The dynamics of invariant object and action recognition in the human visual system

by

Leyla Isik

B.S., Johns Hopkins University, 2010

Submitted to the Program of Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author .... Signature redacted ....
Program of Computational and Systems Biology
May 22, 2015

Certified by... Signature redacted ....
Tomaso Poggio
Professor of Brain and Cognitive Sciences
Thesis Supervisor

Accepted by .... Signature redacted ....
Chris Burge
Professor of Biology and Biological Engineering
Director, Computational and Systems Biology Graduate Program

The dynamics of invariant object and action recognition in the human visual system

by

Leyla Isik

Submitted to the Program of Computational and Systems Biology
on May 22, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

Abstract

Humans can quickly and effortlessly recognize objects, and people and their actions from complex visual inputs. Despite the ease with which the human brain solves this problem, the underlying computational steps have remained enigmatic. What makes object and action recognition challenging are identity-preserving transformations that alter the visual appearance of objects and actions, such as changes in scale, position, and viewpoint. The majority of visual neuroscience studies examining visual recognition either use physiology recordings, which provide high spatiotemporal resolution data with limited brain coverage, or functional MRI, which provides high spatial resolution data from across the brain with limited temporal resolution. High temporal resolution data from across the brain is needed to break down and understand the computational steps underlying invariant visual recognition.

In this thesis I use magenetoencephalography, machine learning, and computational modeling to study invariant visual recognition. I show that a temporal association learning rule for learning invariance in hierarchical visual systems is very robust to manipulations and visual disputations that happen during development (Chapter 2). I next show that object recognition occurs very quickly, with invariance to size and position developing in stages beginning around 100ms after stimulus onset (Chapter 3), and that action recognition occurs on a similarly fast time scale, 200 ms after video onset, with this early representation being invariant to changes in actor and viewpoint (Chapter 4). Finally, I show that the same hierarchical feedforward model can explain both the object and action recognition timing results, putting this timing data in the broader context of computer vision systems and models of the brain. This work sheds light on the computational mechanisms underlying invariant object and action recognition in the brain and demonstrates the importance of using high temporal resolution data to understand neural computations.

Thesis Supervisor: Tomaso Poggio
Title: Professor of Brain and Cognitive Sciences

Acknowledgments

First I would like to thank my advisor Tomaso Poggio. Tommy, before I even came to MIT your work inspired my interest in artificial intelligence and neuroscience. It has been amazing to work with you. Your encouragement gave me the confidence to first try MEG experiments, and take advantage of many other new and exciting opportunities I would not have otherwise. Thank you for everything.

Next I would like to thank my thesis committee: Robert Desimone, Nancy Kanwisher, and Gabriel Kreiman. Your insightful feedback, as well as the excitement and energy you brought to our discussions has had a tremendous impact on me and my work.

I would like to thank all my friends in the Poggio lab for fruitful collaborations and many fun times. Particularly to my wonderful collaborators – Andrea Tacchetti, Ethan Meyers, Joel Z. Leibo, Danny Harari and Yena Han – you are all amazing and I have learned so much from working with each of you. Steve Voinea, Owen Lewis and Youssef Mroueh, thank you for your friendship over the years (and awesome feedback on this thesis). Kathleen Sullivan, Felicia Bishop, Elisa Pompeo, thank you for all of your help and support. Gadi Geiger, thank you for great coffee and many welcome distractions.

I would also like to thank Chris Burge and Jacquie Carota, as well as all of my friends in the CSB program. Thanks to the wonderful GWAMIT community - you all have made my MIT experience so much better. And to my awesome living groups: the Warehouse and Baker House.

I would like to thank the rest of my friends and my family. Dan Saragnese, thank you for encouraging me in all of my endeavors, both large and small. I don't know what I would do without your support. Finally, I would like to thank my parents, Salwa Ammar and Can Isik, for instilling the importance of knowledge and learning in me, as well as my brother Sinan. All of your love and encouragement helps me in everything I do. I could not have done this without you.

Contents

1 Introduction **15**

1.1 State of the field 16

1.1.1 Visual recognition in the brain 16

1.1.2 Timing in the brain 19

1.1.3 Visual recognition with convolutional neural network models . 21

1.2 Background methods 24

1.2.1 MEG 24

1.2.2 Neural decoding analysis 26

1.2.3 HMAX 27

1.3 Contributions 27

1.3.1 Main contributions 27

1.3.2 Organization of this thesis 28

2 Robustness of invariance learning in the ventral stream **31**

2.1 Introduction 32

2.2 Temporal association learning with the cortical model 33

2.2.1 Learning rule 35

2.3 Robustness 36

2.3.1 Training for translation invariance 36

2.3.2 Accuracy of temporal association learning 36

2.3.3 Manipulating the translation invariance of a single cell 37

2.3.4 Individual cell versus population response 39

2.3.5 Robustness of temporal association learning with a population of cells . . . 39
2.4 Discussion . . . 41

3 The dynamics of size- and position-invariant object recognition in the human visual system 45
3.1 Introduction . . . 46
3.2 Materials and Methods . . . 47
3.2.1 Subjects . . . 47
3.2.2 Experimental Procedure . . . 47
3.2.3 Eyetracking . . . 49
3.2.4 MEG recordings and data processing . . . 49
3.2.5 Decoding analysis methods . . . 49
3.2.6 Significance criteria . . . 52
3.2.7 Significance testing with normalized decoding magnitudes . . . 52
3.2.8 Source localization . . . 53
3.2.9 Cortical modeling (HMAX) . . . 53
3.3 Results . . . 54
3.3.1 Fast and robust readout for different types of stimuli . . . 54
3.3.2 Timing of size and position invariant visual representations . . . 57
3.3.3 Varying extent of size and position invariance . . . 58
3.3.4 Combined size- and position-invariance . . . 61
3.3.5 Dynamics of decoded signals . . . 64
3.3.6 Neural sources underlying sensor activity and classification . . . 68
3.3.7 Invariant recognition with a cortical model . . . 68
3.4 Discussion . . . 70
3.5 Appendix - Decoding in source space . . . 74
3.5.1 Methods . . . 75
3.5.2 Results . . . 76
3.5.3 Source estimates at key decoding times . . . 76

3.5.4 Discussion 78

4 Invariant representations for action in the human brain 83

4.1 Introduction 84

4.2 Results 85

4.2.1 Novel invariant action recognition dataset 85

4.2.2 Readout of actions from MEG data is early and invariant . . . 86

4.2.3 Extreme view invariance 89

4.2.4 Recognizing actions with a biologically-inspired hierarchical model 89

4.2.5 Recognizing actor invariant to action 95

4.2.6 The roles of form and motion in invariant action recognition . 96

4.2.7 Neural sources of action recognition 97

4.2.8 Generalization to novel actions with a fixed model 101

4.3 Discussion 102

4.3.1 Fast, invariant action recognition and implications for model architecture 102

4.3.2 Actor invariance and recognition 103

4.3.3 Neural sources underlying invariant action recognition 103

4.3.4 Form and motion in invariant action recognition 104

4.3.5 Conclusion 105

4.4 Methods 105

4.4.1 Action recognition dataset 105

4.4.2 Subjects 106

4.4.3 MEG experimental procedure 106

4.4.4 MEG data acquisition and preprocessing 106

4.4.5 Eyetracking 107

4.4.6 MEG decoding analysis methods 107

4.4.7 Source localization 108

4.4.8 Model 109

4.4.9 Video pre-processing and model classification 110

4.4.10 Model experiments 110
4.5 Supplemental figures 112
4.5.1 C1 model performance on action recognition dataset 112
4.5.2 Effect of eye movement on action decoding 113

5 Conclusions 115

5.1 Summary of findings 115
5.1.1 Temporal association training is a robust method for learning invariance in development 115
5.1.2 Size- and position-invariant object recognition have fast dynamics that match feedforward models of visual cortex 116
5.1.3 Actor- and viewpoint-invariant action representations arise quickly and match a feedforward model of visual cortex 117
5.1.4 Common themes 117
5.2 Future directions 118
5.2.1 Moving towards more naturalistic vision experiments 118
5.2.2 Combining temporal and spatial information in the human brain119
5.3 Conclusion 120

List of Figures

1-1 Timing in the macaque ventral stream 20

1-2 The HMAX model 22

2-1 An illustration of the HMAX model with normal and altered temporal association training 34

2-2 Image recognition with hard-wired model and temporal associate learning model 37

2-3 Manipulating single cell translation invariance through altered visual experience. 38

2-4 Translation invariance for different model population cell sizes 40

3-1 Experimental Task 48

3-2 Decoding accuracy versus time for three different image sets 55

3-3 Decoding parameter optimization 57

3-4 Assessing position and size invariant information 60

3-5 Significant and peak invariant decoding times 62

3-6 Assessing combined size and position invariant information 66

3-7 Dynamics of object decoding 67

3-8 Source localization at key decoding times 69

3-9 Invariant image classification with a cortical model 71

3-10 A left view of source estimates on cortex of subject 1 at 70ms after stimulus onset, the time when image identity can first be decoded, and 150 ms after stimulus onset, the time when size- and position-invariant information can be decoded. Color bar right indicates absolute source magnitude in picoAmpere-meters. 77
3-11 Comparison of decoding in sensor and source space 80
3-12 Decoding within visual ROIs 81

4-1 Novel action recognition dataset 86
4-2 Decoding action 87
4-3 Decoding action invariant to actor and viewpoint 88
4-4 "Extreme" view-invariant decoding 90
4-5 Biologically inspired computational model for action recognition . . . 92
4-6 Model pooling across viewpoints 93
4-7 Model performance on view-invariant action recognition 94
4-8 Decoding actor from MEG data 95
4-9 Model performance on action-invariant actor recognition 96
4-10 The roles of form and motion in MEG action decoding 98
4-11 Model performance on static frames 99
4-12 Neural sources of action recognition 100
4-13 Generalizing to recognize new actions 101
4-14 C1 model performance on action recognition dataset 112
4-15 Decoding with eye tracking movement 113

List of Tables

3.1 Decoding accuracy magnitude, significant time, and peak times with varying amounts of data 63

Chapter 1

Introduction

Humans can process complex visual scenes within a fraction of a second [138, 171, 42, 139, 140]. The main computational difficulty in visual recognition is believed to be transformations that change the low-level representation of objects and actions, such as changes to size, viewpoint, and position in the visual field [32, 5]. The ability to effortlessly discount these transformations and learn from few examples make the human visual system superior to even state of the art computer vision algorithms. Better understanding the neural mechanisms underlying visual recognition in humans will greatly improve artificial intelligence (AI) systems and provide insight into how the brain solves one of the most complex sensory problems.

Studies using single neuron and neural population recordings have provided high spatiotemporal resolution data about object and action recognition, but their coverage is limited to recording from one or two brain regions at a time. fMRI studies have improved our understanding of the spatial organization of the brain and functional regions of interest; however, fMRI its still limited in its temporal resolution.

Understanding *when* different invariant representations are computed across the brain is a crucial component missing from many neuroscience efforts, and will help further our understanding of the algorithms underlying visual recognition. In order to break down the computational steps required to solve the complex problem of visual recognition, it is important to know what the intermediate steps are (to go from pixels/retinal response to invariant representations of objects and actions) and when and in what order they are computed. This thesis focuses on when and how different properties of invariant recognition are computed in the brain, using magnetoencephalography (MEG), machine learning, and computational models of the human visual system.

The remainder of this chapter describes the state of the field for object and action processing in the brain, as well as an overview of state of the art computer vision systems and biologically plausible models of the human visual cortex. Finally, it provides some details on the methods used in this thesis.

1.1 State of the field

1.1.1 Visual recognition in the brain

Visual signals enter the brain via the retina and lateral geniculate nucleus (LGN) and first enter the cortex in the primary visual cortex (V1). The visual cortex has been roughly divided into two pathways - the ventral (or "what") pathway, which processes objects, and the dorsal (or "where") pathway, which processes actions and spatial location [121, 57]. This distinction is somewhat of an oversimplification as there are many interconnections and feedback between areas in both pathways [44, 116]. Despite these complexities, the visual cortex is still considered to be hierarchically organized with the responses of one visual layer, serving as input to the subsequent layer.

Object recognition occurs along the ventral stream. We consider the initial visual response to be occurring in a primarily feedforward manner with connections proceeding from visual area V1 to V2 to V4 to inferior temporal cortex (IT), see Figure 1-1. Invariance and selectivity increase at each layer of the ventral stream [10, 147, 152]. In the dorsal stream, visual signals also originate in V1, where cells are selective for local, directed motion. These motion signals then enter area MT, where cells are selective for motion direction invariant to pattern [122], and neural responses have been closely linked with motion perception [36, 3]. MT then projects to areas MST and FST in the superior temporal sulcus (STS) [175, 31, 79].

Object processing along the ventral stream

Beginning with Hubel and Wiesel's seminal studies in V1 [76], single-cell physiology studies have helped explain many properties of the ventral stream. Hubel and Wiesel showed in the cat and the macaque that cells in V1 fire in response to oriented lines or edges, and are selective for a given orientation [77, 78]. They also discovered two different cell types within V1: simple and complex cells. V1 simple cells are selective (or "tuned") for a given feature (i.e. a line or edge of a given orientation) within a small receptive field. Complex cells receive input from these simple cells and compute an aggregate measure (or "pool") over these features to provide invariance over the union of the simple cells' receptive fields. Due to the retinotopic organization of V1, by pooling over neighboring simple cells, complex cells achieve selectivity to a given feature within a particular spatial region, providing local position invariance (see Figure 1-2 for a model diagram of this process).

Mid-level object and shape representations have been harder to explain than those in V1. Cells in V2 share many properties of shape response with V1 cells and they are also selective to orientation and spatial frequency [73, 74, 8, 38]. It has also been shown that V2 cells are sensitive to higher order statistical dependencies than V1, which may play a role in recognizing image structure [50]. Cells in V4 appear to have more complex shape representations including polar, hyperbolic and Cartesian gratings [54, 55], contours and curvature [129], and complex shape and color [154, 96, 145]. The final layer of the primate ventral stream, IT, is generally divided into a posterior portion and an anterior portion. The posterior portion is selective to object parts and partially invariant, and cells in the anterior portion are selective to complex objects, including faces, and invariant to a wide range of scale and position changes [61, 29, 28, 168, 132, 167].

fMRI studies over the last two decades have revealed similar organization for the human ventral stream [39, 186, 72, 9], and that the transition from low-level statistics to object-level information increases gradually along the ventral stream [58].

In addition, there exist several object-specific regions in visual cortex, which respond preferentially to certain ecologically relevant categories, including faces, places, and bodies [90, 41, 35, 130, 60]. Recent studies have shown that these object-selective regions exist in the macaque, and fMRI-guided physiology has been used to show correspondence with object-selective single cells [173, 84].

Recordings from IT neurons and voxels in object-selective cortex reveal these areas are remarkably invariant to a range of transformations, including not only affine transformations such as changes in size and position [85, 80, 21], but also non-affine transformations such as changes in illumination and viewpoint [113, 59, 111, 51, 16]. All objects undergo affine transformations in the same manner (e.g. they scale and translate in the same manner), so a neural mechanism used to deal with these types of invariance can, in principle, be used for all objects. This is not true of non-affine transformations, such as rotation in depth, which depend on the 3-D structure of an individual object or class of objects, and thus must be learned for each class of objects [180, 13, 105].

Action recognition in the brain

As they do with objects, humans can also recognize actions very quickly, even from very impoverished stimuli. Action recognition, however, has been much less studied than object recognition, and the majority of action recognition studies are focused on experiments with simple, controlled stimuli. These stimuli are mostly static images, which contain only form information, or point light displays (created by placing dots on the joints of a moving human), which contain biological motion information with little to no form information [88].

Previous experiments using such stimuli revealed that neural representations for biological motion exist in the macaque STS [141, 133, 127, 179], and in a posterior region of the STS in humans (pSTS) [62, 63, 177, 14, 130]. The STS is a long sulcus that spans the temporal lobe, and receives input from both the ventral and dorsal streams. The STS has also been implicated in general motion processing, face recognition, and social perception [31, 19, 4, 69, 153, 53].

Actions can also be recognized from static images, and in general this elicits similar activity in motion-selective areas, including MT/MST [97] and STS, as well in the extra striate body area (EBA), a region traditionally considered to be more involved in form processing [120, 110].

In addition to distinguishing biological motion types of motion, fMRI studies have also shown that pSTS can distinguish between different types of actions [178], and can do so in a mirror symmetric manner [64]. Physiology studies have found that neurons in the macaque STS can identify both action invariant to actor and actor invariant to action, showing that the same neural population can be both selective and invariant to each of these features [159].

The computational steps required to go from oriented lines and edges to invariant representations of whole objects, or from directed motions to actions (particularly from natural video stimuli), are still largely unknown. The timing of different stages of visual processing can help explain these underlying neural computations.

1.1.2 Timing in the brain

Most timing information known about the primate ventral stream comes from electrophysiological studies of macaques. Latencies of visual areas along the ventral pathway are known in the macaque, ranging from 40-60ms in V1 to 80-100ms in IT, and timing between each area is approximately 20 ms [126, 155, 172, 80] (see Figure 1-1). These latencies, however, are still largely unknown in humans.

Some recent work using electrocorticography (ECoG) in human patients with pharmacologically intractable epilepsy has provided the rare opportunity to record from the surface of the human brain. These studies have shown that, as in the macaque, human visual signals that are invariant to size and position exist after 100 ms [111]. ECoG methods are limited by the lack of spatial coverage, and the scarcity of eligible subjects.

The majority of human timing results have come from noninvasive studies examining evoked responses in electroencephalography (EEG), and to a lesser extent, MEG. Simon Thorpe showed that humans can solve a visual categorization problem

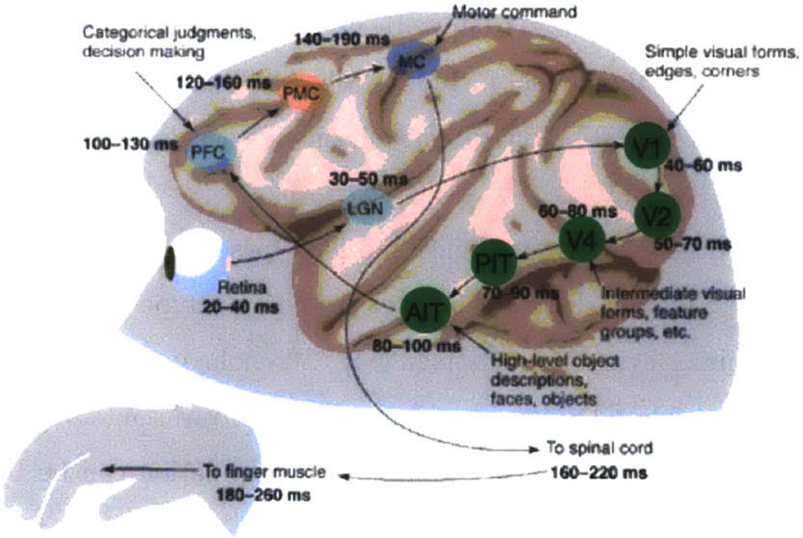


Figure 1-1: A diagram outlining the timing of different steps in the macaque brain to during a rapid object categorization task. Visual areas along the ventral stream are highlighted in green. The latencies of the visual areas along the ventral stream have been measured in the macaque. Figure modified from [172].

(distinguishing between scenes with or without animals) very quickly, with reaction times as fast as 250 ms, and divergence in EEG responses evoked by the two categories occurring at 150ms [171]. In addition, object-specific signals, such as the N170 response to faces, a strong negative potentiation in certain EEG channels 170 ms after image onset in response to faces versus non-faces [15], have a similar latency. It has been shown that the N170 response is also elicited by facial and body movements [187].

Using MEG decoding, others have shown that high-level categorization can be performed, such as distinguishing between two categories of objects (faces vs. cars) [21] or higher-level questions of animacy vs. inanimacy [22, 24], around 150ms. These signals occur late relative to those in the macaque summarized above [126, 155, 172, 80] (Figure 1-1). These results raise the question: what are computational stages in human visual processing, and their timing, that lead to invariant object recognition?

1.1.3 Visual recognition with convolutional neural network models

Hubel and Wiesel's findings in visual cortex, namely that there are simple cells that are feature selective and complex cells that pool over simple cells to build invariance, have inspired a class of computer vision models. These models consist of a hierarchical organization of alternating convolutional or "tuning" layers and subsampling or "pooling" layers. The convolutional layers perform template matching between a given feature (e.g. an oriented Gabor filter, as found in V1 simple cells) and overlapping regions of an input image (analogous to cells' receptive fields in the visual field). The pooling layers perform a max (or other pooling operation) to build invariance over that pooling range (e.g. if a complex cell pools over several simple cells that are selective for the same orientation, that complex cell will be selective for that orientation feature anywhere in its pooling range, invariant to position) and reduce sample complexity (the number of training examples needed to learn a new category) [5, 191] (see Figure 1-2). These models can be broken into two categories, biologically-inspired models and performance-optimizing deep neural networks, based on their architecture and training procedure. (It is important to note that the term "deep neural network" refers to any algorithm that uses many layers of feature extractors and uses the outputs of one layer as input to the next, and not necessarily deep convolutional neural networks, which are one specific type of deep neural network. In this chapter, we will use "deep neural networks" to refer to the specific type of deep convolutional neural networks.)

Biologically-inspired computer vision models

The first type of model, biologically-inspired, was introduced by Fukushima [52]. The models' architecture and parameters are set to closely model properties of cells in the visual cortex, and they are usually trained in an unsupervised manner to mimic natural visual experience [184, 144, 157, 174, 188]. These models have been expanded to recognize actions from videos, by including a temporal component to the templates

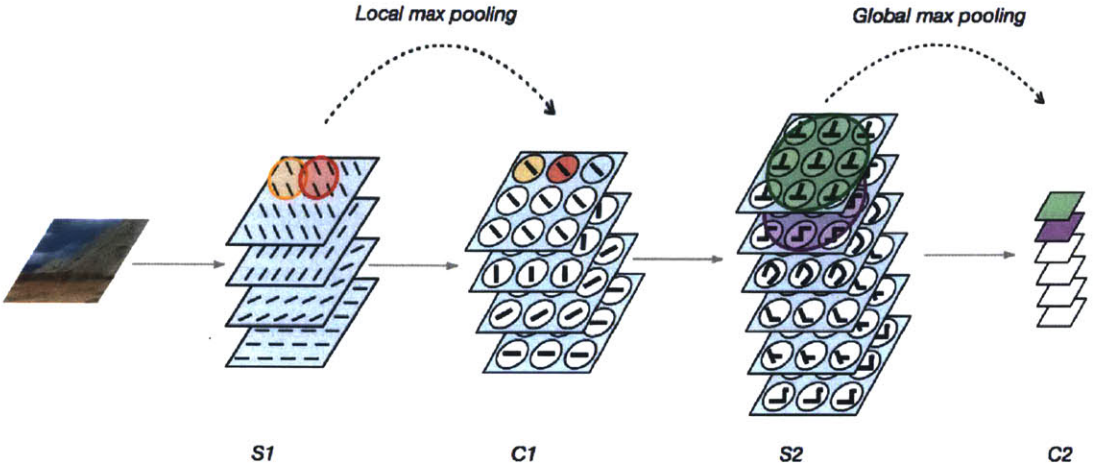


Figure 1-2: An outline of the HMAX model, one type of biologically-inspired convolutional neural network, adapted from [156]. The model consists of alternating layers of simple 'S' cells, which are tuned for a particular feature (oriented lines in the case of S1 and more complex shapes sampled from natural images in S2), and complex 'C' cells which pool over the responses of 'S' cells to build invariance. (The pooling range for a given C1 or C2 cell are highlighted for illustration.) The final C2 feature vector for each image can be compared to the response of a population of IT cells and used to classify images into different categories.

and pooling [56, 87], and to recognize faces invariantly to non-affine transformations (e.g. rotation in depth, changes in background) [105, 109]. A particular instantiation of these Hubel and Wiesel-inspired (HW) models used in this research, known as the HMAX model, is described in section 1.2.3 and Figure 1-2.

In these models, the pooling regions are hard-wired, which raises the question of how these arise in the brain? People have proposed theories for how these invariant representations may be learned in development by taking advantage of the fact that objects typically move in and out of the visual field much slower they transform, and linking temporally adjacent views of the same object [49, 146, 189, 37, 162]. A recent computational theory has suggested that learning how a few objects transform in development and storing "templates" for their transformed views can lead to transformation invariance for novel objects [5]. This theory further explains why these hierarchical networks have achieved such success at object recognition, and states that the main goal of cortical hierarchies is to encode invariant representations of objects. Further, these invariant representations decrease the sample complexity for recognition - i.e. allow the brain to learn from few labeled examples [6, 7].

