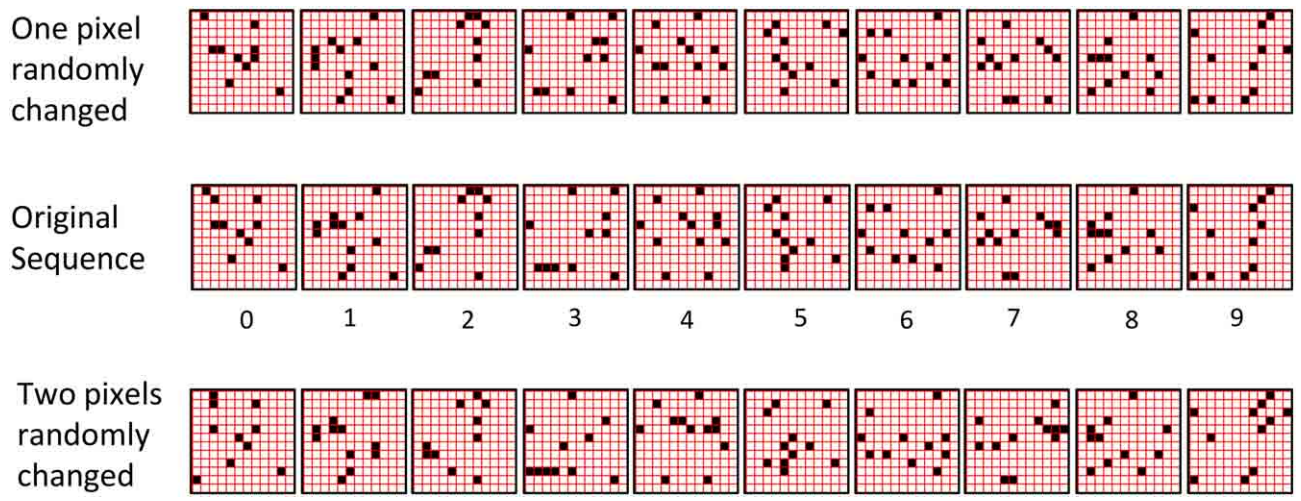


FIGURE III-10 | (Middle row) An example 10-frame training sequence used in Study 3. (Top row) A noisy version of the training sequence in which one pixel was randomly changed in each frame. The resulting frame has two pixel-level differences from the original. **(Bottom Row)** A noisy version of the training sequence in which two pixels were randomly changed in each frame.

Table II-4 | Best-match recognition testing.

Exp	*K*	*Z*	*W*	16.7% (1-pixel-) changed)			33.3% (2-pixels-) changed)		
				S	$R^{*}(x')$	$R^{\Omega}(x')$	***S***	$R^{d*}(x')$	$R^{\Omega}(x')$
1	4	36	6336	2	83.0	67.0	2	83.0	76.0
2	8	72	14,976	5	91.0	86.0	4	98.0	97.0
3	12	108	25,920	8	96.0	96.0	7	94.0	93.0
4	16	144	39,168	10	95.0	94.0	8	92.0	89.0
5	20	180	54,720	11	87.0	84.0	9	90.0	84.0
6	24	216	72,576	12	88.0	84.0	10	86.0	79.0
7	28	252	92,736	13	88.0	84.0	10	89.0	82.0
8	32	288	115,200	15	88.0	86.0	10	91.0	83.0

The R measures (defined in the text) are in %. $Z = Q \times K$ is the total number of L1 units. W is the total number of U and H weights in the model. S is the number of sequences in the training set. All sequences were 10 frames long. $R^{*}(x')$ and $R^{\Omega}(x')$ are averages over the 10 runs of an experiment.

Table II-5 | Detailed frame-by-frame accuracies.

Seq	0	1	2	3	4	5	6	7	8	9	$R^{*}(x')$
1	100	100	100	100	100	100	88.9	100	100	100	99
2	100	100	100	100	55.6	66.7	66.7	66.7	55.6	55.6	77
1	88.9	100	100	100	100	100	100	100	88.9	100	98
2	100	100	100	100	100	100	100	100	100	88.9	99
3	100	100	100	100	100	100	100	100	100	100	100
4	88.9	100	88.9	88.9	100	100	100	100	100	100	97
5	100	100	100	100	100	100	100	100	100	100	100
6	100	88.9	100	100	100	100	100	88.9	100	100	98
7	100	100	100	100	100	100	88.9	100	100	100	99
8	88.9	100	100	100	100	100	100	100	100	100	99
9	100	100	100	100	100	77.8	66.6	22.2	11.1	0	68
10	100	100	100	100	100	100	100	100	100	100	100

All table cells give accuracies as percent. Last column is average of columns indexed 0–9.

fixed, the time it takes to make those Q draws *remains constant* as additional codes (hypotheses) are stored.

What the results in this report say is that that hard-max, i.e., the max-likelihood hypothesis, is returned with probability that can be very close to 1 if the amount of information (i.e., number of hypotheses) stored remains below a soft threshold, and which decreases as we move beyond that threshold. For example, looking at **Table II-5**, we see that for the second experiment (bottom 10 rows), the model chooses the correct, i.e., maximum likelihood, hypothesis on almost all of the 100 frames (moments) of test phase. These are 100 independent decisions, in each of which, all 100 stored hypotheses competed and had some non-zero possibility of being activated. Yet, almost all 100 whole-code-level decisions were correct. And, at the finer scale of the individual CMs, where the actual decision process, albeit a soft decision process, takes place, almost all (861) of the 900 decisions were correct.

Table II-6 shows what happens when we move past or perhaps *through*, the aforementioned soft threshold. In these two experiments, we again used the network with 36,198 weights and the 1-pixel-changed test, but the training set contained 11 sequences (upper 11 rows) and 12 sequences (lower 12 rows), compared to only 10 in the experiment reported in **Table II-5**. For the 11-sequence case, the model still performs very well on six of the sequences, but adding another sequence degrades performance substantially more.

SUMMARY AND CONCLUSION

In this paper, we described the hierarchical and spatiotemporal elaboration of the SDC-based macro/mini-column model of cortical computation described in Rinkus (2010), named Sparsey.

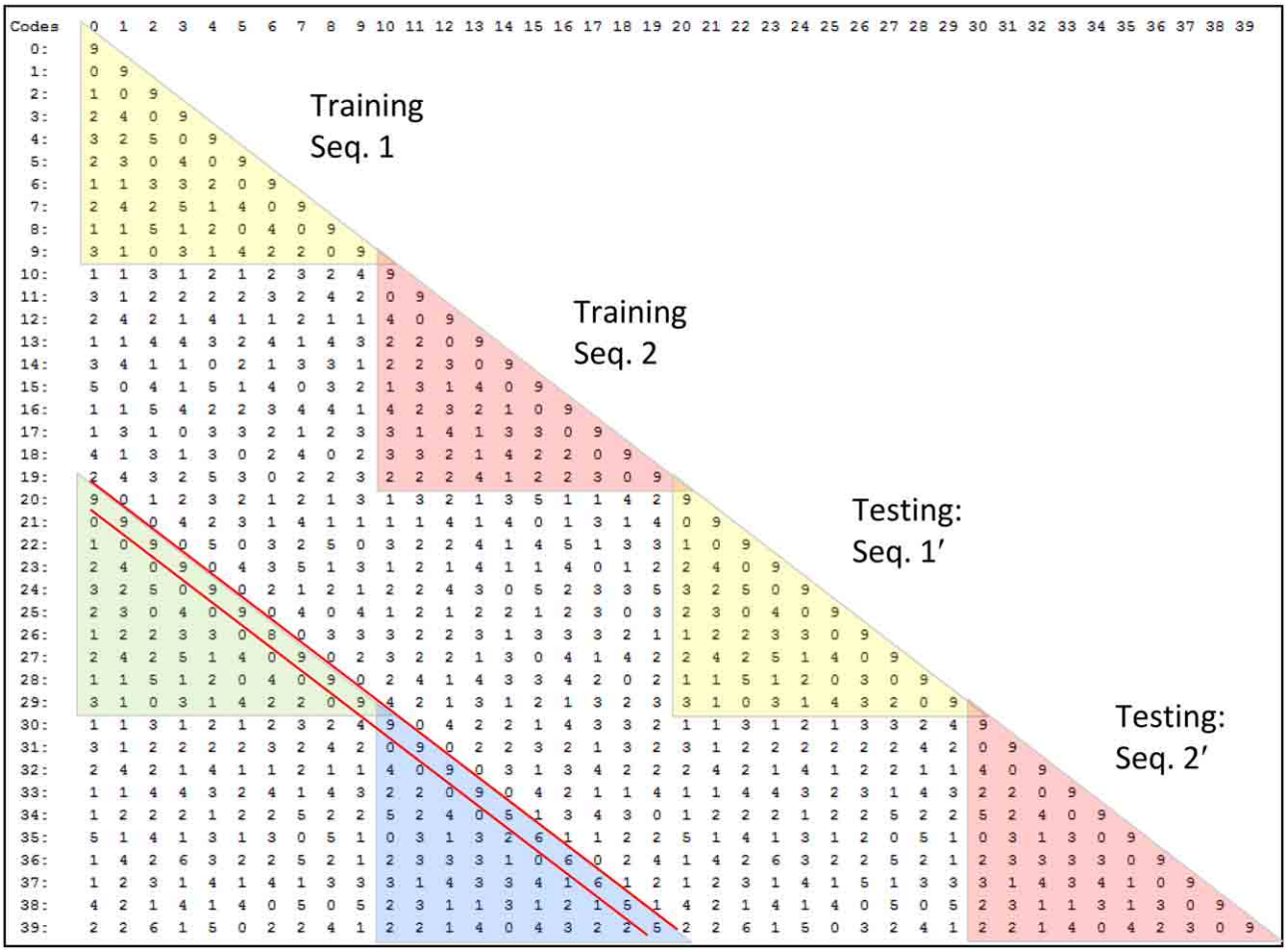


FIGURE III-11 | Pair-wise intersections of all L1 codes assigned in one run of the 1-pixel-changed testing condition for the smallest model tested, which had $Q = 9$, $K = 4$, and 6336 weights.

The notion that *hierarchical representation* is essential to event recognition and intelligence more generally, has been present in models for decades (Fukushima, 1984; Damasio, 1989; Edelman and Poggio, 1991; Riesenhuber and Poggio, 1999; Lucke, 2004; George and Hawkins, 2005; Dean, 2006; Jitsev, 2010) including in the recent "Deep Learning" motif (LeCun and Bengio, 1995; Hinton et al., 2006; Hinton, 2007; Taylor et al., 2010; Le et al., 2011). The representational and processing economy/efficiency of learning and recognition (inference) that is afforded by hierarchical decomposition of concepts/events has been understood (at least implicitly) for thousands of years, e.g., the game of "Twenty Questions," which works because of hierarchical way in which information is organized in our brains.

The hierarchical models noted above and many more all realize the benefit of compositional representation. However, most of those models use localist representations in which, in any given cortical patch, each feature/concept/event is represented by a single unit. In contrast, Sparsey uses sparse distributed codes (SDCs) in every cortical patch. As stated at the outset, the most important distinction between *localism* and SDC is that SDC allows the two essential operations of associative (content-addressable) memory, storing new inputs and retrieving the *best-matching* stored input, to be done *in fixed time for the life of the model*, which is essential for scalability to the huge problem sizes increasingly associated with label, "Big Data." The basis for this fixed-time capability was explained in Section Sparse Distributed Codes vs. Localist Codes.

(1) Because SDCs physically overlap, if one particular SDC, ϕ (and thus, the hypothesis that it represents), stored in a mac is *fully* active, i.e., if all Q of ϕ's cells are active, *then all other codes (and thus, their associated hypotheses) stored in that mac are also simultaneously physically partially active in proportion to the size of their intersections with* ϕ[7].

[7]There is a nuance here. Although we say "all" stored hypotheses physically influence the next time step's decision processes, there may generally be a significant number of hypotheses stored in a mac, which have zero intersection with the current fully active code, ϕ. One might therefore assert that these hypotheses do not physically influence the next time step's decision processes.

Table II-6 | Detailed frame-by-frame accuracies. overloaded case.

Seq	0	1	2	3	4	5	6	7	8	9	$R^*(x')$
1	88.9	88.9	100	100	88.9	100	100	100	100	100	97
2	66.7	77.8	88.9	100	77.8	55.6	22.2	22.2	11.1	0	52
3	100	88.9	100	100	88.9	100	100	100	100	100	99
4	66.7	77.8	77.8	66.7	44.4	33.3	11.1	44.4	0	0	42
5	88.9	100	100	100	100	88.9	77.8	66.7	33.3	33.3	79
6	88.9	100	100	100	100	100	100	100	100	88.9	98
7	66.7	55.6	22.2	22.2	11.1	11.1	33.3	22.2	0	11.1	26
8	66.7	33.3	11.1	11.1	22.2	0	11.1	11.1	22.2	33.3	22
9	100	100	100	100	88.9	100	100	88.9	88.9	88.9	96
10	88.9	100	88.9	88.9	88.9	100	100	100	88.9	66.7	91
11	88.9	100	100	100	100	100	100	100	88.9	88.9	97
1	66.7	66.7	44.4	44.4	11.1	44.4	44.4	11.1	22.2	11.1	37
2	88.9	100	100	100	100	88.9	88.9	77.8	77.8	66.7	89
3	88.9	88.9	100	100	100	100	88.9	100	88.9	66.7	92
4	100	100	88.9	100	88.9	88.9	88.9	88.9	88.9	100	93
5	77.8	77.8	66.7	55.6	44.4	44.4	44.4	22.2	0	0	43
6	100	77.8	88.9	77.8	66.7	66.7	22.2	33.3	44.4	33.3	61
7	88.9	77.8	88.9	88.9	88.9	100	88.9	100	88.9	100	91
8	55.6	100	100	100	88.9	66.7	66.7	44.4	33.3	11.1	67
9	88.9	77.8	100	88.9	77.8	66.7	66.7	44.4	22.2	22.2	66
10	66.7	88.9	77.8	88.9	77.8	77.8	66.7	33.3	55.6	66.7	70
11	100	100	88.9	77.8	55.6	55.6	33.3	11.1	11.1	11.1	54
12	88.9	77.8	77.8	77.8	77.8	77.8	66.7	55.6	88.9	77.8	77

All table cells give accuracies as percent. Last column is average of columns indexed 0–9.

(2) Because the process/algorithm that assigns the codes to inputs (the code selection algorithm, CSA) enforces the *similar-inputs-to-similar-codes* (SISC) property, it follows that all stored inputs (hypotheses) are active with strength in descending order of similarity to (likelihood of) the hypothesis represented by ϕ.

Crucially, since the Q active (spiking) cells represent *all* stored hypotheses (with varying strengths), not just the single most likely hypothesis, ϕ, it follows that *all of these hypotheses physically influence the next time step's decision processes*. Specifically, any stored hypothesis whose code has even one cell in common with ϕ, will physically influence:

(a) the V distributions (and ultimately the ρ distributions) in all CMs of all downstream macs on the next time step, and thus
(b) the resulting likelihood distributions over all the stored hypotheses in each of the downstream macs on the next time step.

While this is true, it still makes sense to say that *all* stored hypotheses are physically influencing subsequent decisions; it's just that the hypotheses having zero intersection with ϕ are so different from ϕ that they are appropriately viewed as having zero likelihood and thus as having no causal influence on subsequent decisions.

We emphasize that the representation of a hypothesis's likelihood (or probability) in our model—i.e., as the *fraction of the its code (of Q cells) that is active*—differs fundamentally from existing representations in which single neurons encode such probabilities in their (typically real-valued) scalar strengths of activation (e.g., firing rates) as described in the recent review of Pouget et al. (2013).

Another way of understanding the advantage of SDC over localism is that an *individual* machine operation on a single unit (cell), and moreover, on a single synapse—e.g., the addition of a synapse's weight into the input summation of a postsynaptic cell—transmits information about *multiple* items (hypotheses) represented in the synapse's presynaptic cell's mac. In stark contrast, in a localist model in which the presynaptic cell represents only one hypothesis, adding the synapse's weight into the input summation of a postsynaptic cell necessarily transmits information only about that *one* hypothesis. We believe this aspect of SDC—which qualifies as an instance of what has been termed *algorithmic*, or *representational*, *parallelism*—to be at the core of the biological brain's remarkable efficiency at processing information.

We also described several other important computational principles/mechanisms used in Sparsey:

1. How a single SDC code active in a mac can simultaneously represent two or more equally likely hypotheses and how information entering that mac on subsequent time steps can pare down the set of equally likely hypotheses (Section Step 5: Compute the number of competing hypotheses that will be active in the mac once the final code for this frame is activated).
2. How an important type of invariance, nonlinear time invariance, can be computed via a "back-off" policy that does not increase the time complexity of recognition (inference) (Section CSA: Retrieval Mode). Essentially, on each frame, a mac computes a series of estimates of the match of the current temporal-context-dependent input (i.e., the current spatiotemporal *moment*) not just to the set of actual moments it experienced during learning (which constitute its explicit spatiotemporal basis), but to a much larger (encompassing) space of variants of the basis moments that were not actually experienced. This is similar in spirit to dynamic time warping (DTW) (Sakoe and Chiba, 1978), but is far more efficient, again because of the underlying algorithmic parallelism.
3. How Sparsey can learn arbitrarily nonlinear and intertwined, i.e., "tangled," classes via supervised learning of associations between codes in different macs (Section Learning arbitrarily complex nonlinear similarity metrics). That categories in the physical world are smooth in the neighborhood around any single exemplar but possibly very nonlinear and intertwined, i.e., "tangled," with other classes at the global scale has been pointed out by many, (e.g., Saul and Roweis, 2002; Bengio, 2007; Bengio et al., 2012). In particular, DiCarlo et al. (2012) state as a next step the need to formally specify what is meant by "untangling local" subspace. We believe that Sparsey addresses this need. First, the CSA's two functions of

storing (learning) and (best-match) retrieval of stored memories, can be viewed as a SISC-respecting content-addressable memory. Thus, individual macs handle the smooth category structure around individual exemplars: i.e., a novel input that is sufficiently similar to a known exemplar should activate an SDC with high intersection with the known exemplar's code and therefore exert similar downstream influence to that which would be produced by the familiar exemplar's code. The global highly nonlinear category structure is untangled by the hierarchy of macs, and specifically, by the ability (strongly subserved by progressive persistence) for multiple arbitrarily different codes in one cortical patch (e.g., one mac or set of macs) to be associated with a single code in another patch.

4. That, during learning, the CSA formally involves two rounds of competition amongst the mac's cells. In the first round, CSA Step 8, the Q cells with the maximal V-values in their respective CMs are determined and must activate (i.e., spike) so that their outputs can be summed and averaged to yield G. In the second round, CSA Step 12, a final winner is chosen in each CM according to the ρ distribution in that CM, i.e., soft max. In general, the second round winners may differ (perhaps substantially, especially when $G \approx 0$) from the first round winners. This hypothesis that the canonical cortical computation involves two rounds of competition is a strong and falsifiable prediction.
5. And, that the concept of feature basis present in Sparsey differs markedly from that present in localist models such as Olshausen and Field (1997), summarized in in **Table II-3**.

A great deal of work remains, particularly in understanding and mechanistically explaining the learning and usage (as in online rapid recognition/inference) of a hierarchy of spatiotemporal features. Even though Sparsey centers around a *single* canonical algorithm/circuit, the CSA [much of which was described (Rinkus, 1996)], the ultimate algorithmic solution of cortex lives in what DiCarlo et al. (2012) term a "very, very large space of details," which will take quite some time to explore, as suggested by Study II (Sections Study 2: Single-trial Learning of Sets of Longer Sequences), which itself only begins to scratch the surface of the myriad parameter interactions that we would like to understand.

ACKNOWLEDGMENT

This work was supported by research contracts under the U.S. Office of Naval Research's Computational Neuroscience Program and DARPA's UPSIDE Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fncom.2014.00160/abstract

REFERENCES

Barkat, T. R., Polley, D. B., and Hensch, T. K. (2011). A critical period for auditory thalamocortical activity. *Nat. Neurosci.* 14, 1189–1196. doi: 10.1038/nn.2882

Barrett, A. B., Billings, G. O., Morris, R. G. M., and van Rossum, M. C. W. (2009). State based model of long-term potentiation and synaptic tagging and capture. *PLoS Comput. Biol.* 5:e1000259. doi: 10.1371/journal.pcbi.1000259

Bart, E., and Hegdé, J. (2012). Invariant object recognition based on extended fragments. *Front. Comput. Neurosci.* 6:56. doi: 10.3389/fncom.2012.00056

Bengio, Y. (2007). "On the challenge of learning complex functions," in *Progress in Brain Research*, Vol. 165, eds P. Cisek, T. Drew, and J. F. Kalaska (Elsevier), 521.

Bengio, Y., Courville, A., and Vincent, P. (2012). *Representation Learning: A Review and New Perspectives*. Technical Report, University of Montreal, Montreal, QC.

Constantinople, C. M., and Bruno, R. M. (2013). Deep cortical layers are activated directly by Thalamus. *Science* 340, 1591–1594. doi: 10.1126/science.1236425

Damasio, A. R. (1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33, 25. doi: 10.1016/0010-0277(89)90005-X

Dean, T. (2006). "Scalable inference in hierarchical generative models," in *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*.

DeAngelis, G. C., Ghose, G. M., Ohzawa, I., and Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons. *J. Neurosci.* 19, 4046–4064.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *J. Neurophysiol.* 69, 1118–1135.

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4, 2051–2062.

Douglas, R. J., Martin, K. A., and Witteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Comput.* 1, 480–488. doi: 10.1162/neco.1989.1.4.480

Douglas, R. J., and Martin, K. A. C. (2004). Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* 27, 419–451. doi: 10.1146/annurev.neuro.27.070203.144152

Edelman, S., and Poggio, T. (1991). Models of object recognition. *Curr. Opin. Neurobiol.* 1, 270. doi: 10.1016/0959-4388(91)90089-P

Feldman, V., and Valiant, L. G. (2009). Experience-induced neural circuits that achieve high capacity. *Neural Comput.* 21, 2715–2754. doi: 10.1162/neco.2009.08-08-851

Frey, U., and Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature* 385, 533–536. doi: 10.1038/385533a0

Fukushima, K. (1984). A hierarchical neural network model for associative memory. *Biol. Cybern.* 50, 105–113. doi: 10.1007/BF00337157

Gauthier, B., Eger, E., Hesselmann, G., Giraud, A.-L., and Kleinschmidt, A. (2012). Temporal tuning properties along the human ventral visual stream. *J. Neurosci.* 32, 14433–14441. doi: 10.1523/JNEUROSCI.2467-12.2012

Gavornik, J. P., and Bear, M. F. (2014). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nat. Neurosci.* 17, 732–737. doi: 10.1038/nn.3683

George, D., and Hawkins, J. (2005). "A hierarchical Bayesian model of invariant pattern recognition in the visual cortex," in *Proceedings of the International Joint Conference on Neural Networks* (Montreal, QC: IEEE).

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Stud. Appl. Math.* 52, 213–257.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychol. Rev.* 87, 1–51. doi: 10.1037/0033-295X.87.1.1

Hecht-Nielsen, R. (2005). Cogent confabulation. *Neural Netw.* 18, 111–115. doi: 10.1016/j.neunet.2004.11.003

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434. doi: 10.1016/j.tics.2007.09.004

Hinton, G. E., Osindero, S., and The, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Jitsev, E. (2010). *On the Self-Organization of a Hierarchical Memory for Compositional Object Representation in the Visual Cortex.* Frankfurt am Main: Goethe University.

Jockel, S. (2009). *Crossmodal Learning and Prediction of Autobiographical Episodic Experiences Using a Sparse Distributed Memory*. Ph.D., University of Hamburg.

Kanerva, P. (1988). *Sparse Distributed Memory*. Cambridge, MA: MIT Press.

Kanerva, P. (1994). "The spatter code for encoding concepts at many levels," in *Proceedings of International Conference on Artificial Neural Networks* (Sorento: Springer-Verlag).

Kanerva, P. (2009). Hyperdimensional computing: an introduction to cmoputing in distributed representation with high-dimensional random vectors. *Cogn. Comput.* 1, 139–159. doi: 10.1007/s12559-009-9009-8

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97, 4296–4309. doi: 10.1152/jn.00024.2007

Knoblich, U., Bouvrie, J., and Poggio, T. (2007). "Biophysical models of neural computation: max and tuning circuits," in *CBCL Technical Report* (Boston, MA: MIT Press).

Kouh, M., and Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput.* 20, 1427–1451. doi: 10.1162/neco.2008.02-07-466

Kreiman, G., Hung, C., Kraskov, A., Quiroga, R. Q., Poggio, T., and Dicarlo, J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron* 49, 433–445. doi: 10.1016/j.neuron.2005.12.019

Le, Q. V., Zou, W., Yeung, S., and Ng, A. Y. I. (2011). "Learning hierarchical spatiotemporal features for action recognition with independent subspace analysis," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC: IEEE), 3361–3368.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (Cambridge, MA: MIT Press), 255–258.