Deep convolutional neural networks

In recent years the second type of neural network models, deep neural networks, have achieved unprecedented human-level performance on a series of vision challenges, in particular, on the object recognition challenge ImageNet, a 1000 category categorization problem which includes thousands of training images per category and over 14 million total images [160, 43, 151]. In 2012, the winner of the ImageNet challenge employed a deep neural network, which outperformed other systems by a unprecedented extent [99]. Today, most high-performing computer vision systems employ these deep neural networks. These convolutional neural networks were pioneered almost 30 years ago [150, 137, 103, 104], and again were inspired loosely by the architecture of brain and the model of Fukushima [52]. While the recent success of these algorithms is undeniable, it seems to be in large part due to the availability of better computing power and the prevalence of large training sets of example images, rather than any changes to the networks architecture [99]. One key distinction with the above biolgically-inspired models is that these networks' architecture and parameters are set for performance rather than biological fidelity. In particular, the weights in the network are learned through "back-propagation" the process of using gradient descent to propagate error signals on training data to update the weights.

As with object recognition, deep convolutional networks are also the top performing computer vision systems on action recognition tasks [101, 93]. In the case of action recognition though, these systems do not achieve the same high, human-like performance as they do with objects (they achieve 40-80% accuracy instead of over 90% achieved on the ImageNet Challenge). It is still unclear if the performance of deep neural networks on action recognition tasks will improve as more large datasets of labeled videos are created.

In addition to their high performance on computer vision tasks, deep neural networks have shown a remarkably high correspondence with primate neural data [192, 18]. The top stages of these networks are able to explain most of the variance in IT recordings, and middle network layers show similarly high correspondence with V4 neurons, both of which have been notably challenging to model. This work suggests that the performance optimizing aspects of these deep neural networks are solving the same optimization as the brain. Surprisingly, this work has even shown that these models predict neural responses even better than models fit with neural data. This is likely due to the fact that there is not enough neural data to constrain such large and complex models.

Despite the recent success of deep neural networks, their basis on the architecture of the brain, and correspondence with neural data, there are several ways in which these algorithms are distinctly different from, and arguably inferior to, human vision. Primarily, these models require thousands of labeled examples to learn a new class of objects, while children and adults can learn a new category of object from only a few labeled examples [20, 114, 190]. Second, while some work has been done to address higher-level tasks (such as image captioning and narration) it is unclear how many tasks beyond visual categorizations can be explained by these models. The goal of many recent neuroscience and AI research efforts is to find new discoveries that will, just as previous neuroscience insights have inspired the above algorithms, further propel progress towards more intelligent and human-like computer vision systems.

1.2 Background methods

1.2.1 MEG

With the complexity of the visual system, one would ideally record simultaneously with high spatiotemporal resolution from across the visual cortex. Unfortunately, this is infeasible with today's recording technologies. Here we focus on measuring high temporal resolution data with broad brain coverage using MEG. As mentioned above, these two aspects in concert have been understudied and can provide great insight into the brain's computations. Specifically, timing data can help explain how (in what order) invariant representations are computed, which can constrain existing and inspire new algorithms for object and action recognition.

MEG is a direct, non-invasive measure of whole-head neural firing with millisecond temporal resolution. EEG and MEG detect the electrical and magnetic fields, respectively, that are produced by synchronous neural firing. MEG requires on the order of 50 million neurons oriented in the same direction to fire synchronously, and thus detects signals primarily from pyramidal neurons aligned perpendicular to the surface of the cortex. Due to the geometry of the cortical surface and resulting magnetic fields, MEG is primarily sensitive to neural sources in the sulci and EEG primarily detects activity primarily from the gyri. Unlike electric fields detected with EEG, magnetic fields are not impeded by the skull and scalp [102], and thus MEG is a less noisy measure of neural activity.

Typical MEG methods involve analyzing event-related fields evoked by a certain stimulus. This is done by presenting a stimulus on the order of 100 times and averaging the value of the magnetic field measured in a given channel over the multiple stimulus repetitions to improve the signal-to-noise ratio. This has led to several discoveries of visual timing hallmarks, described in section 1.1.2 [171, 15]. This procedure, however, has several notable drawbacks, mainly that it requires several stimulus repetitions, is not a direct measure of and is limited by the univariate nature of the analysis.

Source localization methods are often used in conjunction with this event-related field analysis to estimate where the neural sources driving MEG or EEG measurements are located. Source localization, however, is an ill-posed problem. Even when taking into account geometric constraints of cortical shape and orientation of magnetic fields, there are still orders of magnitude more sources one tries to estimate than sensor measurements that are made with MEG. This limits the spatial resolution of MEG. The most widespread way to overcome the ill-posed nature of the problem is to regularize the inverse problem with the L2 norm, leading to the minimum norm estimate [67, 11]. This minimum norm estimate minimizes the total power of the sources, which has little physiological basis. Other methods that impose sparsity of the sources by regularizing with L1 or L1-L2 norm combinations have also been used [66]. Recent results have shown that incorporating fMRI data into source estimation methods can greatly improve the spatial resolution of MEG/EEG measurements [128]. Other new methods have also taken advantage of the dynamic information in MEG to create better temporal models of the inverse problem [100].

Neural decoding analysis is a tool that can be applied to source- or sensor-level data and, unlike the above methods, provides a direct measure of stimulus information present in the data. Neural decoding has been largely unexplored for analyzing MEG visual data.

1.2.2 Neural decoding analysis

Neural decoding analysis uses a machine learning classifier to assess what information about the input stimulus is present in the recorded neural data (for example, what image the subject was looking at). Decoding analysis has the advantage of being able to extract information from the pattern of neural activity across multiple sensors or voxels, and as a result has increased sensitivity over univariate analyses [80, 68, 94, 131].

Decoding analysis, or multi-voxel pattern analysis (MVPA), has been widely used in fMRI visual research and led to great progress in predicting visual cortical response to a range of visual stimuli [68, 89, 70, 94, 185], and even reconstructing visual stimuli purely based on fMRI data [125]. Decoding analysis has also been applied to electrophysiology visual data [80, 119, 193], and to a lesser extent to EEG and MEG visual data. MEG decoding has been used recently to show category-selective signals with some position invariance present at 150 ms [21], and other properties about objects, such as animacy/inanimacy, between 150-200 ms [22, 24].

One main advantage of neural decoding is that it makes it possible to test for the presence of invariant information in the visual signals. By training the classifier on stimuli presented at one condition (a given, position or scale, for example), and testing on a second condition (a different position or scale), it is possible see if the neural information can generalize across that transformation. This technique has been applied to fMRI data [70], electrophysiology data [80, 193], and MEG data [21].

1.2.3 HMAX

HMAX falls in the class of Hubel and Wiesel-inspired models described above in section 1.1.3. One particular instantiation consists of two layers of alternating simple and complex cells. The first simple cell layer, S1, contains features or templates that are Gabor filters of varying scales and orientations. A dot product is computed between each of the Gabor filters at each location in the image (a convolution). In the first complex cell layer, C1, a local maximum is computed across position and scales to provide invariance in these local regions. In the second simple cell layer, S2, the templates are drawn from random natural image patches that are fed through the S1 and C1 layers of the HMAX model, which serve as an intermediate layer feature. In the final C2 layer, pooling is performed for each feature across all sizes and positions to provide global size and position invariance (Figure 1-2). The final C2 vector can then be used as input to a machine learning classifier trained on various tasks, such as object recognition. This model has been shown to match human performance on a feedforward object categorization task where images are masked to prevent feedback processing [157].

Some key questions about this model remain: how biologically faithful is it, namely is there evidence that the brain employs this alternating tuning and pooling beyond V1? And to what extent and on what subset of tasks do these feedforward models explain human visual performance?

1.3 Contributions

1.3.1 Main contributions

In this thesis I show that:

- Visual signals containing information sufficient for object recognition exist as early as 60 ms after stimulus onset in the human visual system.

- Size and position invariance begin at 100 ms after stimulus onset, with invariance to smaller transformations occurring before invariance to larger transformations.

- Action selective visual signals occur as early as 200ms after video onset.

- These early representations for action are also invariant to changes in actor and viewpoint.

- The same feedforward, hierarchical models for visual recognition can explain both these object and action recognition timing results.

1.3.2 Organization of this thesis

Chapter two implements a temporal association learning rule, a popular computational mechanism for learning invariance in development, in the HMAX model. Simulations with this model demonstrate the robustness of this learning mechanism. We find that, as in recent behavioral and physiology studies, we can learn and subsequently disrupt position invariance in single model units through temporal association learning. Despite this single cell disruption, a population of cells remains robust to these manipulations, demonstrating the fidelity of this learning rule across many neurons.

Chapter three describes the dynamics of size- and position-invariant object recognition in the human brain. Using MEG decoding, we show that object identity can be read out in 60ms, with invariance to size and position invariance increasing in stages between 100-150 ms. These results uncover previously unknown latencies for human object recognition, and compelling evidence for a feedforward, hierarchical model of the visual system.

Chapter four describes the dynamics of viewpoint invariant action recognition in the brain and an accompanying computational model. Here we use video stimuli to better examine realistic visual input and understand how the brain uses spatiotemporal information to recognize actions. We show that, like object recognition, action recognition is also fast and invariant, with a representation for action that is invariant to actor and view arising in the brain in around 200 ms. We extend the same class of hierarchical feedforward computational model to account for these MEG results. Finally we show that, like the MEG signals, the model can recognize action invariant to actor and view.

Chapter five concludes the thesis by examining common themes in this work and future directions for using temporal dynamics to understand visual computations in the brain.

Chapter 2

Robustness of invariance learning in the ventral stream

This material in this chapter was published in Frontiers in Computational Neuroscience in 2012 [82]. Joel Z. Leibo and I contributed equally to this work.

Learning by temporal association rules, such as Foldiak's trace rule [49], is an attractive hypothesis that explains the development of invariance in visual recognition. Consistent with these rules, several recent experiments have shown that invariance can be broken by appropriately altering the visual environment. These experiments raise puzzling differences in the effect size and altered training time at the psychophysical [26, 183] versus single cell [107, 108] level. We show a) that associative learning provides appropriate invariance in models of object recognition inspired by Hubel and Wiesel [76], b) that we can replicate the "invariance disruption" experiments using these models with a temporal association learning rule to develop and maintain invariance, and c) that we can thereby explain the apparent discrepancies between psychophysics and singe cells effects. We argue that this mechanism in hierarchical models of visual cortex provides stability of perceptual invariance despite the underlying plasticity of the system, the variability of the visual world and expected noise in the biological mechanisms.

2.1 Introduction

Temporal association learning rules provide a plausible way to learn transformation invariance through natural visual experience [49, 115, 161, 184, 189]. Objects typically move in and out of our visual field much slower than they transform, and based on this difference in time scale the brain learns to group the same object under different transformations. These learning methods are attractive solutions to the problem of invariance development. However, these algorithms have mainly been examined in idealized situations that do not contain the complexities present in the task of learning to see from natural vision or, when they do, ignore the imperfections of a biological learning mechanism. Here we present a model of invariance learning that predicts the invariant object recognition performance of a neural population can be surprisingly robust, even in the face of frequent temporal association errors.

Experimental studies of temporal association and the acquisition of invariance involve putting observers in an altered visual environment where objects change their identity across saccades. Cox *et al.* showed that after a few days of exposure to this altered environment, the subjects mistook one object for another at a specific retinal position, while preserving their ability to discriminate the same objects at other positions [26]. A subsequent physiology experiment by Li and DiCarlo using a similar paradigm showed that individual neurons in primate anterior inferotemporal cortex (AIT) change their selectivity in a position-dependent manner after less than an hour of exposure to the altered visual environment [107]. It is important to note that the stimuli used in the Cox et al. experiment were difficult to discriminate "greeble" objects, while the stimuli used by Li and DiCarlo were easily discriminable, e.g., a teacup versus a sailboat.

This presents a puzzle, if the cells in AIT are really underlying the discrimination task, and exposure to the altered visual environment causes strong neural effects so quickly, then why is it that behavioral effects do not arise until much later? The fact that the neural effects were observed with highly dissimilar objects (the equivalent of an easy discrimination task) while the behavioral effects in the human experiment were only observed with a difficult discrimination task compounds this puzzle.

The physiology experiment did not include a behavioral readout, so the effects of the manipulation on the monkey's perceptual performance is not currently known; however, the human evidence suggests it is highly unlikely that the monkey would really be perceptually confused between teacups and sailboats after such a short exposure to the altered visual environment.

We present a computational model of invariance learning that shows how strong effects at the single cell level do not necessarily cause confusion on the neural population level, and hence do not imply perceptual effects. Our simulations show that a population of cells is surprisingly robust to large numbers of mis-wirings due to errors of temporal association. In accord with the psychophysics literature [26, 183], our model also predicts that the difficulty of the discrimination task is the primary determiner of the amount of exposure necessary to observe a behavioral effect, rather than the strength of the neural effect on individual cells.

2.2 Temporal association learning with the cortical model

We examine temporal feature learning with the HMAX model [157, 144]. The results presented should generalize to other models in the class of Hubel-Wiesel models [52].

The model learns translation invariance, specifically in the S2 to C2 connections, from a continuously translating image sequence, as shown in Figure 2-1, left. During training, an image (face or car) is translated left to right over a time period, which we will call an "association period". During this association period, one C2 cell learns to pool over highly active S2 cells. Correct temporal association should group similar features across spatial locations, as illustrated in Figure 2-1, left. Potential "mis-wiring" effects of a temporally altered image sequence are illustrated in Figure 2-1, right.

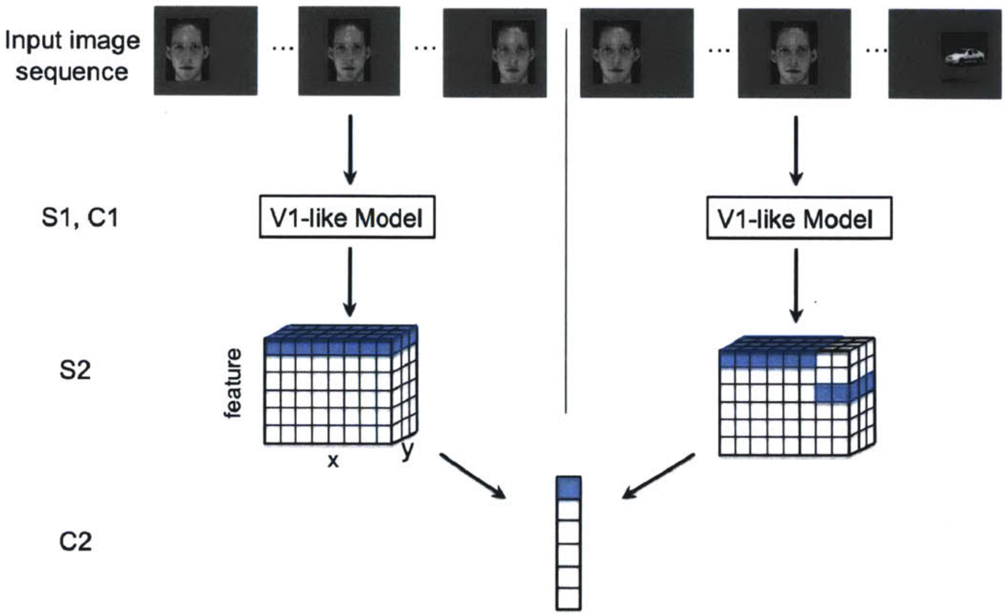


Figure 2-1: An illustration of the HMAX model with two different input image sequences: a normal translating image sequence (left), and an altered temporal image sequence (right). The model consists of four layers of alternating simple and complex cells. **S1 and C1 (V1-like model)**: The first two model layers make up a V1-like model that mimics simple and complex cells in the primary visual cortex. The first simple cell layer, S1, consists of simple orientation-tuned Gabor filters, and the following complex cell layer, C1, performs max pooling over local regions of a given S1 feature. These are identical to the first two model layers in [157]. **S2**: The next simple cell layer, S2, performs template matching between C1 responses from an input image and the C1 responses of stored prototypes (unless otherwise noted, we use prototypes that were tuned to natural image patches). Template matching is performed with a radial basis function, where the responses have a Gaussian-like dependence on the Euclidean distance between the (C1) neural representation of an input image patch and a stored prototype. The RBF response to each template is calculated at various spatial locations for the image (with half overlap). Thus the S2 response to one image (or image sequence) has three dimensions: x and y corresponding to the original image dimensions, and feature the response to each template. **C2**: The final complex cell layer, C2, performs global max pooling over all the S2 units to which it is connected. The S2 to C2 connections are highlighted for both the normal (left) and altered (right) image sequences. To achieve ideal transformation invariance, the C2 cell can pool over all positions for a given feature as shown with the highlighted cells.

2.2.1 Learning rule

In Foldiak's original trace rule, shown in Equation 2.1, the weight of a synapse between an input cell and output cell is strengthened proportionally to the input activity and the trace or average of recent output activity at time t. The dependence of the trace on previous activity decays over time with the δ term [49].

Foldiak trace rule:

$$\begin{aligned}
&\Delta w_{ij}^{(t)} \propto x_j \bar{y}_i^{(t)} \\
&\bar{y}_i^{(t)} = (1-\delta) y_i^{(t-1)} + \delta y_i^{(t)}
\end{aligned} \tag{2.1}$$

In the HMAX model, connections between S and C cells are binary. Additionally, in our training case we want to learn connections based on image sequences of a known length, and thus for simplicity should include a hard time window rather than a decaying time dependence. Thus we employed a modified trace rule that is appropriate for learning S2 to C2 connections in the HMAX model.

Modified trace rule for the HMAX model:

$$\begin{aligned}
&\text{for } t \text{ in } \tau : \\
&\qquad \text{if } x_j > \theta,\ w_{ij} = 1 \\
&\qquad \text{else},\ w_{ij} = 0
\end{aligned} \tag{2.2}$$

With this learning rule, one C2 cell is produced for each association period. The length of the association period is τ.

2.3 Robustness

2.3.1 Training for translation invariance

We model natural invariance learning with a training phase where the model learns to group different representations of a given object based on the learning rule in Equation 2.2. Through the learning rule, the model groups continuously-translating images that move across the field of view over each known association period τ. An example of a translating image sequence is shown at the top, left of Figure 2-1. During this training phase, the model learns the domain of pooling for each C2 cell.

2.3.2 Accuracy of temporal association learning

To test the performance of the HMAX model with the learning rule in Equation 2.2, we train the model with a sequence of training images. Next we compare the learned model's performance to that of the hard-wired HMAX [157] on a translation-invariance recognition task. In standard implementations of the HMAX model, each C2 cell pools all the S2 responses for a given template globally over all spatial locations. This pooling gives the model translation invariance and mimics the outcome of an idealized temporal association process.

We test both models on a face vs. car identification task with 20 faces and 20 cars that contain slight intraclass variation across different translated views[1]. We collect hard-wired C2 units and C2 units learned from temporal sequences of the faces and cars. We then test each model's translation invariance by using a nearest neighbor classifier to compare the correlation of C2 responses for translated objects to those in a given reference position. The accuracy of the two methods (hard-wired and learned from test images) for different amounts of translation is shown in Figure 2-2. The two methods performed equally well, confirming that the temporal associations learned from training yield accurate invariance results.

[1] The training and testing datasets come from a concatenation of two datasets from: http://www.d2.mpi-inf.mpg.de/Datasets/ETH80, and http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

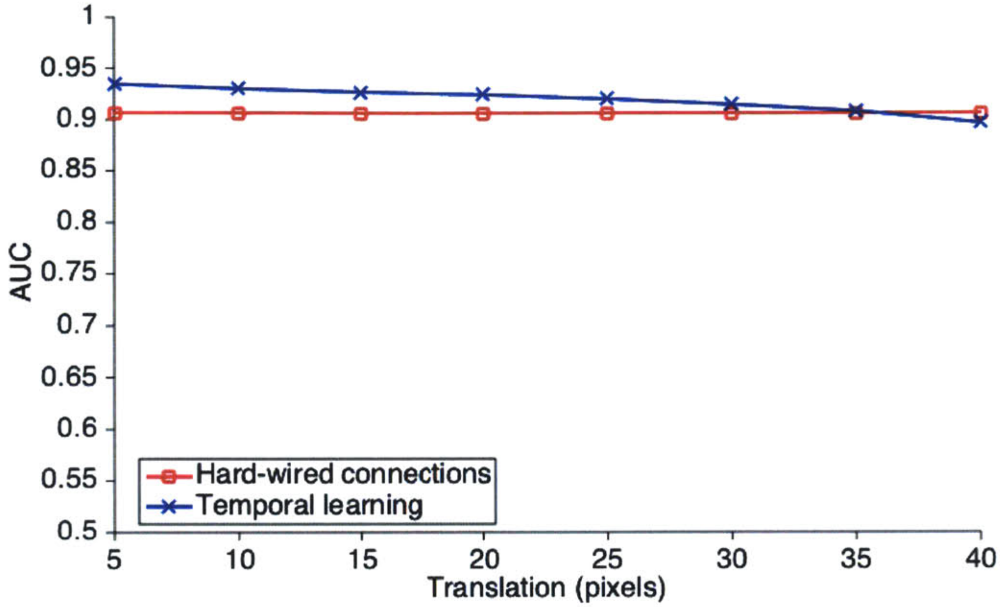
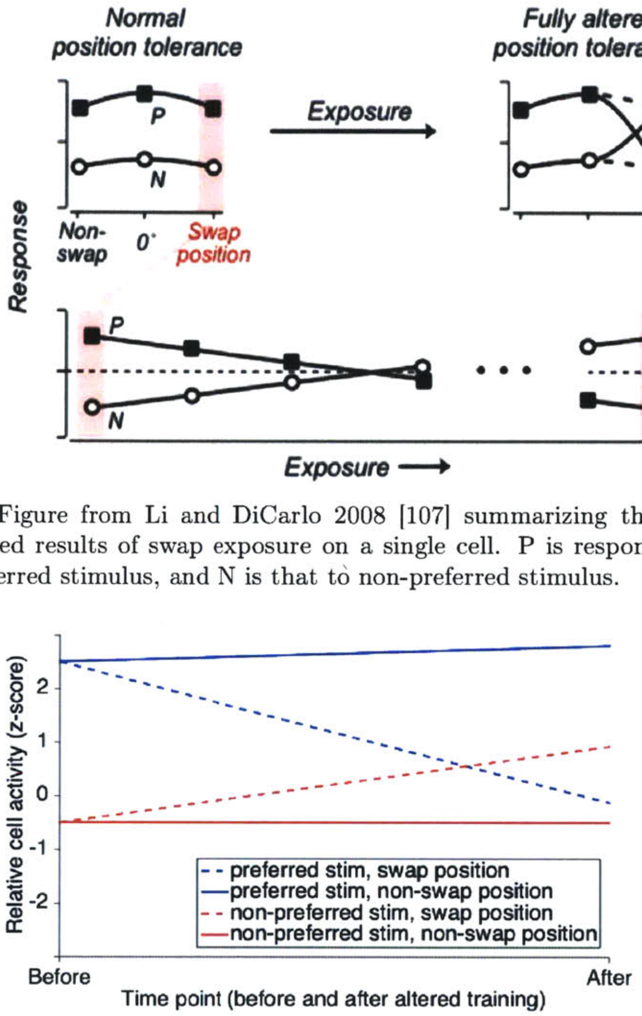


Figure 2-2: The classification accuracy (AUC for ROC curve) for both hard-wired and temporal association learning model plotted for different degrees of translation compared to a reference position with a nearest neighbor classifier. The model was trained and tested on separate training and testing sets, each with 20 car and 20 face images. For temporal association learning, one C2 unit is learned for each association period or training image, yielding 40 learned C2 units. One hard-wired C2 unit was learned from each natural image that cells were tuned to, yielding 10 hard wired C2 units. Increasing the number of hard-wired features has only a marginal effect on classification accuracy.

2.3.3 Manipulating the translation invariance of a single cell

To model the Li and DiCarlo physiology experiments in [107] we perform normal temporal association learning described by Equation 2.2 with a translating image of one face and one car. The S2 units are tuned to (i.e. use templates from) the same face and car images as in the training set to mimic object-specific cells that are found in AIT. Next we select a "swap position" and perform altered training with the face and car images swapped only at that position (see Figure 2-1, top right). After the altered training, we observe the response of one C2 cell, which has a preference for one stimuli over the other (to model single cell recordings), to the preferred and non-preferred objects at the swap position and at a second, non-swap position that was unaltered during training.