Litvak, S., and Ullman, S. (2009). Cortical circuitry implementing graphical models. *Neural Comput.* 21, 3010–3056. doi: 10.1162/neco.2009.05-08-783

Lucke, J. (2004). Hierarchical self-organization of minicolumnar receptive fields. *Neural Netw.* 17, 1377. doi: 10.1016/j.neunet.2004.07.008

Moll, M., and Miikkulainen, R. (1997). Convergence-zone episodic memory: analysis and simulations. *Neural Netw.* 10, 1017–1036. doi: 10.1016/S0893-6080(97)00016-6

Moncada, D., and Viola, H. (2007). Induction of long-term memory by exposure to novelty requires protein synthesis: evidence for a behavioral tagging. *J. Neurosci.* 27, 7476–7481. doi: 10.1523/JNEUROSCI.1083-07.2007

Morris, R. G. M., and Frey, U. (1999). Tagging the Hebb synapse: reply. *Trends Neurosci.* 22, 256. doi: 10.1016/S0166-2236(99)01413-7

Murray, J. F., and Kreutz-Delgado, K. (2007). Visual recognition and inference using dynamic overcomplete sparse learning. *Neural Comput.* 19, 2301–2352. doi: 10.1162/neco.2007.19.9.2301

Nandy, A. S., Sharpee, T. O., Reynolds, J. H., and Mitchell, J. F. (2013). The fine structure of shape tuning in area V4. *Neuron* 78, 1102–1115. doi: 10.1016/j.neuron.2013.04.016

Ohki, K., Chung, S., Ch'Ng, Y. H., Kara, P., and Reid, R. C. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* 433, 597–603. doi: 10.1038/nature03274

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311. doi: 10.1016/S0042-6989(97)00169-7

Oster, M., Douglas, R., and Liu, S.-C. (2009). Computation with spikes in a winner-take-all network. *Neural Comput.* 21, 2437–2465. doi: 10.1162/neco.2009.07-08-829

Pandipati, S., and Schoppa, N. E. (2012). Age-dependent adrenergic actions in the main olfactory bulb that could underlie an olfactory-sensitive period. *J. Neurophysiol.* 108, 1999–2007. doi: 10.1152/jn.00322.2012

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Rachkovskij, D. A. (2001). Representation and processing of structures with binary sparse distributed codes. *IEEE Trans. Knowl. Data Eng.* 13, 261–276. doi: 10.1109/69.917565

Rachkovskij, D. A., and Kussul, E. M. (2001). Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Comput.* 13, 411–452. doi: 10.1162/089976601300014592

Ramirez, A., Pnevmatikakis, E. A., Merel, J., Paninski, L., Miller, K. D., and Bruno, R. M. (2014). Spatiotemporal receptive fields of barrel cortex revealed by reverse correlation of synaptic input. *Nat. Neurosci.* 17, 866–875. doi: 10.1038/nn.3720

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019. doi: 10.1038/14819

Rinkus, G. (1996). *A Combinatorial Neural Network Exhibiting Episodic and Semantic Memory Properties for Spatio-Temporal Patterns*. Ph.D., Boston University.

Rinkus, G. J. (2010). A cortical sparse distributed coding model linking mini- and macrocolumn-scale functionality. *Front. Neuroanat.* 4:17. doi: 10.3389/fnana.2010.00017

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans. Biol. Sci.* 335, 11–21. doi: 10.1098/rstb.1992.0002

Rolls, E. T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Front. Comput. Neurosci.* 6:35. doi: 10.3389/fncom.2012.00035

Rolls, E. T., and Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. Biol. Sci.* 257, 9–15. doi: 10.1098/rspb.1994.0087

Rubio-Garrido, P., Pérez-de-Manzo, F., Porrero, C., Galazo, M. J., and Clascá, F. (2009). Thalamic input to distal apical dendrites in neocortical layer 1 is massive and highly convergent. *Cereb. Cortex* 19, 2380–2395. doi: 10.1093/cercor/bhn259

Rust, N. C., and DiCarlo, J. J. (2010). Selectivity and tolerance ("Invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* 30, 12978–12995. doi: 10.1523/JNEUROSCI.0179-10.2010

Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron* 46, 945–956. doi: 10.1016/j.neuron.2005.05.021

Sadovsky, A. J., and MacLean, J. N. (2014). Mouse visual neocortex supports multiple stereotyped patterns of microcircuit activity. *J. Neurosci.* 34, 7769–7777. doi: 10.1523/JNEUROSCI.0169-14.2014

Sajikumar, S., and Frey, J. U. (2004). Resetting of 'synaptic tags' is time- and activity-dependent in rat hippocampal CA1 *in vitro*. *Neuroscience* 129, 503–507. doi: 10.1016/j.neuroscience.2004.08.014

Sakoe, H., and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* 26, 43–49. doi: 10.1109/TASSP.1978.1163055

Salakhutdinov, R., and Hinton, G. (2012). An efficient learning procedure for deep boltzmann machines. *Neural Comput.* 24, 1967–2006. doi: 10.1162/NECO_a_00311

Saul, L. K., and Roweis, S. (2002). *Think Globally, Fit Locally: Unsupervised Learning of Nonlinear Manifolds*. U. Penn.

Schuldt, C., Laptev, I., and Caputo, B. (2004). *Recognizing Human Actions: A Local SVM Approach*. Cambridge: ICPR.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). *A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex*. Cambridge, MA: AI Memo 2005-036, MIT.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007). "A quantitative theory of immediate visual recognition," in *Progress in Brain Research*, Vol. 165, eds P. Cisek, T. Drew, and J. F. Kalaska (Elsevier), 33–56.

St-Pierre, F., Marshall, J. D., Yang, Y., Gong, Y., Schnitzer, M. J., and Lin, M. Z. (2014). High-fidelity optical reporting of neuronal electrical activity with an ultrafast fluorescent voltage sensor. *Nat. Neurosci.* 17, 884–889. doi: 10.1038/nn.3709

Taylor, G. W., Fergus, R., LeCun, Y., and Bregler, C. (2010). "Convolutional learning of spatio-temporal features," in *European Conference on Computer Vision (ECCV'10)* (Crete).

Theunissen, F. E., and Elie, J. E. (2014). Neural processing of natural sounds. *Nat. Rev. Neurosci.* 15, 355–366. doi: 10.1038/nrn3731

Uusitalo, M. A., Jousmäki, V., and Hari, R. (1997). Activation trace lifetime of human cortical responses evoked by apparent visual motion. *Neurosci. Lett.* 224, 45–48. doi: 10.1016/S0304-3940(97)13445-0

Valiant, L. (2006). A quantitative theory of neural computation. *Biol. Cybern.* 953, 205–211. doi: 10.1007/s00422-006-0079-3

Wiesel, T. N., and Hubel, D. H. (1963). Effects of visual deprivation on morphology and physiology of cell in the cat's lateral geniculate body. *J. Neurophys.* 26, 978–993.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non holographic associative memory. *Nature* 222, 960–962. doi: 10.1038/222960a0

Yu, A. J., Giese, M. A., and Poggio, T. A. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Comput.* 14, 2857–2881. doi: 10.1162/089976602760805313

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 August 2014; accepted: 19 November 2014; published online: 15 December 2014.

Citation: Rinkus GJ (2014) Sparsey^{TM}: event recognition via deep hierarchical sparse distributed codes. Front. Comput. Neurosci. ***8****:160. doi: 10.3389/fncom.2014.00160*

This article was submitted to the journal Frontiers in Computational Neuroscience.

Copyright © 2014 Rinkus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 01 April 2014
doi: 10.3389/fncom.2014.00037

Deformation-specific and deformation-invariant visual object recognition: pose vs. identity recognition of people and deforming objects

Tristan J. Webb*[1] *and Edmund T. Rolls*[1,2]

[1] *Department of Computer Science, University of Warwick, Coventry, UK*
[2] *Oxford Centre for Computational Neuroscience, Oxford, UK*

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Sen Song, Tsinghua University, China
Guy Wallis, University of Queensland, Australia

***Correspondence:**
Edmund T. Rolls, Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK
e-mail: edmund.rolls@oxcns.org

When we see a human sitting down, standing up, or walking, we can recognize one of these poses independently of the individual, or we can recognize the individual person, independently of the pose. The same issues arise for deforming objects. For example, if we see a flag deformed by the wind, either blowing out or hanging languidly, we can usually recognize the flag, independently of its deformation; or we can recognize the deformation independently of the identity of the flag. We hypothesize that these types of recognition can be implemented by the primate visual system using temporo-spatial continuity as objects transform as a learning principle. In particular, we hypothesize that pose or deformation can be learned under conditions in which large numbers of different people are successively seen in the same pose, or objects in the same deformation. We also hypothesize that person-specific representations that are independent of pose, and object-specific representations that are independent of deformation and view, could be built, when individual people or objects are observed successively transforming from one pose or deformation and view to another. These hypotheses were tested in a simulation of the ventral visual system, VisNet, that uses temporal continuity, implemented in a synaptic learning rule with a short-term memory trace of previous neuronal activity, to learn invariant representations. It was found that depending on the statistics of the visual input, either pose-specific or deformation-specific representations could be built that were invariant with respect to individual and view; or that identity-specific representations could be built that were invariant with respect to pose or deformation and view. We propose that this is how pose-specific and pose-invariant, and deformation-specific and deformation-invariant, perceptual representations are built in the brain.

Keywords: VisNet, invariance, object recognition, deformation, pose, inferior temporal visual cortex, trace learning rule

1. INTRODUCTION

When we see a human sitting down, standing up, or walking, we can recognize one of these poses independently of the individual, or we can recognize the individual person, independently of the pose. How might this be achieved in the visual system? Might both types of encoding of visual stimuli be present simultaneously, in different cortical areas? What mechanisms in the visual cortex might be involved?

The same issues arise for deforming objects. If we see a flag deformed by the wind, either blowing out or hanging languidly, we can usually recognize the flag, independently of its deformation. Similarly, we can describe the deformation of an object, for example the flag blowing out or hanging loosely, independently of the identity (e.g., nationality) of the flag.

In general, dealing with deformation in images is difficult for object recognition systems. For example, one approach has used part-based representations to recognize human poses (Yang et al., 2010), but this is unlikely to work for many objects, such as a deforming flag, and relies on accurate recognition of every part, and processing of how the parts are related to each other (Rolls, 2008).

Here we formulate a hypothesis about how the primate including human visual system may be able to implement pose recognition independently with respect to identity; and identity independently of pose, and then test the hypotheses by simulations of a model of the ventral visual cortical pathways, VisNet (Wallis and Rolls, 1997; Rolls and Milward, 2000; Rolls, 2008, 2012).

The hypothesis is that these types of recognition can be implemented by the primate visual system using the temporo-spatial continuity that we hypothesize enables transform invariant representations of objects to be learned. In particular, one hypothesis is that pose identification could be learned under conditions in which large numbers of different people are seen in the same pose, for example sitting down. As different individuals in a sitting crowd are successively fixated and used as input to the ventral visual system, the temporal continuity will be for the pose and not for the individual person, allowing pose-specific representations

to be built that are independent (invariant with respect to) person identity. On another occasion, most of the people successively viewed might be standing up, for example waiting in a bus queue. On another occasion, all the individuals successively fixated might be walking to work. The second hypothesis is that person-specific representations that are independent of pose could be built, in another part of the ventral cortical visual system, when we watch one individual change posture, for example sitting down, then standing up, and then walking. The representation of the identity of another person that is invariant with respect to pose and view could be built using the temporal continuity inherent is seeing another particular person transform through a set of poses and views, etc.

These hypotheses were tested in a simulation of the ventral visual system, VisNet, that uses temporal continuity, implemented in a synaptic learning rule with a short-term memory trace of previous neuronal activity, to learn invariant representations (Rolls, 2012).

2. METHODS

2.1. EXPERIMENTAL DESIGN

The stimuli for the human pose experiment consisted of three individuals (man, woman, and soldier), shown in each of three different poses (standing, sitting, and walking). Each image was shown in 12 different rotational views each 30° apart. To train for pose identification, during training all 36 images had the same pose in succession but with the 36 images otherwise presented in random permuted sequence. One training epoch consisted of showing successively all people and views of one pose, then all people and views of another pose, and then all identities and views of the third pose. This enabled us to test whether VisNet under these circumstances would allocate some neurons to one pose independently of individual and view, other neurons invariantly to the second pose, and other neurons invariantly to the third pose.

To train for recognition of each individual, a training epoch consisted of showing all poses and all views of one individual in a random sequence, then all poses and all views of the second individual in a random sequence, and then all poses and all views of the third individual in a random sequence. This enabled us to test whether VisNet under these circumstances would allocate some neurons to one individual person independently of pose, other neurons to the second individual independently of pose, and other neurons to the third individual. It may be emphasized that the images shown in each of these experiments were identical, and only the order in which they were presented differed.

After training, the trained networks were then tested to determine whether the poses could be identified independently of the person and view transforms; or whether the individual people could be identified independently of the pose and view transforms.

For the flag deformation experiment, there were flags of four individual countries (Holland, Spain, UK, and USA) each shown with five different deformations produced by equally spaced wind values, with each condition shown in two views, from one side, and from the other side. To train to identify the country of the flag, all the deformations and views of the flag of one country were shown in random sequence, then all the transforms of the flag of the second country, etc. To train to identify the deformation (how much the flag drooped because of different wind strengths), one deformation was trained with all images of that deformation, then all images of the second deformation, etc.

After training, the trained networks were then tested to determine whether the particular deformations of the flags produced by each wind speed could be identified independently of the country and view transforms of the flags; or whether the individual countries of each flag image could be identified independently of the deformation produced by the different wind speeds and views.

2.2. STIMULUS CREATION

The images of humans used for training were rendered using Blender software (www.blender.org) to ensure uniform lighting conditions. The models used for rendering were generated from the MakeHuman software (www.makehuman.org). Each model was posed in three variations (standing, sitting, and walking) inside Blender. The camera position in Blender was rotated around each model in 30° increments to produce 12 views of each model in each pose, as illustrated in **Figure 1**. After rendering, each image was converted and scaled to an 8-bit (range: 0–255) grayscale representation, and the pixel intensities were controlled so that the mean value of each model in the front facing standing position was 127. Rendered images were placed on uniform 127 grayscale backgrounds.

The images of flags for different countries (Holland, Spain, UK, and USA) were also created in Blender using its cloth simulation. A force field was placed laterally from the position of the flag to give it a fluttering motion from wind. The wind force was set to five different equally spaced values in the range 0–200 Blender units, chosen so to give a wind effect varying from no wind

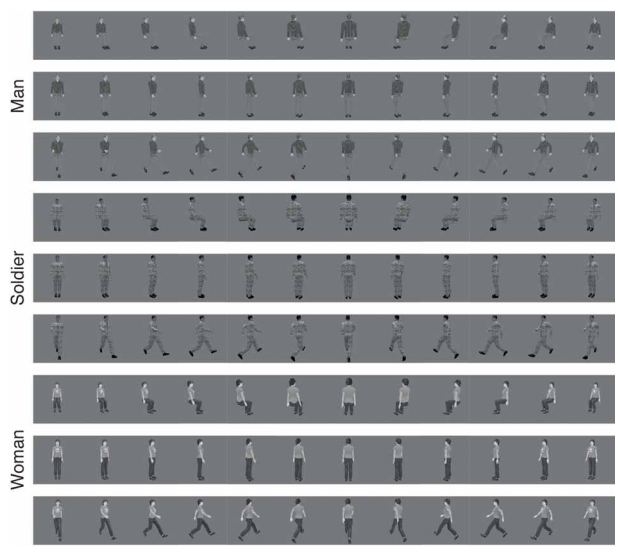


FIGURE 1 | Different views of human stimuli used to train the VisNet model. Each row shows each stimuli in one of three different poses (sitting, standing, and walking) varying across view rotation. The 12 rotations are shown, starting with 0° on the left and proceeding in 30° increments.

to strong wind. Images were rendered with the camera looking straight on to the flag and on the opposite side, as illustrated in **Figure 2**. Rendered images were placed on uniform 127 grayscale backgrounds.

2.3. TRAINING

Training images were presented at the center of the VisNet retina in one of two modes, object or deformation recognition mode. These modes were made distinct so that we could measure either how well the VisNet architecture performs in recognizing stimulus identity (i.e., which person it was) invariantly with respect to deformation and view; and deformation (i.e., which pose it was) invariantly with respect to stimulus identity and view.

In object recognition mode each of the images was grouped depending on the model (man, woman, and soldier for the human objects; or country for the flag objects). Each of the image groups then had each model shown in the 3 different deformations, with 12 rotational views of each deformation. During each epoch of training, using the trace synaptic learning rule, a randomly ordered permutation of the set of all images corresponding to different deformations and views was presented to VisNet. After each group of deformations and views was presented for a single model, the trace values reflecting for each neuron its recent firing rate was reset to 0 before moving on to the next model. (Trace reset speeds learning in VisNet, but is not essential for its operation Rolls and Milward, 2000; Rolls, 2012).

In deformation learning mode the images were grouped based on the different deformations (sitting, standing, and walking poses as the groups for the human objects; or wind speed deformation for the flag objects). For the pose learning of people, each training group consisted of the images of the 3 people in 12 different rotations in the same deformation. Trace learning operated in a similar fashion as above with the trace being reset after each set of a particular pose or deformation.

Simulations were run using 50 training epochs, which was sufficient to enable convergence of the synaptic weights.

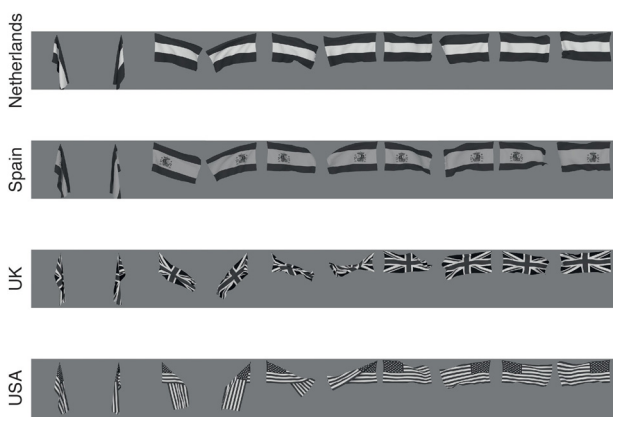


FIGURE 2 | The flag stimuli used to train VisNet. Each flag is shown with different wind forces and rotations. Starting on the left the first pair, both the 0° and 180° views are shown for windspeed 0, and each successive pair is shown for wind force increased by 50 blender units.

2.4. OVERVIEW OF THE VisNet ARCHITECTURE

Fundamental elements of Rolls' 1992 theory for how cortical networks might implement invariant object recognition are described in detail elsewhere (Rolls, 2008, 2012). They provide the basis for the design of VisNet, which is described in the Appendix, and can be summarized as:

- A series of competitive networks, organized in hierarchical layers, exhibiting mutual inhibition over a short range within each layer. These networks allow combinations of features or inputs occurring in a given spatial arrangement to be learned by neurons using competitive learning (Rolls, 2008), ensuring that higher order spatial properties of the input stimuli are represented in the network. In VisNet, layer 1 corresponds to V2, layer 2 to V4, layer 3 to posterior inferior temporal visual cortex, and layer 4 to anterior inferior temporal cortex. Layer one is preceded by a simulation of the Gabor-like receptive fields of V1 neurons produced by each image presented to VisNet (Rolls, 2012).
- A convergent series of connections from a localized population of neurons in the preceding layer to each neuron of the following layer, thus allowing the receptive field size of neurons to increase through the visual processing areas or layers, as illustrated in **Figure 3**.
- A modified associative (Hebb-like) learning rule incorporating a temporal trace of each neuron's previous activity, which, it is suggested (Földiák, 1991; Rolls, 1992, 2012; Wallis et al., 1993; Wallis and Rolls, 1997; Rolls and Milward, 2000), will enable the neurons to learn transform invariances.

2.5. INFORMATION MEASURES OF PERFORMANCE

The performance of VisNet was measured by Shannon information-theoretic measures that are essentially identical to those used to quantify the specificity and selectiveness of the representations provided by neurons in the brain (Rolls and Milward, 2000; Rolls and Treves, 2011; Rolls, 2012). A single cell information measure indicated how much information was conveyed by a single neuron about the most effective stimulus. A multiple cell information measure indicated how much information about every stimulus was conveyed by small populations of neurons, and was used to ensure that all stimuli had some neurons conveying information about them. In the pose or deformation recognition experiments, each stimulus was defined as a particular pose or deformation with all of its identity and view transforms. In the person or object recognition experiments, each stimulus was defined as a particular person or flag with all of its pose or deformation and view transforms. Details are provided in the Appendix.

3. RESULTS

3.1. HUMANS

3.1.1. Recognition of individuals independently of pose

Figure 4 shows the information measured from a network trained in object recognition mode (in this case, recognition of the individual person) using three human individuals in three different poses (deformations). There were 12 views of each individual in each of the three poses or deformations. **Figure 4A** shows how a

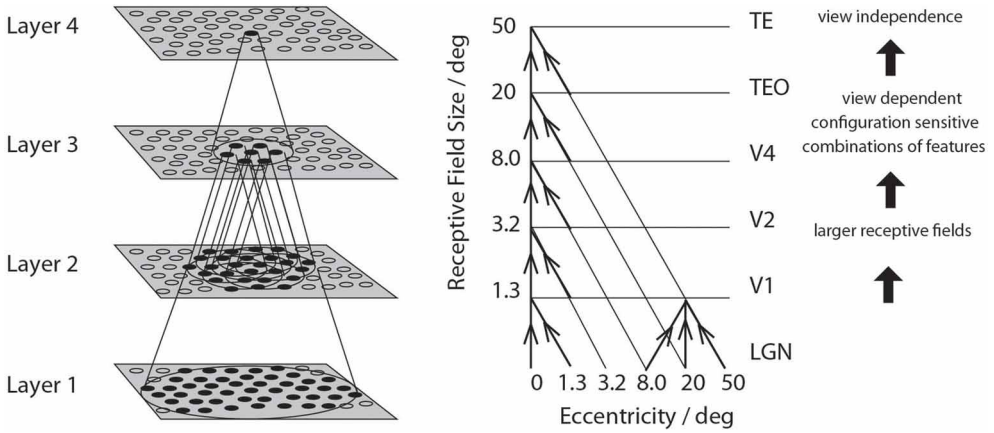


FIGURE 3 | Convergence in the visual system. Right—as it occurs in the brain. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex (IT). Left—as implemented in VisNet. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina.

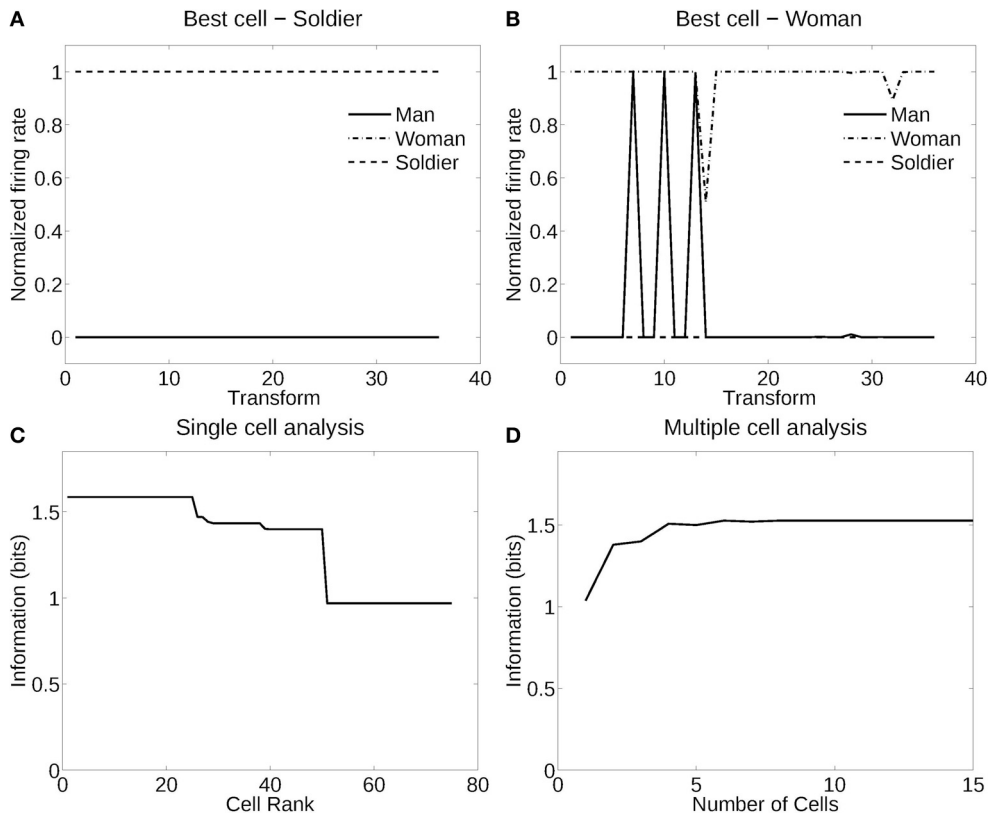


FIGURE 4 | Information analysis of the network trained to recognize human stimuli. (A) Firing rate response of the best single cell responding to an individual, the soldier, independently of poses and views. The transforms vary fastest over views. Thus transforms 1–12 are all views of pose 1, followed by all views of pose 2, etc. **(B)** Firing rate response of another single cell responding to an individual, the woman, across most poses and views, and not responding to most poses and views of the two other individuals. **(C)** A sorted ranking of the information for the set of 25 single cells with the highest information for each stimulus. **(D)** The multiple cell information of the network using the set of five best cells for each stimuli.

typical well trained neuron, as measured by the single cell information analysis, responded to one individual in all the different poses (deformations) at different views. The neuron responded to all views of all poses of the Soldier, and to no images of the other two individuals. The single cell information was 1.59 bits, which indicates perfect selectivity with responses to all transforms of one individual, and no responses to any other individual. (1.59 bits is $\log_2$ of the number of stimuli, in this case the three different

people). **Figure 4B** shows another neuron that responded to most views of the Woman, but to some views of the Man. The single cell information for this neuron was 1.5 bits. The single cell information for the 75 most selective cells was high, as shown in **Figure 4C**. The multiple cell information was measured at 1.55 bits (as shown in **Figure 4D**), and corresponded to 96% correct. VisNet had thus learned to recognize the individual people independently of their pose and view transforms when trained for identity. The trace rule was important in achieving this result, for when training was with a purely associative (Hebbian) learning rule (Rolls, 2012), the multiple cell information was measured at 0.42 bits and corresponded to 52% correct.