(a) Figure from Li and DiCarlo 2008 [107] summarizing the expected results of swap exposure on a single cell. P is response to preferred stimulus, and N is that to non-preferred stimulus.

(b) The response of a C2 cell tuned to a preferred object before (time point 1) and after (time point 2) altered visual training where the preferred and non-preferred objects were swapped at a given position. To model the experimental paradigm used in [107, 108, 26, 183], training and testing were performed on the same altered image sequence. The C2 cell's relative response (Z-score) to the preferred and non-preferred objects at both the swap and non-swap positions are plotted.

Figure 2-3: Manipulating single cell translation invariance through altered visual experience.

As shown in Figure 2.3.3 the C2 preference has switched at the swap position: the response for the preferred object at the swap position (but not the non-swap position) is lower after training, and the C2 response to the non-preferred object is higher at the swap position. As in the physiology experiments performed by Li and DiCarlo, these results are object and position specific.

2.3.4 Individual cell versus population response

In the previous section we modeled the single cell results of Li and DiCarlo, namely that translation invariant representations of objects can be disrupted by a relatively small amount of exposure to altered temporal associations. However, single cell changes do not necessarily reflect whole population or perceptual behavior and no behavioral tests were performed on the animals in this study.

A cortical model with a temporal association learning rule provides a way to model population behavior with swap exposures similar to the ones used by Li and DiCarlo [107, 108]. A C2 cell in the HMAX model can be treated as analogous to an AIT cell (as tested by Li and DiCarlo), and a C2 vector as a population of these cells. We can thus apply a classifier to this cell population to obtain a model of behavior or perception.

2.3.5 Robustness of temporal association learning with a population of cells

We next model the response of a population of cells to different amounts of swap exposure, as illustrated in Figure 2-1, right. The translating image sequence with which we train the model replicates visual experience, and thus jumbling varying amounts of these training images is analogous to presenting different amounts of altered exposure to a test subject as in [107, 108]. These disruptions also model the mis-associations that may occur with temporal association learning due to sudden changes in the visual field (such as light, occlusions, etc), or other imperfections of the biological learning mechanism. During each training phase we randomly swap different face and car images in the image sequences with a certain probability, and observe the effect on the response of a classifier applied to a population of C2 cells. The performance, as measured by area under the ROC curve (AUC), versus different neural population sizes (number of C2 cells) is shown in Figure 2-4 for several amounts of altered exposure. We measured altered exposure by the probability of flipping a face and car image in the training sequence.

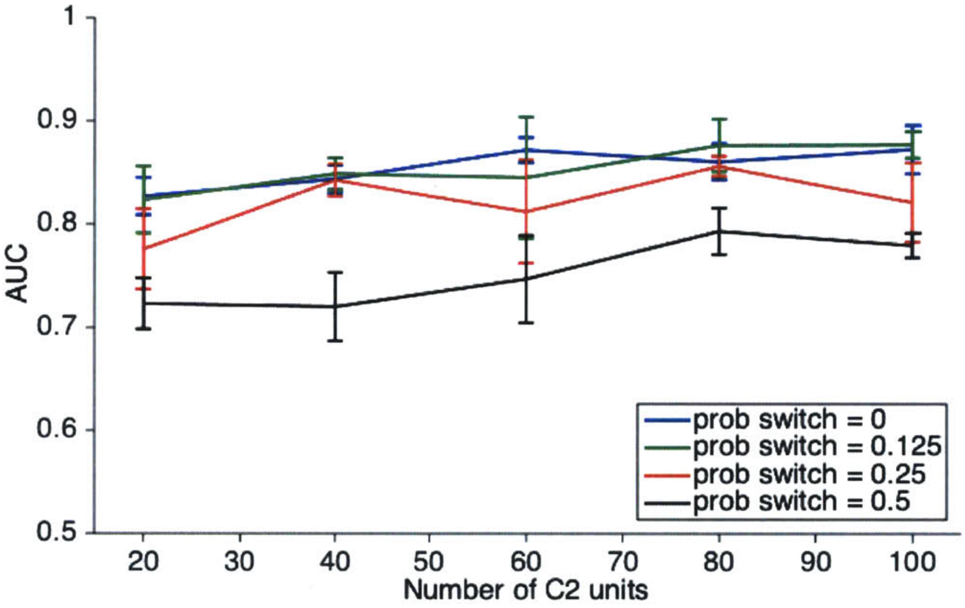


Figure 2-4: Results of a translation invariance task (+/- 40 pixels) with varying amounts of altered visual experience. To model the experimental paradigm used in [107, 108, 26, 183], training and testing were performed on the same altered image sequence. The accuracy (AUC for ROC curve) with a nearest neighbor classifier compared to center face for a translation invariance task versus the number of C2 units. Different curves have a different amount of exposure to altered visual training as measured by the probability of swapping a car and face image in training. The error bars show +/- one standard deviation.

A small amount of exposure to altered temporal training (0.125 probability of flipping face and car) has negligible effects, and the model under this altered training performs as well as with normal temporal training. A larger amount of exposure to altered temporal training (0.25 image flip probability) is not significantly different than perfect temporal training, especially if the neural population is large enough. With enough C2 cells, each of which is learned from a temporal training sequence, the effects of small amounts of jumbling in training images are insignificant. Even with half altered exposure (0.5 image flip probability), if there are enough C2 cells then classification performance is still fairly high (Figure 2-4). This is likely because with similar training (multiple translating faces or cars), redundant C2 cells are formed, creating robustness to association errors that occurred during altered training. Similar redundancies are likely to occur in natural vision. This indicates that in natural learning, mis-wirings do not have a strong effect on learning translation invariance, particularly with familiar objects or tasks.

2.4 Discussion

We use a cortical model inspired by Hubel and Wiesel [76], where translation invariance is learned through a variation of Foldiak's trace rule [49] to model the visual response to altered temporal exposure. We first show that this temporal association learning rule is accurate by comparing its performance to that of a similar model with hard-wired translation invariance [157]. This extends previous modeling results by Masquelier *et al.* [115] for models of V1 to higher levels in the visual recognition architecture. Next, we test the robustness of translation invariance learning on single cell and whole population responses. We show that even if single cell translation invariance is disrupted, the whole population is robust enough to maintain invariance despite a large number of mis-wirings.

The results of this study provide insight into the evolution and development of transformation invariance mechanisms in the brain. It is unclear why a translation invariance learning rule, like the one we modeled, by [26, 107, 108], would remain active after development. We have shown that the errors associated with a continuously active learning rule are negligible, and thus it may be simpler to leave these processes active than to develop a mechanism to turn them off.

Extending this logic to other transformations is interesting. Translation is a *generic* transformation; all objects translate in the same manner, so translation invariance, in principle, can be learned during development for all types of objects. This is not true of "non-generic" or *class-specific* transformations, such as rotation in depth, which depends on the 3-D structure of an individual object or class of objects [148, 106, 105]. For example, knowledge of how 2-D images of faces rotate in depth can be used to predict how a new face will appear after a rotation. However, knowledge of how faces rotate is not useful for predicting the appearance of non-face objects after the same 3-D transformation. Many transformations are class-specific in this sense (e.g. changes in illumination, which depend on both 3-D structure and material properties of objects [105]). One hypothesis as to why invariance-learning mechanisms remain active in the mature visual system could be a continuing need to learn and refine invariant representations for more objects under non-generic transformations.

Disrupting rotation in depth has been studied in psychophysics experiments. Wallis and Bulthoff showed that training subjects with slowly morphing faces, disrupts viewpoint invariance after only a few instances of altered training [183]. This effect occurs with a faster time course than observed in the translation invariance experiments [26]. One possible explanation for this time discrepancy is that face processing mechanisms are higher-level than those for the "greeble objects" and thus easier to disrupt. However, we conjecture that the strong, fast effect has to do with the type of transformation rather than the specific class of stimuli.

Unlike generic transformations, class-specific transformations cannot be generalized between objects with different properties. It is even possible that we learn non-generic transformations of novel objects through a memory-based architecture that requires the visual system to store each viewpoint of a novel object. Therefore, it is logical that learning rules for non-generic transformations should remain active as we are exposed to new objects throughout life.

In daily visual experience we are exposed more to translations than rotations in depth, so through visual development or evolutionary mechanisms there may be more cells dedicated to translation-invariance than rotation-invariance. We showed that the size of a population of cells has a significant effect on its robustness to altered training, see Figure 4. Thus rotation invariance may also be easier to disrupt, because there could be fewer cells involved in this process.

Two plausible hypotheses both point to rotation (class-specific) versus translation (generic) being the key difference between the Wallis and Bulthoff and Cox *et al.* experiments. We conjecture that if an experiment controlled for variables such as the type and size of the stimulus, class-specific invariances would be easier to disrupt than generic invariances.

This study shows that despite unavoidable disruptions, models based on temporal association learning are quite robust and therefore provide a promising solution for learning invariance from natural vision. These models will also be critical in understanding the interplay between the mechanisms for developing different types of transformation invariance.

Chapter 3

The dynamics of size- and position-invariant object recognition in the human visual system

The material in this chapter was published in the Journal of Neurophysiology in 2014 [83].

The human visual system can rapidly recognize objects despite transformations that alter their appearance. The precise timing of when the brain computes neural representations that are invariant to particular transformations, however, has not been mapped in humans. Here we employ magnetoencephalography (MEG) decoding analysis to measure the dynamics of size- and position-invariant visual information development in the ventral visual stream. With this method we can read out the identity of objects beginning as early as 60 ms. Size- and position-invariant visual information appear around 125 ms and 150 ms, respectively, and both develop in stages, with invariance to smaller transformations arising before invariance to larger transformations. Additionally, the MEG sensor activity localizes to neural sources that are in the most posterior occipital regions at the early decoding times and then move temporally as invariant information develops. These results provide previously unknown latencies for key stages of human invariant object recognition, as well as new and compelling evidence for a feed-forward hierarchical model of invariant object recognition where invariance increases at each successive visual area along the ventral stream.

3.1 Introduction

Humans can identify objects in complex scenes within a fraction of a second ([138, 171]. The main computational difficulty in object recognition is believed to be identifying objects across transformations that change the photoreceptor-level representation of the object, such as position in the visual field, size, and viewpoint [32]. Invariance to these transformations increases along the ventral visual pathway [112, 10, 147, 152], and the latencies of the visual areas along this pathway (from V1 to IT) are known in the macaque [126, 155, 172, 80]. For instance, position and size invariance is found in macaque IT at about 100 ms. In humans, electroencephalography (EEG) studies have shown that neural signals containing object category information can be found at 150ms or later [15, 171, 95], however, the timing and steps to develop the invariant object representations that drive this categorization are still unknown. To understand the timing of invariant object recognition in humans, we use a technique called neural decoding analysis (also known as multivariate pattern analysis, or readout). Neural decoding analysis applies a machine learning classifier to assess what information about the input stimulus (e.g., what image the subject was looking at) is present in the recorded neural data. This technique is widely used in functional magnetic resonance imaging (fMRI) [70] and brain-machine interfaces [34], and has also been applied to electrophysiology data [80, 119], EEG data [134, 135, 142], and MEG motor [182] and semantic data [163]. These analyses, however have only been applied to visual data in a few instances [65, 21, 22]. MEG provides high temporal resolution, whole-head neural signals, making it a useful tool to study the different stages of invariant object recognition throughout the brain. Using MEG decoding we could identify the precise times when neural signals contain object information that is invariant to position and size. We also examined the dynamics of these signals with high temporal accuracy and estimated their underlying neural sources. Finally, we compared the timing data uncovered here to a feed-forward model of invariant object recognition in the ventral stream. These results allow us to draw conclusions about when and where key stages of invariant object recognition occur, and provide insight into the computations the brain uses to solve complex visual problems.

3.2 Materials and Methods

3.2.1 Subjects

Eleven subjects (three female) age 18 or older with normal or corrected to normal vision took part in the experiment. The MIT Committee on the Use of Humans as Experimental approved the experimental protocol. Subjects provided informed written consent before the experiment. One subject (S1) was an author and all others were unaware of the purpose of the experiment.

3.2.2 Experimental Procedure

In this experiment, subjects performed a task unrelated to the images presented. The images were presented in two image blocks and the fixation crossed changed color (red, blue or green) when the first image was presented, then changed to black during the inter-stimulus interval, and then turn a second color when the second image was presented. The subjects' task was to report if the color of the fixation cross was the same or different at the beginning and end of each two image (Figure 3-1), and thus helped ensure that they maintained a center fixation while both images were presented (this was also verified for two subjects with eye tracking, see below).

To evaluate the robustness of the MEG decoding methods, three subjects (S1, S2, S3) were each shown a different dataset of images presented at one size and position. Subject S1 was shown 25 scene images (from stanford.edu/fmriscenes/resources.html) presented in the center of the visual field at a size of 4x6 degrees of visual angle, Subject S2 was shown 26 black letters on white background presented in the center

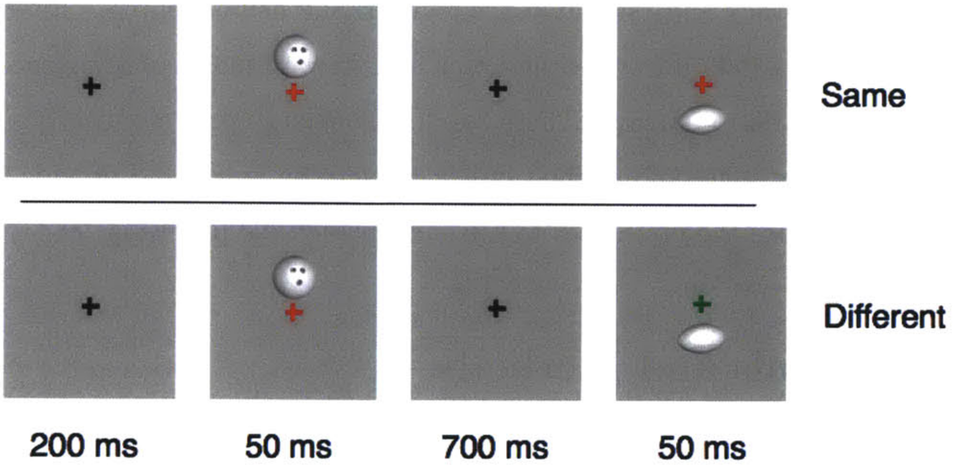


Figure 3-1: Experimental task. In order to keep their gaze at the center of the screen, the subjects' task was to report if the color of the fixation cross was the same or different at the beginning and end of each two image. a) Illustrates a trial where the fixation cross is the same color (red) at beginning and end, and b) illustrates a trial where the fixation cross changes color (from red to green) between beginning and end. The fixation cross changed color when the images were on the screen and was black between stimulus presentations.

of the visual field at a size of 5x5 degrees of visual angle, and Subject S3 was shown 25 isolated objects on a gray background, presented in the center of the visual field at a size of 5x5 degrees of visual angle (Figure 3-2, right). To study size- and position-invariance, eight subjects (S4-S11) were shown the same subset of six images from the isolated objects dataset, presented at three sizes (two, four and six-degrees of visual angle in diameter) in the center of the visual field, and three six-degree diameter images shown at three positions (centered, and +/- three degrees vertically).

Images were presented for 48ms with 704 ms inter-stimulus interval. Image order was randomized for each experiment, and each stimulus was repeated 50 times. All images were presented in grayscale on a 48cm x 36 cm screen, 140 cm away from the subject, thus the screen occupied 19 x 14 degrees of visual angle.

3.2.3 Eyetracking

To verify that the above subjects maintain central fixation, eyetracking was performed during MEG recordings for two subjects (S9, S10) with the Eyelink 1000 eye tracker from SR Research. A 9-point calibration was used at the beginning of each experiment. We discarded trials that were greater than two degrees away from the mean eye position, which we used as center to account for calibration errors, or that contained artifacts such as blinks. 6% of trials were rejected for subject S9 and 11% were discarded for subject S10. Discarding data did not have a significant effect on decoding, so the data shown contains all trials for each subject.

3.2.4 MEG recordings and data processing

The MEG scanner used was an Elekta Neuromag Triux with 102 magnetometers at 204 planar gradiometers, and the MEG data was sampled at 1000 Hz. The MEG data were pre-processed using Brainstorm software [166]. First the signals were filtered using Signal Space Projection for movement and sensor contamination [170]. The signals were also band-pass filtered from 2-100 Hz with a linear phase FIR digital filter to remove external and irrelevant biological noise, and the signal is mirrored to avoid edge effects of band-pass filtering. Recent studies have shown that high-pass filtering may lead to artifacts that affect evoked response latencies in MEG/EEG data [1, 149]. To ensure that the high-pass filter threshold did not affect our results, we performed one set of analyses with a 0.01 Hz high-pass filter threshold, and observed no noticeable difference in the latency or shape of decoding accuracy.

3.2.5 Decoding analysis methods

Decoding analyses were performed with the Neural Decoding Toolbox [117], a Matlab package implementing neural population decoding methods. In this decoding procedure, a pattern classifier was trained to associate the patterns of MEG data with the stimulus conditions that were present (the identity of the image shown) when the MEG recording were made. The amount of information in the MEG signal was evaluated by testing the accuracy of the classifier on a separate set of test data. In our analyses, data from both magnetometers and gradiometers were used as features that were passed to the pattern classifier (we found both types of sensors had information that contributed to increasing the decoding performance). We also averaged the MEG in 5 ms non-overlapping bins (i.e. each sensor's activity was averaged within each 5 ms time window) prior to beginning the decoding procedure.

All decoding analyses were performed with a cross-validation procedure where the classifier is trained on a subset of the data and then the classifier's performance is evaluated on the held-out test data. Our recordings consisted of 50 repetitions of each stimulus condition (see 'Experimental Procedures' above). For each decoding run, data from these 50 trials were divided into 5 sets of 10 trials, and the data from each set of 10 trials were averaged together. We were also able to decode without this averaging (using single trials), but found that averaging trials led to an increase in the signal to noise ratio of our results (see Figure 3-3). This gave rise to five cross-validation splits. The classifier was trained on 4 of these splits (80% of the data) and then tested on the remaining split (20% of the data), and the procedure was repeated 5 times leaving out each cross-validation split.

In each training phase of the decoding procedure, the mean and standard deviation of the each sensor over the entire time series was used to Z-score normalize the data. Additionally, an analysis of variance (ANOVA) test was applied to the training data to select the 25 sensors at each time point that are most selective for image identity (those sensors with the lowest p-values determined by an F-test). The test data was then z-score normalized using the mean and standard deviation learned from the training data, and only the top 25 sensors that had the lowest p-values were used when testing the classifier. The pattern of the most selected sensors was very localized to the occipital portion of the sensor helmet beginning 60ms after stimulus onset (Supplemental Video 1).

Decoding analyses were performed using a maximum correlation coefficient classifier. This classifier computes the correlation between each test vector x^* and a vector $\bar{x}_i$ that is created from taking the mean of the training vectors from class i. The test point is assigned the label of the class of the training data with which it is maximally correlated. This can be formulated as:

$$i^* = argmax_i(corr(x*, \bar{x_i}))$$

The classification accuracy is reported as the percentage of correct trials classified in the test set averaged over all cross-validation splits. This decoding procedure was repeated for 50 decoding runs with different training and test cross-validation splits being generated on each run, and the final decoding accuracy reported is the average decoding accuracy across the 50 runs. For more details on the decoding procedure, and to view the code used for these analyses, please visit http://www.readout.info.

The decoding parameters, including number of stimulus repetitions, number of trials averaged, number of sensors used, bin width, and classifier used in decoding were chosen to maximize a signal to noise ratio (SNR), defined as the peak decoding accuracy divided by the standard deviation during the baseline period. Using data from the initial 3 subjects on the 25 image discrimination tasks (Figure 3-2), we found good SNR values for most of these parameter settings (Figure 3-3(a)-(e)). The results showed 50 stimulus repetitions were more than sufficient to provide good SNR, and that averaging 10 trials and selecting 25 features led to a clear increase in decoding performance. In addition, small bin size not only led to an increase in decoding performance, but also allowed us to interpret our results with finer temporal resolution. Next, we performed the decoding analysis using several different classifiers (correlation coefficient, support vector machine, and regularized least squares with linear and Gaussian kernels), and found that classifier choice did not affect decoding accuracy (Figure 3-3(f)). Consequently, in order to have the clearest results possible to examine the effects of interest, we use 50 stimulus repetitions, the average 10 trials, the 25 most selective features, 5 ms bin width, and a correlation coefficient classifier for subsequent invariance analyses.

3.2.6 Significance criteria

We assessed significance using a permutation test. To perform this test, we generated a null distribution by running the full decoding procedure 200 times using data with randomly shuffled labels with 10 cross-validation split repetitions used on each run. Decoding results performing above all points in the null distribution for the corresponding time point were deemed significant with p<0.005 (1/200). The first time decoding reached significantly above chance ("significant time") was defined as the point when accuracy was significant for two consecutive time bins. We chose this significance criterion was selected such that no spurious correlations in the baseline period were deemed significant. This criterion was met for all decoding experiments, except one subject in one position-invariance condition (S7, train down/test up condition) whose data was still included in our analyses.

3.2.7 Significance testing with normalized decoding magnitudes

To examine the effect of decoding magnitude on significance time, we also performed a procedure to approximately normalize the peak decoding accuracy across trials. We then repeated this significance testing to see the latencies across different conditions with normalized magnitudes. To normalize the decoding magnitude for different conditions, we included less data for those conditions with higher decoding accuracy: if the peak decoding magnitude was above .7 for one condition or pair of conditions (in the case of invariance conditions, the average of each train and test pair was considered) we performed decoding with 20% of data collected, if the peak decoding magnitude was between 0.6-0.7 we performed decoding with 30% of data collected, and if the peak decoding magnitude was between 0.44-0.6 we performed decoding with 50% of the data collected. After this normalization procedure, peak decoding accuracy for all conditions fell within the same narrow range of 33%-43%. Decoding analysis was still performed with five cross-validation splits, and all the data in each split (3 trials for those conditions using 30% of data, and 5 trials for those conditions using 50% of data) was still averaged at each cross validation run. All other decoding parameters were kept the same. This procedure adjusted the peak decoding magnitudes for each condition so they were between the 0.33-0.44 desired range.

3.2.8 Source localization

We used the minimum norm estimate (MNE) distributed source modeling method, which finds the set of sources along the cortical surface that minimizes the total power of the sources [67], for three subjects (S9-S11) using Brainstorm software. MNE was performed using the cortical orientation constraints and with the default SNR value (signal-to-noise ratio of power of data) of 3. The sources were estimated on the colin27 standard brain template [75]. (Head positions for S1-S8 were not measured in the scanner, so they were excluded from this analysis.) A head model was generated for each subject's head position using the overlapping spheres method. A full noise covariance matrix from the 233 ms baseline period of 1530 visual presentations was generated for each subject, and used in the MNE algorithm. Sources that were $p<0.001$ significant (based on a t-test versus the baseline period, Bonferroni corrected for multiple comparisons) were selected.