3.1.2. Recognition of pose independently of individual

Figure 5 shows the performance of VisNet when trained in deformation recognition mode to identify the pose independently of the individual person (object) and its view. **Figure 5A** shows how a typical well-trained neuron, as measured by the single cell information analysis, which responded to almost all views and all individuals in one pose (sitting). The single cell information was 1.59 bits. **Figure 5B** shows how another neuron responded to the majority of views and individuals in another pose (standing). The single cell information was 1.5 bits. The single cell information for the 75 most selective cells was high, as shown in **Figure 5C**. The multiple cell information was measured at 1.55 bits (as shown in **Figure 5D**), and corresponded to 96% correct. VisNet had thus learned to recognize the pose independently of the identity of the person or the view when trained for pose. The trace rule was important in achieving this result, for when training was with a purely associative (Hebbian) learning rule (Rolls, 2012), the multiple cell information was measured at 0.41 bits and corresponded to 56% correct.

3.2. FLAG OBJECTS

3.2.1. Recognition of flag country independently of deformation (windspeed)

Figure 6 shows the information measured from a network trained in object recognition mode to recognize four different flags independently of five deformations and two views. **Figure 6A** shows how a typical well-trained neuron, as measured by the single cell information analysis, responded to one flag (USA) in all the different deformations in the different views, and to none of the other flags. The single cell information was 2.0 bits (i.e., $\log_2$ of the number of flag countries). The single cell information for the 100 most selective cells was 2.0 bits (perfect discrimination), as shown in **Figure 6B**. The multiple cell information was measured

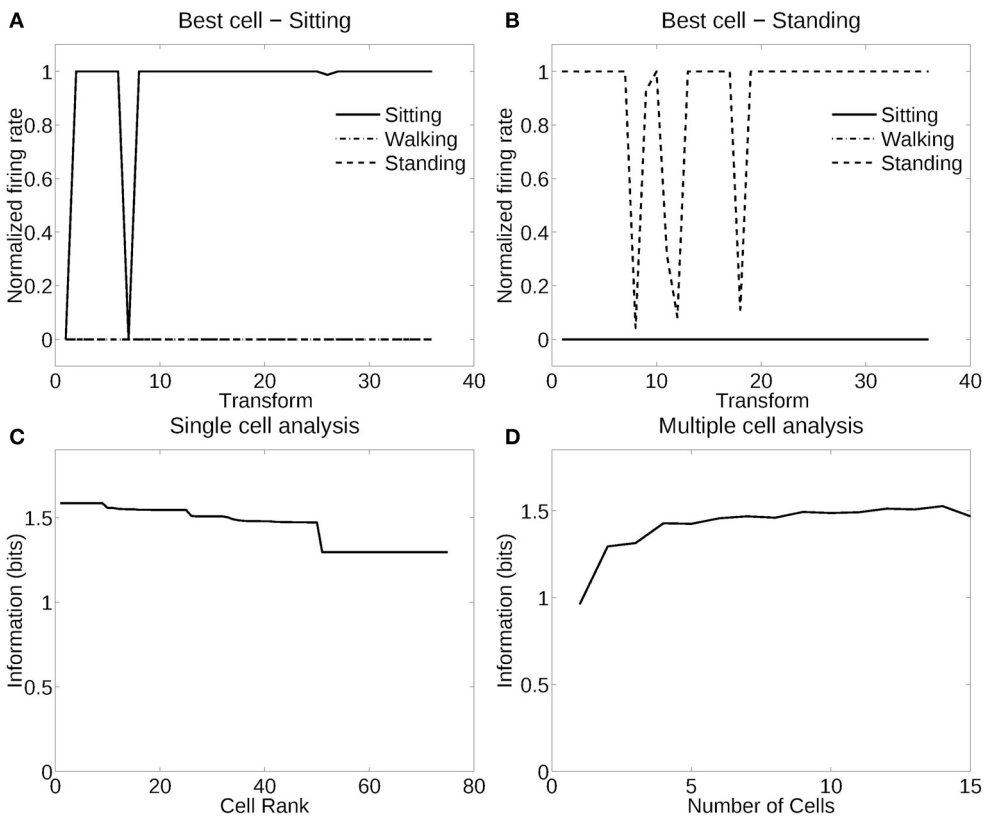


FIGURE 5 | Information analysis of the network trained to recognize human poses invariantly with respect to individual and view. (A) Firing rate response of a single cell with responses to the pose of sitting almost invariantly with respect to the 3 individuals and 12 views. Transforms vary fastest over views. **(B)** Firing rate response of a single cell with responses to the pose of standing almost invariantly with respect to the 3 individuals and 12 views. **(C)** A sorted ranking of the information for the set of 25 single cells with the highest information for each pose. **(D)** The multiple cell information of the network using the set of five best cells for each pose.

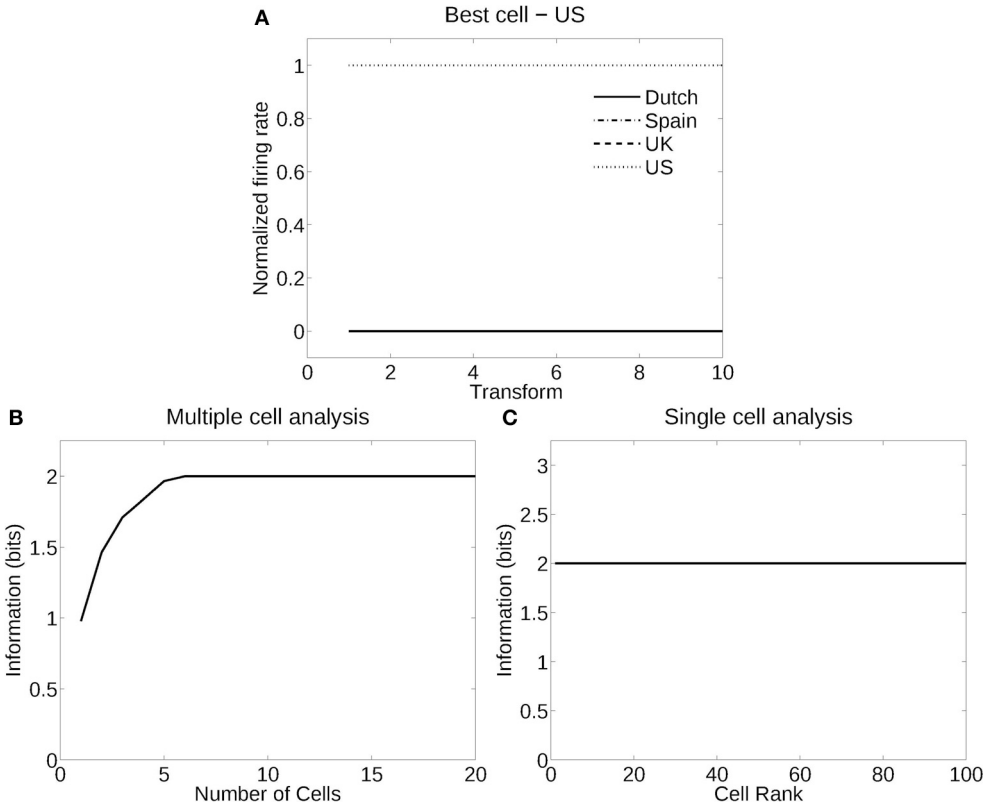


FIGURE 6 | Information analysis of the network trained to recognize flag countries invariantly with respect to deformation produced by different windspeeds and views. (A) Firing rate response of a single cell with responses to the USA flag invariantly with respect to the five windspeed deformations and two views. Transforms vary fastest over views. **(B)** A sorted ranking of the information for the set of 25 single cells with the highest information for the flag of each country. **(C)** The multiple cell information of the network using the set of five best cells for each pose.

at 2.0 bits (as shown in **Figure 6C**), and corresponded to 100% correct. VisNet had thus learned to recognize the individual flags for each country independently of their deformation and view transforms when trained for identity.

3.2.2. Recognition of windspeed (deformation) independently of flag country

Figure 7 shows the analysis for a network trained in deformation recognition mode to recognize five deformations each produced by a different windspeed, but independently of flag country and view. **Figure 7A** shows how a typical well trained neuron, as measured by the single cell information analysis, responded to one deformation (windspeed parameter 150) in the flags of all four countries and two views, and almost not at all to any other deformation across all countries and views. The single cell information was 2.32 bits (i.e., $\log_2$ of the number of deformation types). The single cell information for many of the 125 most selective cells was 2.32 bits (perfect discrimination), as shown in **Figure 7B**. The multiple cell information was measured at 2.32 bits (as shown in **Figure 7C**), and corresponded to 100% correct. VisNet had thus learned to recognize the deformation independently of the identity of the flag or the view when trained for deformation. In this case, VisNet had learned to recognize effectively the wind speed by the deformation it produced, independently of the country and view of each flag.

3.3. FLAG CAPACITY

The deformation invariant recognition of flags described above was obtained with a set of four flags (each with five deformations each with two views, as illustrated in **Figure 2**). On that task, performance was 100% correct. We tested how well VisNet would perform when the number of different flags in the set on which VisNet was trained and tested was increased. To perform this investigation, 24 more flags were constructed (of the NATO countries, and the NATO flag), each with the same set of deformations and views illustrated in **Figure 2**. Four of this further set of flags are illustrated in **Figure 8**. For training and testing with a given number of flags, random subsets of the flags and 60 training epochs were used. As shown in **Figure 9**, it was found that performance remained close to 100% correct for up to eight flags. The performance with higher numbers of flags was as follows: 10 flags = 92%; 15 flags = 86%; 20 flags = 79%.

3.4. POSE GENERALIZATION TO NEW HUMAN STIMULI

We tested the ability of VisNet to identify human poses invariantly with respect to person and with respect to view using stimuli

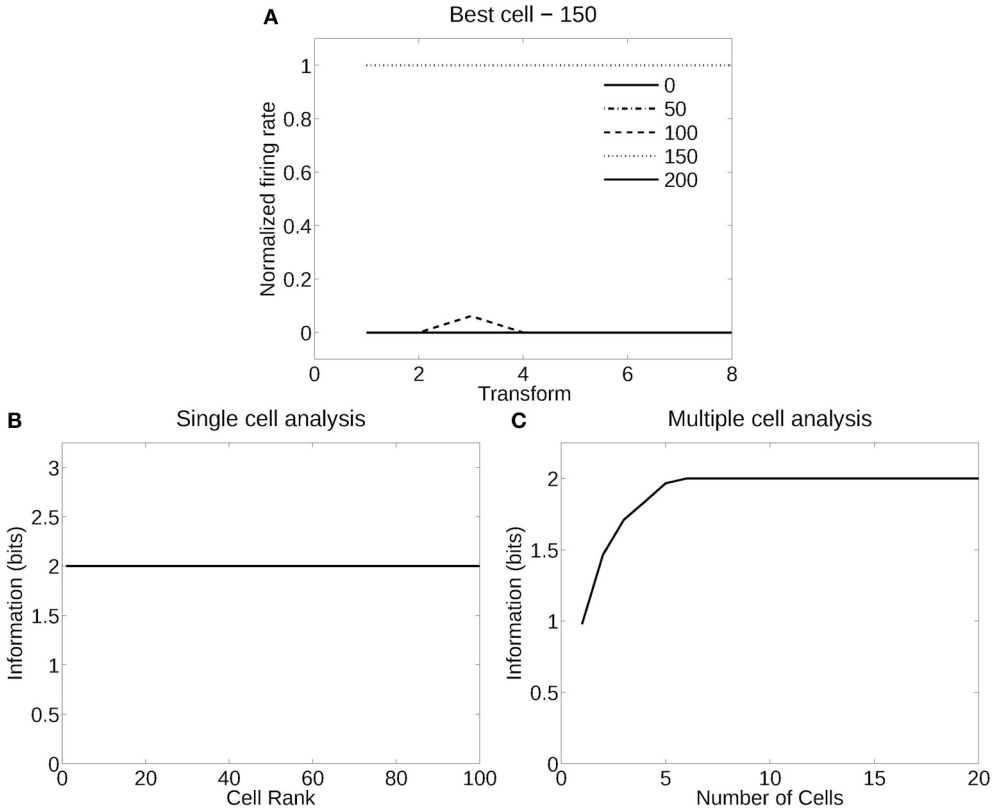


FIGURE 7 | Information analysis of the network trained to recognize flag deformation invariantly with respect to flag country (four) and views (two). (A) Firing rate response of a single cell with responses to the windspeed with deformation parameter 150 invariantly with respect to the four country flags and two views. Transforms vary fastest over views. **(B)** A sorted ranking of the information for the set of 25 single cells with the highest information for each of the five deformations (windspeeds). **(C)** The multiple cell information of the network using the set of five best cells for each deformation (windspeed).

it had not been trained with. This was thus a cross-validation assessment of pose identification. To perform the cross-validation training and testing, three more human characters were created using the same methods as described in section 2.2, so that we could perform cross-validation training and testing on the network using six different individuals. The network was set up in deformation recognition mode as described above, that is, one of the poses formed a group the images of which were presented in a permuted sequence so that the trace rule could learn about a single pose. The group of images contained all 5 training individuals in all 12 views, and these images were permuted. After one pose group had been trained within an epoch, each of the other two pose groups was trained, to complete an epoch. The three poses were, as before, sitting, standing, and walking. Trace learning operated in a similar fashion as above with the trace being reset after every group. The network was then tested with all of the views and poses of the remaining individual person, and the output of layer 4 of the network was classified using a pattern associator that had been trained with the five training poses, see section A.1.5 . The 15 single cells comprised of the 5 cells with the highest single cell information for each of the three poses were used as the input for training the pattern associator, which was then tested using the firing of the same 15 cells to the poses and views of the sixth, untrained, individual, to test how well the pose of that untrained individual was identified. The cross-validation training was perform in this leave-one-out protocol, training with five objects and testing with one.

In this cross-validation investigation, VisNet was able to correctly classify a pose with 76% accuracy, where chance was 33% accuracy. These results were found to be highly significantly different from chance with $p < 10^{-37}$ using a standard binomial test. The correct classification rate for the pose of different individuals was between 30% and 92%, with a standard deviation of 26%.

In a control comparison, the performance on the same task using an untrained network was 19% correct. Thus the good performance indicating pose recognition invariant with respect to the individual and view described above was only obtained when VisNet was trained to perform the pose-recognition task.

4. DISCUSSION

The new hypothesis about how pose is learned is that spatio-temporal continuity in the synaptic training rule in a network architecture designed to incorporate many of the properties of the hierarchy of ventral visual cortical areas can allow neurons specific to a pose and invariant with respect to individual and view to be learned, when there is continuity during training in

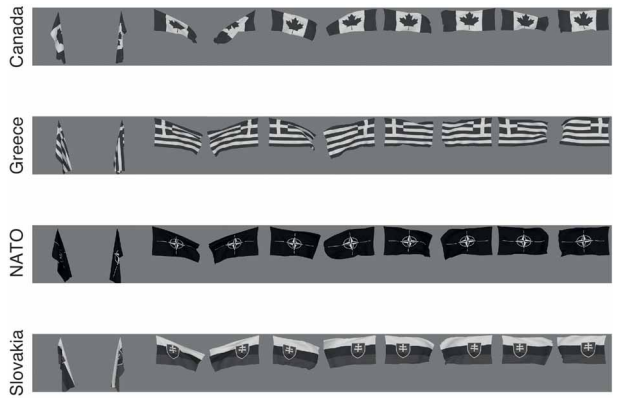


FIGURE 8 | Four more of the set of 24 more flag stimuli used to train VisNet to test how many flags could be recognized independently of deformation (see text). Each flag is shown with different wind forces and rotations. Starting on the left the first pair, both the 0° and 180° views are shown for windspeed 0, and each successive pair is shown for wind force increased by 50 blender units.

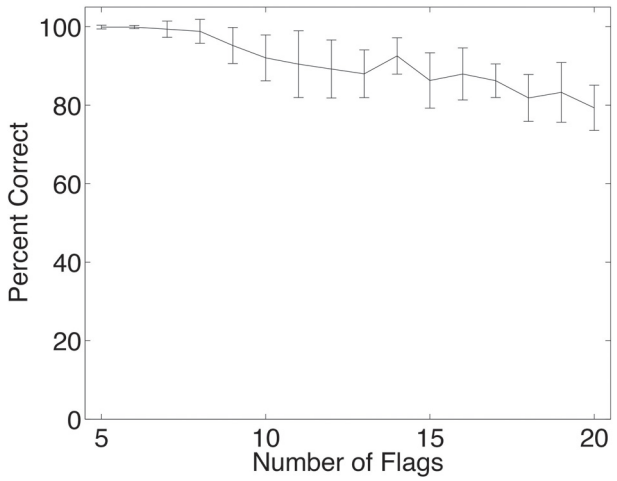


FIGURE 9 | Performance of the network when trained and tested in deformation invariant object recognition for different numbers of objects, in this case flags each in five deformations and two views. The set of neurons used were the five cells with the highest single cell information for each country. The mean and standard deviation of the percent correct are shown, taken from 20 trials, with each trial using a different random subset of the 28 NATO flags. In all trials, the network was trained with 60 epochs in each layer.

pose. This hypothesis was confirmed by the simulation results. A similar hypothesis about how deformation-specific recognition of objects invariantly with respect to the identity and view of the object could be learned using temporal continuity was also confirmed by the simulation results.

The new hypothesis about how person identity can be learned is that spatio-temporal continuity in the synaptic training rule in the same network architecture can allow neurons specific to an individual person and invariant with respect to pose and view to be learned, when there is continuity during training in the individual person being seen. This hypothesis was confirmed by the simulation results. A similar hypothesis about how individual recognition of specific objects invariantly with respect to the deformation and view of the object can be learned using temporal continuity was also confirmed by the simulation results.

In addition, it was found that the capacity of the system allowed for more objects to be recognized independently of deformation. In addition, we found that the functional architecture of VisNet allowed pose recognition to occur for untrained individual people in a cross-validation experiment, showing domain generality of pose recognition across people.

This research provides a mechanism for leaning both pose-specific and pose invariant representations in the visual cortical areas. Some evidence for pose-specific representations are the face expression selective neurons in the cortex in the anterior part of the superior temporal sulcus, which can respond to a particular face expression, independently of the individual person (Hasselmo et al., 1989a). Some evidence for individual-specific representations are the individual-selective neurons in the cortex in the gyrus of the inferior temporal visual cortex, which can respond to a particular individual, independently of the face expression (Hasselmo et al., 1989a). Further evidence for pose-specific neurons is that some neurons in the temporal visual cortical areas respond to face view (e.g., the right profile) relatively independently of the individual person (Perrett et al., 1985; Hasselmo et al., 1989b); and that other neurons respond for example to people walking (Barraclough et al., 2006).

The learning described here is made possible by use of a learning rule with a trace of previous neuronal activity, allowing neurons to learn from the temporal statistics of objects in the natural world as they transform continuously in time. We developed this hypothesis (Földiák, 1991; Rolls, 1992, 1995, 2012; Wallis et al., 1993) into a model of the ventral visual system that can account for translation, size, view, lighting, and rotation invariance (Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000, 2002, 2008; Rolls and Stringer, 2001, 2006, 2007; Elliffe et al., 2002; Perry et al., 2006, 2010; Stringer et al., 2006, 2007; Rolls, 2008, 2012). Consistent with the hypothesis, we have demonstrated these types of invariance (and spatial frequency invariance) in the responses of neurons in the macaque inferior temporal visual cortex (Rolls et al., 1985, 1987, 2003; Rolls and Baylis, 1986; Hasselmo et al., 1989b; Tovee et al., 1994; Booth and Rolls, 1998). Moreover, we have tested the hypothesis by placing small 3D objects in the macaque's home environment, and showing that in the absence of any specific rewards being delivered, this type of visual experience in which objects can be seen from different views as they transform continuously in time to reveal different views leads to single neurons in the inferior temporal visual cortex that respond to individual objects from any one of several different views, demonstrating the development of view-invariance learning (Booth and Rolls, 1998). (In control experiments, view invariant representations were not found for objects that had not been viewed in this way). The learning shown by neurons in the inferior temporal visual cortex can take just a small number of trials (Rolls et al., 1989). The finding that temporal contiguity in the absence of reward is sufficient to lead to view invariant object representations in the inferior temporal visual

cortex has been confirmed (Li and DiCarlo, 2008, 2010, 2012). The importance of temporal continuity in learning invariant representations has also been demonstrated in human psychophysics experiments (Perry et al., 2006; Wallis, 2013). Some other simulation models are also adopting the use of temporal continuity as a guiding principle for developing invariant representations by learning (Wiskott and Sejnowski, 2002; Wiskott, 2003; Wyss et al., 2006; Franzius et al., 2007), and the temporal trace learning principle has also been applied recently (Isik et al., 2012) to HMAX (Riesenhuber and Poggio, 2000; Serre et al., 2007), which nevertheless does not produce representations similar to those found in the inferior temporal visual cortex (Rolls, 2012).

The findings described in this paper demonstrate a mechanism by which neurons that respond to pose independently of individual person identity could be formed, and also how neurons that respond to identity independently of pose could be formed. The natural world conditions that could provide the appropriate conditions for these two types of representation to be formed include the following. To learn pose independently of identity the natural world might consist of large numbers of individuals all in the same pose, for example all standing up (perhaps in a queue), or all sitting down (for example in a theatre or stadium). As the eyes moved over scenes of this type, the natural environment would provide the conditions of temporal continuity for pose to be learned independently of identity. To learn identity independently of pose, appropriate environmental conditions might include looking at a single person while that person alters pose, from perhaps lying down, then sitting, and then standing up. This leads to the interesting prediction that neurons that encode pose independently of identity might be more likely to be close to parts of the temporal lobe visual cortex where the representations are of large-scale, such as scenes; whereas neurons sensitive to identity independently of pose might be more likely to be found close to cortical areas where single objects are represented, such as faces. In any case, self-organizing topological maps would be likely to be formed so that these two types of representation would be somewhat separated into different cortical regions or neuronal clusters (Rolls, 2008). Further segregation might occur because some poses such as walking are associated with movement, and thus representations of such poses might be close to the temporal lobe visual cotical areas with movement-related neurons (Baylis et al., 1987; Hasselmo et al., 1989b; Barraclough et al., 2006).

ACKNOWLEDGMENTS

We acknowledge the use of Blender software (http://www.blender.org) to render the 3D objects, MakeHuman software (http://www.makehuman.org) to create human character models, and the Blend Swap Open Source 3D model repository (http://www.blendswap.com) for some other of the models used. Flag textures were downloaded from the Wikimedia commons (http://www.wikimedia.org).

REFERENCES

Abbott, L. F., Rolls, E. T., and Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cereb. Cortex* 6, 498–505. doi: 10.1093/cercor/6.3.498

Barraclough, N. E., Xiao, D., Oram, M. W., and Perrett, D. I. (2006). The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images. *Prog. Brain Res.* 154, 135–148. doi: 10.1016/S0079-6123(06)54007-5

Baylis, G. C., Rolls, E. T., and Leonard, C. M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res.* 342, 91–102. doi: 10.1016/0006-8993(85)91356-3

Baylis, G. C., Rolls, E. T., and Leonard, C. M. (1987). Functional subdivisions of temporal lobe neocortex. *J. Neurosci.* 7, 330–342.

Booth, M. C. A., and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523. doi: 10.1093/cercor/8.6.510

Buhmann, J., Lange, J., von der Malsburg, C., Vorbrüggen, J. C., and Würtz, R. P. (1991). "Object recognition in the dynamic link architecture: parallel implementation of a transputer network," in *Neural Networks for Signal Processing*, ed B. Kosko (Englewood Cliffs, NJ: Prentice Hall), 121–159.

Daugman, J. (1988). Complete discrete 2D-Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust. Speech Signal Process.* 36, 1169–1179. doi: 10.1109/29.1644

Deco, G., and Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vis. Res.* 44, 621–644. doi: 10.1016/j.visres.2003.09.037

De Valois, R. L., and De Valois, K. K. (1988). *Spatial Vision*. New York, NY: Oxford University Press.

DeWeese, M. R., and Meister, M. (1999). How to measure the information gained from one symbol. *Network* 10, 325–340. doi: 10.1088/0954-898X/10/4/303

Elliffe, M. C. M., Rolls, E. T., and Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biol. Cybernet.* 86, 59–71. doi: 10.1007/s004220100284

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 193–199. doi: 10.1162/neco.1991.3.2.194

Földiák, P. (1992). Models of sensory coding. *Technical Report CUED/F–INFENG/TR 91*, Department of Engineering, University of Cambridge, Cambridge.

Franco, L., Rolls, E. T., Aggelopoulos, N. C., and Jerez, J. M. (2007). Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biol. Cybernet.* 96, 547–560. doi: 10.1007/s00422-007-0149-1

Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* 3:e166. doi: 10.1371/journal.pcbi.0030166

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36, 193–202. doi: 10.1007/BF00344251

Garthwaite, J. (2008). Concepts of neural nitric oxide-mediated transmission. *Eur. J. Neurosci.* 27, 2783–3802. doi: 10.1111/j.1460-9568.2008.06285.x

Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989a). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.* 32, 203–218. doi: 10.1016/S0166-4328(89)80054-3

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., and Nalwa, V. (1989b). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Exp. Brain Res.* 75, 417–429. doi: 10.1007/BF00247948

Hawken, M. J., and Parker, A. J. (1987). Spatial properties of the monkey striate cortex. *Proc. R. Soc. Lond. B* 231, 251–288. doi: 10.1098/rspb.1987.0044

Hestrin, S., Sah, P., and Nicoll, R. (1990). Mechanisms generating the time course of dual component excitatory synaptic currents recorded in hippocampal slices. *Neuron* 5, 247–253. doi: 10.1016/0896-6273(90)90162-9

Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517. doi: 10.1037/0033-295X.99.3.480

Isik, L., Leibo, J. Z., and Poggio, T. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. *Front. Comput. Neurosci.* 6:37. doi: 10.3389/fncom.2012.00037

Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.* 18, 959–971. doi: 10.1109/34.541406

Li, N., and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321, 1502–1507. doi: 10.1126/science.1160028

Li, N., and DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67, 1062–1075. doi: 10.1016/j.neuron.2010.08.029

Li, N., and DiCarlo, J. J. (2012). Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *J. Neurosci.* 32, 6611–6620. doi: 10.1523/JNEUROSCI.3786-11.2012

Malsburg, C. V. D. (1973). Self-organization of orientation-sensitive columns in the striate cortex. *Kybernetik* 14, 85–100. doi: 10.1007/BF00288907

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335, 817–820. doi: 10.1007/BF00288907

Montague, P. R., Gally, J. A., and Edelman, G. M. (1991). Spatial signalling in the development and function of neural connections. *Cereb. Cortex* 1, 199–220. doi: 10.1093/cercor/1.3.199

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol.* 15, 267–273. doi: 10.1007/BF00275687

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., and Benson, P. J. (1991). Viewer–centered and object centered coding of heads in the macaque temporal cortex. *Exp. Brain Res.* 86, 159–173. doi: 10.1007/BF00231050

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, D., et al. (1985). Visual cells in temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. Lond. Ser. B* 223, 293–317. doi: 10.1098/rspb.1985.0003

Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vis. Res.* 46, 3994–4006. doi: 10.1016/j.visres.2006.07.025

Perry, G., Rolls, E. T., and Stringer, S. M. (2010). Continuous transformation learning of translation invariant representations. *Exp. Brain Res.* 204, 255–270. doi: 10.1007/s00221-010-2309-0

Pollen, D., and Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science* 212, 1409–1411. doi: 10.1126/science.7233231

Rhodes, P. (1992). The open time of the NMDA channel facilitates the self-organisation of invariant object responses in cortex. *Soc. Neurosci. Abstr.* 18, 740.

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci. Suppl.* 3, 1199–1204. doi: 10.1038/81479

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans. R. Soc.* 335, 11–21. doi: 10.1098/rstb.1992.0002

Rolls, E. T. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behav. Brain Res.* 66, 177–185. doi: 10.1016/0166-4328(94)00138-6

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27, 205–218. doi: 10.1016/S0896-6273(00)00030-1

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes of primates including humans. *Neuropsychologia* 45, 124–143. doi: 10.1016/j.neuropsychologia.2006.04.019

Rolls, E. T. (2008). *Memory, Attention, and Decision-Making. A Unifying Computational Neuroscience Approach*. Oxford: Oxford University Press.

Rolls, E. T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Front. Comput. Neurosci.* 6:1–70. doi: 10.3389/fncom.2012.00035

Rolls, E. T. (2014). *Emotion and Decision-Making Explained*. Oxford: Oxford University Press.

Rolls, E. T., Aggelopoulos, N. C., and Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *J. Neurosci.* 23, 339–348.

Rolls, E. T., and Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp. Brain Res.* 65, 38–48. doi: 10.1007/BF00243828

Rolls, E. T., Baylis, G. C., and Hasselmo, M. E. (1987). The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. *Vis. Res.* 27, 311–326. doi: 10.1016/0042-6989(87)90081-2

Rolls, E. T., Baylis, G. C., Hasselmo, M., and Nalwa, V. (1989). "The representation of information in the temporal lobe visual cortical areas of macaque monkeys," in *Seeing Contour and Colour*, eds J. Kulikowski, C. Dickinson, and I. Murray (Oxford: Pergamon).

Rolls, E. T., Baylis, G. C., and Leonard, C. M. (1985). Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. *Vis. Res.* 25, 1021–1035. doi: 10.1016/0042-6989(85)90091-4

Rolls, E. T., and Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

Rolls, E. T., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.* 12, 2547–2572. doi: 10.1162/089976600300014845

Rolls, E. T., and Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network* 12, 111–129. doi: 10.1088/0954-898X/12/2/302

Rolls, E. T., and Stringer, S. M. (2006). Invariant visual object recognition: a model, with lighting invariance. *J. Physiol.* 100, 43–62. doi: 10.1016/j.jphysparis.2006.09.004

Rolls, E. T., and Stringer, S. M. (2007). Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural Comput.* 19, 139–169. doi: 10.1162/neco.2007.19.1.139

Rolls, E. T., and Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. R. Soc. B* 257, 9–15. doi: 10.1098/rspb.1994.0087

Rolls, E. T., and Treves, A. (1998). *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Rolls, E. T., and Treves, A. (2011). The neuronal encoding of information in the brain. *Prog. Neurobiol.* 95, 448–490. doi: 10.1016/j.pneurobio.2011.08.002

Rolls, E. T., Treves, A., and Tovee, M. J. (1997a). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp. Brain Res.* 114, 149–162. doi: 10.1007/PL00005615

Rolls, E. T., Treves, A., Tovee, M., and Panzeri, S. (1997b). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.* 4, 309–333. doi: 10.1023/A:1008899916425

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Spruston, N., Jonas, P., and Sakmann, B. (1995). Dendritic glutamate receptor channel in rat hippocampal CA3 and CA1 pyramidal neurons. *J. Physiol.* 482, 325–352.

Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybernet.* 94, 128–142. doi: 10.1007/s00422-005-0030-z

Stringer, S. M., and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Netw.* 13, 305–315. doi: 10.1016/S0893-6080(00)00017-4

Stringer, S. M., and Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3D objects. *Neural Comput.* 14, 2585–2596. doi: 10.1162/089976602760407982

Stringer, S. M., and Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Netw.* 21, 888–903. doi: 10.1016/j.neunet.2007.11.004

Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network* 18, 161–187. doi: 10.1080/09548980701556055

Sutton, R. S., and Barto, A. G. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170. doi: 10.1037/0033-295X.88.2.135

Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. *J. Neurophysiol.* 72, 1049–1060.

Van Essen, D., Anderson, C. H., and Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science* 255, 419–423. doi: 10.1126/science.1734518

Wallis, G. (2013). Toward a unified model of face and object recognition in the human visual system. *Front. Psychol.* 4:497. doi: 10.3389/fpsyg.2013.00497

Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194. doi: 10.1016/S0301-0082(96)00054-8

Wallis, G., Rolls, E. T., and Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *Int. Joint Conf. Neural Netw.* 2, 1087–1090.

Wiskott, L. (2003). Slow feature analysis: a theoretical analysis of optimal free responses. *Neural Comput.* 15, 2147–2177. doi: 10.1162/089976603322297331

Wiskott, L., and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14, 715–770. doi: 10.1162/089976602317318938

Wyss, R., Konig, P., and Verschure, P. F. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol.* 4:e120. doi: 10.1371/journal.pbio.0040120

Yang, W., Wang, Y., and Mori, G. (2010). "Recognizing human actions from still images with latent poses," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA: IEEE), 2030–2037. doi: 10.1109/CVPR.2010.5539879

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 November 2013; accepted: 12 March 2014; published online: 01 April 2014.

Citation: Webb TJ and Rolls ET (2014) Deformation-specific and deformation-invariant visual object recognition: pose vs. identity recognition of people and deforming objects. Front. Comput. Neurosci. ***8****:37. doi: 10.3389/fncom.2014.00037*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Webb and Rolls. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

A.1 THE ARCHITECTURE OF VisNet

Fundamental elements of Rolls' 1992 theory for how cortical networks might implement invariant object recognition are described in detail elsewhere (Rolls, 2008, 2012). They provide the basis for the design of VisNet, and can be summarized as:

- A series of competitive networks, organized in hierarchical layers, exhibiting mutual inhibition over a short range within each layer. These networks allow combinations of features or inputs occurring in a given spatial arrangement to be learned by neurons, ensuring that higher order spatial properties of the input stimuli are represented in the network.
- A convergent series of connections from a localized population of cells in preceding layers to each cell of the following layer, thus allowing the receptive field size of cells to increase through the visual processing areas or layers.
- A modified Hebb-like learning rule incorporating a temporal trace of each cell's previous activity, which, it is suggested (Földiák, 1991; Rolls, 1992, 2012; Wallis et al., 1993; Wallis and Rolls, 1997), will enable the neurons to learn transform invariances.

A.1.1 The trace rule

The learning rule implemented in the VisNet simulations utilizes the spatio-temporal constraints placed upon the behavior of "real-world" objects to learn about natural object transformations. By presenting consistent sequences of transforming objects the cells in the network can learn to respond to the same object through all of its naturally transformed states, as described by Földiák (1991), Rolls (1992), Wallis et al. (1993), Wallis and Rolls (1997), and Rolls (2012). The learning rule incorporates a decaying trace of previous cell activity and is henceforth referred to simply as the "trace" learning rule. The learning paradigm we describe here is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons, including position, size, view, lighting, and spatial frequency (Rolls, 1992, 2000, 2008, 2012; Rolls and Deco, 2002).

Various biological bases for this temporal trace have been advanced as follows: (The precise mechanisms involved may alter the precise form of the trace rule which should be used. Földiák, 1992 describes an alternative trace rule which models individual NMDA channels. Equally, a trace implemented by extended cell firing should be reflected in representing the trace as an external firing rate, rather than an internal signal).

- The persistent firing of neurons for as long as 100–400 ms observed after presentations of stimuli for 16 ms (Rolls and Tovee, 1994) could provide a time window within which to associate subsequent images. Maintained activity may potentially be implemented by recurrent connections between as well as within cortical areas (Rolls and Treves, 1998; Rolls and Deco, 2002; Rolls, 2008). [The prolonged firing of inferior temporal cortex neurons during memory delay periods of several seconds, and associative links reported to develop between stimuli presented several seconds apart (Miyashita, 1988) are on too long a time scale to be immediately relevant to the present theory. In fact, associations between visual events occurring several seconds apart would, under *normal* environmental conditions, be detrimental to the operation of a network of the type described here, because they would probably arise from different objects. In contrast, the system described benefits from associations between visual events that occur close in time (typically within 1 s), as they are likely to be from the same object].
- The binding period of glutamate in the NMDA channels, which may last for 100 ms or more, may implement a trace rule by producing a narrow time window over which the *average* activity at each presynaptic site affects learning (Hestrin et al., 1990; Földiák, 1992; Rhodes, 1992; Rolls, 1992; Spruston et al., 1995).
- Chemicals such as nitric oxide may be released during high neural activity and gradually decay in concentration over a short time window during which learning could be enhanced (Montague et al., 1991; Földiák, 1992; Garthwaite, 2008).

The trace update rule used in the baseline simulations of VisNet (Wallis and Rolls, 1997) is equivalent to both Földiák's used in the context of translation invariance (Wallis et al., 1993) and to the earlier rule of Sutton and Barto (1981) explored in the context of modeling the temporal properties of classical conditioning, and can be summarized as follows:

$$\delta w_j = \alpha \bar{y}^{\tau} x_j \tag{1}$$

where

$$\bar{y}^{\tau} = (1 - \eta)\, y^{\tau} + \eta \bar{y}^{\tau - 1} \tag{2}$$

and

x_j: jth input to the neuron.
$\bar{y}^{\tau}$: Trace value of the output of the neuron at time step τ.
w_j: Synaptic weight between jth input and the neuron.
y: Output from the neuron.
α: Learning rate.
η: Trace value. The optimal value varies with presentation sequence length.

At the start of a series of investigations of different forms of the trace learning rule, Rolls and Milward (2000) demonstrated that VisNet's performance could be greatly enhanced with a modified Hebbian trace learning rule (Equation 3) that incorporated a trace of activity from the preceding time steps, with no contribution from the activity being produced by the stimulus at the current time step. This rule took the form

$$\delta w_j = \alpha \bar{y}^{\tau - 1} x_j^{\tau}. \tag{3}$$

The trace shown in Equation (3) is in the postsynaptic term. The crucial difference from the earlier rule (see Equation 1) was that the trace should be calculated up to only the preceding timestep, with no contribution to the trace from the firing on the current trial to the current stimulus. This has the effect of updating

the weights based on the preceding activity of the neuron, which is likely given the spatio-temporal statistics of the visual world to be from previous transforms of the same object (Rolls and Milward, 2000; Rolls and Stringer, 2001). This is biologically not at all implausible, as considered in more detail elsewhere (Rolls, 2008, 2012), and this version of the trace rule was used in this investigation.

The optimal value of η in the trace rule is likely to be different for different layers of VisNet. For early layers with small receptive fields, few successive transforms are likely to contain similar information within the receptive field, so the value for η might be low to produce a short trace. In later layers of VisNet, successive transforms may be in the receptive field for longer, and invariance may be developing in earlier layers, so a longer trace may be beneficial. In practice, after exploration we used η values of 0.6 for layer 2, and 0.8 for layers 3 and 4. In addition, it is important to form feature combinations with high spatial precision before invariance learning supported by a temporal trace starts, in order that the feature combinations and not the individual features have invariant representations (Rolls, 2008, 2012). For this reason, purely associative learning with no temporal trace was used in layer 1 of VisNet (Rolls and Milward, 2000).

The following principled method was introduced to choose the value of the learning rate α for each layer. The mean weight change from all the neurons in that layer for each epoch of training was measured, and was set so that with slow learning over 15–50 trials, the weight changes per epoch would gradually decrease and asymptote with that number of epochs, reflecting convergence. Slow learning rates are useful in competitive nets, for if the learning rates are too high, previous learning in the synaptic weights will be overwritten by large weight changes later within the same epoch produced if a neuron starts to respond to another stimulus (Rolls, 2008). If the learning rates are too low, then no useful learning or convergence will occur. It was found that the following learning rates enabled good operation with the 100 transforms of each of 4 stimuli used in each epoch in the present investigation: Layer 1 $\alpha = 0.05$; Layer 2 $\alpha = 0.03$ (this is relatively high to allow for the sparse representations in layer 1); Layer 3 $\alpha = 0.005$; Layer 4 $\alpha = 0.005$.

To bound the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is explicitly normalized (a method similarly employed by Malsburg (1973) which is commonly used in competitive networks Rolls, 2008). An alternative, more biologically relevant implementation, using a local weight bounding operation which utilizes a form of heterosynaptic long-term depression (Rolls, 2008), has in part been explored using a version of the Oja (1982) rule (see Wallis and Rolls, 1997).

A.1.2 The network implemented in VisNet

The network itself is designed as a series of hierarchical, convergent, competitive networks, in accordance with the hypotheses advanced above. The actual network consists of a series of four layers, constructed such that the convergence of information from the most disparate parts of the network's input layer can potentially influence firing in a single neuron in the final layer—see **Figure 3**. This corresponds to the scheme described by many researchers (Rolls, 1992, 2008; Van Essen et al., 1992, for example) as present in the primate visual system—see **Figure 3**. The forward connections to a cell in one layer are derived from a topologically related and confined region of the preceding layer. The choice of whether a connection between neurons in adjacent layers exists or not is based upon a Gaussian distribution of connection probabilities which roll off radially from the focal point of connections for each neuron. (A minor extra constraint precludes the repeated connection of any pair of cells). In particular, the forward connections to a cell in one layer come from a small region of the preceding layer defined by the radius in **Table A1** which will contain approximately 67% of the connections from the preceding layer. **Table A1** shows the dimensions for the research described here, a (16x) larger version than the version of VisNet used in most of our previous investigations, which utilized 32 × 32 neurons per layer. For the research on view and translation invariance learning described here, we decreased the number of connections to layer 1 neurons to 100 (from 272), in order to increase the selectivity of the network between objects. We increased the number of connections to each neuron in layers 2–4 to 400 (from 100), because this helped layer 4 neurons to reflect evidence from neurons in previous layers about the large number of transforms (typically 100 transforms, from 4 views of each object and 25 locations) each of which corresponded to a particular object.

Figure 3 shows the general convergent network architecture used. Localization and limitation of connectivity in the network is intended to mimic cortical connectivity, partially because of the clear retention of retinal topology through regions of visual cortex. This architecture also encourages the gradual combination of features from layer to layer which has relevance to the binding problem, as described elsewhere (Rolls, 2008, 2012).

A.1.3 Competition and lateral inhibition

In order to act as a competitive network some form of mutual inhibition is required within each layer, which should help to ensure that all stimuli presented are evenly represented by the neurons in each layer. This is implemented in VisNet by a form of lateral inhibition. The idea behind the lateral inhibition, apart from this being a property of cortical architecture in the brain, was to prevent too many neurons that received inputs from a similar part of the preceding layer responding to the same activity patterns. The purpose of the lateral inhibition was to ensure that different receiving neurons coded for different inputs. This is important in reducing redundancy (Rolls, 2008). The lateral inhibition is conceived as operating within a radius that was similar to that of the region within which a neuron received converging

Table A1 | VisNet dimensions.

	Dimensions	# Connections	Radius
Layer 4	128 × 128	400	48
Layer 3	128 × 128	400	36
Layer 2	128 × 128	400	24
Layer 1	128 × 128	100	24
Input layer	256 × 256 × 16	–	–

inputs from the preceding layer (because activity in one zone of topologically organized processing within a layer should not inhibit processing in another zone in the same layer, concerned perhaps with another part of the image). The lateral inhibition used in this investigation used the parameters for σ shown in **Table A3**.

The lateral inhibition and contrast enhancement just described are actually implemented in VisNet2 (Rolls and Milward, 2000) and VisNetL (Perry et al., 2010) in two stages, to produce filtering of the type illustrated elsewhere (Rolls, 2008, 2012). The lateral inhibition was implemented by convolving the activation of the neurons in a layer with a spatial filter, I, where δ controls the contrast and σ controls the width, and a and b index the distance away from the center of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum\limits_{a \neq 0, b \neq 0} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \quad (4)$$

This is a filter that leaves the average activity unchanged.

The second stage involves contrast enhancement. A sigmoid activation function was used in the way described previously (Rolls and Milward, 2000):

$$y = \mathrm{f}^{\mathrm{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \quad (5)$$

where r is the activation (or firing rate) of the neuron after the lateral inhibition, y is the firing rate after the contrast enhancement produced by the activation function, and β is the slope or gain and α is the threshold or bias of the activation function. The sigmoid bounds the firing rate between 0 and 1 so global normalization is not required. The slope and threshold are held constant within each layer. The slope is constant throughout training, whereas the threshold is used to control the sparseness of firing rates within each layer. The (population) sparseness of the firing within a layer is defined (Rolls and Treves, 1998; Franco et al., 2007; Rolls, 2008; Rolls and Treves, 2011) as:

$$a = \frac{\left(\sum_i y_i/n\right)^2}{\sum_i y_i^2/n} \quad (6)$$

where n is the number of neurons in the layer. To set the sparseness to a given value, e.g., 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer.

The sigmoid activation function was used with parameters (selected after a number of optimization runs) as shown in **Table A2**.

Table A2 | Sigmoid parameters for the runs with 25 locations by Rolls and Milward (2000).

Layer	1	2	3	4
Percentile	99.2	98	88	95
Slope β	190	40	75	26

In addition, the lateral inhibition parameters are as shown in **Table A3**.

A.1.4 The input to VisNet

VisNet is provided with a set of input filters which can be applied to an image to produce inputs to the network which correspond to those provided by simple cells in visual cortical area 1 (V1). The purpose of this is to enable within VisNet the more complicated response properties of cells between V1 and the inferior temporal cortex (IT) to be investigated, using as inputs natural stimuli such as those that could be applied to the retina of the real visual system. This is to facilitate comparisons between the activity of neurons in VisNet and those in the real visual system, to the same stimuli. In VisNet no attempt is made to train the response properties of simple cells in V1, but instead we start with a defined series of filters to perform fixed feature extraction to a level equivalent to that of simple cells in V1, as have other researchers in the field (Fukushima, 1980; Buhmann et al., 1991; Hummel and Biederman, 1992), because we wish to simulate the more complicated response properties of cells between V1 and the inferior temporal cortex (IT). The elongated orientation-tuned input filters used accord with the general tuning profiles of simple cells in V1 (Hawken and Parker, 1987) were computed by Gabor filters. Each individual filter is tuned to spatial frequency (0.0626–0.5 cycles/pixel over four octaves); orientation (0°–135° in steps of 45°); and sign (±1). Of the 100 layer 1 connections, the number to each group in VisNetL is as shown in **Table A4**. Any zero D.C. filter can of course produce a negative as well as positive output, which would mean that this simulation of a simple cell would permit negative as well as positive firing. The response of each filter is zero thresholded and the negative results used to form a separate anti-phase input to the network. The filter outputs are also normalized across scales to compensate for the low frequency bias in the images of natural objects.

The Gabor filters used were similar to those used previously (Deco and Rolls, 2004). Following Daugman (1988) the receptive fields of the simple cell-like input neurons are modeled by 2D-Gabor functions. The Gabor receptive fields have five degrees of freedom given essentially by the product of an elliptical Gaussian and a complex plane wave. The first two degrees of freedom are the 2D-locations of the receptive field's center; the third is the size of the receptive field; the fourth is the orientation of the

Table A3 | Lateral inhibition parameters for the 25-location runs.

Layer	1	2	3	4
Radius, σ	1.38	2.7	4.0	6.0
Contrast, δ	1.5	1.5	1.6	1.4

Table A4 | VisNet layer 1 connectivity.

Frequency	0.5	0.25	0.125	0.0625
# Connections	74	19	5	2

The frequency is in cycles per pixel.

boundaries separating excitatory and inhibitory regions; and the fifth is the symmetry. This fifth degree of freedom is given in the standard Gabor transform by the real and imaginary part, i.e., by the phase of the complex function representing it, whereas in a biological context this can be done by combining pairs of neurons with even and odd receptive fields. This design is supported by the experimental work of Pollen and Ronner (1981), who found simple cells in quadrature-phase pairs. Even more, Daugman (1988) proposed that an ensemble of simple cells is best modeled as a family of 2D-Gabor wavelets sampling the frequency domain in a log-polar manner as a function of eccentricity. Experimental neurophysiological evidence constrains the relation between the free parameters that define a 2D-Gabor receptive field (De Valois and De Valois, 1988). There are three constraints fixing the relation between the width, height, orientation, and spatial frequency (Lee, 1996). The first constraint posits that the aspect ratio of the elliptical Gaussian envelope is 2:1. The second constraint postulates that the plane wave tends to have its propagating direction along the short axis of the elliptical Gaussian. The third constraint assumes that the half-amplitude bandwidth of the frequency response is about 1–1.5 octaves along the optimal orientation. Further, we assume that the mean is zero in order to have an admissible wavelet basis (Lee, 1996).

In more detail, the Gabor filters are constructed as follows (Deco and Rolls, 2004). We consider a pixelized gray-scale image given by a $N \times N$ matrix $\Gamma_{ij}^{\text{orig}}$. The subindices ij denote the spatial position of the pixel. Each pixel value is given a gray level brightness value coded in a scale between 0 (black) and 255 (white). The first step in the preprocessing consists of removing the DC component of the image (i.e., the mean value of the gray-scale intensity of the pixels). (The equivalent in the brain is the low-pass filtering performed by the retinal ganglion cells and lateral geniculate cells. The visual representation in the LGN is essentially a contrast invariant pixel representation of the image, i.e., each neuron encodes the relative brightness value at one location in visual space referred to the mean value of the image brightness). We denote this contrast-invariant LGN representation by the $N \times N$ matrix Γ_{ij} defined by the equation

$$\Gamma_{ij} = \Gamma_{ij}^{\text{orig}} - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}^{\text{orig}}. \tag{7}$$

Feedforward connections to a layer of V1 neurons perform the extraction of simple features like bars at different locations, orientations and sizes. Realistic receptive fields for V1 neurons that extract these simple features can be represented by 2D-Gabor wavelets. Lee (1996) derived a family of discretized 2D-Gabor wavelets that satisfy the wavelet theory and the neurophysiological constraints for simple cells mentioned above. They are given by an expression of the form

$$G_{pqkl}(x, y) = a^{-k} \Psi_{\Theta_l} \left(a^{-k} \left(x - 2p \right), a^{-k} \left(y - 2q \right) \right) \tag{8}$$

where

$$\Psi_{\Theta_l} = \Psi \left(x \cos \left(l\Theta_0 \right) + y \sin \left(l\Theta_0 \right), -x \sin \left(l\Theta_0 \right) + y \cos \left(l\Theta_0 \right) \right), \tag{9}$$

and the mother wavelet is given by

$$\Psi \left(x, y \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{8}\left(4x^2 + y^2\right)} \left[e^{i\kappa x} - e^{-\frac{\kappa^2}{2}} \right]. \tag{10}$$

In the above equations $\Theta_0 = \pi / L$ denotes the step size of each angular rotation; l the index of rotation corresponding to the preferred orientation $\Theta_l = l\pi / L$; k denotes the octave; and the indices pq the position of the receptive field center at $c_x = p$ and $c_y = q$. In this form, the receptive fields at all levels cover the spatial domain in the same way, i.e., by always overlapping the receptive fields in the same fashion. In the model we use $a = 2$, $b = 1$, and $\kappa = \pi$ corresponding to a spatial frequency bandwidth of one octave. We used symmetric filters with the angular spacing between the different orientations set to 45°; and with four filter frequencies spaced one octave apart starting with 0.5 cycles per pixel, and with the sampling from the spatial frequencies set as shown in **Table A4**.

Cells of layer 1 receive a topologically consistent, localized, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial frequency and receives a constant number of inputs.