3.2.9 Cortical modeling (HMAX)

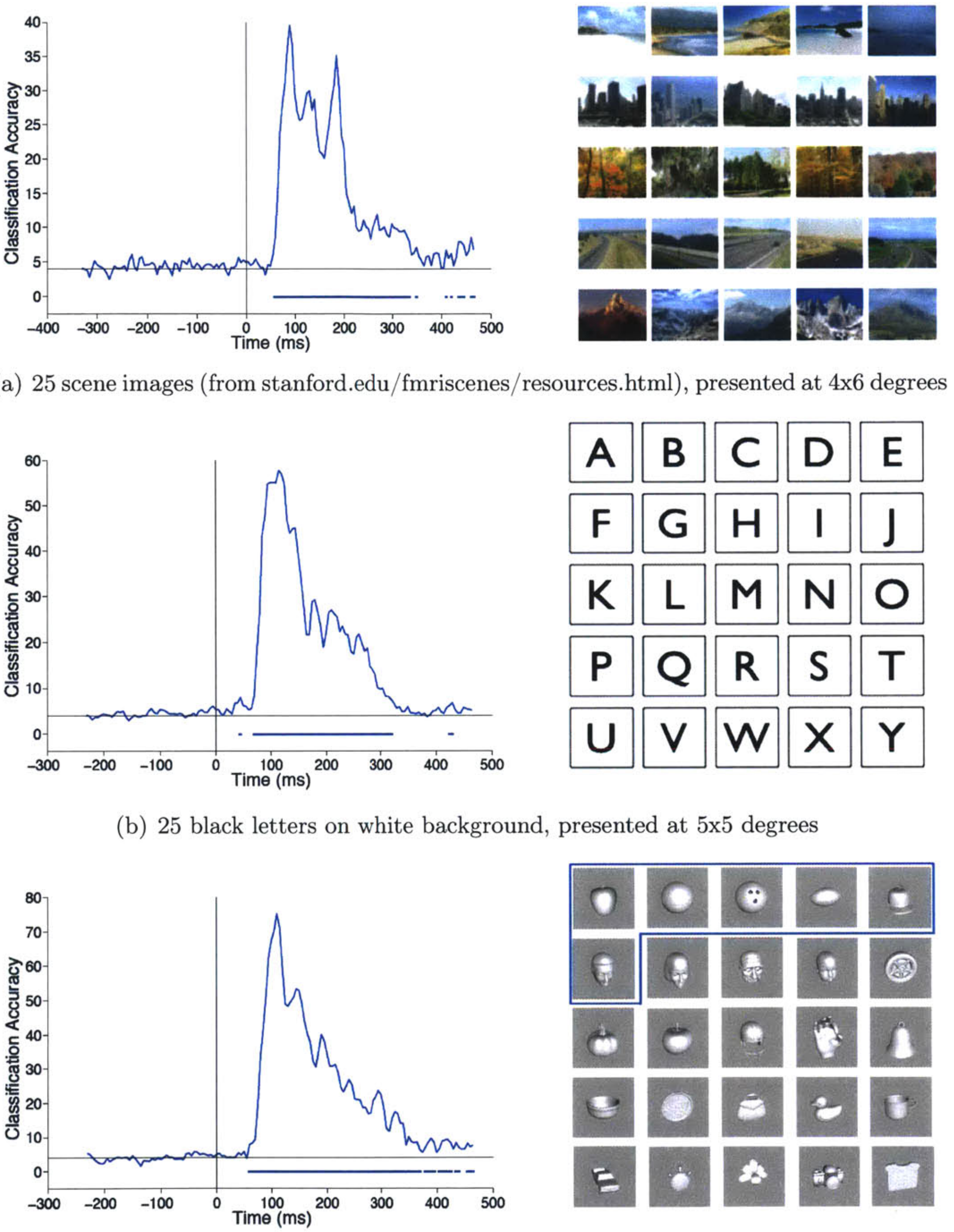
To model the MEG invariant decoding results, we tested the HMAX model [157]. The model consists of alternating layers of simple units and complex units. Simple cells perform a template matching operation between its inputs and stored templates (in the first layer these templates are oriented Gabor functions, similar to those found in primary visual cortex) to build selectivity, and complex cells perform a pooling operation over local regions (here we use max pooling) to build invariance. HMAX was implemented using the Cortical Network Simulator GPU-based framework [123]. The HMAX parameters used were the same as in [157]. 1000 model units were randomly sampled at each model layer, and used as the feature vector for classification. As in the decoding procedure, a correlation coefficient classifier was used to classify the same image across two different sizes or positions, at each model layer. This procedure was repeated ten times and results were averaged.

3.3 Results

3.3.1 Fast and robust readout for different types of stimuli

To examine whether we could extract visual information from MEG signals, we first decoded the identity of the presented images. Three subjects were each shown a different stimulus set, which consisted of either images of scenes, images of letters, or images of isolated objects (Figure 3-2 (a)-(c), right), while MEG signals were recorded from 306 sensors covering the full head. The stimulus sets each had 25 images, and each image was shown 50 times to each subject. We trained a correlation coefficient classifier to discriminate between the different images based on the subject's MEG data. The MEG signals were averaged over 5ms time bins, and data from 10 different trials were averaged together. The classifier was trained and tested separately on each time bin and the 25 most selective sensors were chosen in training for each time point (see 3.2.5). These decoding parameters were chosen maximize signal to noise in the recordings (Figure 3-3).

For each image set and subject, we could reliably decode the identity of the 25 different images in the set. Decoding was significantly above chance (based on a $p<0.005$ permutation test) from 60-335 ms after stimulus presentation for scene images, 70-325 ms for letter images, and 60-370 ms for object images (Figure 3-2(a)-(c), left). The peak decoding accuracies ranged from 38-70% correct (chance accuracy is 4%), showing that we were able to reliably extract information from MEG signals from a large range of different stimulus sets.

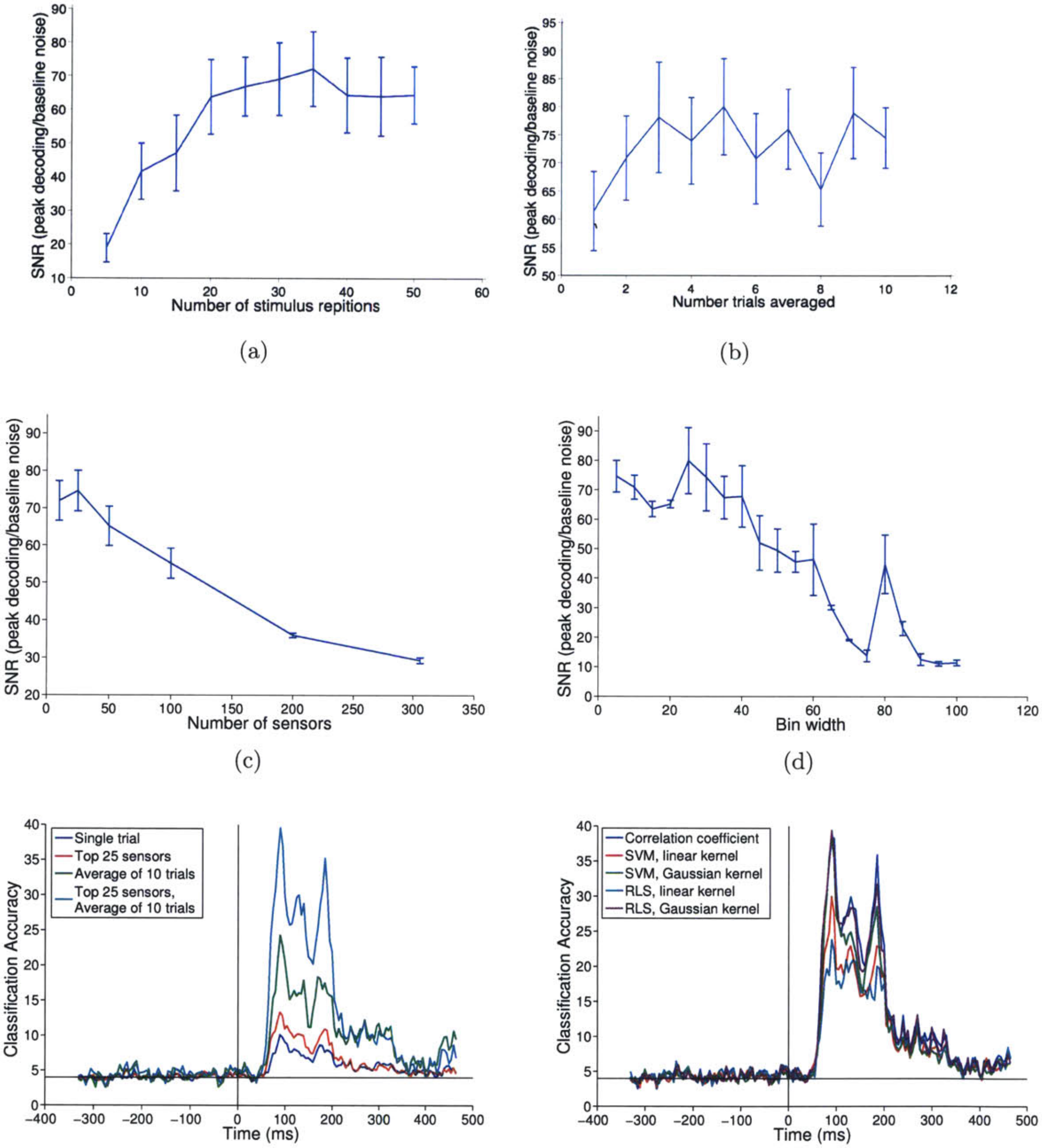


(a) 25 scene images (from stanford.edu/fmriscenes/resources.html), presented at 4x6 degrees

(b) 25 black letters on white background, presented at 5x5 degrees

(c) 25 isolated objects on a gray background, presented at 5x5 degrees (thumbnail images in blue box indicate the subset used in subsequent invariance experiments).

Figure 3-2: Decoding accuracy versus time for three different image sets. Time zero corresponds to the time of stimulus onset. Each image set was run on a separate date with a separate subject. Please note the change in scale for classification accuracy (y-axis) across the three sub-plots. The horizontal line indicates chance performance. The bars at the bottom of each plot indicate when decoding was significantly above chance ($p<0.005$, permutation test).


SNR (peak decoding/baseline noise)
Number of stimulus repitions
0 10 20 30 40 50 60
10 20 30 40 50 60 70 80 90
SNR (peak decoding/baseline noise)
Number trials averaged
0 2 4 6 8 10 12
50 55 60 65 70 75 80 85 90 95
SNR (peak decoding/baseline noise)
Number of sensors
0 50 100 150 200 250 300 350
20 30 40 50 60 70 80 90
SNR (peak decoding/baseline noise)
Bin width
0 20 40 60 80 100 120
0 10 20 30 40 50 60 70 80 90
Classification Accuracy
Time (ms)
Single trial
Top 25 sensors
Average of 10 trials
Top 25 sensors, Average of 10 trials
−400 −300 −200 −100 0 100 200 300 400 500
0 5 10 15 20 25 30 35 40
Classification Accuracy
Time (ms)
Correlation coefficient
SVM, linear kernel
SVM, Gaussian kernel
RLS, linear kernel
RLS, Gaussian kernel
−400 −300 −200 −100 0 100 200 300 400 500
0 5 10 15 20 25 30 35 40

(a)

(b)

(c)

(d)

(e)

(f)

Figure 3-3: Parameter optimization. The effect of a) number of stimulus repetitions used in decoding (using single trial data, he top 25 features, and 5 ms bin width), b) number of trials averaged (using 50 stimulus repetitions, the top features, and 5 ms bin width), c) number of sensors used in decoding (using 50 stimulus repetitions, the average of 10 trials, and 5 ms bin width), and d) bin width (using 50 stimulus repetitions, the average of 10 trials, and the top 25 sensors) on signal to noise ratio (SNR). SNR is measured by the peak decoding height divided by the baseline noise (standard deviation of the decoded signal before stimulus onset). SNR data is averaged for three subjects (S1-S3) on three different data sets (25 scenes, 26 letters, and 25 isolated objects), and the error bars show standard error from the mean. e) The combined effects of different number of trials averaged and number of sensors used in decoding on decoding accuracy versus time for one subject (S1). f) The effect of classifier on decoding accuracy versus time for one subject (S1).

3.3.2 Timing of size and position invariant visual representations

Once we established that we could decode basic visual information from MEG signals, we then tested whether we could detect visual representations that are invariant to image transformations. To do this we presented a subset of six of the isolated object images (shown in Figure 2c, right in blue box) at various sizes and positions to eight different subjects. We presented large images (6x6 degrees of visual angle) centered and in the upper and lower halves of the visual field (+/- 3 degrees vertically), and presented centered images at medium and small sizes (4x4 and 2x2 degrees of visual angle, respectively). To make sure any invariant information we extracted was not due to eye-movements we used a brief presentation time of less than 50 ms and randomized position and size of the image. Humans require at least 80-100 ms to make a saccadic eye movement [45, 17], thus presenting images for only 50ms in a random position ensured subjects would not be able to saccade to peripheral images. Eye position was also measured for two of the eight subjects (see 3.2.3).

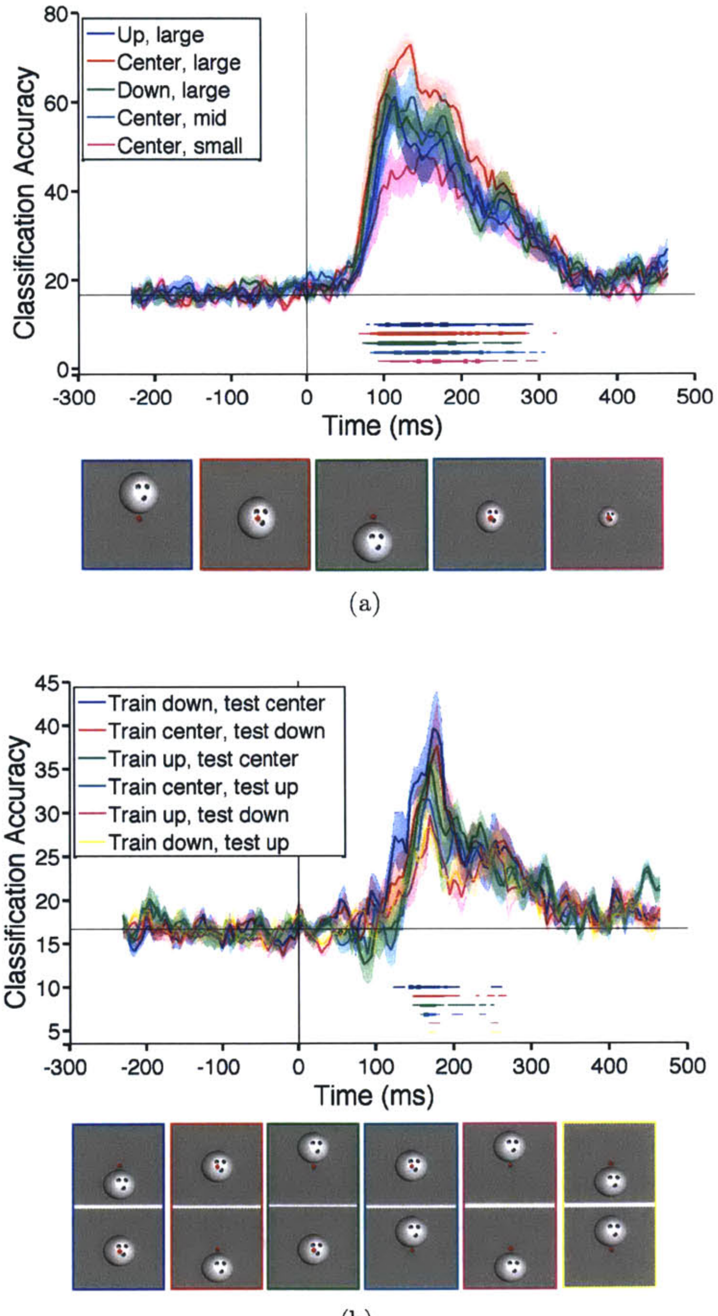
As a first check to make sure that we could extract similar visual information from this new stimulus set, we decoded the identity of the images at each of the five different position and size conditions (Figure 3-4(a)). The results showed a similar time course as the larger image sets in (Figure 3-2(b)), indicating that our initial results generalized to both the new stimulus set and the larger number of subjects.

We next sought to detect position-invariant information by training the classifier on data from images presented at one position and testing it on images presented at a second position. This technique allowed us to detect when common neural representations arose between images of the same object presented at two different positions - i.e., representations that are invariant to position. Using this method, we detected position invariant visual signals for the six different position comparisons beginning at 150 ms on average (Figure 3-4(b)). Similarly, to detect size-invariant visual signals, we trained the classifier on data from images presented at one size and tested it on data from images presented at a second size for six different size comparison cases (Figure 3-4(c)). On average, size-invariant information was first detected around 125 ms. These results provide previously unknown human latencies for size and position invariant object recognition, which are consistent across subjects. Additionally they uncover a potential latency difference between size and position invariant processing, which may have interesting implications for how and where in the visual pathway these two types of transformations are processed.

3.3.3 Varying extent of size and position invariance

To quantify when non-invariant, position-invariant and size-invariant information rises and peaks, we looked at the first time decoding rose significantly ($p<0.005$ permutation test) above chance for two consecutive 5ms time bins, and the time when decoding reached peak performance. The non-invariant information appeared at 80 ms and peaked at 135 ms, on average, which was before the size-invariant information (125 ms appearance, 170 ms peak) and the position-invariant information (150 ms appearance, 180 ms peak) (Figure 3-5(a)).

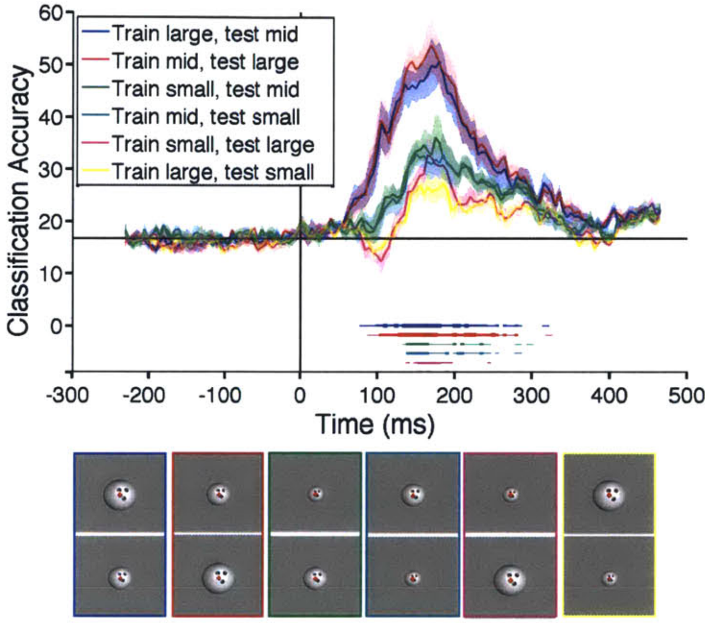
We also looked at the individual invariant decoding conditions, which showed that position and size invariant information developed in stages, with the decoding signals from the smaller transformed cases rising before signals from the larger transformed cases (Figure 3-5(b)). The three-degree position-invariance cases (lower/centered and centered/upper) both developed before the six-degree position-transformation cases



Classification Accuracy
80
60
40
20
0
-300
-200
-100
0
100
200
300
400
500
Time (ms)
Up, large
Center, large
Down, large
Center, mid
Center, small
45
40
35
30
25
20
15
10
5
Train down, test center
Train center, test down
Train up, test center
Train center, test up
Train up, test down
Train down, test up

(a)

(b)



(c)

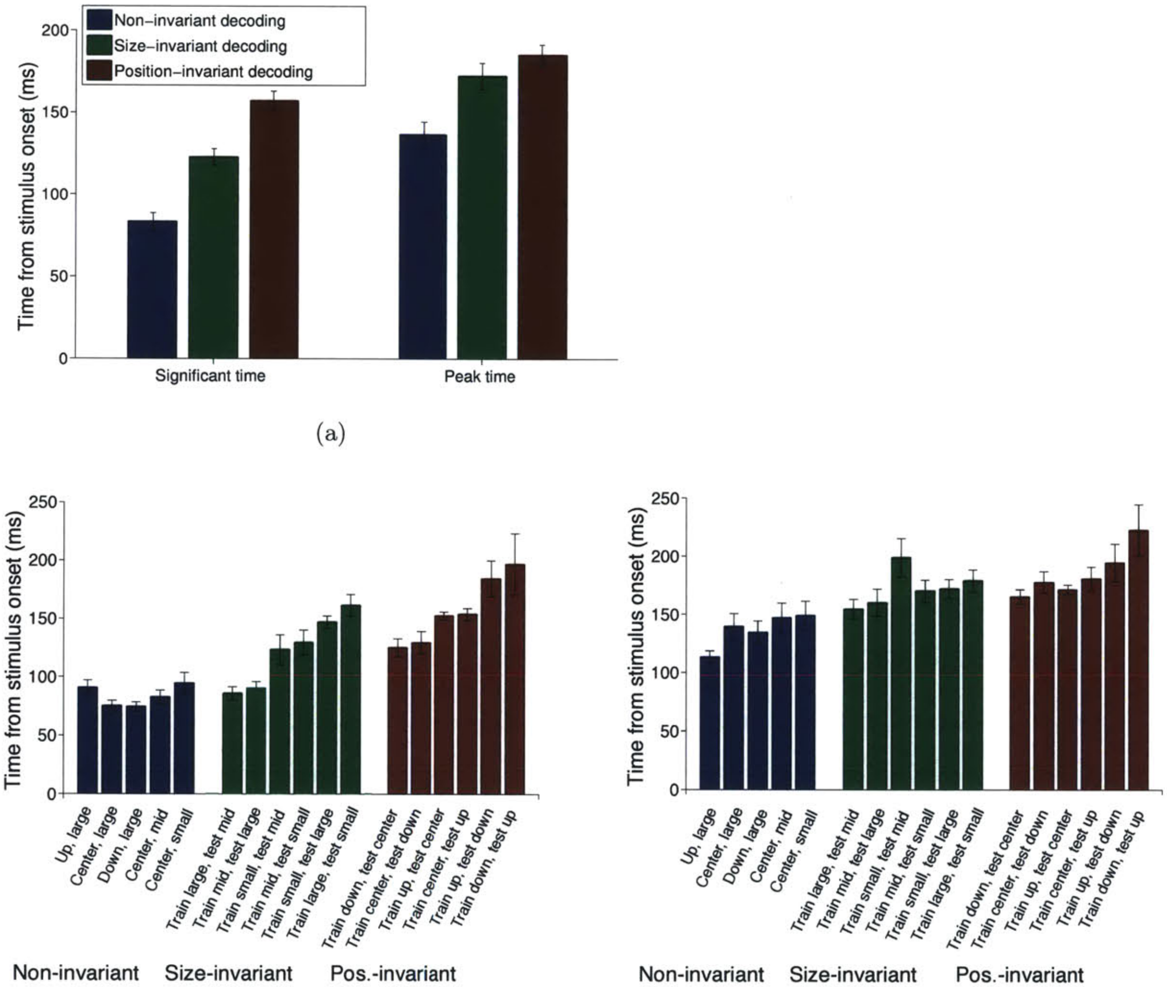
Figure 3-4: Assessing position and size invariant information. Six different images of isolated objects (Figure 3-2c, right in blue box) were presented at three different positions (centered, and +/- three degrees vertically) and three different sizes (two, four and six degrees in diameter). The different training and test conditions are illustrated using a bowling ball-like object, one of the six images used in this experiment. Classification accuracy versus time is plotted for a) average of subjects' results to five non-invariant conditions (illustrated below plot); b) average of subjects' results to six position-invariant conditions; c) average of subjects' results to six size-invariant conditions. Please note the change in scale for classification accuracy (y-axis) across the four sub-plots. The horizontal line indicates chance performance. Error bars represent the standard error across mean accuracy per subject. The bars below each plot indicate when decoding was significantly above chance ($p<0.005$, permutation test) for four (thinnest line), six (middle line), or all eight (thickest line) subjects for each condition, indicated by the color of the bar.

(lower/upper). A similar order was true of the size invariant cases with the 2.25x area increase appearing first (larger/middle), followed by the 4x increase (middle/small), and finally the 9x size increase (large/small). A similar trend is true when you examine peak times, however, there is much less spread in these latencies as most signals tend to peak around the same time (Figure 3-5(c)). This modular development indicates that size and position invariant signals are being computed in stages by a hierarchical system that increases invariance in a feed forward manner.

An alternative possibility is that the difference in decoding latencies is an artifact of the different magnitudes of decoding accuracy across conditions. In general, conditions with higher peak decoding accuracy also had shorter latency, and its possible that these conditions could surpass the level of noise sooner thus with significance testing appear to have shorter latencies only due to their higher magnitude. To test this possibility, we normalized the decoding magnitudes for different decoding conditions by including only a fraction of the MEG data for the conditions with higher accuracy (Table 1). By including 20-50% of the data for certain conditions (please see 3.2.7) we were able to approximately normalize the decoding magnitudes across condition. Importantly, there was little effect on decoding latency, and the decoding order shown in Figure 3-5 still held for normalized decoding magnitudes.

3.3.4 Combined size- and position-invariance

Using these results, we were also able to look at combined size- and position-invariant visual processing, by performing decoding across the two types of transformations: training with centered, large images and testing with small or medium images presented in the upper and lower halves of the visual field, and vice versa (Figure 3-6). In two cases (center/small versus up/large, and center/small versus down/large, Figure 6a-b), the corresponding size-invariant and position-invariant decoding had similar magnitude, but in the two other cases (center/medium versus up/large, and center/medium versus down/large, Figure 3-6(c)-(d)), the corresponding size-invariant



(a)

(b) Onset time

(c) Peak time

Figure 3-5: Significant and peak invariant decoding times. Significant and peak decoding times averaged across subjects for (a) the mean of all non-invariant, position-invariant and size-invariant conditions, (b) significant decoding time for each individual condition, and (c) peak decoding time for each individual condition. Significant decoding times indicate the first time decoding is significantly above chance ($p<0.005$, permutation test) for two consecutive 5ms time bins. Peak decoding time is the maximum decoding performance over the entire time window. Error bars represent standard error across subjects.

Condition	Decoding magnitude	Sig. time (ms)	Peak time (ms)	Proportion data in normalization	Norm. decoding magnitude	Norm. sig. time (ms)	Norm. peak time (ms)
Up, large	0.66	91	113	0.30	0.40	101	125
Center, large	0.78	75	140	0.20	0.42	88	137
Down, large	0.69	75	135	0.30	0.42	91	111
Center, mid	0.70	83	147	0.30	0.42	96	148
Center, smal	0.59	95	149	0.50	0.41	98	148
Train large, test mid	0.58	86	155	0.50	0.38	102	161
Train mid, test large	0.69	90	160	0.50	0.38	102	161
Train small, test mid	0.44	123	199	0.50	0.33	127	183
Train mid, test small	0.41	130	170	1.00	0.41	130	170
Train small, test large	0.39	147	172	1.00	0.39	147	172
Train large, test small	0.34	161	179	1.00	0.34	161	179
Train down, test center	0.47	125	165	0.50	0.37	120	164
Train center, test down	0.43	130	178	1.00	0.43	130	178
Train up, test center	0.42	153	171	1.00	0.42	153	171
Train center, test up	0.37	154	181	1.00	0.37	154	181
Train up, test down	0.34	185	195	1.00	0.34	184.50	195
Train down, test up	0.33	178	223	1.00	0.33	178	223

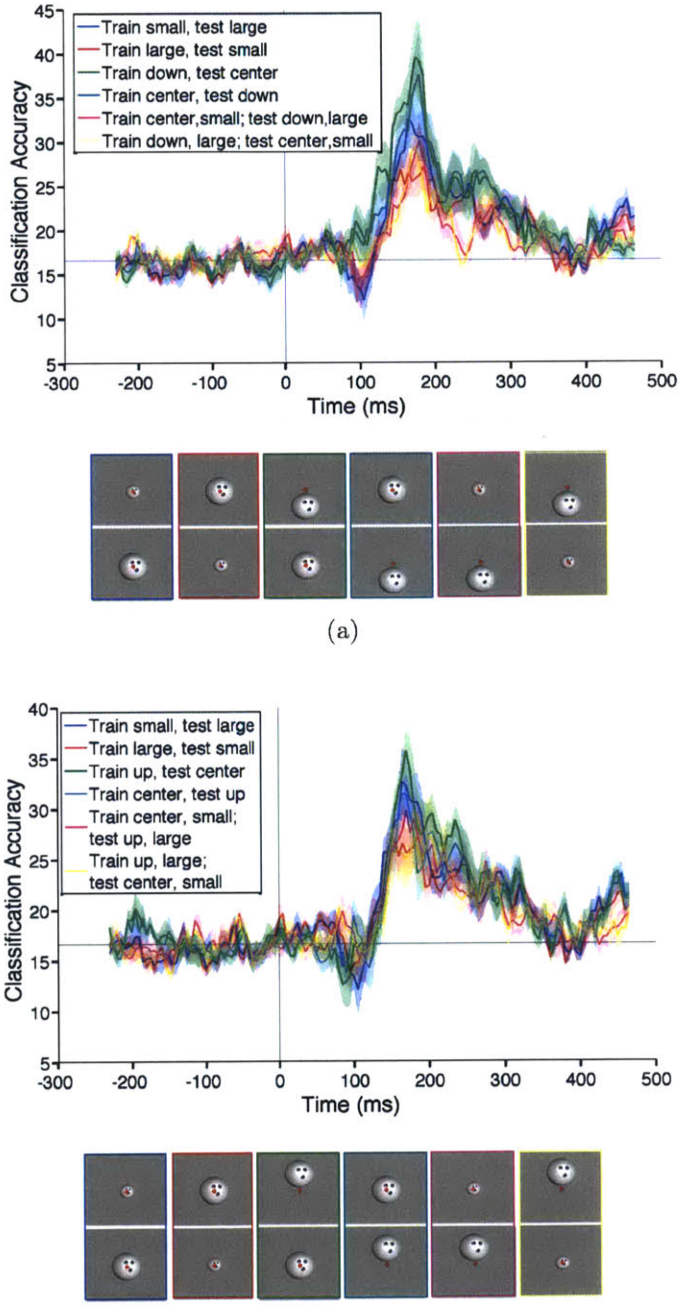
Table 3.1: The above table summarizes the average magnitude of the peak of decoding accuracy, significant time, and peak time for the different size and position conditions using all data (columns 2-4). For those conditions with highest decoding accuracy, a fraction of the total data (column 5) was used to normalize peak decoding accuracy (see 3.2.7), and the modified peak accuracy, significant time and peak time with a fraction of the data are also shown (columns 6-8). Latency values for the normalized decoding values (columns 7-8) are very similar to those from decoding performed with all data (columns 3-4), suggesting that different latencies are not due only to different magnitudes of decoding performance.

decoding occurred much earlier and with larger magnitude than the position invariant decoding. In all four cases the combined size- and position-invariant decoding had similar magnitude and latency to the corresponding position-invariant decoding. This suggests that the slower and lower accuracy transformation, in this case position, limits combined size- and position-invariant decoding, and thus visual processing.

3.3.5 Dynamics of decoded signals

We examined the similarity in the decoded signals at different times by performing a temporal-cross-training analysis (TCT analysis) [119, 118]. In TCT analysis, a classifier is trained with data from one time point, and then tested on data from different trials that were taken either from the same time point or from a different time point. This method yielded a matrix of decoding accuracies for each training and test time bin, where the rows of the matrix indicate the times when the classifier was trained, and the columns indicate the times when the classifier was tested. The diagonal entries of this matrix are the same results as plotted in Figure 3-2, where the classifier was trained and tested with data from the same time points, and again show that there is high decoding accuracy from about 70ms-300ms after stimulus onset.

Additionally, this new analysis showed very low classification accuracy when the classifier was trained and tested at different time points (off-diagonal elements), indicating that different patterns of MEG sensor activity contained object information at different time points in an experimental trial (Figure 3-7(a)). The same pattern was true for a position-invariant case (Figure 3-7(b)) and a size-invariant case (Figure 3-7(c)) with the six-object image. The width of the well-decoded window along the diagonal is 20-50 ms wide, indicating that the neural signal is highly dynamic. Further analysis showed that these dynamics are not due to information moving to different sensors, but instead to the information in a given set of sensors changing over time (Figure 3-7(d)). It is important to note that each sensor coarsely samples several brain areas, so these results do not speak directly to the specific regions driving the decoding.

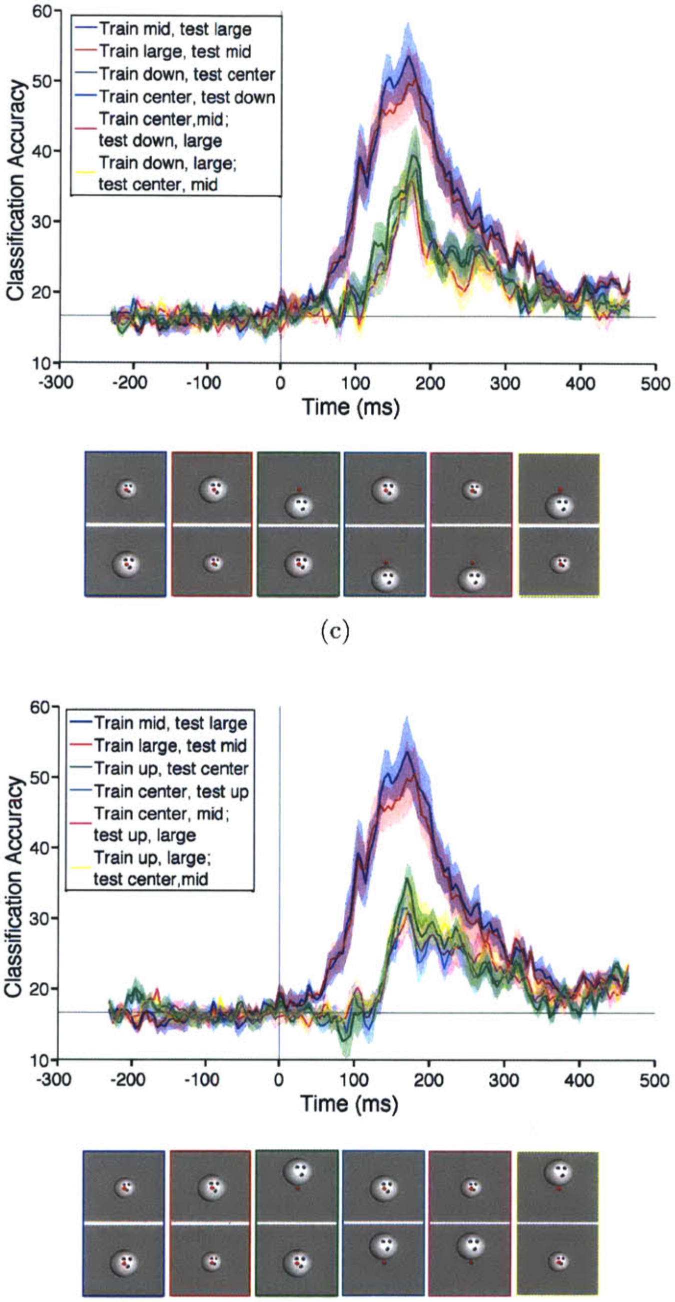


Classification Accuracy
Time (ms)
45
40
35
30
25
20
15
10
5
-300
-200
-100
0
100
200
300
400
500
Train small, test large
Train large, test small
Train down, test center
Train center, test down
Train center,small; test down,large
Train down, large; test center,small

(a)

Classification Accuracy
Time (ms)
40
35
30
25
20
15
10
5
-300
-200
-100
0
100
200
300
400
500
Train small, test large
Train large, test small
Train up, test center
Train center, test up
Train center, small; test up, large
Train up, large; test center, small

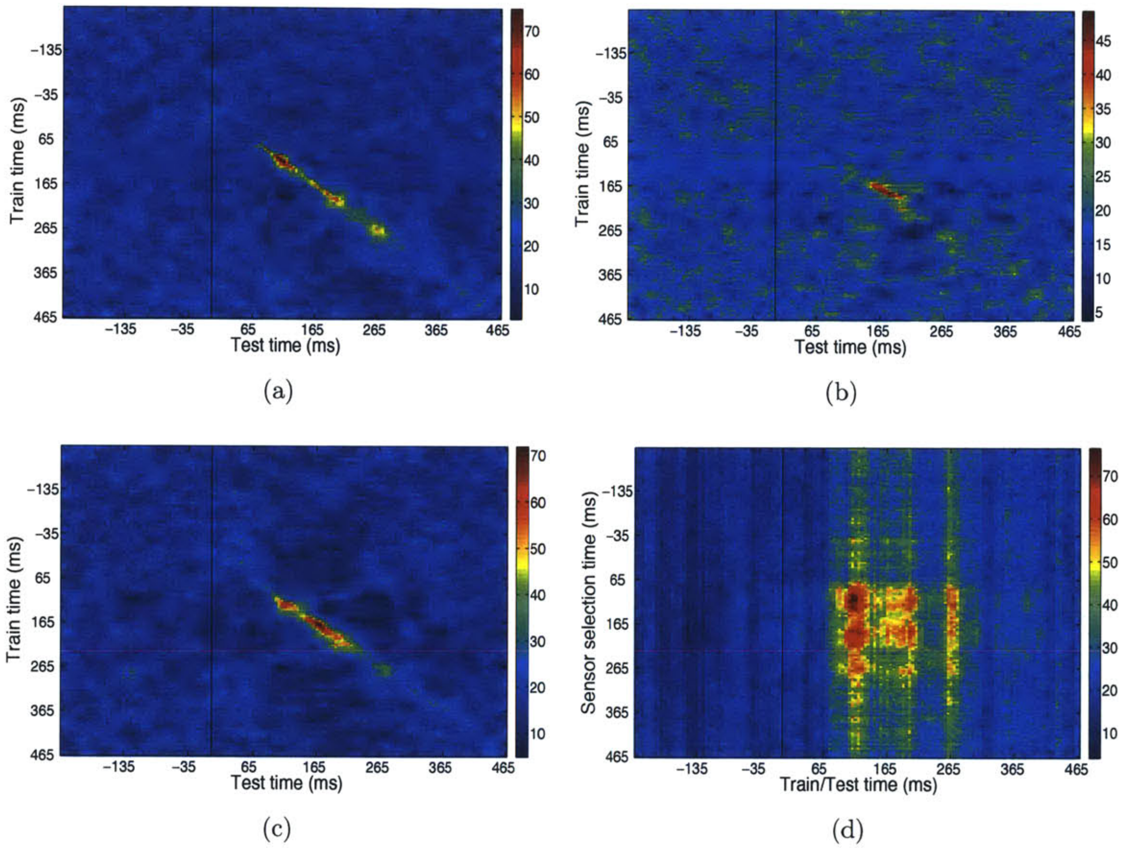
(b)



(c)

(d)

Figure 3-6: Assessing combined size and position-invariant information. Six different images of isolated objects (Figure 3-2(c), right in blue box) were presented at three different positions (centered, and +/- three degrees vertically) and three different sizes (two, four and six degrees in diameter). The different training and test conditions are illustrated using a bowling ball-like object, one of the six images used in this experiment. Classification accuracy versus time is plotted for individual size-invariant (red, blue traces), position-invariant (green, cyan traces) and the corresponding combination of size and position-invariant (pink, yellow traces) decoding in each subplot (a-d).



(a)

(b)

(c)

(d)

Figure 3-7: Dynamics of object decoding. Temporal cross-training matrices showing the decoding results for training the classifier at one point in time and testing the classifier at a second point in time. The color bar on right indicates classification accuracy for one subject on: (a) six-object images as in Figure 3-2 (presented three degrees below center), (b) six-object images decoded invariant to position (train at center, test at three degrees below), and (c) six-object images decoded invariant to size (train at 4 degree diameter, test at 6 degree diameter). High decoding accuracies are only achieved when the classifier is trained and tested within the same time window, indicating that the object information is contained in unique patterns across the sensors at different time points. d) Matrix of classification accuracy (for same subject on six-object dataset) with 25 sensors selected at one time (y-axis), and decoding (training and testing) performed at all other times using these sensors (x-axis). Color bar on right indicates classification accuracy for the experiment.

3.3.6 Neural sources underlying sensor activity and classification

To understand which brain areas were behind the high decoding performance, we used a distributed source localization algorithm to determine where the primary neural sources are located at key decoding times (see Section 3.2.8). We measured head position in the scanner for three subjects during the six-image invariance experiment. We examined the sources for images presented at each individual position and size, as well as for an average of all image presentations across all positions and sizes, shown in (Figure 3-8). Sources for the individual conditions looked similar to the overall average.

When identity-specific information first appears in most subjects, at 70ms, the strongest neural sources were localized in the occipital lobe near early visual areas (Figure 3-8, top row). When both size and position invariant information is present in the signal, at 150ms, the neural sources were located more temporally, further down the ventral visual stream (Figure 3-8, bottom row). The strongest sources at each time point are a good indication of the brain region carrying visual information (see Section 3.5), and indicate that very occipital areas are driving early decoding, while later visual areas contain size and position invariant visual information.

3.3.7 Invariant recognition with a cortical model

To make sure that low level visual features could not account for the invariance results, we tested a hierarchical model of object recognition, HMAX [157], on our six-object dataset to compare with our experimental invariance results. The model, which is inspired by Hubel and Wiesel's findings in V1 [76], consists of alternating layers of simple cells that build selectivity and complex cells that build invariance. Each stage of the model yields a set of features that models the representations contained in different brain regions in the ventral visual processing stream. To test whether features from different stages of the model could account for the invariant decoding results, we applied the model to the same six-object image set presented at the same

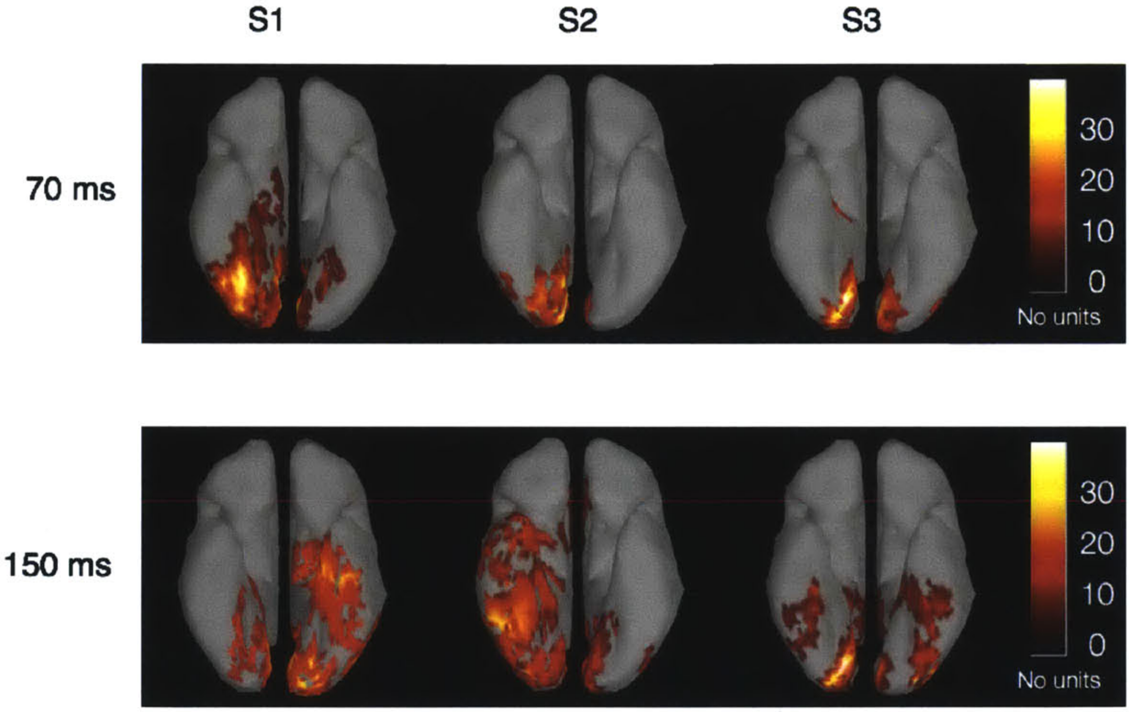


Figure 3-8: Source localization at key decoding times. Source localization results for MEG signals of three subjects (left, center, right) on a standard cortex at (top row) 70 ms, when decoding first rises significantly above chance, and (bottom row) 150 ms, when position and size invariant decoding both rise significantly above chance. Ventral view is presented. Color bar at right indicates magnetic poles strength in picoAmpere-meters. Sources are thresholded to only show source activity that is significantly above chance with $p<0.001$ significance criteria based on a t-test versus the baseline period, Bonferroni for multiple comparisons (see 3.2.8).

sizes and positions and then applied a classification analysis to the different layers of model outputs that was analogous to MEG invariance analyses.

The results showed that a V1-like model, consisting of the first pair of simple/complex cell layers, was not able to achieve above-chance performance on the size and position invariance-decoding task. A mid-level visual model, consisting of an additional layer of simple/complex cells, however, could classify smaller transformed images with above chance performance. The final model output, which modeled cells in anterior inferior temporal cortex and employed global tuning/pooling, was able to classify the transformed images with high performance for each invariance case (Figure 3-9). The model results show a sequential order of invariance (smaller transformations before larger transformations), which is similar to the MEG experimental results. This data provides further evidence that a feed forward, hierarchical model can account for the timing of experimental invariance results, suggesting that the timing may be directly related to the location of the invariance computations.

3.4 Discussion

While it is widely believed that the ventral visual processing stream is involved in object recognition, how this pathway builds up representations that are invariant to visual transformations is still not well understood. Here we addressed this issue by comparing the time course of invariance to two types of transformations, position and size, in the human brain. Using MEG decoding we were able to see the temporal flow of invariant information much more clearly than was possible using conventional analyses.

We detected image identity information as early as 60 ms, and size and position-invariant visual signals at 125 and 150 ms, respectively. The timing of the initial identity decoding is similar to the latency of macaque V1, which is around 60 ms. Additionally, the timing for size and position invariant information is close to the latencies of size and position-invariant signals in macaque IT, which first occur around 100 ms [80]. The slightly longer latencies seen in our study are likely due to the fact

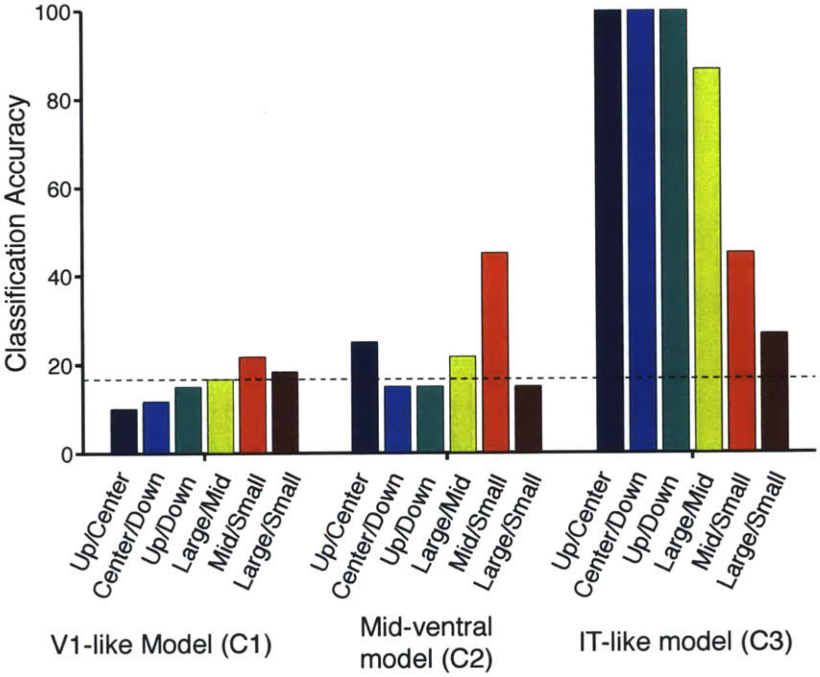


Figure 3-9: Invariant image classification with a cortical model. Classification accuracy for various stages of HMAX model output for three position invariance cases (images in the upper-half of the visual field and centered, centered and lower-half of the visual field, and upper and lower-halves of the visual field) and three size invariance cases (large and mid-size images, mid-size and small images, and large and small image). Each model layer (C1, C2, C3) consists of an additional pair of simple and complex cell layers that perform tuning and pooling operations. See [157] for model and parameter details. Dashed horizontal line (at 16.67% accuracy) indicates chance performance.

that human brains are larger, which is believed to lead to longer neural latencies [172]. Unlike previous physiology studies of invariant object recognition, which are limited in the number of brain regions they can record from, we were able to see a clear latency difference between the initial identity signal and size and position invariant information.

The source localization results showed that neural activity moved to more ventral regions when invariant information developed at 150 ms (Figures 3-8). While one potential criticism is that there is a fair amount of variation in the sources across subjects, and source localization algorithms taking into consideration structural and functional MRI data may provide a finer picture of where in the brain invariance computations occur [67], these results do show a clear progression in each subject where activity appears to move down the ventral stream. These source localization results combined with timing data and our results showing that it was not possible to decode invariant information from a V1-like model (Figure 3-9), all suggest that early visual areas are driving the initial identity decoding, and later visual areas are computing the invariant representations.

The timing between neural events recorded through EEG/MEG and behavioral reaction times for visual tasks has not always been consistent in the literature. For example, humans can distinguish between scenes with or without animals with saccades that are as fast as 120 ms [95], yet the earliest differences between EEG event related potentials (ERPs) on this task were not observed until 150ms after stimulus onset [171]. Similarly, the N170 ERP (a negative potential observed in certain channels at 170 ms) response to faces [15] also occurs late relative to behavioral reaction times and latencies in the macaque. A reason for this discrepancy might be that ERP analysis is too coarse a method to capture the earliest components of object related information. By using decoding based methods, we are able to see discriminative visual signals at significantly earlier latencies in humans, also see [111, 21]. In the study by Carlson et al., the authors found similar latencies for position-invariant MEG decoding using a categorization task. They were able to categorize faces, cars, and face and car textures as early as 105-135 ms post-stimulus onset. Interestingly, in contrast with our results, Carlson et al. did not detect a difference in latency between their position-invariant and non-invariant decoding conditions. This discrepancy may due to the fact that the authors used a categorization task, which requires generalization (unlike our identification task) and may occur when the neural signals already show a certain degree of invariance. A recent study by the same group shows that more abstract categorization has a longer decoding latency [22], supporting this explanation. With our experimental paradigm, we were able to see a clear range of latencies from the initial non-invariant identity signal to size- and position-invariant neural signals, which help to frame previous human timing results.