A.1.5 Measures for network performance

Information theory measures. A neuron can be said to have learned an invariant representation if it discriminates one set of stimuli from another set, across all transforms. For example, a neuron's response is translation invariant if its response to one set of stimuli irrespective of presentation is consistently higher than for all other stimuli irrespective of presentation location. Note that we state "set of stimuli" since neurons in the inferior temporal cortex are not generally selective for a single stimulus but rather a subpopulation of stimuli (Baylis et al., 1985; Abbott et al., 1996; Rolls et al., 1997a; Rolls and Treves, 1998, 2011; Rolls and Deco, 2002; Franco et al., 2007; Rolls, 2007, 2008). We used measures of network performance (Rolls and Milward, 2000) based on information theory and similar to those used in the analysis of the firing of real neurons in the brain (Rolls, 2008; Rolls and Treves, 2011). A single cell information measure was introduced which is the maximum amount of information the cell has about any one object independently of which transform (here position on the retina and view) is shown. Because the competitive algorithm used in VisNet tends to produce local representations (in which single cells become tuned to one stimulus or object), this information measure can approach $\log_2 N_S$ bits, where N_S is the number of different stimuli. Indeed, it is an advantage of this measure that it has a defined maximal value, which enables how well the network is performing to be quantified. Rolls and Milward (2000) also introduced a multiple cell information measure used here, which has the advantage that it provides a measure of whether all stimuli are encoded by different neurons in the network. Again, a high value of this measure indicates good performance.

For completeness, we provide further specification of the two information theoretic measures, which are described in detail by Rolls and Milward (2000) (see Rolls, 2008 and Rolls and Treves, 2011 for an introduction to the concepts). The measures assess

the extent to which either a single cell, or a population of cells, responds to the same stimulus invariantly with respect to its location, yet responds differently to different stimuli. The measures effectively show what one learns about which stimulus was presented from a single presentation of the stimulus at any randomly chosen location. Results for top (4th) layer cells are shown. High information measures thus show that cells fire similarly to the different transforms of a given stimulus (object), and differently to the other stimuli. The single cell stimulus-specific information, $I(s, R)$, is the amount of information the set of responses, R, has about a specific stimulus, s (see Rolls et al., 1997b and Rolls and Milward, 2000). $I(s, R)$ is given by

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{11}$$

where r is an individual response from the set of responses R of the neuron. For each cell the performance measure used was the maximum amount of information a cell conveyed about any one stimulus. This (rather than the mutual information, $I(S, R)$ where S is the whole set of stimuli s), is appropriate for a competitive network in which the cells tend to become tuned to one stimulus. ($I(s, R)$ has more recently been called the stimulus-specific surprise (DeWeese and Meister, 1999; Rolls and Treves, 2011). Its average across stimuli is the mutual information $I(S, R)$).

If all the output cells of VisNet learned to respond to the same stimulus, then the information about the set of stimuli S would be very poor, and would not reach its maximal value of $\log_2$ of the number of stimuli (in bits). The second measure that is used here is the information provided by a set of cells about the stimulus set, using the procedures described by Rolls et al. (1997a) and Rolls and Milward (2000). The multiple cell information is the mutual information between the whole set of stimuli S and of responses R calculated using a decoding procedure in which the stimulus s' that gave rise to the particular firing rate response vector on each trial is estimated. (The decoding step is needed because the high dimensionality of the response space would lead to an inaccurate estimate of the information if the responses were used directly, as described by Rolls et al. (1997a) and Rolls and Treves (1998)). A probability table is then constructed of the real stimuli s and the decoded stimuli s'. From this probability table, the mutual information between the set of actual stimuli S and the decoded estimates S' is calculated as

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')} \tag{12}$$

This was calculated for the subset of cells which had as single cells the most information about which stimulus was shown. In particular, in Rolls and Milward (2000) and subsequent papers, the multiple cell information was calculated from the first five cells for each stimulus that had maximal single cell information about that stimulus, that is from a population of 35 cells if there were seven stimuli (each of which might have been shown in for example 9 or 25 positions on the retina).

Pattern association decoding. The output of the inferior temporal visual cortex reaches structures such as the orbitofrontal cortex and amygdala, where associations to other stimuli are learned by a pattern association network with an associative (Hebbian) learning rule (Rolls, 2008, 2014). We therefore used a one-layer pattern association network (Rolls, 2008) to measure how well the output of VisNet could be classified into one of the objects. The pattern association network had four output neurons, one for each object. The inputs were the ten neurons from layer 4 of VisNet for each of the four objects with the best single cell information, making 40 inputs to each neuron. The network was trained with the Hebb rule:

$$\delta w_{ij} = \alpha y_i x_j \tag{13}$$

where δw_{ij} is the change of the synaptic weight w_{ij} that results from the simultaneous (or conjunctive) presence of presynaptic firing x_j and postsynaptic firing or activation y_i, and α is a learning rate constant that specifies how much the synapses alter on any one pairing. The pattern associator was trained for one trial on the output of VisNet produced by every transform of each object.

Performance on the test images extracted from the scenes was tested by presenting an image to VisNet, and then measuring the classification produced by the pattern associator. Performance was measured by the percentage of the correct classifications of an image as the correct object.

This approach to measuring the performance is very biologically appropriate, for it models the type of learning thought to be implemented in structures that receive information from the inferior temporal visual cortex such as the orbitofrontal cortex and amygdala (Rolls, 2008, 2014). The small number of neurons selected from layer 4 of VisNet might correspond to the most selective for this stimulus set in a sparse distributed representation (Rolls, 2008; Rolls and Treves, 2011). The method would measure whether neurons of the type recorded in the inferior temporal visual with good view and position invariance are developed in VisNet. In fact, an appropriate neuron for an input to such a decoding mechanism might have high firing rates to all or most of the view and position transforms of one of the stimuli, and smaller or no responses to any of the transforms of other objects, as found in the inferior temporal cortex for some neurons (Hasselmo et al., 1989b; Perrett et al., 1991; Booth and Rolls, 1998). Moreover, it would be inappropriate to train a device such as a support vector machine of even an error correction perceptron on the outputs of all the neurons in layer 4 of VisNet to produce four classifications, for such learning procedures, not biologically plausible (Rolls, 2008), could map the responses produced by a multilayer network with untrained random weights to obtain good classifications.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 25 March 2015
doi: 10.3389/fncom.2015.00035

Competition improves robustness against loss of information

*Arash Kermani Kolankeh, Michael Teichmann and Fred H. Hamker**

Department of Computer Science, Chemnitz University of Technology, Chemnitz, Germany

A substantial number of works have aimed at modeling the receptive field properties of the primary visual cortex (V1). Their evaluation criterion is usually the similarity of the model response properties to the recorded responses from biological organisms. However, as several algorithms were able to demonstrate some degree of similarity to biological data based on the existing criteria, we focus on the robustness against loss of information in the form of occlusions as an additional constraint for better understanding the algorithmic level of early vision in the brain. We try to investigate the influence of competition mechanisms on the robustness. Therefore, we compared four methods employing different competition mechanisms, namely, independent component analysis, non-negative matrix factorization with sparseness constraint, predictive coding/biased competition, and a Hebbian neural network with lateral inhibitory connections. Each of those methods is known to be capable of developing receptive fields comparable to those of V1 simple-cells. Since measuring the robustness of methods having simple-cell like receptive fields against occlusion is difficult, we measure the robustness using the classification accuracy on the MNIST hand written digit dataset. For this we trained all methods on the training set of the MNIST hand written digits dataset and tested them on a MNIST test set with different levels of occlusions. We observe that methods which employ competitive mechanisms have higher robustness against loss of information. Also the kind of the competition mechanisms plays an important role in robustness. Global feedback inhibition as employed in predictive coding/biased competition has an advantage compared to local lateral inhibition learned by an anti-Hebb rule.

Keywords: competition, lateral inhibition, Hebbian learning, independent component analysis, non-negative matrix factorization, predictive coding/biased competition, occlusion, information loss

OPEN ACCESS

Edited by:
Ales Leonardis, University of Birmingham School of Computer Science, UK

Reviewed by:
Michael W. Spratling, King's College London, UK
Xin Tian, Tianjin Medical University, China

****Correspondence:***
Fred H. Hamker, Fakultät für Informatik, Professur Künstliche Intelligenz, Straße der Nationen 62, 09107 Chemnitz, Germany
fred.hamker@informatik.tu-chemnitz.de

Received: *31 March 2014*
Accepted: *03 March 2015*
Published: *25 March 2015*

Citation:
Kermani Kolankeh A, Teichmann M and Hamker FH (2015) Competition improves robustness against loss of information. Front. Comput. Neurosci. 9:35. doi: 10.3389/fncom.2015.00035

1. Introduction

Several different learning approaches have been developed to model early vision, particularly at the level of V1 (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Hoyer and Hyvärinen, 2000; Falconbridge et al., 2006; Rehn and Sommer, 2007; Wiltschut and Hamker, 2009; Spratling, 2010; Zylberberg et al., 2011). In many of the works, the proposed characteristics of the visual system have been considered as optimization objectives and thus as criteria for measuring the efficiency of coding. Several kinds of optimization objectives, like sparseness of activity (Olshausen and Field, 1996; Hoyer, 2004) or independence (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998) have been used for this purpose. One major criterion for evaluation of those models is their ability to develop oriented, band-pass receptive fields and the similarity of the distribution of receptive fields to observed ones

in the macaque (Ringach, 2002). Although the match to biological data can be considered as one important criterion, further criteria are required to evaluate different approaches.

The visual system has the remarkable capability of robustness, or invariance, against different kinds of variances like, shift, rotation, scaling, occlusion, etc. of objects. This invariance is likely gradually achieved over different hierarchical levels, but robustness can be explained also in the form of information coding on the level of a single layer. This means, also units like V1 simple-cells show robustness against typical deformations of their preferred stimuli. In this work we have focused on the robustness under loss of information in the form of occlusion. Since typical forms of perturbations locally effecting V1-cells can be different lightning conditions—like reflections or flares; unclear media like soiled glasses, windows, heated air; or covered objects like the view through a fence—we define occlusion here as the random removal of visual information.

To investigate the role of different interactions, in fact competition, we compare four methods implementing different competition and learning strategies: Fast independent component analysis (FastICA) (Hyvärinen and Oja, 1997; Hoyer and Hyvärinen, 2000), non-negative matrix factorization with sparseness constraint (NMFSC) (Hoyer, 2004), predictive coding/biased competition (PC/BC) (Spratling, 2010), and a Hebbian neural network (further called HNN) with lateral inhibition based on Teichmann et al. (2012). Each method is capable of learning V1 simple-cell like receptive fields from natural images. FastICA was chosen as a method which tries to find new representations of data with minimal dependency between components without employing any kind of competition in the neural dynamics, but it enforces independent components via the learning rule. NMFSC uses a top-down, subtractive inhibition of the inputs to compute the outputs. NMFSC also keeps the output activity sparse on a desired, predefined level leading to unspecific competitive dynamics. PC/BC (Spratling, 2010) tries to find components minimizing the reconstruction error by a global error minimization employing inhibitory feedback connections. All of the above algorithms minimize a reconstruction error. While ICA minimizes a substractive reconstruction error, NMFSC (Hoyer, 2004) and PC/BC (Spratling, 2010) use divisive updating rules for the weight matrix that are derived from minimizing the Kullback-Leibler divergence (Lee and Seung, 1999). HNN uses Hebbian learning to learn the feedforward weights and anti-Hebbian learning to learn lateral inhibitory connections. The units compete via these lateral connections and suppress competing neurons locally based on the learned relations.

To evaluate the different algorithms trained all methods on the train set of the MNIST hand written digit dataset and measured their recognition accuracy on the occluded MNIST test set. The recognition accuracy was measured by feeding the activity patterns to a linear classifier. Here, the interesting aspect of each method was not its best accuracy in recognizing the classes, but its robustness in recognizing objects when the input was distorted, that is the change of the performance dependent on the level of occlusion.

2. Materials and Methods

2.1. Dataset and Preprocessing

We use the MNIST handwritten digit dataset[1] to evaluate all methods. The dataset consists of 60,000 training images and 10,000 test images. All are centered, size normalized (28 × 28 pixel), and have black (i.e., zero) background. We downscale the images to 12 × 12 using the MATLAB (2013a) function imresize() by the factor of 0.40 with default parameters (i.e., bicubic interpolation). This matches the original configuration of the HNN input for learning V1 like receptive fields (Wiltschut and Hamker, 2009). In order to simulate the function of the early visual system up to the Lateral Geniculate Nucleus (LGN), which transfers signals from the eyes to V1, we whitened the images using the same method as in Olshausen and Field (1997). The whitened image contains positive and negative values. The positive part and the absolute values of the negative part of each whitened image were reshaped to vectors and concatenated to form a 288-dimensional input vector. The positive part resembles the on-center receptive fields of the Lateral Geniculate Nucleus (LGN) cells and the negative part the off-center receptive fields (Wiltschut and Hamker, 2009).

We used a partially occluded test set to study the effect of loss of information on classification: the original non-occluded of MNIST and different occluded versions of it. A test set is formed by applying a particular occlusion level on all images in the original MNIST test set. That is, in each version, the level of occlusion was the same for all digits, although the position of the occluded pixels was generated randomly for each digit. The occluded test sets had an amount of 5–60%, in steps of 5%, occluded pixels. Only digit pixel and no background pixels were occluded. Occlusions were produced by randomly setting non-zero pixel values to zero before whitening an image (**Figure 1**). Since we are testing on all test sets we will further use the term "test set" to denote all of these test images. No occlusion was applied to the train set.

2.2. Models and Training

In this section we will give a short introduction in the main principles and the training of the used methods. To facilitate comparison all methods are using 288 units. For our simulations we used software provided by the respective authors.

2.2.1. Fast Independent Component Analysis

In fast independent component analysis (FastICA; Hyvärinen and Oja, 1997), the goal is finding statistically independent components of the data by maximizing neg-entropy. Neg-entropy is a measure of non-gaussianity and non-gaussianity is in direct relation with independence; the more non-Gaussian the activity distributions, the more independent are the components. The problem can be stated as

$$x = Vy$$

or

$$y = Wx$$

[1]http://yann.lecun.com/exdb/mnist/

where V is the mixing matrix and W its inverse, x is the input vector and y is the vector of sources or components which should be independent. ICA, as a generative method, tries to generate the inputs as a sum of components y weighted by the weights of the mixing matrix V. In FastICA matrices V and W are found in an optimization process which maximizes neg-entropy of the activities.

After W was determined on the (non-occluded) MNIST train set, we used W to calculate the output on the occluded test set by calculating $y_o = Wx_o$, where x_o stands for the occluded input and y_o for the corresponding output activities. Thus, the FastICA method has no competitive mechanism effecting the output, its just applying a linear transformation matrix on the input.

2.2.2. Non-Negative Matrix Factorization with Sparseness Constraint

In non-negative matrix factorization with sparseness constraint (NMFSC; Hoyer, 2004), the goal is to factorize the matrix of the input data in non-negative components and non-negative source matrices, imposing more biological plausibility in comparison to FastICA, as neuron responses are non-negative. NMFSC approaches the matrix of components V to satisfy $X \approx V \otimes Y$. Where Y is the matrix of output vectors and X the matrix of corresponding input vectors. Y and V are calculated while approaching the objective of reducing the difference between the input X and its reconstruction $V \otimes Y$:

$$V \leftarrow V \otimes (XY^T) \oslash (VYY^T) \tag{1}$$

Where $\otimes$ means element-wise multiplication and $\oslash$ element-wise division. One could say that the term $(XY^T) \oslash (VYY^T)$ is actually the modulated input which is used to update V. In some literature it is interpreted as a divisive form of feedback inhibition (Kompass, 2007; Spratling et al., 2009). This method, introduced by Lee and Seung (1999), tries to minimize the difference between the distributions of the input and its reconstruction based on Kullback-Leibler divergence.

In some other works this process is done by adding the subtractive difference between the input and its reconstruction to V. One could call both subtractive different and the divisive modulated input the inhibited input which is used for learning (Spratling et al., 2009).

The advantage of NMFSC to pure non-negative matrix factorization (NMF; Lee and Seung, 1999) is that the sparseness of the computed activities Y can be set to a desired level. An increase in the sparseness shifts the code from global to more local features (Hoyer, 2004). However, NMFSC deviates from a multiplicative update of the output Y and uses a subtractive one

$$Y \leftarrow Y - \mu V^T(VY - X)$$

Thus, the nodes compete with each other using a top-down, subtractive inhibition of their input. In order to obtain the desired level of sparseness a projection step is applied by keeping V fixed and looking for the closest Y which could both optimally cause to low reconstruction error and satisfy the sparseness constraint (for details see Hoyer, 2004, pp. 1462–1463). NMFSC also allows to control the sparseness of V, but this feature is not used by us.

To obtain the best classification accuracy, we tested four different sparseness levels (0, meaning no constraint; 0.75; 0.85; and 0.95). We found that 0.85 sparseness gives the best results

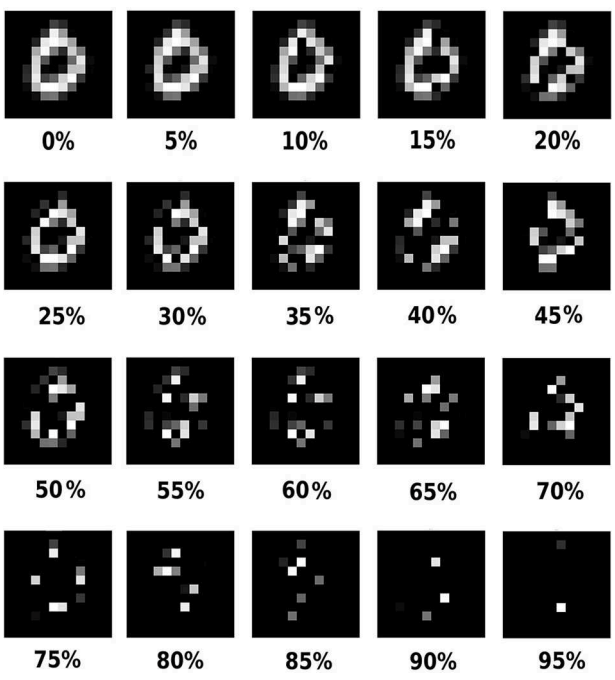


FIGURE 1 | An example of the input digits under 0-95% occlusions.

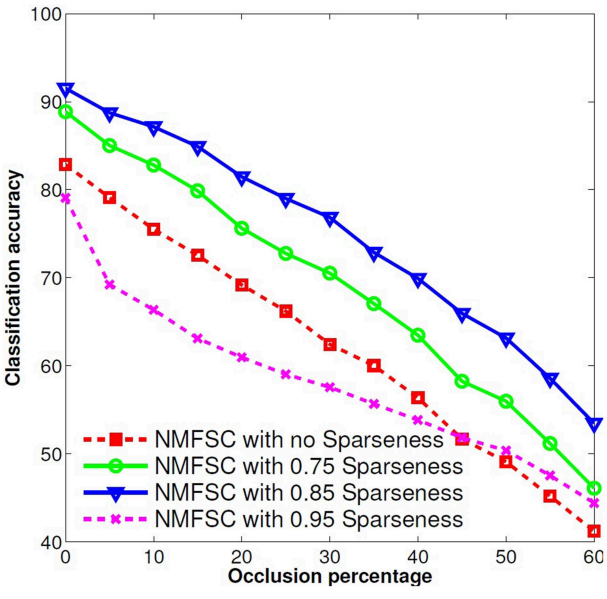


FIGURE 2 | Effect of different sparseness levels on the robustness of NMFSC, using the occluded MNIST test set. A sparseness of 0.85 shows the best robustness. Very high or no sparseness reduces the performance.

(**Figure 2**). The same sparseness level was found by Hoyer (2004) as the best level to learn Gabor-like filters from natural images. Hoyer defines the sparseness level as the relation of the L_1 norm to the L_2 norm. Where a sparseness of zero denotes the densest output vector, this is when all outputs are equally active, and of one denotes the sparsest vector, when just one output is active. For equation and an illustration of different degrees of sparseness please see (Hoyer, 2004, pp. 1460–1461). After we have trained NMFSC on the train set, we used V to calculate the output on the occluded test set. For this, we kept the obtained V fixed and ran the optimization process for Y, approaching the predefined sparseness level for Y while trying to reduce the reconstruct error.

2.2.3. Predictive Coding/Biased Competition

In predictive coding/biased competition (PC/BC; Spratling, 2010), like in the two other generative models, the goal is finding components so that the output can resemble the input with minimal error. This method uses divisive input modulation (DIM), introduced in Spratling et al. (2009), which is in turn based on NMF. The modifications, in comparison to NMF, are mainly two. First, it is on-line, while NMF is a batch method. Second, in contrast to NMF which uses the component weight matrix both for computing the output and reconstructing the input, DIM considers two sets of weight matrices; feedforward for producing the output and feedback for producing the reconstruction of the input. The two weight matrices differ just in the form of normalization, which makes the method more powerful than NMF in the case of overlap and occlusion (Spratling et al., 2009). In PC/BC, the inputs are inhibited by being divided by their reconstruction. This is done explicitly in the units called error units. The error units basically do the same job as the term $(XY^T) \oslash (WYY^T)$ in (Equation 1) in NMF. Their activity is described as following:

$$e = x \oslash (\epsilon_1 + V^T y)$$

where x is the input vector, y is the output vector, V is the feedback weight matrix, and ϵ_1 is a small value to avoid division by zero. The inhibited input from the error units is used for both producing the output and updating the weights. Thus, PC/BC uses in both cases a multiplicative updating, whereas NMFSC uses a subtractive one for the output.

To calculate the output the inhibited input is used:

$$y \leftarrow (\epsilon_2 + y) \otimes We$$

where ϵ_2 is a random small number which prevents the output from being zero, W is the feedforward weight matrix and e is the activity vector of the error units. Based on the output activities y and the error units e the weights are adopted as following:

$$W \leftarrow W \otimes \{(1 + \beta y)[e^T - 1]\}$$

where β is the learning rate. If the input and its reconstruction are equal, the error will be equal to unity and, thus, the weights will not change.

The input inhibition of PC/BC affects, besides the weight development, the output. Strong units suppressing weaker ones by removing their representation from the input. This is done in several iteration of updating the error units by the received reconstruction of the output units. This iterative process leads to a low reconstruction error and provides the competitive mechanism of PC/BC.

We trained PC/BC on 100,000 randomly, and potentially repeatedly, chosen digits from the 60,000 images of the MNIST train set and saved the weights for later calculating the outputs on the test set. Therefore, each image of the test set was presented for 200 iterations to the final network to achieve convergence of the outputs.

2.2.4. The Hebbian Neural Network

Finally, we use a Hebbian neural network (HNN), employing the well accepted mechanisms of rate based threshold linear neurons and Hebbian learning. A set of neurons in one layer receive feedforward input and lateral inhibitory connections being the source of competition between the neurons. The connection strengths are learned using a Hebbian learning rule for the feedforward connections and an anti-Hebbian one for the lateral connections (Földiák, 1990; Wiltschut and Hamker, 2009; Teichmann et al., 2012). For simulation, we use a slightly modified version of the one previously published by Teichmann et al. (2012). To learn the feedforward weights, the model employs a set of different mechanisms like covariance learning with Oja normalization (Oja, 1982), regulated by an activity dependent homeostatic term (Teichmann et al., 2012). It uses calcium traces of the neuron activity instead of activities for learning. However, we use a fast trace so the model works similar to an activity based model (see Appendix for further model details).

Since Teichmann et al. (2012) demonstrated the model to learn V1 complex-cell properties we verified that the model used here, if trained on natural images, learns simple-cell receptive fields (**Figure 3**) to fulfill our main criteria for model selection.

In this kind of network, inhibitory lateral connections are the source of competition between units. During the learning process the lateral weights develop proportional to the correlated firing between units, leading to strong inhibition between units that are often coactive. Hence, units in the HNN tend to reduce coactivity in the training phase and thus build a sparse representation of the input. Consequently, each unit uses the stored knowledge in the lateral weights to suppress potentially competing units.

We trained the network on 200,000 randomly chosen digits from the train set. During training each image is been presented to the network for 100 time steps (ms) to allow for a convergence of the dynamics. After learning, we keep the weights fixed and use this network to obtain the responses on the images of the test set.

2.3. Classification

As a criterion for robustness, we considered the accuracy of a classifier on the top of each method. The idea behind was that the classifier would indicate by its performance drop to classify the digits if some information is lost. Thus, a method with a more stable representation should have less accuracy decrease under increasing levels of occlusion. We have decided to use a simple linear classifier as it is assumed that also the neural processing in the brain should facilitate linear classification (DiCarlo and Cox, 2007). To measure the accuracy of classification we use Linear

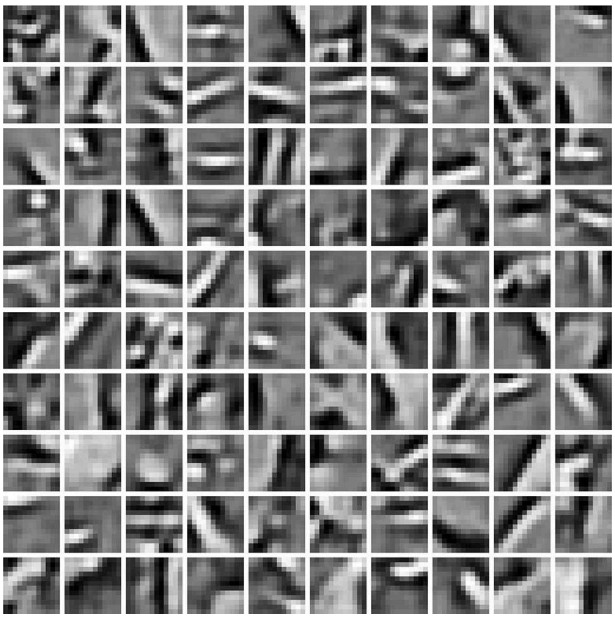


FIGURE 3 | Gabor-like receptive fields learned from natural images by the Hebbian neural network.

Discriminant Analysis (LDA) on the output of the methods on the test set. That is, we used the MATLAB (2013a) function classify() with default parameters (linear discriminant function). The classifier is trained using the output of the respective method on the train set.