Our timing results also showed that both size and position invariance developed in a sequential order, meaning that smaller transformations were decoded before larger transformations. This sequential development is consistent with a hierarchical, feed-forward visual model where receptive fields pool at each successive visual layer to first create local invariance and then build invariance over a larger area. We tested this theory with a biologically inspired object recognition system, which employs this feed-forward hierarchical architecture, known as HMAX [157] (Figure 3-9). HMAX performance had a similar trend to the order of the MEG experimental results: an early visual model could not decode stimuli invariant to size or position with above chance accuracy, a mid-level visual model could decode small transformations with above chance accuracy, and an IT-like model could decode all transformations with above chance accuracy. These results give new and compelling evidence that such a feed-forward hierarchy is a plausible model for invariant object recognition in the human ventral stream.

The order and timing information presented here have valuable applications not only for constraining models of the visual system, but also for answering more complex algorithmic questions about invariant object recognition, for example: do different types of invariance arise at different times in the ventral visual pathway? These results allow us to directly compare varying extents of these two transformations, position and size. The shorter latencies for size-invariant decoding, suggest that size-invariance may begin to develop before position-invariance. However, it was not the case that all size-invariance cases arose before position-invariant cases. The timing difference between the two types of invariance is being driven largely by the early rise of the smallest size-invariant shift (between 4 degree and 6 degree images). Additionally, it is difficult to directly compare the "extent" of two different types of transformations. For example, how does a 2-degree linear size increase compare to a 2-degree translation? Our results, however, do suggest that both size and position invariance develop in several areas along the ventral stream and appear significantly later than the initial identity signal.

The MEG decoding methods outlined in this study are a powerful tool to examine the dynamics of visual processing. Unlike conventional methods examining evoked responses, which require recordings from 50 or more stimulus repetitions to be averaged, decoding analysis is sensitive enough to detect visual signals by averaging only a few trials, or even from single trial data (Figure 3-3). The results and decoding methods presented here serve as a framework to examine an extended range of transformations, which should help lead to a real computational understanding of invariant object recognition.

3.5 Appendix - Decoding in source space

This section includes follow-up experiments performed to examine the information in our source estimates shown in Figure 3-8. This work appeared in the proceedings of the 2013 NIPS workshop on machine learning and interpretations for neuroimaging (MLINI).

Above we showed that size- and position-invariant visual signals can be decoded from MEG sensor-level data, and that the underlying neural sources localized to regions along the ventral stream. Further examining these source localization results would allow us to answer more precise anatomical questions about invariant object recognition, but these interpretations may be limited by the accuracy of the source localization. Here we compare MEG decoding analysis using features in sensor and source space in the same size- and position-invariant visual decoding task in order to both assess the promise of decoding in source space and attempt to gain a better spatiotemporal profile of invariant object recognition in humans.

We compare decoding in sensor and source space using data from the same experimental paradigm described above in Section 3.2. We assess if there is relevant stimulus information in the most active source estimates from across the brain, as well as investigate decoding in individual anatomically defined regions of interest (ROIs) to examine how invariant visual information evolves across the human ventral stream.

3.5.1 Methods

Two additional subjects participated in the MEG experiment described in figure 3-1, and we collected structural MRI for both subjects.

Source localization

Source localization was performed using the Minimum Norm Estimate (MNE) distributed source localization method, which finds the set of sources that minimizes the total power (L2 norm) of the sources [67]. Structural MRIs were collected for both subjects, and cortical reconstruction and volumetric segmentation was performed with the Freesurfer image analysis suite [27, 47, 46]. We estimated 15,000 sources constrained to each subject's cortical surface. Source localization was performed using fixed orientation constraints, with the default signal-to-noise ratio (proportional to inverse of the regularizer) of 3.

Sensor and source feature selection

In figure 3-3, we showed that using between 10 and 50 sensors at each time point from the above ANOVA feature selection procedure provides optimal signal to noise ratio for decoding with sensor data. Here we are selecting 30 (approximately 10% of the total) features at each time point. To compare decoding in source space with the sensor-level data, we also perform feature selection to downsample the large number of sources. We can assess the information in the active sources, by choosing the top 1500 (10% of the total) most active sources at each time point, calculated for the average of all image presentations, and performing source level decoding with only sources in these locations. While the source locations were chosen based on aggregate data, decoding was performed as described in the above section with individual trials in each cross-validation split.

Cortical Parcellation

Cortical parcellation of each subject's MRI was performed automatically using Freesurfer. We examined how the visual signals evolve in different brain areas, by decoding in four visual regions of interest: V1 and V2 (defined by the Brodman Area atlas in Freesurfer), and the occipital inferior and temporal inferior regions (both gyri and sulci defined by the Destrieux Atlas [48] in FreeSurfer). The four regions are illustrated on the cortex of Subject 1 in Figure 3-12(a).

3.5.2 Results

3.5.3 Source estimates at key decoding times

As in Figure 3-8, but now using the subjects' own anatomy rather than a common reference brain, we show the sources at two key decoding times, 70 ms: the time when stimulus information can first be decoded, and 150 ms: the time when size- and position-invariant signals can be decoded, shown in Figure 3-10. At both time points, sources in the occipital lobe are highly active, consistent with the visual task the subjects performed. Additionally, at 150 ms, there are more active sources that have spread further down the temporal lobe, near later visual areas.

Decoding with top sensors and sources

The decoding results using the 30 (approximately 10%) most selective sensors and 1500 (10%) most active sources as classifier features are shown in Figure 3-11. We see that source space features have similar performance to the sensor-level features

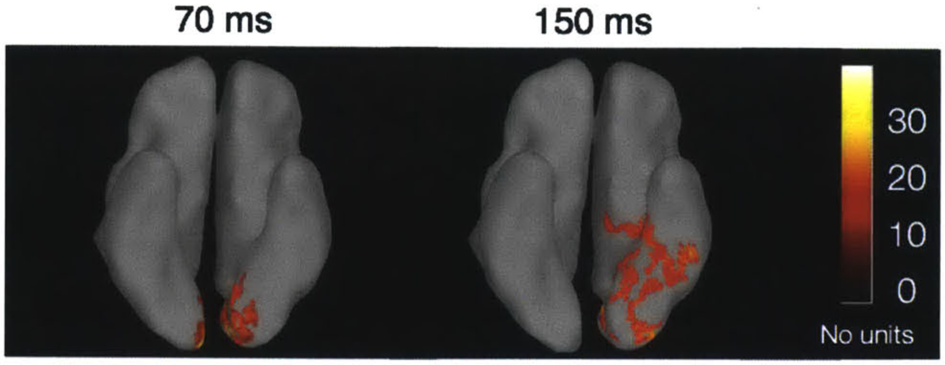


Figure 3-10: A left view of source estimates on cortex of subject 1 at 70ms after stimulus onset, the time when image identity can first be decoded, and 150 ms after stimulus onset, the time when size- and position-invariant information can be decoded. Color bar right indicates absolute source magnitude in picoAmpere-meters.

for the non-invariant, size- and position-invariant conditions even though, unlike the top sensors, they were not explicitly chosen to contain stimulus identity information. These results indicate that the most active sources contain important invariant visual information.

Decoding within anatomically defined ROIs

The decoding results in four visual ROIs: V1, V2, occipital inferior and temporal inferior regions are shown in Figure 3-12(a). Decoding results for the non-invariant, size-invariant, and position-invariant conditions are shown for these four brain regions in Figure 3-12.

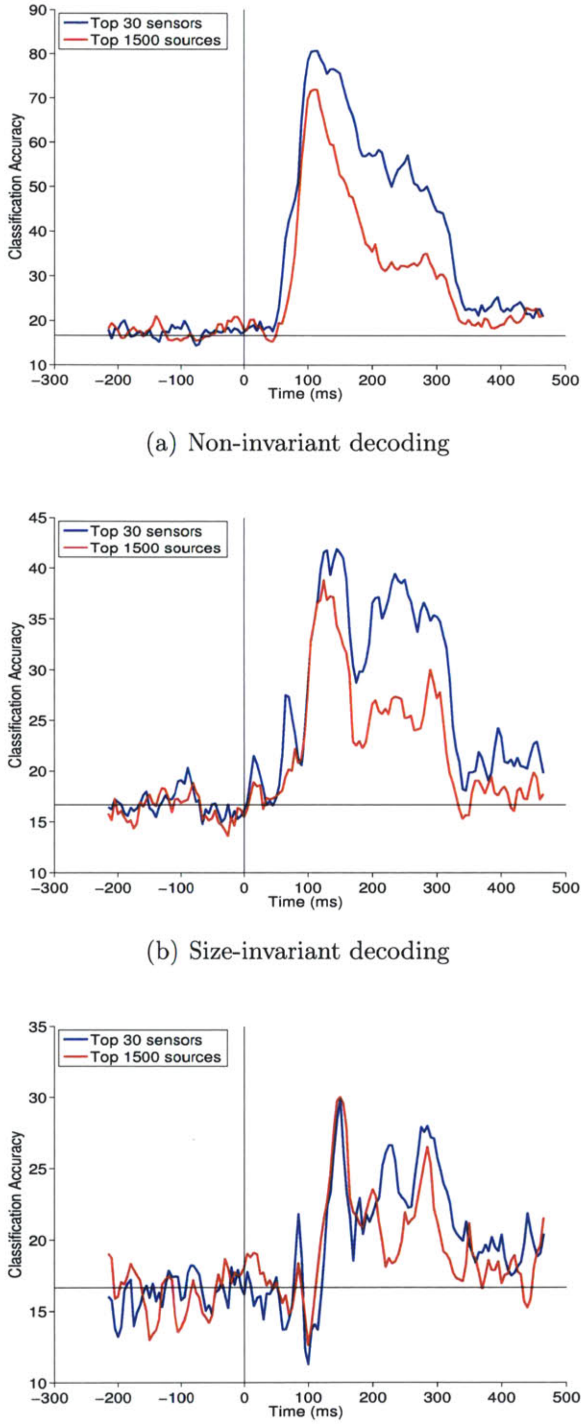
For the non-invariant decoding conditions, information appears to progress in a feedforward manner throughout the different visual regions: V1 and V2 have the highest accuracy and an earlier onset latency, while the two later visual areas have longer latencies and slightly lower decoding performance, Figure 3-12(b). For the position invariant decoding, the latest visual region, temporal inferior, has slightly higher accuracy, but all four regions appear to have similar latency for both the size- and position-invariant condition, Figures 3-12(c) and 3-12(d). Additionally, its important to note that regions sampled outside of the temporal and occipital lobes did not contain any relevant visual information (results not shown), supporting the accuracy of the source localization at a coarse level.

3.5.4 Discussion

In this work we performed source localization using MNE for two subjects performing a visual object recognition task. We showed that the most active sources moved further down the temporal lobe as invariant information developed, and that these sources contained invariant visual information. Both of these results support the accuracy of the source estimates. Finally, we decoded within different visual regions of interest to gain a spatiotemporal profile of how invariant visual signals evolve in the ventral stream.

The non-invariant decoding results in Figure 3-12(b) show a distinction between different visual areas, and a progression of visual stimulus information in a feedforward manner along the ventral stream. The size- and position-invariant results in Figures 3-12(c) and 3-12(d), however, show that invariant information develops in all visual areas at the same time and with similar accuracy. Additionally, although the source estimates have become more active further down the temporal lobe at 150 ms, there are still active sources in the early, occipital regions. This may represent an actual spread of invariant visual information across these regions, or, is more likely due to the fact that the source estimates are not resolved at a fine enough spatial scale to distinguish between adjacent cortical regions. So although the non-invariant results show a feedforward progression of visual information, the spatial resolution does not seem high enough to draw definitive conclusions about feedforward versus feedback processing in our size- and position-invariant visual tasks. It is possible that a more sparse source localization method utilizing the L1 norm [176], or a combination of the L1 and L2 norms [128], as well as methods incorporating spatial and temporal smoothness constraints [100], may provide more precise anatomical estimates and be better suited to answer these questions.

In this work we were able to evaluate the MNE source estimates on different levels: we affirmed that the most active sources were along the occipital and temporal lobes, that they had relevant invariant visual information, and that we could see some distinctions between visual information in nearby brain regions. These results show a coarse picture of how invariant visual information travels through the ventral stream, and provide a framework for future studies to answer visual processing questions at a finer anatomical level with more precise source estimates.

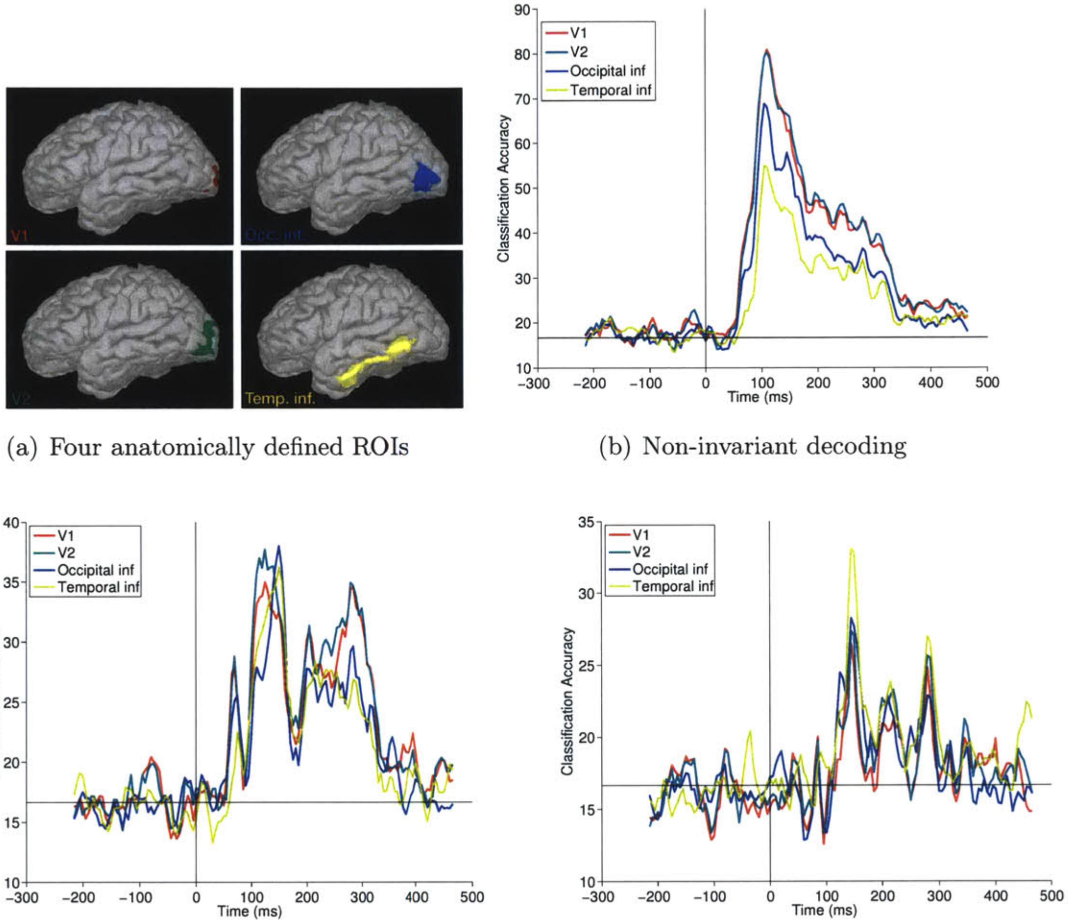


(a) Non-invariant decoding

(b) Size-invariant decoding

(c) Position-invariant decoding

Figure 3-11: Comparison of decoding with most selective sensors (blue) and most active sources (red) for (a) non-invariant, (b) size-invariant and (c) position-invariant conditions. (Please note the different scales of the y-axes in a-c.)



(a) Four anatomically defined ROIs

(b) Non-invariant decoding

(c) Size-invariant decoding

(d) Position-invariant decoding

Figure 3-12: Decoding within different ROIs for (a) non-invariant, (b) size-invariant, and (c) position-invariant conditions. The regions of interest are highlighted with their corresponding from (a). (Please note the different scales on the y-axes in b-d.)

Chapter 4

Invariant representations for action in the human brain

A version of this chapter is in preparation to be submitted for publication. Andrea Tacchetti and I contributed equally to this work.

The human brain rapidly parses a constant stream of visual input. Most visual neuroscience studies, however, focus on responses to static images. We use magnetoencephalography (MEG) decoding and a computational model to study invariant action recognition in videos. We created a well-controlled, naturalistic dataset to study action recognition across different views and actors. Actions, like objects, can be decoded from MEG data in under 200 ms, and this early representation is invariant to changes in both actor and viewpoint. We developed a biologically inspired computer vision model, extending hierarchical models of object recognition. This model can achieve viewpoint invariance by pooling across views, through the same mechanism as it achieves invariance to affine transformations by pooling across position and scale. These results provide a temporal map of the first few hundred milliseconds of human action recognition, and a mechanistic explanation of the computations underlying invariant action recognition.

4.1 Introduction

As a social species, humans rely on the ability to recognize the actions of others as a crucial part of their everyday lives. We can quickly and effortlessly extract action information from rich dynamic stimuli, despite variation in the appearance of these actions due to transformations such as changes in position, size, viewpoint and actor. The computations underlying this process, however, are still poorly understood, as evidence by the fact that humans still drastically outperform state of the art computer vision algorithms on action recognition tasks [101, 93].

Several studies have examined which regions in the brain are involved in processing actions and biological motion. In humans and nonhuman primates, the extrastriate body area (EBA) has been implicated in recognizing human form and action [35, 120, 110], and the superior temporal sulcus (STS) has been implicated in recognizing action and biological motion [141, 133, 127, 179]. In humans, the posterior portion of the STS (pSTS) in particular has been found to be involved in recognizing biological motion [62, 63, 177, 14, 130]. fMRI BOLD responses in this region are selective for action from biological motion [178] and can also recognize action in a mirror-symmetric manner [64]. Recent studies have also shown that neurons in macaque STS recognize actor invariant to action and action invariant to actor [159].

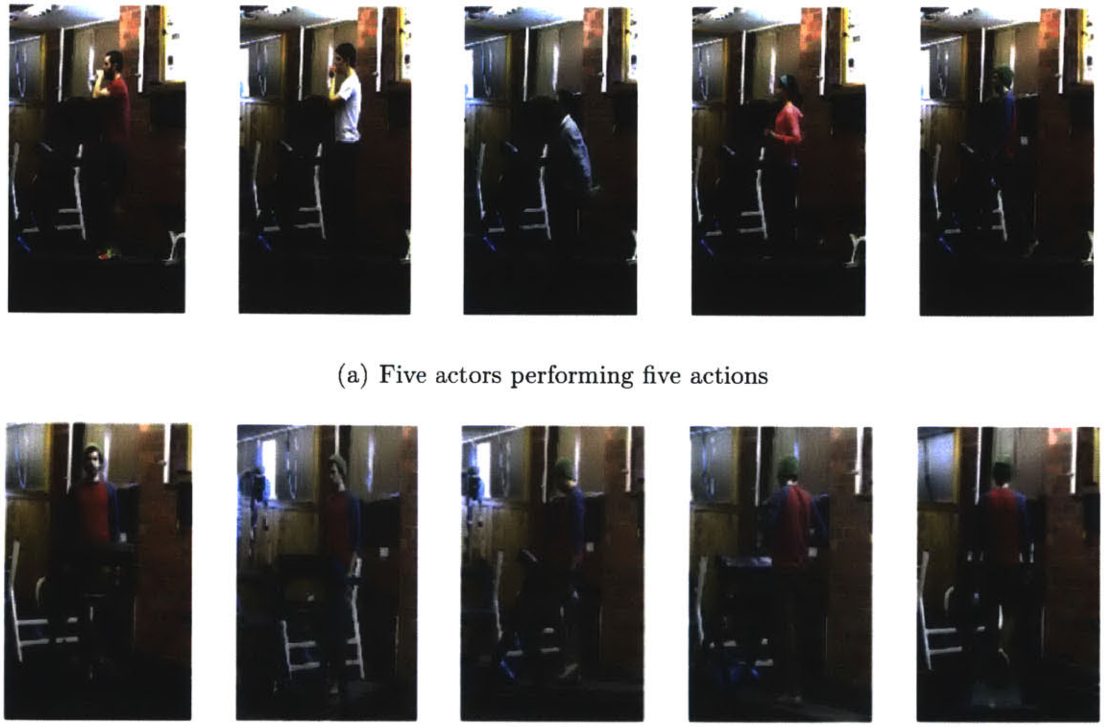
These studies all focus on the neural response to simple artificial stimuli, and do not provide information about the underlying computations across the brain or their timing. We hope to better understand invariant action recognition by first, looking at responses to natural movies, rather than simple artificial stimuli, second, understanding the dynamics of neural processing to help elucidate the underlying neural computations, and finally, implementing these insights into a biologically-inspired computational model. Here we use magnetoencephalography (MEG) decoding analysis and a computational model of the visual cortex, to understand when and how different computations are carried out to perform actor and view invariant action recognition in the visual system.

We filmed a realistic yet controlled dataset to study action recognition invariant to actor and viewpoint, and examine the effects of form and motion on invariant action recognition. We showed with MEG decoding that actions are recognized very quickly (in under 200 ms after the video onset) and this early representation is invariant to non-affine transformations (view and actor). These MEG data localizes to neural sources in ventral and dorsal stream, including the pSTS, consistent with previous physiology and fMRI studies. We next used these insights to extend a computational and theoretical framework for invariant object recognition to recognize actions from videos in a manner that is also invariant to actor and viewpoint on the same dataset. We showed that, despite being invariant to actor, there are still neural representations for actor identity in MEG signals, and the same class of computational models can perform both actor-invariant action recognition and action-invariant actor recognition. We also showed using behavioral data, MEG, and the model that both form and motion are important for action recognition.

4.2 Results

4.2.1 Novel invariant action recognition dataset

To study the effect of changes in view and actor on action recognition, we filmed a dataset of five actors performing five different actions (drink, eat, jump, run and walk) on a treadmill from five different views (0, 45, 90, 135, and 180 degrees from the front of the actor/treadmill) [Figure 4-1]. The dataset was filmed on a fixed, constant background. To avoid low-level object/action confounds the actors held the same objects (an apple and a water bottle) in each video, regardless of the action they performed. This ensures that the main variations between videos are the action, actor, and view, and allows controlled testing of different hypotheses of invariant recognition. The videos were cut into two-second clips that each included at least one cycle of each action, and started at random points in the cycle (for example, a jump may start mid air or on the ground). The dataset includes 26 two-second clips for each actor, action, and view, for a total of 3250 video clips. The dataset allows



(a) Five actors performing five actions

(b) At five different views

Figure 4-1: Dataset consisted of five actors, performing five actions (drink, eat, jump, run and walk), on a fixed position in the visual field (with a treadmill) and fixed background across five different views (0, 45, 90, 135, and 180 degrees). To avoid low-level confounds, the actors held the same objects in each hand (a water bottle and an apple), regardless of action.

testing of actor/view invariant action recognition, with few low-level confounds. A motion energy model (C1 layer of the model described below) cannot distinguish action invariant to view (Figure 4-14).

4.2.2 Readout of actions from MEG data is early and invariant

Five subjects viewed the above dataset while their neural activity was recorded in a MEG scanner. We use decoding analysis, which applies a linear machine learning classifier to discriminate stimuli based on the neural response they elicit, to analyze the MEG signals. By repeating the decoding procedure at each 5ms time window, we can also see when different types of stimulus information are present in the brain.

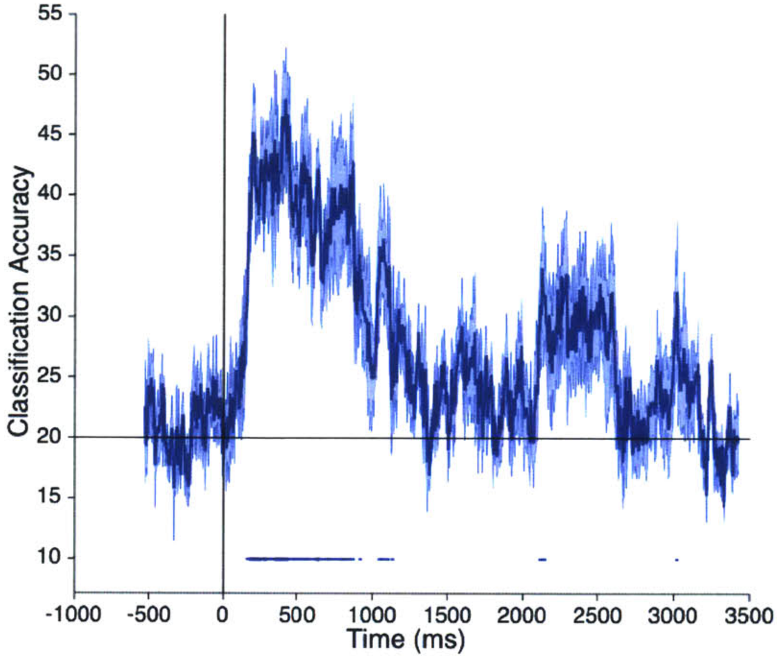


Figure 4-2: Action can be decoded from subjects' MEG data as early as 200 ms after stimulus onset (time 0). Results are from the average of five different subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with $p<0.01$ permutation test, with the thickness of the line indicating if the significance holds for 3, 4 or all 5 subjects.

Action can be read out from the subjects' MEG data as early as 200ms after the video starts (after only about 6 frames of each two-second video) [Figure 4-2].

We can test if these MEG signals are invariant to actor by training the machine learning classifier on data from subjects viewing videos of four actors and testing the classifier on the fifth held out actor. Similarly, we can verify that the MEG signals are invariant to view by training the classifier on data from subjects viewing actions performed at four views and testing the classifier on the fifth held out view. View and actor invariant MEG signals have a similar accuracy and latency to the case without any variation [Figure 4-3]. These MEG results suggest that the brain quickly computes a representation for action that is invariant to both the actor performing the action and the viewpoint at which the action is recorded.

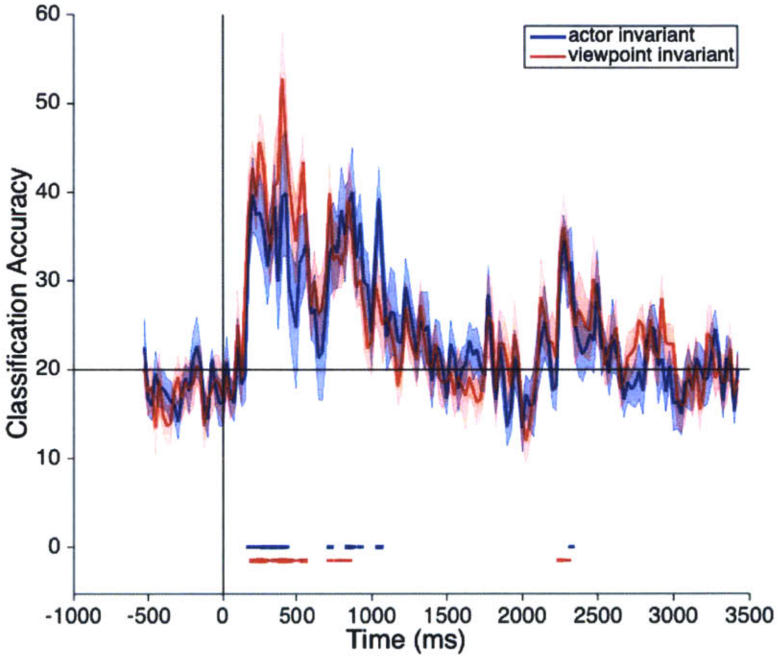


Figure 4-3: Action can be decoded invariant to actor (train classifier on four actors, test on fifth held-out actor), or view (train classifier on four views, test on fifth held-out view). Results are from the average of five different subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with $p<0.01$ permutation test, with the thickness of the line indicating if the significance holds for 3, 4 or all 5 subjects.

4.2.3 Extreme view invariance

It is possible that the actor and view invariant action decoding look so similar to the case without variation, because the classifier has a large range of viewpoint variation in the training data. In other words, perhaps generalizing to a fifth view is not very challenging for a classifier that was trained on four other views. To examine if the neural signals can perform a more extreme viewpoint generalization task, we recorded MEG data from five additional subjects viewing videos at two views (0 degrees and 90 degrees). We then decoded by training only on one view (0 degrees or 90 degrees), and testing on a second view (0 degrees or 90 degrees). Even with a more limited training set, MEG signals can generalize across view. There is no difference in the latency between the across views case (train on 0 and test 90, or train on 90 and test on 0) and the within view case (train and test at 0, or train and test at 90) [Figure 4-4], suggesting that the early action recognition signals are indeed view invariant.

4.2.4 Recognizing actions with a biologically-inspired hierarchical model

We extended a hierarchical feedforward model of visual cortex for object recognition from static images [52, 144, 157, 147, 123] to recognize actions from videos. This system is organized hierarchically: the sensory input goes through a layer of computation, and the output of this layer serves as input for the following layer. Within each layer many functional units (cells) perform the same computation on different portions of the input (e.g. match a template to a specific region of the visual field, a process analogous to a cell firing when a prefered stimulus is in its receptive field). This hierarchical model is inspired by Hubel and Wiesel's findings in primary visual cortex, and is constructed by alternating layers of simple and complex cells [77]. In simple cell layers, each unit computes a measure of similarity between a portion of the input (e.g. an image patch, or a clip of a video) and a pre-stored template (e.g. an oriented Gabor patch) to build selectivity. In complex cell layers, units pool over the output of simple cell units that store the same transformed template to build

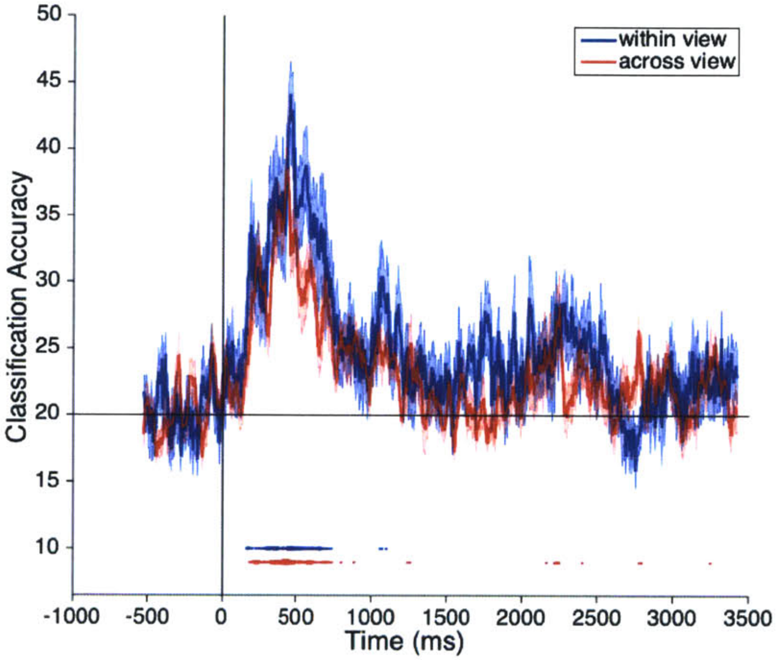


Figure 4-4: Action can be decoded with "extreme" view invariance: train and test on same view ('within-view' condition), or train on one view (0 degrees or 90 degrees) and test on second view ('across view' condition). Results are from the average of five different subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with p<0.01 permutation test, with the thickness of the line indicating if the significance holds for 3, 4 or all 5 subjects.

invariance.

The model we describe here extends this architecture to video stimuli, by adding a temporal component to simple cell templates and complex cell pooling regions. Furthermore, by pooling over cells that contain templates that are rotated in depth (in addition to the traditional pooling over position and scale), our model computes a response that is invariant to this transformation.

Our model consists of two simple-complex layers pairs. At the first simple layer (S1), tuning functions are moving Gabor-like stimuli that model the receptive fields found in primate V1 and MT [2, 158, 122, 124]. The first complex layer (C1) applies local max pooling to the S1 output and its output serves as input to a second simple layer (S2). Templates in S2 layers are sampled randomly from the C1 responses of training videos 4-5.

To build invariance to viewpoint, the model's C2 units compute the max of the response elicited in all cells whose tuning function come from videos containing the same actor performing the same action across different views [Figure 4-6 (a)]. Many theories and experimental evidence have suggested how this wiring across views is learned in development (cite Foldiak, Wiskott/Sejnowski, Wallis and Bulthoff). We compare this experimental model to an unstructured control model, which contains the same templates, but where action is not taken into account in the pooling scheme and instead each C2 cell pools over a random, unstructured set of S2 cell templates [Figure 4-6 (b)].

Both the experimental and control models can recognize action within one view (82+/-7% and 79+/-5% accuracy, respectively). The model with structured pooling provides significantly better accuracy on the view-invariant action recognition task (49 +/-5% vs. 36+/-5% accuracy) [Figure 4-7], suggesting this structured pooling across different views of each action is important for achieving view invariance. In addition, the model is always tested on videos from a held-out actor, so, like the MEG data, the model can also recognize actions invariant to actor.

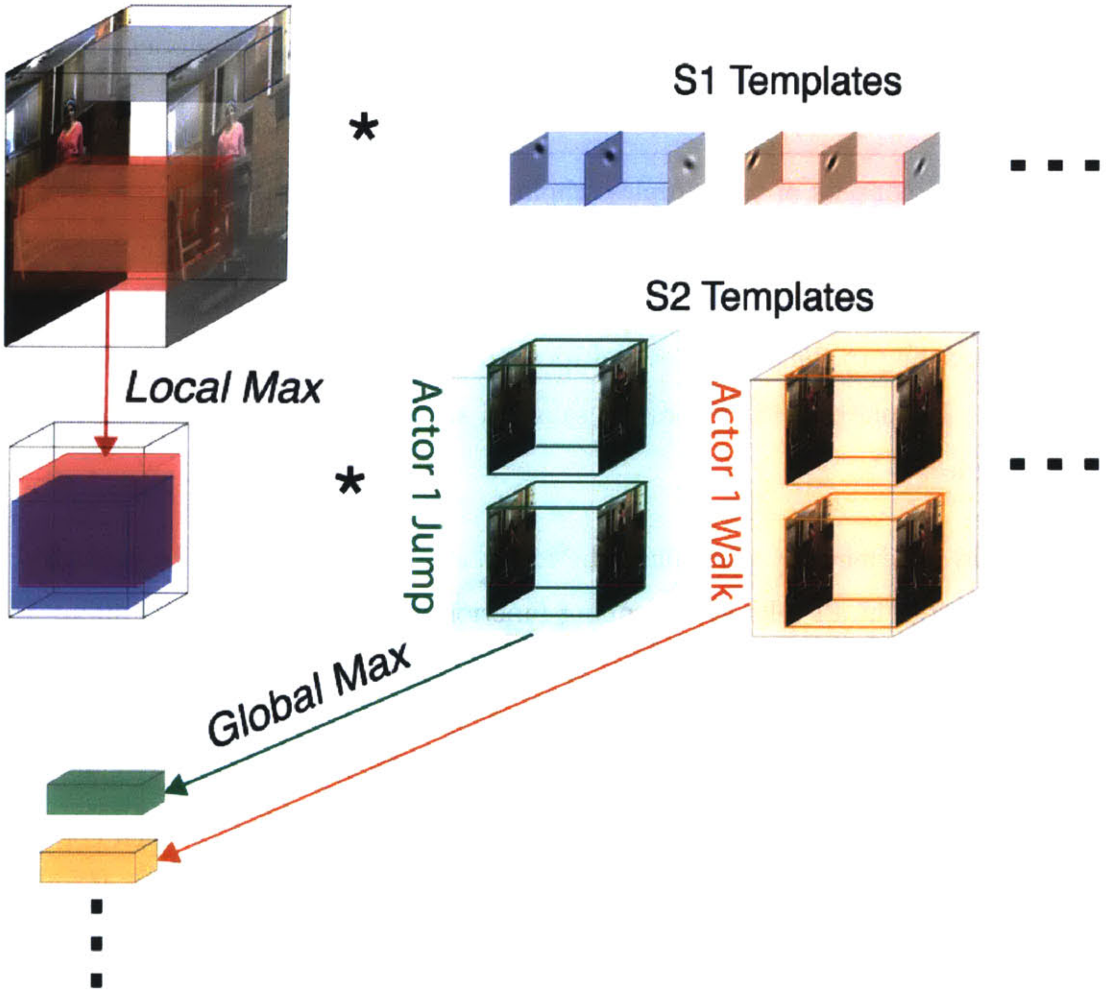
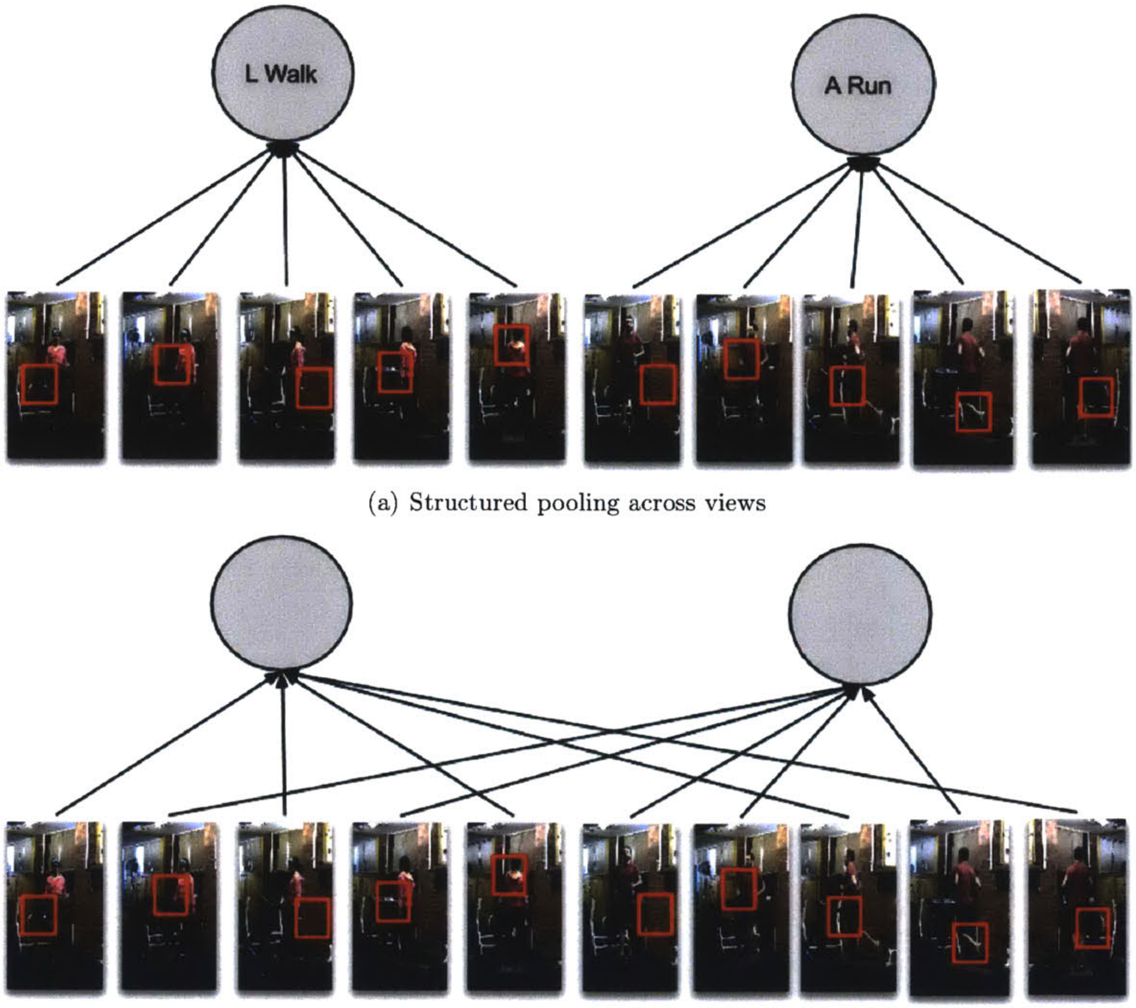


Figure 4-5: An input video is convolved with the S1 Gabor templates. At the C1 layer, a local max pooling is applied across position. At the S2 layer, the inputs are convolved with filters sampled from training videos that are passed through the S1-C1 model layers. At the final layer a global max across positions and views is computed.



(a) Structured pooling across views

(b) Random pooling across views

Figure 4-6: To build invariance to viewpoint, the model's C2 units pool over S2 units whose templates come from videos containing the same actor performing the same action across different views. We keep track of the which video each template comes from so that we can enforce structure in the wiring between simple and complex cells at the S2-C2 pooling stage. We compare this experimental model (a) to an unstructured control model (b), which contains the same templates, but where each C2 cell pools over a random, unstructured set of S2 cell templates.

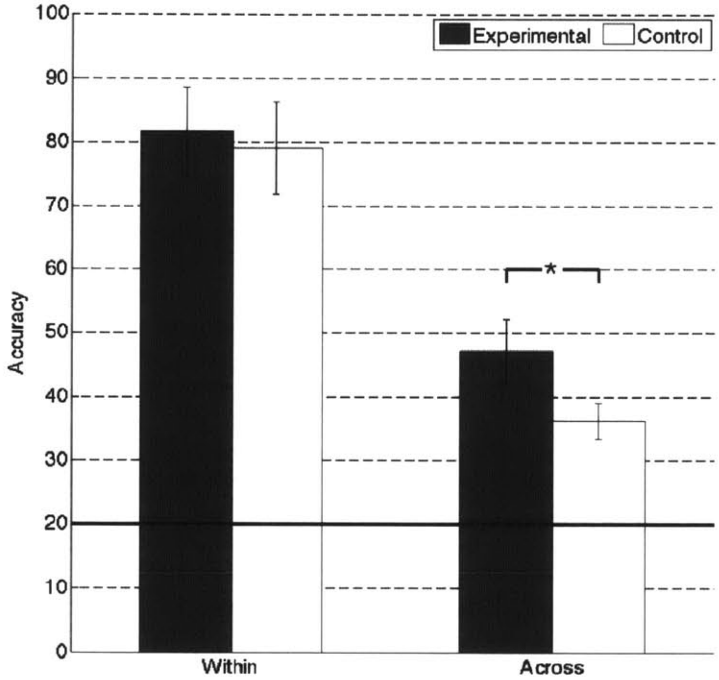


Figure 4-7: The model can recognize action when trained and tested on the same view ('within-view' condition), or trained on one view (0 degrees or 90 degrees) and tested on second view ('across view' condition). The Experimental model employs structured pooling as described in Figure 3B, top, and the Control model employs random C2 pooling as described in Figure 3B, bottom. Error bars indicated standard deviation across model runs. Horizontal line indicates chance performance (20%). Asterisk indicates a statistically significant difference with $p<0.01$.

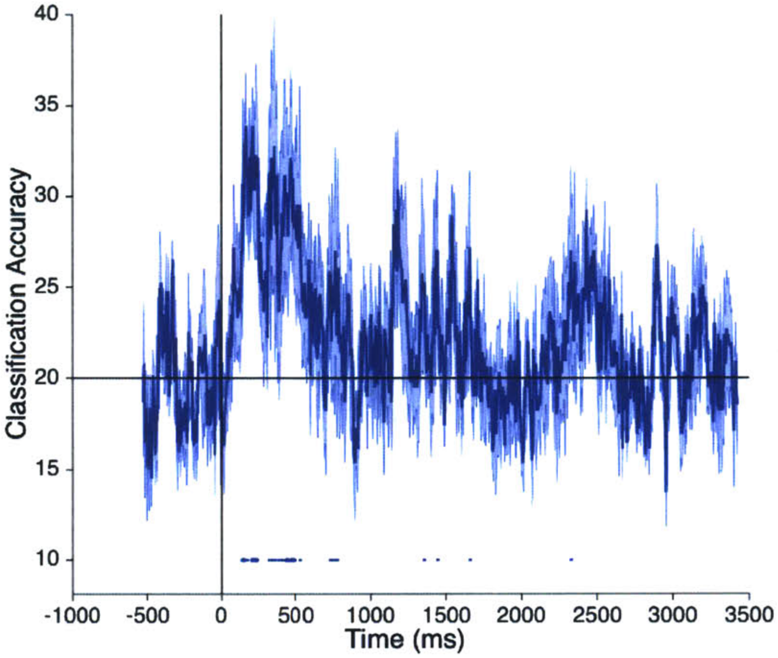


Figure 4-8: We can decode actor from the subjects performing the above action recognition experiment, even though they are explicitly doing a task to discount actor. Results are from the average of five subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with p<0.01 permutation test, with the thickness of the line indicating if the significance holds for 3, 4 or all 5 subjects.

4.2.5 Recognizing actor invariant to action

Both MEG and model data can recognize action invariant to actor. This raises the question: can information about actor still be extracted from this data? With the same MEG data used for action recognition, we can decode actor at a similar latency to when we can decode action [Figure 4-8] . It is important to note that subjects were doing an action recognition task, and this attentional effect on action may explain the lower decoding accuracy for actor.

Similarly, the model can also recognize actors invariantly to action [Figure 4-9]. Much like in the viewpoint invariance case, a structured wiring pattern from simple to complex cells helps recognition. In this case, all responses elicited in cells whose tuning function came from the same action and viewpoint (regardless of the

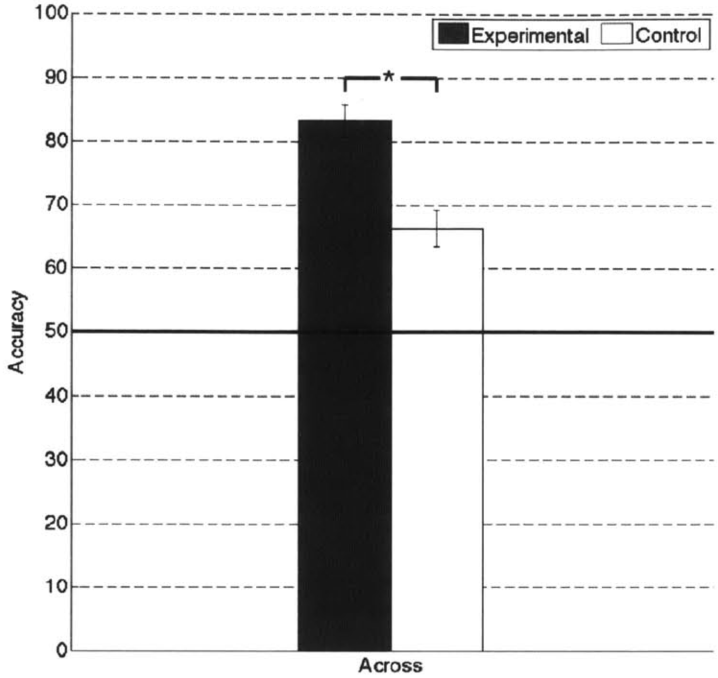


Figure 4-9: The same model class was tested to also recognize actions invariant to actor. In the Experimental model, the S2 cell responses whose tuning function came from the same action and viewpoint (regardless of the actor) were pooled by the same complex cell. This model is again compared to the above-described control model where the wiring between simple and complex cells is randomized in [Figure 3B, bottom]. Error bars indicated standard deviation across model runs. Horizontal line indicates chance performance (20%). Asterisk indicates a statistically significant difference with p<0.01.

actor) were routed to a single complex cell. This model is again compared to the above-described control model where the wiring between simple and complex cells is randomized in [Figure 4-6]. Although we can recognize action across actors, actor information is still represented in brain and model.

4.2.6 The roles of form and motion in invariant action recognition

To test the effect of form and motion on action recognition, we used two limited stimulus sets. The first 'Form' stimulus set consisted of one static frame from each video (no motion information). The second 'Motion' stimulus set, consisted of point light figures that are comprised of dots on each actor's head, arm joints, torso, and leg joints and move with the actor's joints (limited form information) [88].