2.4. Visualization of Weights and Receptive Fields

Obviously, if a method is able to learn a superior representation of the data, it will have a better robustness to the other ones. We visualize the weight matrices of all methods to get an insight into how the data are processed. If the methods share a similar character in their weight organization it can be assumed that this feedforward part of the processing shares similarities. Hence, the differences in the robustness of the methods have to come from the competitive mechanisms. Further, we can look at the receptive field shapes[2] of the units, as the competition is typically not changing their overall shapes, indeed the inhibitory effects between units are considered.

Hence, we used two approaches for visualizing the receptive fields. One was representing the weight matrix of a unit as gray-scale images. As the weight matrices correspond to the on-center and off-center inputs, we subtract this two parts from each other (Wiltschut and Hamker, 2009). The strength of each of these weights was shown as the intensity of a pixel in the image, where white denotes the maximum weight, gray denotes zero, and black the minimum weight. As an alternative, to visualize the receptive fields, we used reverse correlation. In order to obtain the optimal stimulus of a unit, we weighted images containing 90 random dots in front of black (zero) background with their corresponding outputs from a single unit. The average of the result was shown as the receptive field. This way we could observe to which input parts each unit is sensitive, regarding the competition between the units. In other words the resulting matrices visualize the correlation between the input and output values of each unit.

[2]Receptive fields are here defined as a map of regions in the image where a unit is excited or inhibited if a stimulus is there (cf. Hubel and Wiesel, 1962).

3. Results

3.1. Learned Receptive Fields

In order to verify if the models represent the input data in a comparable way, we visualize the weight vectors and receptive fields of 100 units for each model (cf. Section 2.4). To visualize the weight vectors of the Hebbian neural network (HNN), we have used the feedforward weight matrices showing the driving stimulus of the neurons (**Figure 4A**). For FastICA, we visualize the mixing matrix V (**Figure 4C**). The V matrix of basis vectors is visualized for NMFSC (**Figure 4E**). In PC/BC, we show the feedforward matrices (**Figure 4G**). For each method we also show the receptive fields estimated by reverse correlation (**Figures 4B,D,F,H**), being not much different from the visualization of the weight matrices. All methods develop receptive fields with holistic forms of digits. Indeed, in NMFSC not all units show digit like shapes which may result from the chosen level of sparseness as mentioned in the methods.

3.2. Classification Accuracy Under Occlusion

To investigate the differences in robustness to increasing levels of occlusions in the input, we have measured the classification accuracy of all methods and the raw data on the test set. We repeated the experiments 10 times with each algorithm under different starting conditions, i.e., randomly initialized weights. We do not show the error bars as they are zero for FastICA and NMFSC as they are deterministic and have been low for PC/BC and the HNN. We observed (**Figure 5**) that FastICA does not improve the classification accuracy to that of the raw data. NMFSC causes a super-linear decrease of classification accuracy with respect to the linear increase of occlusion. PC/BC shows the highest robustness against occlusion. The robustness of the HNN is higher than NMFSC and lower than PC/BC. The methods having more "advanced" competitive mechanisms perform better under increasing occlusions.

To further investigate the influence of the competitive mechanisms we turn them off for PC/BC and the HNN. This is, setting the lateral inhibitory connections to zero for the HNN, and using only the first iteration step of PC/BC. The training of the classifier is repeated for these modified models. Both HNN and PC/BC will cause a very low performance even worse than the raw data when their competitive mechanisms are not used (**Figure 6**). Meaning that the competitive mechanism has a substantial influence on the accuracy under occlusion and the pure feedforward processing is not enough have robust recognition results.

3.3. Effect of Occlusion on Activity Pattern

It is obvious that the activity pattern as a function of the input changes by increasing the occlusion in the input. The question is

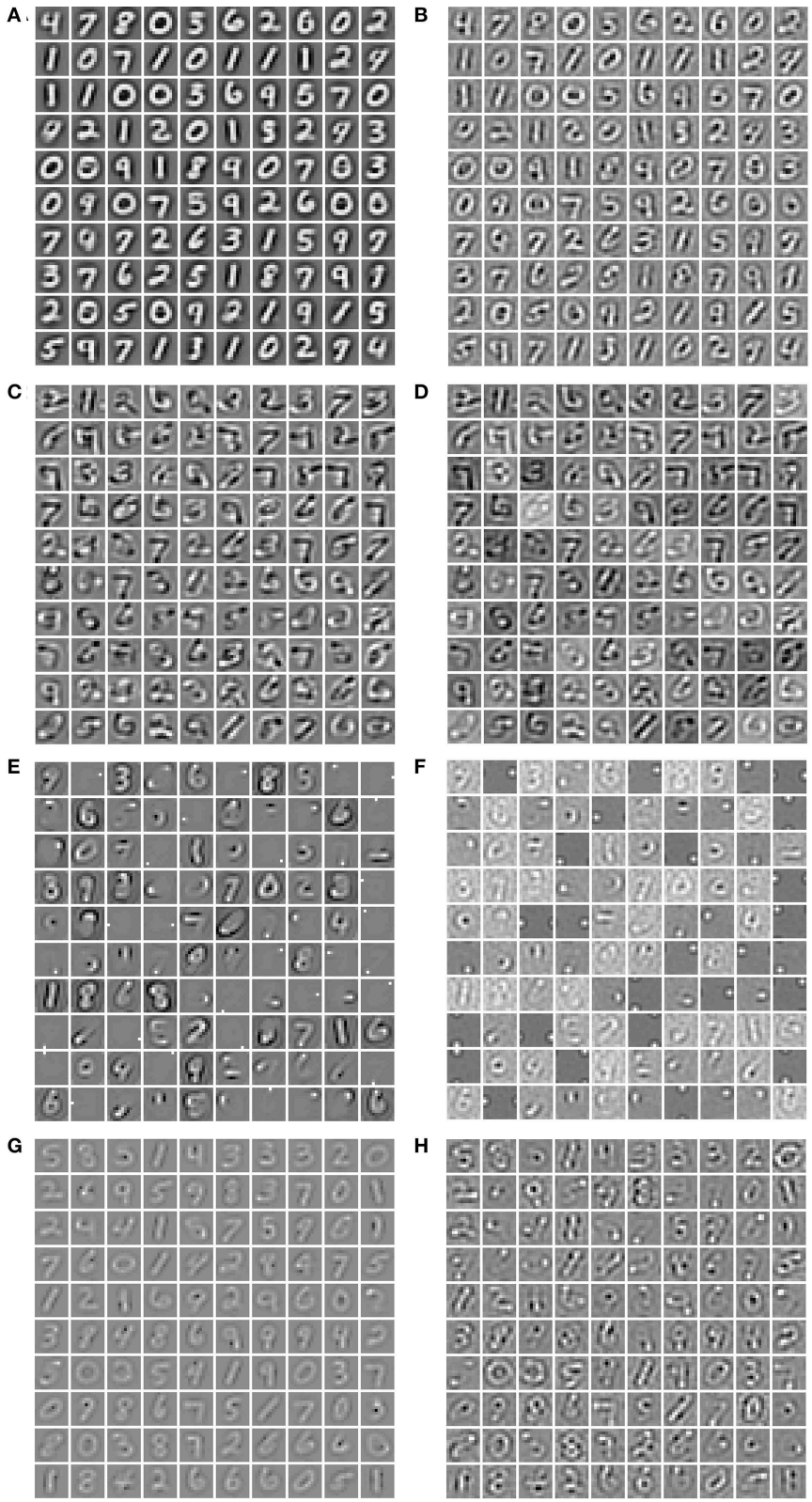


FIGURE 4 | Visualization of the feedforward weights and the receptive fields of 100 units, after training. Off-weights where subtracted from on-weights and each plot is scaled so that white denotes the maximum value and black the minimum. **(A)** The feedforward weight matrices of the HNN and **(B)** its reverse correlation. **(C)** The component matrices of FastICA and **(D)** its reverse correlation. **(E)** The component matrices of NMFSC and **(F)** its reverse correlation. **(G)** The feedforward weights of PC/BC and **(H)** its reverse correlation.

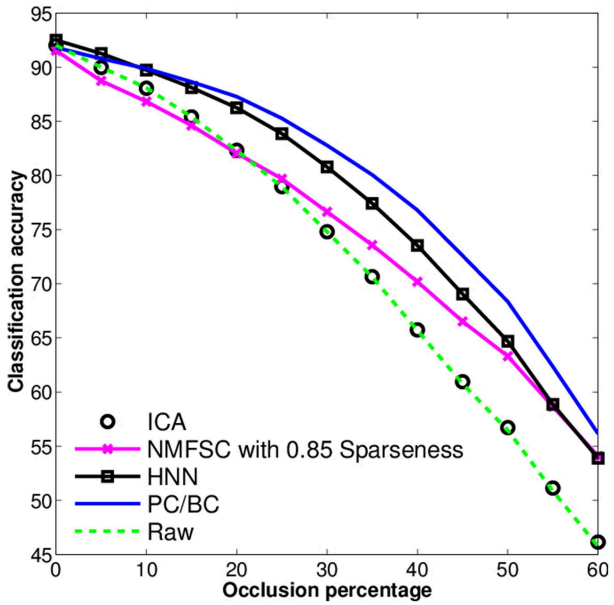


FIGURE 5 | Classification accuracy on the output of FastICA, NMFSC, HNN, and PC/BC, using the occluded MNIST test set. Methods using competitive mechanisms show better robustness.

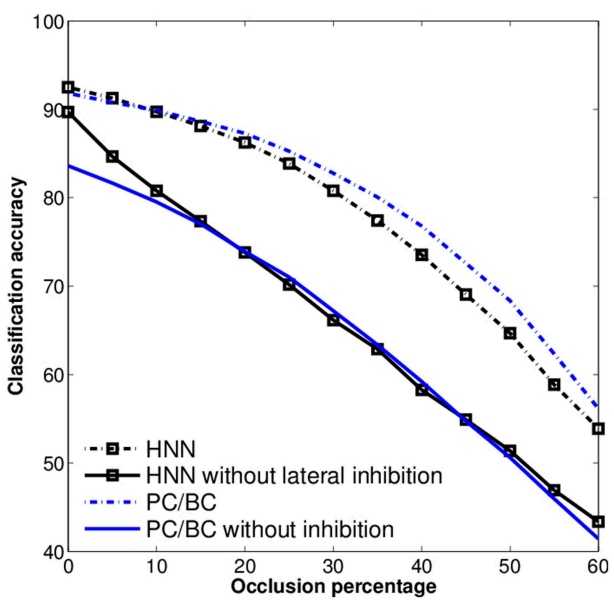


FIGURE 6 | Robustness of the HNN and PC/BC with and without inhibition, using the occluded MNIST test set. Without the inhibitory connections the models show a sharp drop in performance against loss of information (occlusion).

how stable the activity patterns of a method are when the occlusion in the input is increased. This is basically the same question as how much the classification accuracy is robust under loss of information. In **Figures 7–10** the activity patterns corresponding to three random inputs under 0, 20, and 40% occlusion are illustrated. As one can see in NMFSC, HNN, and PC/BC the activity patterns corresponding to non-occluded input and low occluded (20%) are comparable. In FastICA, though, the activity patterns are not easily comparable as ICA by nature produces very dense activity patterns. The activity pattern of FastICA on the (non-occluded) train set have a mean sparseness (Hoyer, 2004) of 0.41, which is, in comparison with NMFSC with 0.89, HNN with 0.80, and PC/BC with 0.89 sparseness, quite dense. However, in all methods the activity pattern loses its original form when occlusion is increased.

To measure how stable the activity patterns of a method are, for different levels of occlusion, we used the cosine of the angle between the non-occluded and the occluded activity vector. We calculate the cosine on the test set with 20 and 40% occlusion (**Table 1**) and found that methods showing a more robust recognition accuracy also having a lesser turn in their activity vector. Exceptionally, the HNN shows a more stable code than PC/BC based on this measure.

3.4. Selective Inhibition in the Hebbian Neural Network

To investigate the selectivity of inhibition in the HNN, we study the relation between the strength of the lateral connections and the similarity of the feedforward weights of a neuron to its laterally connected neurons by visualizing the feedforward weights of the laterally connected neurons sorted by the strength of the outgoing lateral connections. Therefore, we randomly select 10 neurons (left side) and plot the weights of the laterally connected neuron (**Figure 11**). As one can see, the shape of the feedforward weights of neurons being strongly inhibited are more similar to the weights of the inhibiting neuron than the ones which are lesser inhibited. This is, neurons have the strongest inhibition to neurons representing similar digits, mostly from the same class, followed by other classes sharing many similarities. Being expected as the strength of the inhibition is relative to the correlation of the neurons.

4. Discussion

We observed that the competitive mechanisms in the considered methods, FastICA, NMFSC, PC/BC, and HNN, have direct effect on their robustness under loss of information. Results showed that all methods have developed receptive fields similar to digit shapes and so the methods should be comparable. Apparently, this similarity itself cannot be used as a criterion for robustness against loss of information (occlusion). We observe that the receptive fields of FastICA are more similar to digits than the most of NMFSC, although, NMFSC shows a better accuracy under occlusion. However, without using its competition mechanism it behaves worser than FastICA. Further, HNN and PC/BC have the most clear receptive fields and the highest performances, indeed, without the competitive mechanism their accuracy drops lower than FastICA and NMFSC. Also the recognition accuracies of PC/BC and HNN with and without competition can not be explained by differences in the receptive field shapes. Without competition HNN behaves slightly better than PC/BC, whereas with competition PC/BC shows better accuracy. This means, the

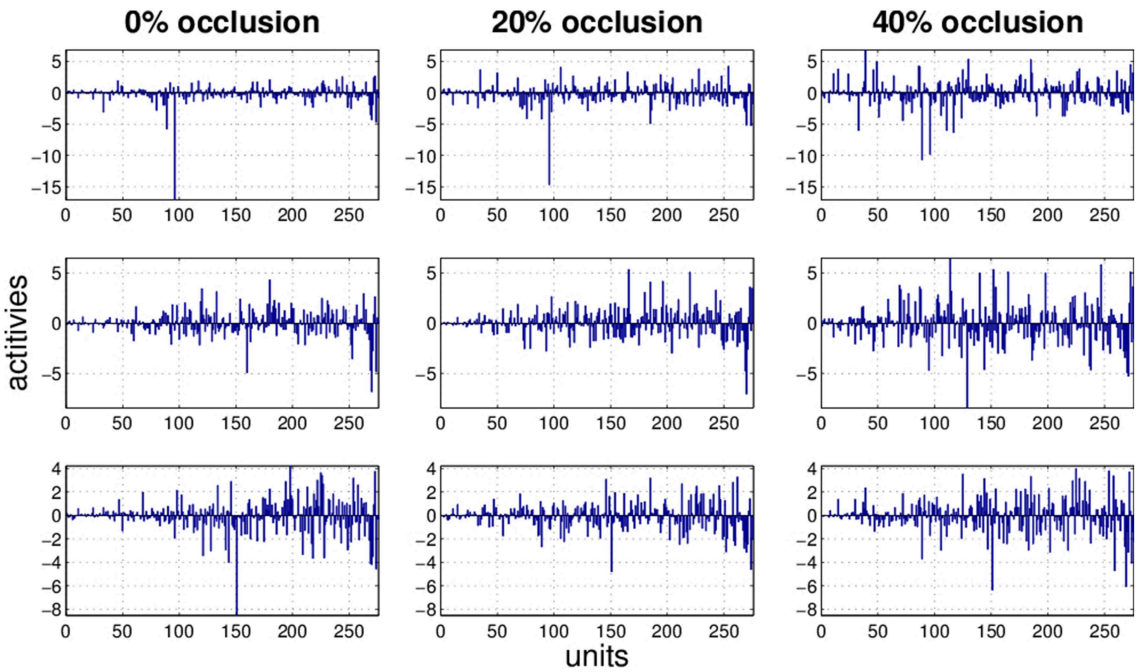


FIGURE 7 | Three examples (row) how the activity pattern vary, under 0, 20, and 40% of occlusion (column) in FastICA.

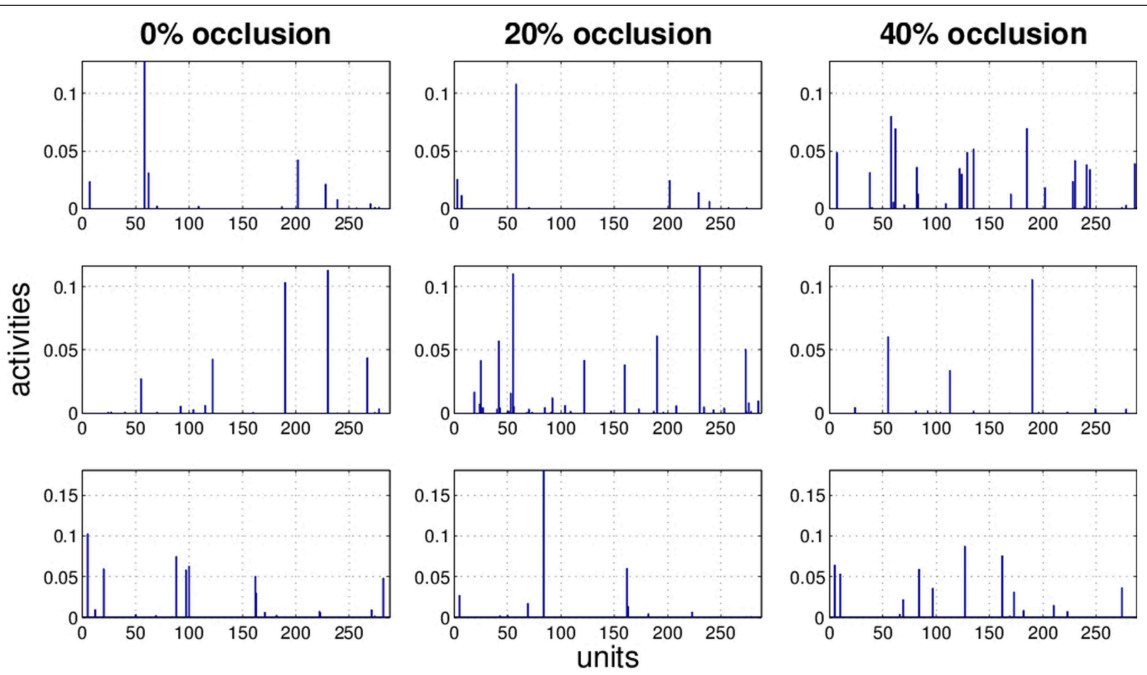


FIGURE 8 | Three examples (row) how the activity pattern varies, under 0, 20, and 40% of occlusion (column) in NMFSC.

receptive field quality alone does not cause the observed higher robustness.

Without occlusions no method shows a strong superiority in the accuracy, indeed, they show clear differences when the input is distorted. Some models are more stable when the input is occluded. This stability is in line with the results of the classification accuracy. While the HNN shows the least change in the cosine between its population responses with and without

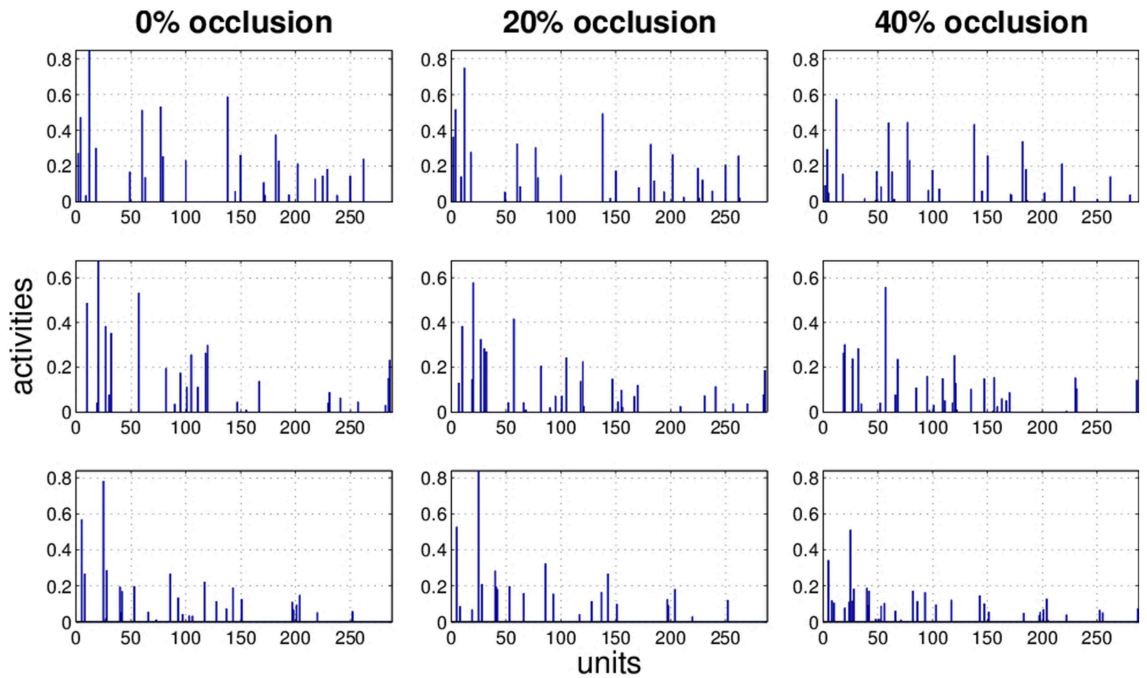


FIGURE 9 | Three examples (row) how the activity pattern varies, under 0, 20, and 40% of occlusion (column) in HNN.

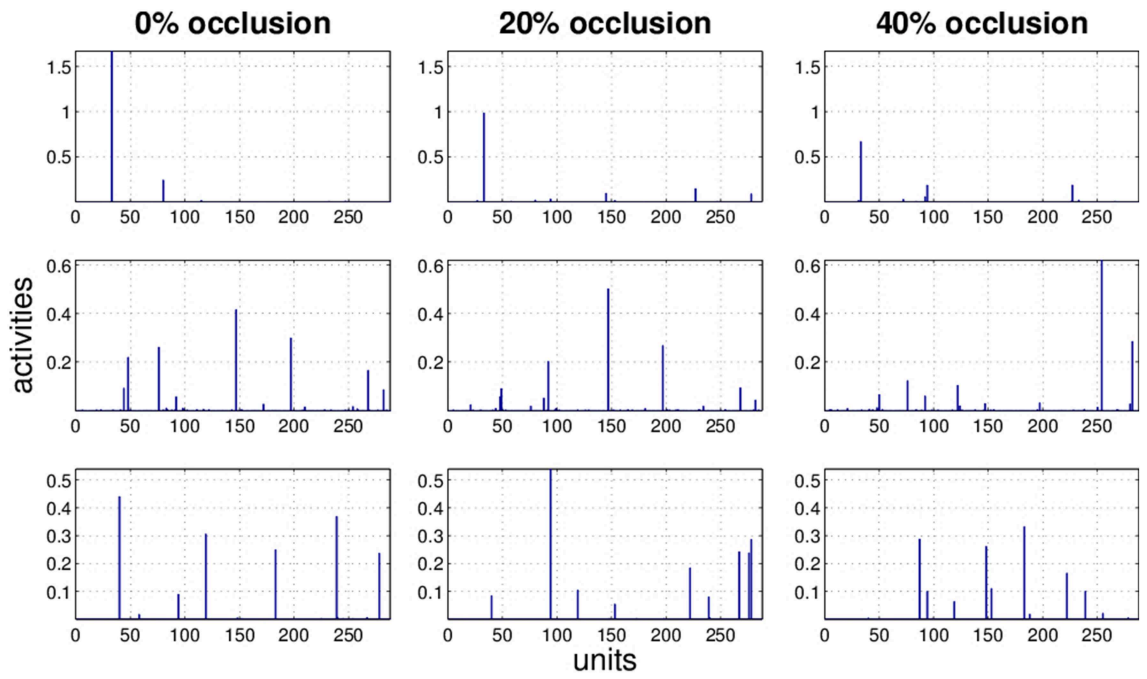


FIGURE 10 | Three examples (row) how the activity pattern varies, under 0, 20, and 40% of occlusion (column) in PC/BC.

occlusion, its classification accuracy is a bit weaker than the one of PC/BC for larger occlusions. The two dominant methods in this study, the HNN and PC/BC, employ different mechanisms for competition. These mechanisms help the systems to selectively inhibit the output of other units or respectively their input. In order to observe how much competition enhances robustness under occlusion, we have evaluated the classification performance when the competitive mechanisms were turned

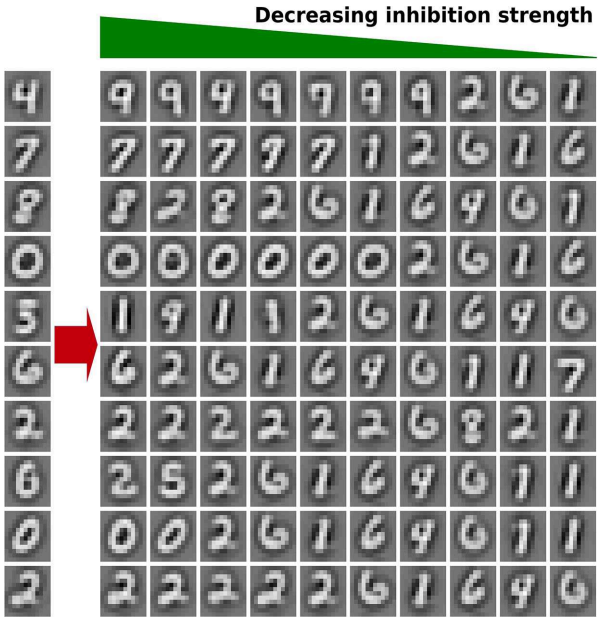


FIGURE 11 | Selective inhibition in the HNN. On the left side the feedforward weights of 10 randomly chosen neurons are illustrated. Right of each neuron, the weights of 10 neurons receiving inhibition from this neuron are plotted, sorted from left to right by descending lateral weight strength (inhibition). The illustration shows that neurons having more similar feedforward weights are more inhibited than neurons having less similar weights.

TABLE 1 | Cosine between non-occluded and occluded activity patterns, calculated on the test set with having particular occlusion levels.

	20% Occlusion	40% Occlusion
FastICA	0.65	0.46
NMFSC	0.71	0.61
PC/BC	0.78	0.61
HNN	0.87	0.76

A cosine of 1 denotes an equal direction and 0 denotes an orthogonal one. The stability of the activity patterns conform the results for the recognition accuracy, except the HNN shows a higher stability.

off. When the mechanisms are off, PC/BC and the HNN show a very low performance in the robustness to occlusions, as NMFSC without using the sparseness constraint. So obviously, the feedforward processing is not enough to obtain a sufficiently differentiated output and it can be assumed that competition is playing an essential role in the robustness of these systems.

We also observed that methods benefiting from a competitive mechanism are superior to FastICA, having no competitive mechanism on the output computation. FastICA linearly transfers the input space into a new space with least dependent components. When facing an image, FastICA produces a dense set of activities to describe the image in the new space. NMFSC without sparseness constraint acts as FastICA. However, when a reasonable level of sparseness is set for the activities of NMFSC it outperforms FastICA. The reason is that the sparseness constraint omits the appearance of redundant information to some extent. Indeed, a too sparse representation can remove some useful information and resulting in reduced accuracy. However, NMFSC acts weaker than the Hebbian neural network and PC/BC which may depend on the subtractive updating rule for output competition. Moreover, the optimal sparseness level is practically impossible, since a priori knowledge about the number features for an optimal representation is needed (Spratling, 2006). Also having this knowledge does not have to lead to an optimal result as different classes often need different amounts of features.

Among the three generative models FastICA, NMFSC, and PC/BC, PC/BC has been the superior model in this experiment. It uses a multiplicative updating rule to calculate the output activity. It finds the best matching units and removes their representations from the input of the other units, producing a sparse output while approaching a minimal reconstruction error. This online error minimization is realized by iteratively updating the error units representing the local elements of the reconstruction error and driving the output units. The HNN also has a competitive mechanism according to which the best matching units suppress other ones. In contrast to PC/BC, which tries to minimize the reconstruction error, the Hebbian Neural Network, as a whole, does not approach any explicit objective. It only exploits the knowledge from the training phase about the coactivity of units in order to suppress them. Thus, stronger units suppress potentially confusing weaker ones. That is, in HNN each unit is competing with other units based on its learned, local inhibitory weights, whereas PC/BC is actively using its distributed representation of the reconstruction error to minimize a global error signal. This may be the reason for the slight advantage of PC/BC against HNN for larger occlusions.

We conclude, that in order to achieve high robustness against loss of information in object recognition, one should focus on improving the competitive mechanism. Competition between units seems to play an important role in preventing the system from producing redundant activities. The experiments give also evidence that the cortical mechanisms of competition, as lateral inhibition, are the source of its robust recognition performance, even on single layer level. Similar effects to our V1 based evaluation can be found in deeper models of the visual cortex ventral stream, where even inhibitory lateral connections play an important role in robustness to occlusions (O'Reilly et al., 2013).

Funding

This work has been supported by the German Academic Exchange Service (DAAD) and the German Research Foundation (DFG GRK1780/1). The publication costs of this article were funded by the German Research Foundation/DFG (INST 270/219-1) and the Chemnitz University of Technology in the funding program Open Access Publishing.

References

Bell, A. J., and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vis. Res.* 37, 3327–3338. doi: 10.1016/S0042-6989(97)00121-1

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

Falconbridge, M. S., Stamps, R. L., and Badcock, D. R. (2006). A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images. *Neural Comput.* 18, 415–429. doi: 10.1162/089976606775093891

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* 237, 55–56.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* 5, 1457–1469. Available online at: http://dl.acm.org/citation.cfm?id=1044709

Hoyer, P. O., and Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network* 11, 191–210. doi: 10.1088/0954-898X/11/3/302

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hyvärinen, A., and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Comput.* 9, 1483–1492. doi: 10.1162/neco.1997.9.7.1483

Kompass, R. (2007). A generalized divergence measure for nonnegative matrix factorization. *Neural Comput.* 19, 780–791. doi: 10.1162/neco.2007.19.3.780

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *J. Math. Biol.* 15, 267–273.

Olshausen, B., and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1007/BF00275687

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311–3325. doi: 10.1016/S0042-6989(97)00169-7

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., and Jilk, D. J. (2013). Recurrent processing during object recognition. *Front. Psychol.* 4:124. doi: 10.3389/fpsyg.2013.00124

Rehn, M., and Sommer, F. T. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J. Comput. Neurosci.* 22, 135–146. doi: 10.1007/s10827-006-0003-9

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88, 455–463. doi: 10.1152/jn.00881.2001

Spratling, M. W. (2006). Learning image components for object recognition. *J. Mach. Learn. Res.* 7, 793–815. Available online at: http://dl.acm.org/citation.cfm?id=1248575

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *J. Neurosci.* 30, 3531–3543. doi: 10.1523/JNEUROSCI.4911-09.2010

Spratling, M. W., De Meyer, K., and Kompass, R. (2009). Unsupervised learning of overlapping image components using divisive input modulation. *Comput. Intell. Neurosci.* 2009:381457. doi: 10.1155/2009/381457

Teichmann, M., Wiltschut, J., and Hamker, F. H. (2012). Learning invariance from natural images inspired by observations in the primary visual cortex. *Neural Comput.* 24, 1271–1296. doi: 10.1162/NECO_a_00268

Turrigiano, G. G., and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.* 5, 97–107. doi: 10.1038/nrn1327

van Hateren, J. H., and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci.* 265, 359–366. doi: 10.1098/rspb.1998.0303

Wiltschut, J., and Hamker, F. H. (2009). Efficient coding correlates with spatial frequency tuning in a model of V1 receptive field organization. *Vis. Neurosci.* 26, 21–34. doi: 10.1017/S0952523808080966

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Comput. Biol.* 7:e1002250. doi: 10.1371/journal.pcbi.1002250

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Kermani Kolankeh, Teichmann and Hamker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix

Computing Neural Activity in the HNN

As in Teichmann et al. (2012), the membrane potential m_j of a neuron j is calculated as the sum of each neuron's pre-synaptic input values r_i^{Input} weighted by the corresponding synaptic strengths w_{ij}. The resulting sum was decreased by the amount of inhibition from other layer neurons r_k, weighted by the corresponding synaptic strengths of the lateral connections c_{kj} (Equation A1) and a non-linearity function f (Equation A2). In contrast to Teichmann et al. (2012), we add an additional decay term $\bar{r}_j$ which mimic the intrinsic adaption of the firing threshold (Turrigiano and Nelson, 2004) based on the temporal activity of the neuron (Equation A3). The resulting activation r_j of a neuron j is calculated using the top half rectified membrane potential $(m_j)^+$. Further, we apply a saturation term for high membrane potentials to avoid unrealistic high activations (Equation A4).

The change of the activity is described by

$$\tau_m \frac{\partial m_j}{\partial t} = \sum_i w_{ij} r_i^{Input} - \sum_{k,k\neq j} f\left(c_{kj} r_k\right) - \bar{r}_j - m_j \tag{A1}$$

with $\tau_m = 10$ and the non-linearity function

$$f(x) = d_{nl} \cdot \log\left(\frac{1+x}{1-x}\right) \tag{A2}$$

and the temporal activity

$$\tau_{\bar{r}} \frac{\partial \bar{r}_j}{\partial t} = r_j - \bar{r}_j \tag{A3}$$

with $\tau_{\bar{r}} = 10000$ and the transfer function

$$r_j = \begin{cases} 0.5 + \frac{1}{1+e^{-3.5(m_j-1)}} & \text{if } m_j > 1 \\ (m_j)^+ & \text{else} \end{cases} \tag{A4}$$

Changes in Neural Learning of the HNN

As in Teichmann et al. (2012), we use Oja's constraint (Oja, 1982) for normalizing the length of the weight vector, preventing an infinite increase of weights. In contrast to Oja, each neuron can have an individual weight vector length. Differently to our previous work, we calculate the factor α, determining the length, so that its change is based only on the squared membrane potential minus a fixed average membrane potential $\beta = \frac{1}{288}$. The value β is defined as 1 divided by amount of neurons in the layer (Equation A5).

$$\tau_\alpha \frac{\partial \alpha_j}{\partial t} = m_j^2 - \beta \tag{A5}$$

Since the learning rule was originally proposed for learning complex cells, the learning is based on calcium traces, following the activity of a neuron, to allow exploiting the temporal structure of the input. As this is not needed for learning simple cells or handwritten digits, we are using here a short time constant of $\tau_{Ca} = 10$ for the calcium trace. The time constant is chosen that short to turn off the influence of previous stimuli on the learning result. Besides, we have shown in Teichmann et al. (2012) that using this time constant in a setup for learning complex cells, causes a huge amount of the cells with simple cell properties. All other parameters in this model are chosen as in Teichmann et al. (2012).

frontiers in
COMPUTATIONAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 17 July 2014
doi: 10.3389/fncom.2014.00071

From image processing to computational neuroscience: a neural model based on histogram equalization

Marcelo Bertalmío*

Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Luis M. Martinez, Spanish National Research Council, Spain
Michael E. Rudd, University of Washington, USA

***Correspondence:**
Marcelo Bertalmío, Departament de Tecnologies de l'Informació i les Comunicacions, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain
e-mail: marcelo.bertalmio@upf.edu

There are many ways in which the human visual system works to reduce the inherent redundancy of the visual information in natural scenes, coding it in an efficient way. The non-linear response curves of photoreceptors and the spatial organization of the receptive fields of visual neurons both work toward this goal of efficient coding. A related, very important aspect is that of the existence of post-retinal mechanisms for contrast enhancement that compensate for the blurring produced in early stages of the visual process. And alongside mechanisms for coding and wiring efficiency, there is neural activity in the human visual cortex that correlates with the perceptual phenomenon of lightness induction. In this paper we propose a neural model that is derived from an image processing technique for histogram equalization, and that is able to deal with all the aspects just mentioned: this new model is able to predict lightness induction phenomena, and improves the efficiency of the representation by flattening both the histogram and the power spectrum of the image signal.

Keywords: neural model, Wilson-Cowan equation, efficient coding, redundancy reduction, contrast enhancement, lightness induction

1. INTRODUCTION

The human visual system works in many ways in order to efficiently encode the visual information coming from natural environments, reducing its inherent redundancy, as proposed in the seminal work of Barlow (1961) (see Olshausen and Field, 2000 for a review). For instance, while natural scenes have luminance distributions which are very lopsided, with a high peak and a very rapid fall-off, photoreceptors encode this information with signals that have a much more even distribution: indeed, photoreceptors perform histogram equalization, as demonstrated by Laughlin (1981). And the receptive fields of visual neurons, both retinal and post-retinal, compensate the $1/f^2$ decay of the power spectrum of natural images, whitening the spectrum of the resulting signal and thus minimizing interpixel redundancies and increasing coding efficiency (see Atick, 1992; Dan et al., 1996 where the existence of whitening at the local geniculate nucleus is demonstrated for natural images).

Apart from efficiency in coding, another very important aspect is that of biological efficiency in terms of *wiring*. The resolution of retinal mosaics is limited by the number of axons that can pass through the optic nerve, which acts as a bottleneck (Olshausen, 2003). But the visual system is able to achieve a visual acuity beyond the limit imposed by the number of photoreceptors at the retina: in their classical paper on contrast constancy, Georgeson and Sullivan (1975) suggest that there are cortical mechanisms for contrast enhancement that compensate for the blurring produced in early stages of the visual process. Very recently Martinez et al. (2014) have confirmed that contrast enhancement takes place at the lateral geniculate nucleus (LGN) and, remarkably, the authors point out that this contrast enhancement procedure is very much alike the common techniques used in image processing.

Alongside mechanisms for coding and wiring efficiency, there is neural activity in region $V1$ of the human visual cortex that correlates with the perceptual phenomenon of lightness induction, as proven by Pereverzeva and Murray (2008). The term *lightness induction* or *achromatic induction* designates the visual phenomenon by which the perceived reflectance of an object depends on its surround. It can take the form of *lightness contrast*, when the object's lightness shifts away from that of its surroundings: a dark object on a light background appears even darker, or a light object in a dark surround becomes even lighter. The reverse is called *lightness assimilation*, in which case the appearance of the object shifts in the direction of the lightness of its surround. As pointed out by Shevell (2003), lightness assimilation occurs in situations of high spatial frequency while lightness contrast is associated with relatively lower spatial frequencies.

Our contribution in this paper is to propose a neural activity model, a partial differential equation (PDE) in the form of a Wilson-Cowan equation (Wilson and Cowan, 1972), which takes care simultaneously of the four aspects mentioned above: it performs histogram equalization, spectrum whitening, contrast enhancement, and it also predicts lightness induction. The proposed model is based on a state of the art method for color and contrast enhancement from the image processing literature, so we start the following section reviewing some key image processing concepts.

2. IMAGE PROCESSING FOR CONTRAST ENHANCEMENT

2.1. HISTOGRAM EQUALIZATION

Histogram equalization is a classical, very basic image processing technique dating at least to the early 1970s (see Pratt, 2007 and references therein), aiming at enhancing the contrast and

improving the appearance of images by way of re-distributing their levels uniformly accross the available range. In this sense, an image would be optimal if its histogram were flat or "equalized," meaning that all the range is used and all levels are represented by the same amount of pixels. Therefore, when an image has a flat histogram its cumulative histogram is simply a ramp, and this allows for a very straightforward computation for the histogram equalization procedure: assuming we are working on a graylevel image in the range [0,1], we have to substitute each level g in the original image by the value of its normalized cumulative histogram, $H(g)$. The solution is computed very fast using a look-up table (LUT). An example result can be seen in **Figure 1** (notice that, while the range has been expanded and the resulting image has a more even histogram, it's not actually uniform).

While in **Figure 1** histogram equalization improves the visual appearance of the image, **Figure 2** shows an example where the image is actually made worse, which Pratt (2007) points out is often the case when the image is overexposed, as it is here. This is aggravated by the fact that the equalization procedure is a one-shot technique, that only produces a final result, without any "in-between," so if the resulting image shows any type of unpleasant artifact there is nothing to do about it. This issue was addressed by Sapiro and Caselles (1997), who proved that the minimization of the energy functional

$$E(I) = 2\sum_x \left(I(x) - \frac{1}{2}\right)^2 - \frac{1}{AB}\sum_x \sum_y |I(x) - I(y)| \qquad (1)$$

produces an image I with a flat histogram. The range of I is [0, 1], x, y are pixels and A, B are the image dimensions. While the result of histogram equalization is very often unsatisfactory and can't be altered, Sapiro and Caselles (1997) propose to start with an input image I_0 and apply to it step after step of the minimization of Equation 1, letting the user decide when to stop. If the user lets the minimization run to convergence, she'll get the same result as with a LUT, but otherwise a better result can be obtained if the iterative procedure stops before the appearance of severe artifacts. The squared differences in the first term of Equation 1 and the absolute differences in the second one are required to ensure that the minimization yields an image with equalized histogram, see Sapiro and Caselles (1997) for details. The energy in Equation 1 can be interpreted as the difference between two positive and competing terms,

$$E(I) = D(I) - C(I). \qquad (2)$$

The first term measures the dispersion around the average value of $\frac{1}{2}$, as in the *gray world* hypothesis for color constancy, stating that our visual system estimates the illuminant as one half the average of the colors of the scene, an observation made by Judd (1940, 1979a) and formalized by Buchsbaum (1980). The second term measures the contrast as the sum of the absolute value of the pixel differences.



FIGURE 1 | Left: image and associated histogram. **Right:** after histogram equalization.



FIGURE 2 | Left: original image. **Right:** after histogram equalization.

2.2. PERCEPTUALLY-BASED CONTRAST ENHANCEMENT

The abovementioned measure of contrast is global, not local, i.e., the differences are computed regardless of the spatial locations of the pixels. This is not consistent with how we *perceive* contrast, which is in a localized manner, at each point having neighbors exert a higher influence than far-away points. Using the concepts introduced by the popular perceptually-based color correction method ACE of Rizzi et al. (2003), the authors of Bertalmío et al. (2007) propose an adapted version of the functional of Equation 1 that complies with some very basic visual perception principles, namely those of locality, color constancy and *white patch* (the latter stating that the brightest spot in the image is perceived as white, an observation that is often attributed, incorrectly, to the Retinex theory of Land (1977), but which has a long history that dates back at least to the works of Helmholtz, as explained by Judd (1979b,c)):

$$E(I) = \frac{\alpha}{2}\sum_x \left(I(x) - \frac{1}{2}\right)^2 - \gamma \sum_x \sum_y w(x,y)|I(x) - I(y)| + \frac{\beta}{2}\sum_x (I(x) - I_0(x))^2, \qquad (3)$$

where w is a distance function such that its value decreases as the distance between x and y increases, I_0 is the original image and α, β and γ are positive weights (which can be chosen so as to guarantee the white patch property, see Bertalmío et al., 2007 for details). The gradient descent equation for the functional in Equation 3 is the following, and its numerical implementation is essentially equivalent to the method of Rizzi et al. (2003):

$$I_t(x) = -\alpha\left(I(x) - \frac{1}{2}\right) + \gamma \sum_y w(x,y) sgn(I(x) - I(y)) - \beta(I(x) - I_0(x)). \qquad (4)$$

Starting from $I = I_0$, we iterate Equation 4 until we reach a steady state, which will be the result of this algorithm.

By minimizing the energy in Equation 3 we are locally enhancing contrast (second term) and promoting color constancy by discounting the illuminant (first term), while preventing the image from departing too much from its original values (third term). We could also say that the minimization of Equation 3 approximates *local* histogram equalization.

The method of Bertalmío et al. (2007) has several good properties:

1. It yields very good color constancy results, being able to remove strong color casts and to deal with non-uniform illumination (a challenging scenario for most color constancy algorithms, see Bertalmío, 2014).
2. It increases the dynamic range of the image (i.e., it tends to "flatten" its histogram).
3. It has a very good local contrast enhancement performance, producing results without halos, spurious colors or any other kind of visual artifact.
4. It can deal with both underexposed and overexposed pictures.
5. It reproduces visual perception phenomena such as simultaneous contrast and the Mach Band effect.

But regarding color constancy, there is also a very interesting and close connection with the classical approach of Retinex. In their kernel-based Retinex (KBR) formulation, Bertalmío et al. (2009) take all the essential elements of the Retinex theory of Land (1977) (channel independence, the ratio reset mechanism, local averages, non-linear correction) and propose an implementation that is intrinsically 2D, and therefore free of the issues associated with the 1D paths used in the original Retinex algorithm. The results obtained with this algorithm comply with all the expected properties of Retinex (such as performing color constancy while being unable to deal with overexposed images) but don't suffer from the usual shortcomings such as sensitivity to noise, appearance of halos, etc. In Bertalmío et al. (2009) it is proven that there isn't any energy that is minimized by the iterative application of the KBR algorithm, and this fact is linked to its limitations regarding overxposed pictures. Using the analysis of contrast performed by Palma-Amestoy et al. (2009), Bertalmío et al. (2009) are able to determine how to modify the basic KBR equation so that it can also handle overexposed images, and the resulting, modified KBR equation turns out to be essentially the gradient descent of the energy given by Equation 3. In other words, the method of Bertalmío et al. (2007) can be seen as an iterative application of Retinex, although in a modified version that allows to produce good results also in the case of overexposed images.

3. A NEW NEURAL MODEL

3.1. CONNECTION WITH NEUROSCIENCE

The activity of a population of neurons in the region $V1$ of the visual cortex evolves in time according to the Wilson-Cowan equations (see Wilson and Cowan, 1972, 1973; Bressloff et al., 2002). Treating $V1$ as a planar sheet of nervous tissue, the state $a(r, \phi, t)$ of a population of cells with cortical space coordenates $r \in \mathbb{R}^2$ and orientation preference $\phi \in [0, \pi)$ can be modeled with the following PDE (Bressloff et al., 2002):

$$\frac{\partial a(r,\phi,t)}{\partial t} = -\alpha a(r,\phi,t) + \mu \int_0^{\pi} \int_{\mathbb{R}^2} \omega(r,\phi \| r',\phi') \sigma(a(r',\phi',t)) dr' d\phi' + h(r,\phi,t), \quad (5)$$

where α, μ are coupling coefficients, $h(r, \phi, t)$ is the external input (visual stimuli), $\omega(r, \phi \| r', \phi')$ is a kernel that decays with the differences $|r - r'|, |\phi - \phi'|$ and σ is a sigmoid function. If we ignore the orientation ϕ and assume that the input h is constant in time, it can be shown that Equation 5 is closely related to the gradient descent Equation 3, where neural activity a plays the role of image value I, sigmoid function σ behaves as the derivative of the absolute value function, and the visual input h is the initial image I_0. This connection was already pointed out by Bertalmío et al. (2007), and Bertalmío and Cowan (2009) use it to argue that the Wilson-Cowan equations could therefore be the gradient descent of a certain energy, and also that there would appear to be a physical substrate at the cortex for the Retinex theory.

3.2. LIGHTNESS INDUCTION

Looking closely at Equation 4, we can see that the spatial arrangement of the image data plays no role in it. Therefore, we can expect that the local contrast enhancement procedure of Bertalmío et al. (2007) will always produce lightness contrast, not assimilation, since as we mentioned earlier assimilation is linked to high spatial frequencies (Shevell, 2003). **Figure 3** confirms this: (**Figure 3A**) produces lightness assimilation, because all gray bars have the same value but they are perceived darker when surrounded by black and lighter when surrounded by white; on the other hand, the result produced by Bertalmío et al. (2007)

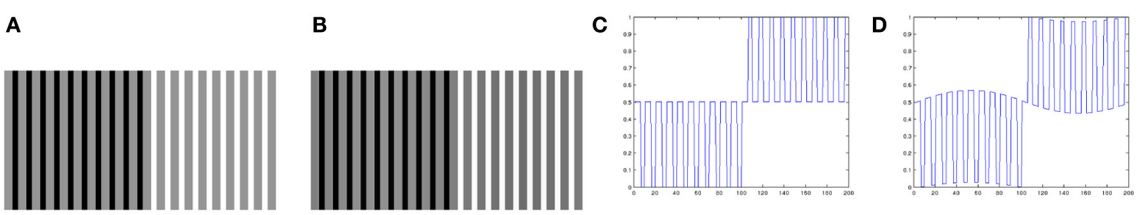


FIGURE 3 | (A) Original image, example of lightness assimilation: the gray bars have all the same value but appear different over black and white backgrounds. **(B)** Result of applying the model of Bertalmío et al. (2007) to image **(A)**. **(C)** Profile of a line from image **(A)**. **(D)** Profile of a line from image **(B)**: notice how the model of Bertalmío et al. (2007) actually emulates lightness contrast rather than assimilation.

in **Figure 3B** actually emulates lightness contrast rather than assimilation, as the line profiles in **Figures 3C,D** show.

Rudd (2010) studies lightness induction using a disk-and-ring (DAR) display for matching experiments, see **Figure 4** (left). The intensity of the background B, of the left ring R_M and of the disk on the right D_T is kept constant; the intensity of the right ring R_T is modified, and the observer has to adjust the intensity of the left disk D_M so as to match the appearance of the right disk D_T. Using the model of Bertalmío et al. (2007), the predicted value of D_M as a function of the varying R_T can be computed, and it is shown in **Figure 4** (right). We can see that as R_T increases D_M always decreases, so according to this model we should only have lightness contrast in this situation. But the data from the perceptual experiments of Rudd (2010) says otherwise, see **Figure 5**: as R_T increases D_M also increases (lightness assimilation) until R_T reaches some value, beyond which D_M decreases (lightness contrast). These plots are well approximated by parabolas and, as the ring widths become larger, the resulting parabolas have their curvature decrease, implying that "assimilation is more likely to be observed with narrow surrounds" (Rudd, 2010).

3.3. PROPOSED MODEL

In order to overcome the intrinsic limitations of Bertalmío et al. (2007) with respect to lightness induction, we should introduce spatial frequency in the energy functional. We propose a new model consisting in the following PDE, a modification of the gradient descent Equation (4):

$$I_t(x) = -\alpha(I(x) - \mu(x)) + \gamma(1 + (\sigma(x))^c) \sum_y w(x, y) sgn(I(x) - I(y)) - \beta(I(x) - I_0(x)), \quad (6)$$

where $\mu(x)$ is the mean average of the original image data computed over a neighborhood of x, $\sigma(x)$ is the standard deviation of the image data computed over a small neighborhood of x, and the exponent c is a positive constant. The differences with respect to Equation 4 are that now the average in the first term is no longer global (the 1/2 value of Equation 4) but local, and that the weight for the second term is no longer a constant, but it changes both spatially and with each iteration, according to the local standard deviation σ: if the neighborhood over which it is computed is sufficiently small, standard deviation can provide a simple estimate of spatial frequency. But also, the standard deviation is commonly used in the vision literature as an estimate of local contrast. We have this contrast $\sigma(x)$ raised to a power c, and this is also the case with other neural models where a power law is applied to the contrast, as we will briefly discuss later.

Again, this is a Wilson-Cowan type of neural activity model, where I_0 is the visual input. We take I_0 as a non-linear modification of the radiance stimulus, e.g., I_0 could be the result of applying the Naka-Rushton equation, which models photoreceptor responses (see Shapley and Enroth-Cugell, 1984), to the radiance stimuli. As we did with Equation 4, we start with an image $I = I_0$ and iterate Equation 6 until convergence, obtaining a result which we'll see is able to predict perceptual phenomena as well as improve the efficiency of the representation.

4. THE PROPOSED MODEL PREDICTS INDUCTION AND IMPROVES EFFICIENCY

4.1. PREDICTING LIGHTNESS INDUCTION

Using this new model, now we can qualitatively predict the results of Rudd (2010). We fix B, D_T, R_M and for each value of R_T we

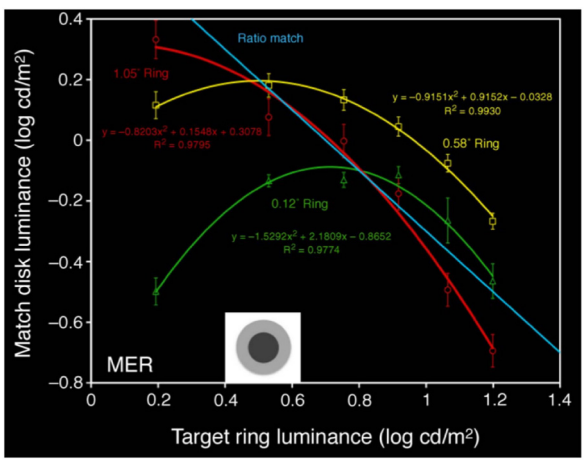


FIGURE 5 | Value of D_M as a function of R_T: results of perceptual matches for different ring widths, from Rudd (2010).

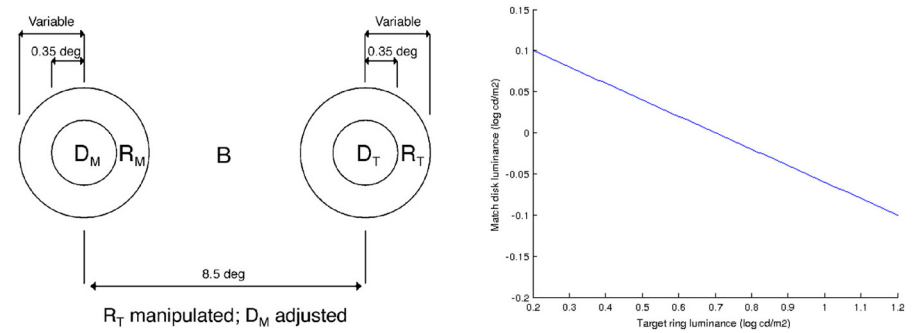


FIGURE 4 | Left: diagram of the disk-and-ring display used by Rudd (2010), taken from that paper. **Right:** value of D_M as a function of R_T, predicted with the model of Bertalmío et al. (2007).

find the steady state of Equation 6 at the center of the right disk, at the middle of the right ring and at the middle of the left ring: these would be our model's predictions of the perceived values for D_T, R_T and R_M. Next, we compute the difference between the first two values and add it to the third, yielding the prediction of the perceived (lightness) value for D_M, from which we can recover the actual luminance value D_M using again Equation 6 (see Appendix for implementation details). From these results we can derive the following conclusions, that corroborate the findings of the perceptual experiments of Rudd (2010):

- As shown in **Figure 6** (left), the predicted match luminance plots are no longer linear but quadratic, with an initial lightness assimilation regime for low values of R_T followed by a lightness contrast part.
- The curvature of these parabolas decreases with increasing ring width.
- The previous experiments were for a *double-decrement* display: $B > R_T > D_T$. If we now make the ring have a value in between disk and background, $B > R_T < D_T$, the plot curvature remains negative as in the *double-decrement* case, see **Figure 6** (right). As reported by Rudd (2010), this behavior can not be predicted with some other neural models, like that of Rudd and Arrington (2001), with which the curvature changes sign in this situation.

We can also predict lightness assimilation in the previous example of the alternating gray bars of **Figure 3**, as we now show in **Figure 7**.

It is interesting to note that the shape of the curves in **Figure 6** does vary with extent of the neighborhood over which the standard deviation is computed, as **Figure 8** shows: when the neighborhood covers disk, ring and some background we have an inverted parabola as before (red curve), but if we decrease the neighborhood size so that it only covers disk and ring but no background then the parabola concavity is reversed (green curve), and if the neighborhood is further reduced so that it only covers the disk then the curve is no longer parabolic but linear.

Finally, we may point out that some recent models which also predict lightness induction based on neural attributes of the visual system can be found in Otazu et al. (2008) and Penacchio et al. (2013).

4.2. EFFICIENCY: REDUNDANCY REDUCTION AND CONTRAST ENHANCEMENT

In this section we argue that the proposed model of Equation 6, which as we have seen has the form of a Wilson-Cowan equation, performs local contrast enhancement and is closely related to basic image processing techniques, is a good candidate for a neural model providing the contrast constancy effects described by Georgeson and Sullivan (1975) and Martinez et al. (2014). But furthermore we will now see how this new neural model, applied to signals already encoded by photoreceptors, further improves efficiency by reducing redundancy: flattening the histogram and whitening the power spectrum.

Figure 9A shows a high dynamic range (HDR) image or *radiance map*, linearly scaled to the range [0, 1]. Clearly this kind of

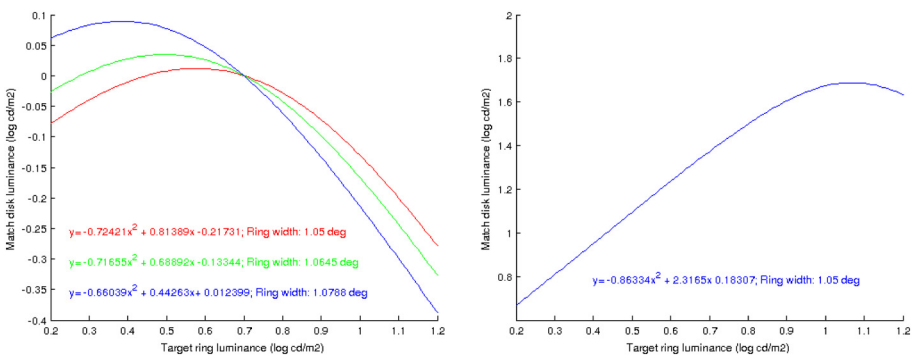


FIGURE 6 | Left: predicted value of D_M as a function of R_T, using the proposed model, for different ring widths; the plots are well approximated by parabolas, whose curvature decreases as the ring width increases. **Right:** prediction when the disk is a luminance increment with respect to the ring; the sign of the curvature remains negative.

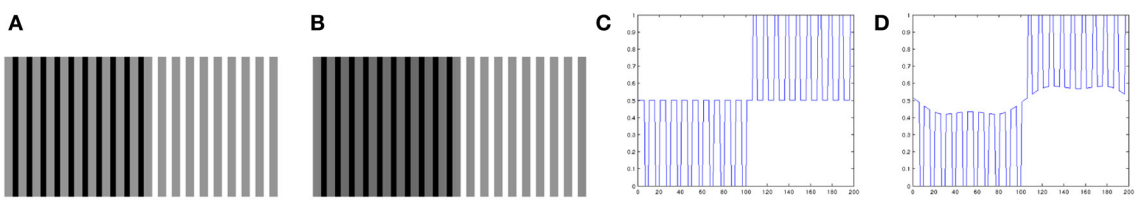


FIGURE 7 | (A) Original image, example of lightness assimilation: the gray bars have all the same value but appear different over black and white backgrounds. **(B)** Result of applying the proposed model to image **(A)**. **(C)** Profile of a line from image **(A)**. **(D)** Profile of a line from image **(B)**: notice how the proposed model is capable of emulating lightness assimilation.

mapping is useless, which is a way to explain the need for light adaptation and gain control mechanisms in our photoreceptors. **Figure 9B** shows the result of applying the Naka-Rushton equation to the previous HDR image. **Figure 9C** shows the result of applying our proposed model to **Figure 9B**. As we can see, the original radiance image has a very lopsided histogram, which is made considerably more uniform by applying the Naka-Rushton equation and even more flat if we apply our proposed method to the Naka-Rushton output. Local contrast is clearly enhanced as well, see for instance the window frames, the book cases behind the windows, etc. For the implementation details we refer the reader to the Appendix.

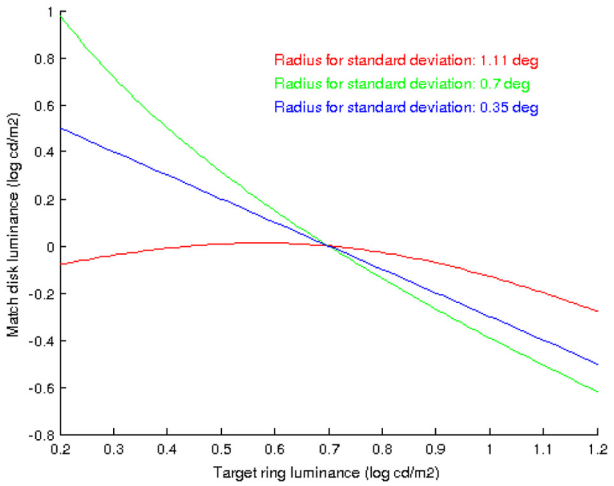


FIGURE 8 | The shape of the curves that predict the value of D_M as a function of R_T, using the proposed model, depend on the extent of the neighborhood over which the standard deviation is computed. In red: the neighborhood covers disk, ring and some background. In green: the neighborhood covers disk and ring but no background. In blue: the neighborhood covers just the disk.

Figure 10A shows the result of applying the Naka-Rushton equation to a high dynamic range image. **Figure 10B** shows the result of applying the model of Bertalmío et al. (2007) to **Figure 10A** (this is roughly equivalent to the tone mapping approach proposed by Ferradans et al. (2011) in an image processing/computer graphics context). **Figure 10C** shows the result of applying our proposed model to **Figure 10A**. **Figure 10D** compares the power spectrum of the three previous images. We can see that our model improves spectrum whitening over the other two results. In this image the contrast enhancement is more subtle but still noticeable, especially in the interior of the tree-trunk and on the leaves and grass in the foreground.

An interesting aspect is given by the constant c in Equation 6 and its relationship to the whitening of the power spectrum. Given a Naka-Rushton output, we compute the rotational average of its power spectrum which, in log-log coordinates, can be fit by a line with a certain slope. We do the same for the output of our model, that has been applied to the Naka-Rushton output using some value for c, and obtain a new linear fit with a new slope. Let us estimate the "increase in whitening" provided by our model as the difference between these two slopes, call it W. The value of W is a function of the constant c used in our model. If we now vary c in the interval $[0, 1]$ we can plot the resulting function $W(c)$, as shown in **Figure 11**. Disregarding the spikes for low values of c, we can see that there is an optimum value for c, with which we can obtain the maximum power spectrum whitening that our model can provide. In our model c is the power to which we raise the local standard deviation $\sigma(x)$, and this standard deviation is one of the possible measures that are commonly used to estimate local

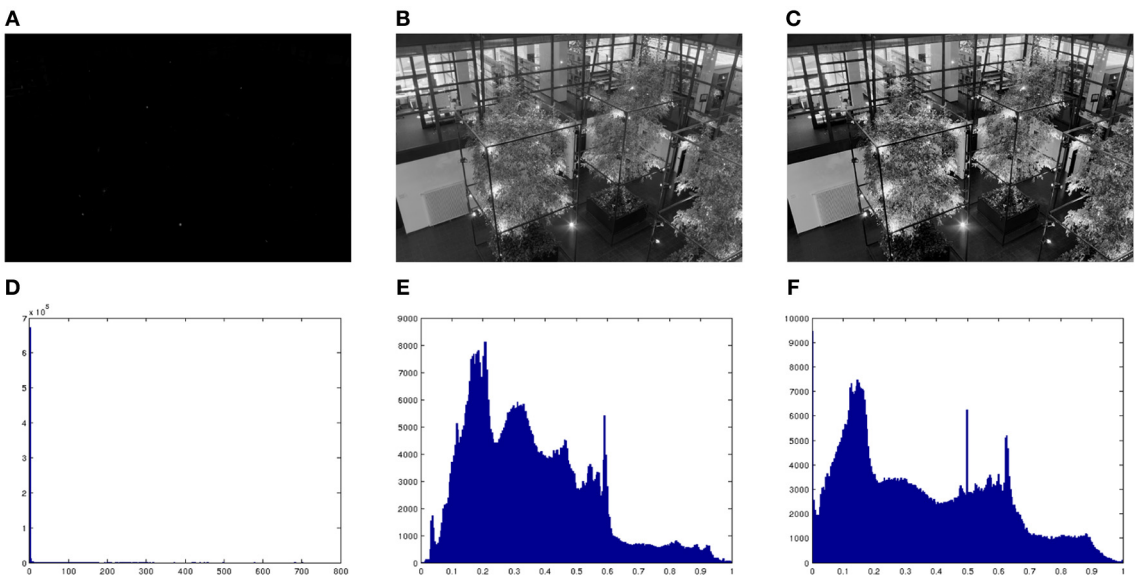


FIGURE 9 | (A) High dynamic range (HDR) image, linearly scaled. **(B)** Result of applying the Naka-Rushton equation to the HDR image. **(C)** Result of applying proposed model to image **(B)**. **(D)** Histogram of **(A)**. **(E)** Histogram of **(B)**. **(F)** Histogram of **(C)**. Original image courtesy of Max Planck Institute.

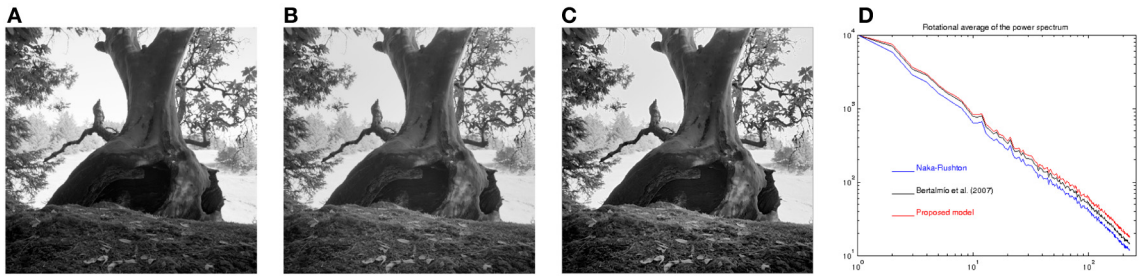


FIGURE 10 | (A) Result of applying the Naka-Rushton equation to a high dynamic range image. **(B)** Result of applying the model of Bertalmío et al. (2007) to image **(A)**. **(C)** Result of applying proposed model to image **(A)**. **(D)** Power spectrum of images **(A–C)**. Original image property of Industrial Light and Magic.

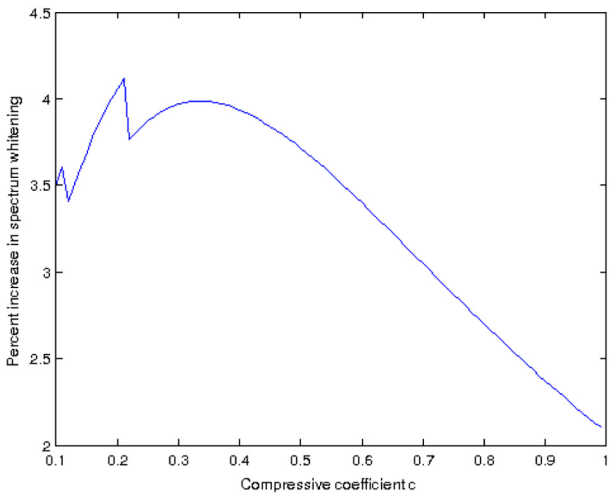


FIGURE 11 | Increase in spectrum whitening as a function of the constant *c* in Equation 6, for the image in Figure 10.

contrast. Because of this reason we are currently investigating the possible connections of our model with the works of Mante et al. (2008) and Kay et al. (2013), since both of them apply a static power-law non-linearity to the contrast. In particular, Kay et al. (2013) computed the value of the exponent of this power law and found that while it varies accross the visual cortex, it is in the range [0, 0.35] with a value of around 1/3 in the case of $V1$: this is all consistent with the tests we have performed so far for different high dynamic range images.

5. CONCLUSION AND FUTURE WORK

In this paper we have proposed a neural model, in the form of a Wilson-Cowan equation, that is derived from an image processing technique for local histogram equalization. This new model is able to predict lightness induction phenomena, and improves the efficiency of the representation by flattening both the histogram and the power spectrum of the image signal and increasing local contrast.

We are very much interested in finding evidence of neural responses following our proposed model. Our method performs contrast enhancement, so we would like to explore whether there is any relationship with the work of Martinez et al. (2014), who have very recently confirmed that contrast enhancement takes place at the LGN and is much alike the common techniques used in image processing. Our model has a term where a power law is applied to the contrast, and we can optimize the exponent of this power law so as to maximize the whitening of the spectrum; for the limited tests that we have performed so far, our results appear to be in agreement with what is reported by Mante et al. (2008) and Kay et al. (2013), so we also want to investigate possible connections with those works. And as immediate future work, we will extend our formulation to the color case in order to predict color induction as well.

Last but not least, we believe we can use our proposed model to go back to some image processing and computer vision applications, which could benefit from the insights gained in the visual neuroscience domain. In particular, we are currently working in extending this new model for problems such as tone mapping, gamut mapping and computational color constancy.

ACKNOWLEDGMENTS

This work was supported by European Research Council, Starting Grant ref. 306337, by the ICREA Acadèmia Award and by Spanish grants ref. TIN2011-15954-E and ref. TIN2012-38112.

REFERENCES

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network* 3, 213–251. doi: 10.1088/0954-898X/3/2/009

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sens. Commun.* 217–234.

Bertalmío, M. (2014). *Image Processing for Cinema*. Boca Raton, FL: CRC Press; Taylor and Francis.

Bertalmío, M., Caselles, V., and Provenzi, E. (2009). Issues about retinex theory and contrast enhancement. *Int. J. Comput. Vis.* 83, 101–119. doi: 10.1007/s11263-009-0221-5

Bertalmío, M., Caselles, V., Provenzi, E., and Rizzi, A. (2007). Perceptual color correction through variational techniques. *IEEE Trans. Image Process.* 16, 1058–1072. doi: 10.1109/TIP.2007.891777

Bertalmío, M., and Cowan, J. (2009). Implementing the Retinex algorithm with Wilson-Cowan equations. *J. Physiol.* 103, 69–72. doi: 10.1016/j.jphysparis.2009.05.001

Bressloff, P., Cowan, J., Golubitsky, M., Thomas, P., and Wiener, M. (2002). What geometric visual hallucinations tell us about the visual cortex. *Neural Comput.* 14, 473–491. doi: 10.1162/089976602317250861

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *J. Frankl. Instit.* 310, 337–350. doi: 10.1016/0016-0032(80)90058-7

Dan, Y., Atick, J. J., and Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* 16, 3351–3362.

Ferradans, S., Bertalmío, M., Provenzi, E., and Caselles, V. (2011). An analysis of visual adaptation and contrast perception for tone mapping. *IEEE Trans. Patt. Analy. Mach. Intelli.* 33, 2002–2012. doi: 10.1109/TPAMI.2011.46

Georgeson, M., and Sullivan, G. (1975). Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol.* 252, 627–656.

Judd, D. (1940). Hue, saturation, and lightness of surface colors with chromatic illumination. *J. Opt. Soc. Am.* 30, 2–32. doi: 10.1364/JOSA.30.000002

Judd, D. (1979a). "The unsolved problem of color perception," in *Contributions to Color Science*, Vol. 545, ed D. L. MacAdam (Washington, DC: National Bureau of Standards), 516–522.

Judd, D. (1979b). "Appraisal of Land's work on two-primary color projections," *Contributions to Color Science*, Vol. 545, ed D. L. MacAdam (Washington, DC: National Bureau of Standards), 471–486.

Judd, D. (1979c). "Color appearance," *Contributions to Color Science*, Vol. 545, ed D. L. MacAdam (Washington, DC: National Bureau of Standards), 539–564.

Kay, K., Winawer, J., Mezer, A., and Wandell, B. (2013). Compressive spatial summation in human visual cortex. *J. Neurophysiol.* 110, 481–494. doi: 10.1152/jn.00105.2013

Land, E. (1977). The Retinex theory of color vision. *Sci. Am.* 237, 108–128. doi: 10.1038/scientificamerican1277-108

Laughlin, S. (1981). A simple coding procedure enhances a neurons information capacity. *Z. Naturforsch* 36:51.

Mante, V., Bonin, V., and Carandini, M. (2008). Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron* 58, 625–638. doi: 10.1016/j.neuron.2008.03.011

Martinez, L., Molano-Mazón, M., Wang, X., Sommer, F., and Hirsch, J. (2014). Statistical wiring of thalamic receptive fields optimizes spatial sampling of the retinal image. *Neuron* 81, 943–956. doi: 10.1016/j.neuron.2013.12.014

Olshausen, B. (2003). "Principles of image representation in visual cortex," in *The Visual Neurosciences*, Vol. 2, eds L. M. Chalupa and J. S. Werner (Cambridge, MA: MIT Press), 1603–1615.

Olshausen, B., and Field, D. (2000). Vision and the coding of natural images. *Am. Sci.* 88, 238–245. doi: 10.1511/2000.3.238

Otazu, X., Vanrell, M., and Alejandro Parraga, C. (2008). Multiresolution wavelet framework models brightness induction effects. *Vis. Res.* 48, 733–751. doi: 10.1016/j.visres.2007.12.008

Palma-Amestoy, R., Provenzi, E., Bertalmío, M., and Caselles, V. (2009). A perceptually inspired variational framework for color enhancement. *IEEE Trans. Patt. Analy. Mach. Intell.* 31, 458–474. doi: 10.1109/TPAMI.2008.86

Penacchio, O., Otazu, X., and Dempere-Marco, L. (2013). A neurodynamical model of brightness induction in v1. *PLoS ONE* 8:e64086. doi: 10.1371/journal.pone.0064086

Pereverzeva, M., and Murray, S. (2008). Neural activity in human v1 correlates with dynamic lightness induction. *J. Vis.* 8:8. doi: 10.1167/8.15.8

Pratt, W. (2007). *Digital Image Processing, 4th Edn*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0470097434

Rizzi, A., Gatta, C., and Marini, D. (2003). A new algorithm for unsupervised global and local color correction. *Patt. Recognit. Lett.* 24, 1663–1677. doi: 10.1016/S0167-8655(02)00323-9

Rudd, M. (2010). How attention and contrast gain control interact to regulate lightness contrast and assimilation: a computational neural model. *J. Vision* 10:40. doi: 10.1167/10.14.40

Rudd, M., and Arrington, K. (2001). Darkness filling-in: a neural model of darkness induction. *Vis. Res.* 41, 3649–3662. doi: 10.1016/S0042-6989(01)00216-4

Sapiro, G., and Caselles, V. (1997). Histogram modification via differential equations. *J. Diff. Eq.* 135, 238–266. doi: 10.1006/jdeq.1996.3237

Shapley, R., and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. *Prog. Retin. Res.* 3, 263–346. doi: 10.1016/0278-4327(84)90011-7

Shevell, S. K. (ed.). (2003). "Color appearance," *The Science of Color* (Oxford, UK: Elsevier Science Ltd.), 149–190.

Wilson, H., and Cowan, J. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1–24. doi: 10.1016/S0006-3495(72)86068-5

Wilson, H., and Cowan, J. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biol. Cybern.* 13, 55–80.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 April 2014; accepted: 26 June 2014; published online: 17 July 2014.
Citation: Bertalmío M (2014) From image processing to computational neuroscience: a neural model based on histogram equalization. Front. Comput. Neurosci. ***8****:71. doi: 10.3389/fncom.2014.00071*
This article was submitted to the journal Frontiers in Computational Neuroscience.
Copyright © 2014 Bertalmío. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

In this section we give the implementation details for the results reported in the paper.

For the examples in **Figure 6** we computed explicitly the steady state solution of Equation 6, with $\alpha = \beta = \gamma = 1$, the second term of the equation with a weight $\gamma(1 + 5\sigma^c)$, and a compression constant of $c = 1/3$ (**Figure 6**, left) or $c = 0.75$ (**Figure 6**, right).

For the example in **Figure 7** we have adapted the fast numerical implementation of Bertalmío et al. (2007), with a polynomial approximation of degree 7 for the sign function, time step $\Delta t = 0.15$ and the stopping condition being fulfilled when the difference between the images of the current and the previous iteration falls below 0.1%. The original image is of size 100×200 and the parameter values were: $\alpha = \beta = \gamma = 1$, $c = 1/3$, stencil size for the computation of the standard deviation $\sigma(x) : 21 \times 21$, effective radius of the locality kernel $w(x, y) : 75$, and of the neighborhood over which we compute the mean average $\mu(x) : 19$.

For the examples in **Figures 9**, **10** we have also used the same fast numerical implementation of Bertalmío et al. (2007), where now the stopping condition is fulfilled when the difference between the images of the current and the previous iteration falls below 0.5%. The parameter values were: $\alpha = \beta = \gamma = 1$, $c = 1/3$, stencil size for the computation of the standard deviation $\sigma(x) : 3 \times 3$, effective radius of the locality kernel $w(x, y)$, effective radius of the neighborhood over which we compute the mean average $\mu(x)$: 1/3 of the numbers of rows or columns, whichever is larger.

ADVANTAGES OF PUBLISHING IN FRONTIERS



FAST PUBLICATION
Average 90 days from submission to publication



COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive



RESEARCH NETWORK
Our network increases readership for your article



OPEN ACCESS
Articles are free to read, for greatest visibility



TRANSPARENT
Editors and reviewers acknowledged by name on published articles



GLOBAL SPREAD
Six million monthly page views worldwide



COPYRIGHT TO AUTHORS
No limit to article distribution and re-use



IMPACT METRICS
Advanced metrics track your article's impact



SUPPORT
By our Swiss-based editorial team

frontiers

EPFL Innovation Park · Building I · 1015 Lausanne · Switzerland
T +41 21 510 17 00 · info@frontiersin.org · frontiersin.org