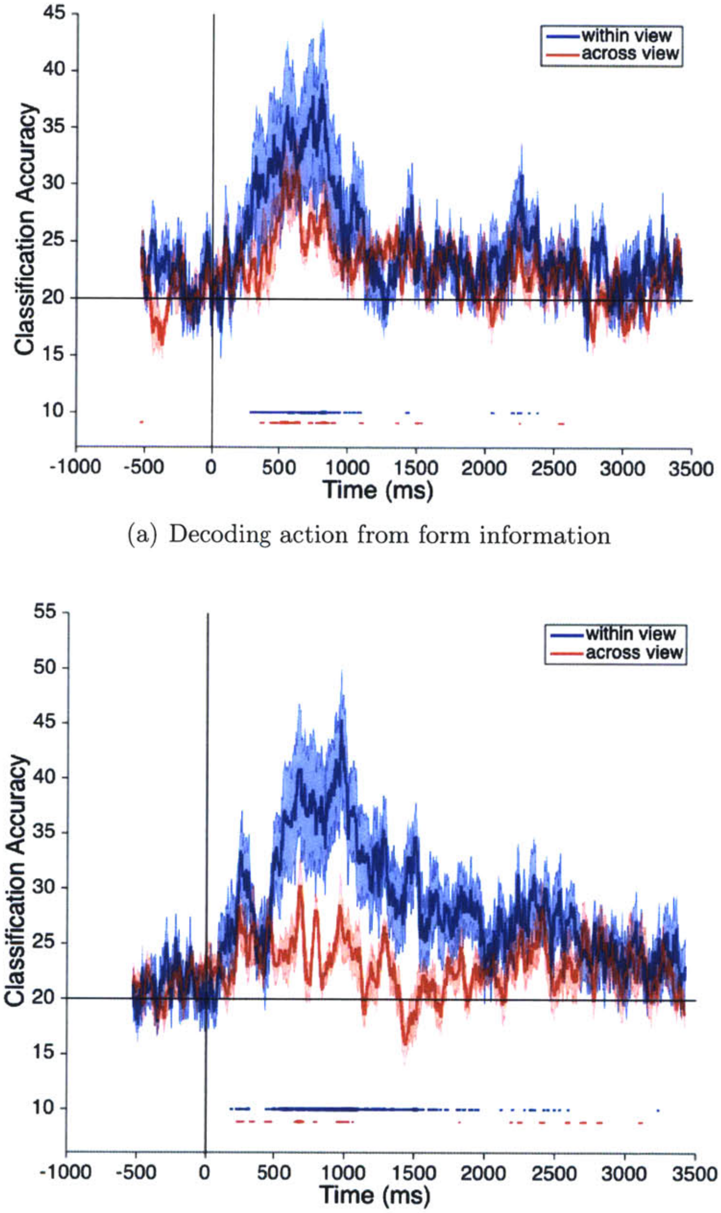
Five subjects viewed each of the form and motion datasets in the MEG. We could decode action within view in both datasets. Decoding performance across view, however, was significantly lower than the case of full movies [Figure 4.2.6]. In addition, subjects' behavioral performance dropped from 92% correct with full movies to 76% correct on the 'Form' dataset and 78% on the 'Motion' dataset, suggesting that the lack of motion information hinders recognition and this recognition deficit is reflected particularly in the MEG results.

We examined the effects of form and motion with our model by testing both stimulus sets on a model trained on full videos. While it is still possible to classify correctly which action was performed, performance was significantly lower than in the case where full videos were used 4-11.

4.2.7 Neural sources of action recognition

To understand where in the brain these invariant action signals originate, we performed source localization on three subjects who performed the extreme view invariance experiment. To reconstruct their neural activity, we computed the minimum norm estimate (MNE) [67] constrained by their structural MRI.

About 100 ms after video onset, the neural sources are localized to very early visual regions. Around 200-225 ms when decoding first goes significantly above chance, the neural sources have moved along the ventral and dorsal streams (Figure 4-12). Notably, we see activity in all three subjects in the posterior STS, an area implicated in fMRI and electrophysiology as being responsible for biological motion processing and invariant action recognition. By 500 ms the power in most sources has attenuated. There is little source activity outside of ventral and dorsal streams. The pattern of neural activity at key decoding times suggests that the actor and view invariant action recognition signals we are detecting is driven largely by the visual system, including both the ventral and dorsal streams and the pSTS.



(a) Decoding action from form information

(b) Decoding action from motion information

Figure 4-10: (a) Action can also be decoded invariantly to view from static images. (b) Action can be decoded from biological motion only (point light walker stimuli). Results are each from the average of five different subjects. Error bars represent standard error across subjects. Horizontal line indicates chance decoding (20%). Lines at bottom of plot indicate significance with $p<0.01$ permutation test, with the thickness of the line indicating if the significance holds for 3, 4 or all 5 subjects.

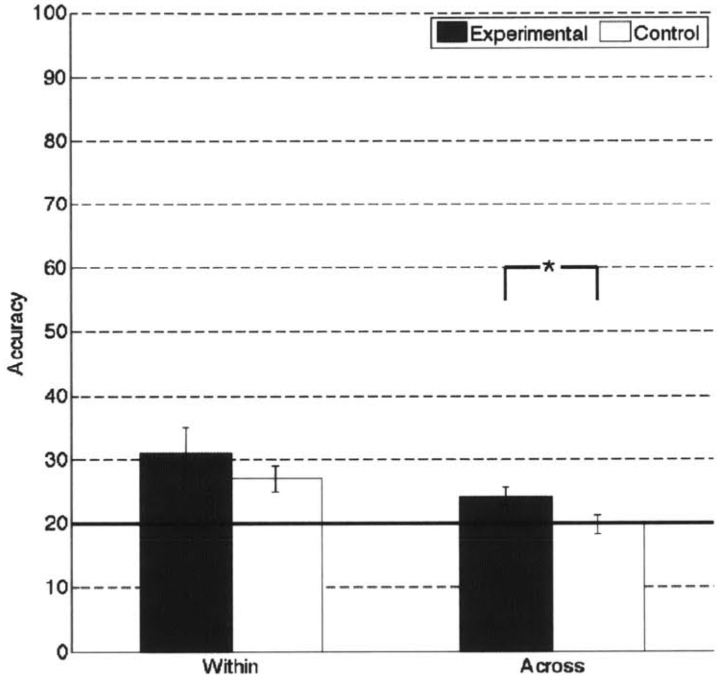


Figure 4-11: The model can recognize action from static frames, but the performance is much lower than with full videos. The Experimental model employs structured pooling as described in Figure 3B, top, and the Control model employs random C2 pooling as described in Figure 3B, bottom. Error bars indicated standard deviation across model runs. Horizontal line indicates chance performance (20%). Asterisk indicates a statistically significant difference with p<0.01.

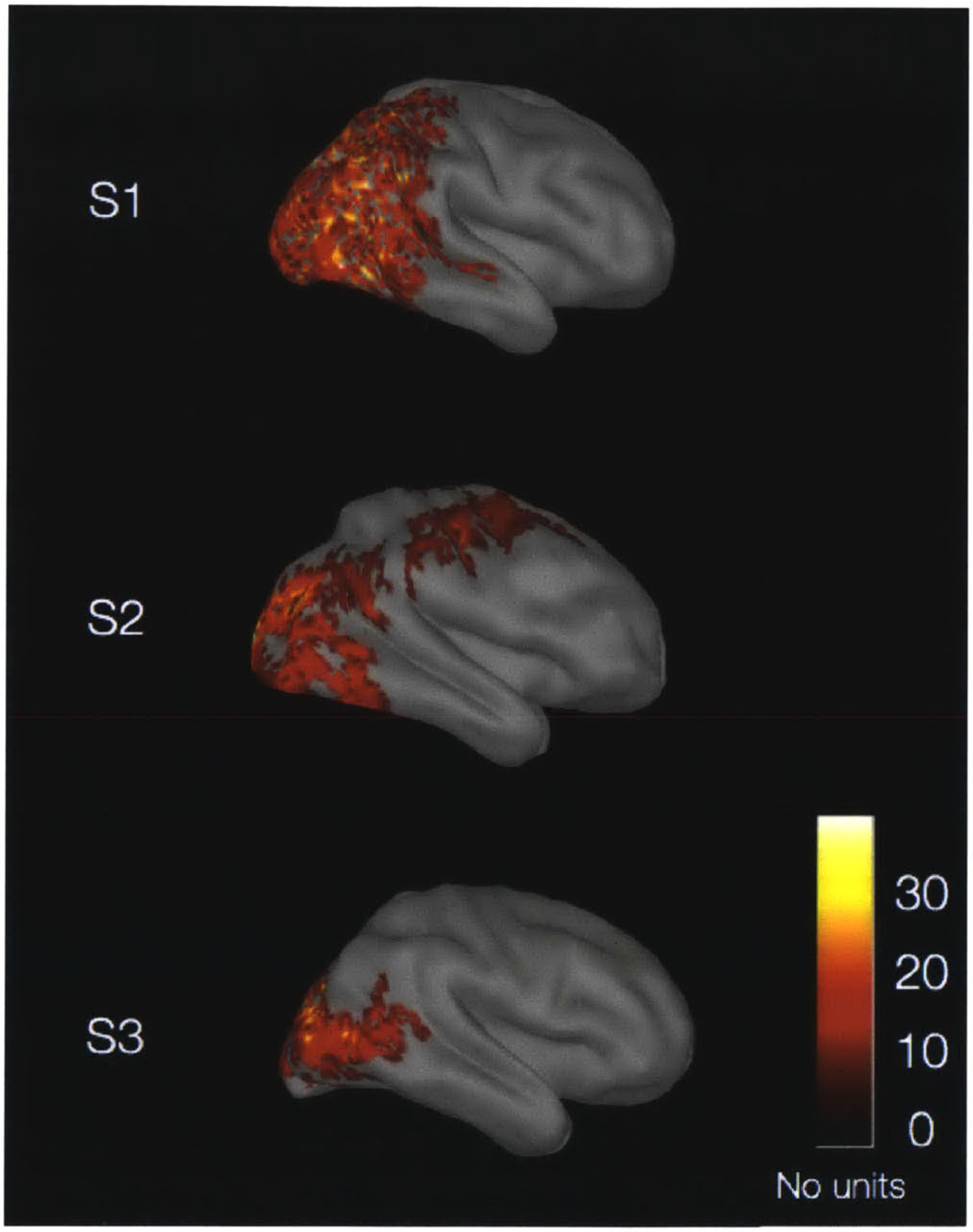


Figure 4-12: Neural sources for three subjects performing "extreme" view experiment at 200 ms after movie onset (initial decoding peak). Minimum Norm Estimates, with $p<0.0001$ significant sources (determined by t-test, corrected for multiple comparisons) versus baseline period. Sources are localized primarily to ventral and dorsal streams, including activity in pSTS.

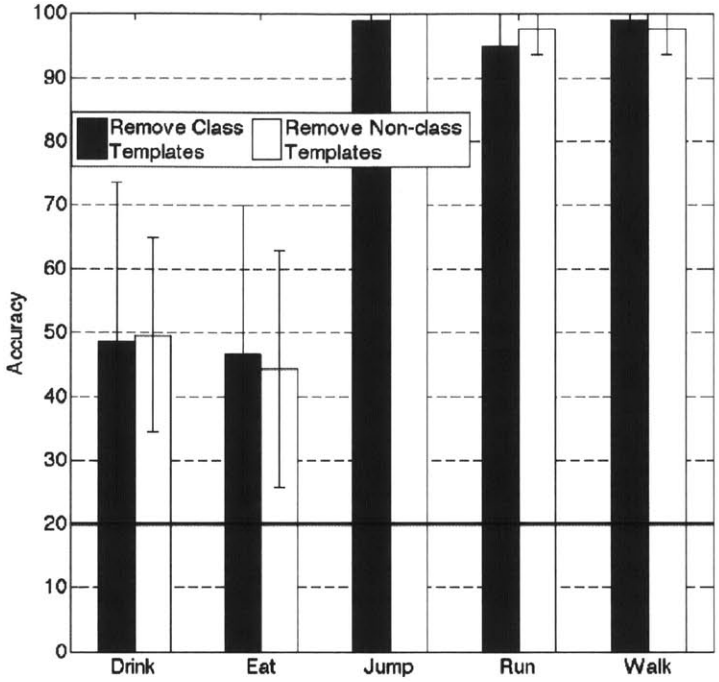


Figure 4-13: The model was trained using templates from four of the five actions to see if templates for a given action are required to recognize that action. The figure reports the classification accuracy (y-axis), with the standard deviation across model runs, for each action when templates for that action (light gray bars) or templates for one of the four other actions were removed (dark gray bars). For each action, performance between the case removing class or non-class templates is similar, suggesting that templates for a given action are not required to recognize that action, and the model can generalize to recognize actions for which it does not have stored templates..

4.2.8 Generalization to novel actions with a fixed model

To see if templates for a specific action are required to recognize that action, we performed a set of experiments where we fixed the model to contain templates sampled from videos of only four actions, and then tested the model's ability to recognize all five actions. The model's recognition of each action was similar, regardless of which templates the model had stored [Figure 4-13]. This suggests that it is plausible for such a model with stored templates to be able to generalize to new actions even after development (when templates are formed and stored) is over.

4.3 Discussion

4.3.1 Fast, invariant action recognition and implications for model architecture

We analyzed the dynamics of invariant action recognition in the human brain to find that action recognition occurs as early as 200 ms after a video begins. This early neural representation is invariant to changes in actor and position. These timing results provide compelling evidence that these computations are performed in feedforward manner, and interestingly that invariant representations for action are computed at same time as non-invariant representations. This seems to be in contrast to object recognition where invariance increases at subsequent layers in the ventral stream [10, 147, 152] and over time [83].

Action recognition occurs on a similar fast time scale, but slightly later than, object recognition in humans [111, 21, 22, 83, 24]. It is possible that since actions require information about both form and motion, higher level visual features are required for even basic action recognition than simple image discriminations (which are based on low-level features like lines and edges), and therefore the early action representation is already invariant.

We used these neural insights to develop a feedforward cortical model that performs action recognition invariantly to actor and view (non-affine transformations). The computations underlying the model's invariance to complex transformations are performed in the same model layer and using the same pooling mechanism as size and position (affine transformations). Our modeling results offer a computational explanation of the underlying neural mechanisms that lead to this fast and invariant response in visual cortex. In particular, our model showed that a simple-complex cell architecture [5, 6], extends to video stimuli and complex transformations. The model architecture is inspired by [56] and is a direct extension to non-affine transformations of the model proposed by Jhuang *et al.* [87]. The idea of building invariance to non-affine transformations by simply pooling over them was proposed for face 3D rotation in artificial stimuli in [105], the work presented here extends that framework for the first time to realistic videos.

The highest performing computer vision systems on action recognition tasks are deep convolutional neural networks, which have a similar architecture to our model, but more layers and parameters that are tuned for performance on a given classification task using backpropogation [93]. Our model, in contrast, is developed to have biologically faithful parameters and mimic human visual development. This modeling effort is primarily concerned with describing the neural data, and providing an interpretable architecture to explain viewpoint invariance, rather than optimizing for absolute performance gains.

4.3.2 Actor invariance and recognition

Singer and Sheinberg showed that the same population of neurons was both selective and invariant to actor and action (cite Singer/Sheinberg 2010). Previous object recognition studies have shown that transformation invariance does not eliminate information about the discounted transformation from even high-level neural representations [33]. Indeed here we show that despite the fact that subjects were performing an actor-invariant action recognition task, information about actor identity is still present in the neural signals at a similar latency to the action recognition signals. In addition, the same class of model can perform both an actor-invariant action recognition task and action-invariant actor recognition task, showing good fidelity with biological data. Despite the fact that actor can be discounted to recognize actions invariantly, this information is still represented in the brain and model.

4.3.3 Neural sources underlying invariant action recognition

Our source localization results show that neural activity during key decoding times is localized primarily in the ventral and dorsal streams. These results, however, suffer from the limited spatial resolution of MEG due to the ill-posed nature of the source localization problem [67]. Without any ground truth measure of the underlying neural activity for this task, these results provide only an estimate of the spatial activation in the brain. A network of brain regions involved in processing action, including both form [110, 120] and motion areas [62, 63, 177, 14, 130, 64, 178] has been mapped in humans using fMRI. This work provides rough agreement with those neural locations, and helps put into context the timing of different steps for form and motion in invariant action recognition.

4.3.4 Form and motion in invariant action recognition

As shown previously [88, 62, 177, 159, 141, 133, 127, 179], we found that biological motion and form are each enough alone to recognize actions, however decoding and model performance for the viewpoint invariant decoding drops to almost chance when either form or motion information is removed. This is also reflected in a slight drop in behavioral performance.

For the form and motion experiments, the model was trained on full videos and likewise, the S2 templates were sampled from the original naturalistic videos. This was done to mimic the visual experience of humans, who mostly experience actions from full video stimuli and due to the limited size of the reduced datasets. Given this training though, it is perhaps unsurprising that the model did not perform as well on the form or motion datasets.

While these limited data sets afford more experiment control, it is worth considering if they are the best way to understand the neural mechanisms underlying action recognition. Humans can indeed recognize action from diminished stimuli, but here we show it elicits different neural response than full video stimuli, particularly in the case of viewpoint invariant recognition. Moving toward more naturalistic stimuli, possibly in conjunction with controlled experiments with form or motion-only data, is thus important to understand the full range of neural responses to human action recognition.

4.3.5 Conclusion

This work highlights the advantages of using natural video stimuli, as well as the importance of using timing data to constrain the computational steps of invariant recognition and implementing these insights into an interpretable and biologically-plausible model. The results also fit into the broader context of models that have achieved wide success for cortical modeling and computer vision. Close interchange between artificial intelligence and neuroscience efforts may help move towards a deeper understanding of more realistic perception of humans actions.

4.4 Methods

4.4.1 Action recognition dataset

We filmed a dataset of five actors performing five actions (run, walk, jump, eat and drink) from five views (0, 45, 90, 135, and 180 degrees from the front) on a treadmill in front of a fixed background. By using a treadmill we avoided having actors move in and out of frame during the video. To avoid low-level object confounds, the actors held a water bottle and an apple in each hand, regardless of the action they performed. Each action was filmed for 52 seconds, and then cut into 26 two-second clips at 30 fps.

For single frame dataset, single frames that were as unambiguous as possible for action identity were hand selected (special attention was paid to actions eat and drink and occluded views). For the motion point light dataset, the videos were put on Amazon Mechanical Turk and workers were asked to label 15 points on each actors on every single frame: center of head, shoulders, elbows, hands, torso, hips, knees, and ankles. The spatial median of three independent labeling of each frame was used to increase the signal to noise ratio. The time series for each of the 15 points was independently low-passed to reduce the high frequency artifacts introduced by the single-frame labeling we used

4.4.2 Subjects

Twenty subjects age 18 or older with normal or corrected to normal vision took part in the experiment. The MIT Committee on the Use of Humans as Experimental approved the experimental protocol. Subjects provided informed written consent before the experiment. One subject (S5) was an author, and all others were unfamiliar with the experiment and its aims.

4.4.3 MEG experimental procedure

In the first experiment, five subjects were shown 125 two-second image clips (one for each of five actors, actions, and views), each presented 10 times. In the second experiment, five subjects were shown 50 two-second video clips (one for each of five actors, actions, and two views, 0 and 90 degrees), each presented 20 times. In the third experiment, five subjects were shown 50 static images, which were single frames from the videos in Experiment 2, for 2 seconds presented 20 times each. In the fourth experiment, five subjects were shown 10 two-second video clips, which consisted of point-light walkers traced along one actor's videos in experiment two, presented 100 times each.

In each experiment, subjects performed an action recognition task, where they were asked after a random subset of videos or images (for each of 125 videos in experiment one, and twice for each of the fifty each videos or images in experiments two through four) what action was portrayed in the previous image or video. The purpose of this behavioral task was to ensure subjects were attentive and assess behavioral performance on the various datasets. The button order for each action was randomized each trial to avoid systematic motor confounds in the decoding.

4.4.4 MEG data acquisition and preprocessing

The MEG data was collected using an Elekta Neuromag Triux scanner with 102 magnetometers at 204 planar gradiometers. The MEG data were sampled at 1,000 Hz. The signals were pre-processed using and preprocessed using Brainstorm software

[166]. First the signals were filtered using temporal Signal Space Separation (tSSS) with Elekta Neuromag software. Next, Signal Space Projection (SSP) was applied for movement and sensor contamination, and band-pass filtering from 0.1-100 Hz to remove external and irrelevant biological noise were applied using Brainstorm software [170].

4.4.5 Eyetracking

To verify that the subjects' eye movement could not account for the action discrimination, eye tracking was performed during MEG recordings for Experiment 1 (subjects S1-S5 viewing five actors performing five actions at five views) with the Eyelink 1000 eye tracker from SR Research. A nine-point calibration was used at the beginning of each experiment. We then performed decoding using the position data for the left and right eye, and found that decoding performance was not significantly above chance for more than two consecutive 5ms time bins, much below the significance threshold outlined for decoding (Figure 4-15).

4.4.6 MEG decoding analysis methods

MEG decoding analyses were performed with the Neural Decoding Toolbox [118], a Matlab package implementing neural population decoding methods. In this decoding procedure, a pattern classifier was trained to associate the patterns of MEG data with the identity of the action (or actor) in the presented image or video. The stimulus information in the MEG signal was evaluated by testing the accuracy of the classifier on a separate set of test data.

Decoding analysis was performed using a cross validation to assess the classifier accuracy where the classifier was trained on on 80% of data and tested on the held out 20%. To improve signal to noise, we averaged the different trials for each stimulus in a given cross validation split. Z-score normalized and performed sensor selection (used an ANOVA to choose sensors selective for stimulus identity with $p<0.05$ significance based on F-test) based on the data in the training set only. Decoding analyses were performed using a maximum correlation coefficient classifier, which computes the correlation between each test vector and a mean training vector that is created from taking the mean of the training data from a given class. Each test point is assigned the label of the class of the training data with which it is maximally correlated.

We averaged the data in each sensor into 50 ms overlapping bins with a 5 ms step size. We repeated the above decoding procedure in each time bin to assess the decoding accuracy versus time. Decoding accuracy is reported as the average percent correct of the test set data across all cross validation splits.

We assessed significance using a permutation test. To perform this test, we generated a null distribution by the decoding procedure for 100 time bins using data with randomly shuffled labels with 10 cross-validation split repetitions used on each run. Decoding results performing above all points in the null distribution for the corresponding time point were deemed significant with $P < 0.01$ (1/100). The first time decoding reached significantly above chance was defined as the point when accuracy was significant for five consecutive time bins. This significance criterion was selected such that no spurious correlations in the baseline period were deemed significant.

See Chapter 3 for more decoding methods details.

4.4.7 Source localization

We collected structural MRIs for three subjects (S6-S8) who performed the 'extreme invariance' (Experiment 2). We used Freesurfer software [27, 47, 46] to reconstruct each subject's MRI, and used this as input to the Minimum Norm Estimate (MNE) source localization algorithm. We used the Minimum Norm Estimate (MNE) source reconstruction in Brainstorm. An overlapping spheres head model was constructed using each subject's MRI, a full noise covariance matrix was computed for the data from 200 ms prior to stimulus onset, source reconstruction was constrained by a loose orientation constraint with an orientation parameter of 0.2. Sources that are significantly above baseline period ($p<0.0001$ with t-test Bonferroni corrected for multiple comparisons) are displayed.

4.4.8 Model

The model was written using the CNS: Cortical Network Simulator [123] and is composed of 4 layers. The input video is scaled down, preserving the aspect ratio, with the largest spatial dimension being 128px. A total of three scaled replicas of each video is run through the model in parallel; the scaling is by a factor of 1/2.

The first layer is composed of a regular grid of simple cells placed 1px apart (no sub-sampling), the tuning functions for these units are Gabor receptive fields that move in time while they change phase (as described in [2, 158]). Cells have spatially square receptive field of size 7, 9 and 11 px, extend for 3 4 and 5 frames and compute the dot product between the input and their template. The Gabor filters in each receptive field move exclusively in the direction orthogonal to the spatial modulation at 3 speeds, linearly distributed between 4/3 and 4 pixels per frame.

The second layer is a grid of complex cells that compute the maximum of their afferent simple cells. Cells are placed 2 units apart in both spatial dimensions (sub-sampling by a factor of 2) and every unit in the time dimension. Complex cells at the C1 level have spatial receptive fields of 4 simple cells and span 2 scales with one scale overlap, bringing the number of scaled versions from 3 to 2.

The third layer is composed of a grid of simple cells that compute the dot product between their input and a stored template. The templates at this level are sampled randomly from the training set (and never from the test set). We sample 512 different templates uniformly distributed across classes and across videos within each class. The cells span 9, 17 and 25 units in space and 3, 7 and 11 units in time.

The fourth layer, C2, is composed of complex units that compute the maximum of their inputs and cells pool across all positions and scales. The wiring between simple and complex cells at the C2 layer is described by a matrix with each column corresponding to a complex cell and having a list of indices for the simple cells; in the structured models these correspond to transformations, in control models, the rows of this matrix are scrambled. S2 template sizes are always pooled independently from one another. The output of the C2 layers is concatenated over time and cells and serves as input to a supervised machine learning classifier.

4.4.9 Video pre-processing and model classification

We used non-causal temporal median filtering background subtraction for all videos [136]. All classification experiments for the model were carried out using the Gaussian Kernel Regularized Least Squares classification pipeline available in [165]. Both the kernel bandwidth and the regularization parameter were chosen using leave-one-out cross validation.

4.4.10 Model experiments

For each of the experiments reported the computer vision model was an instance of the general architecture outlined above and the training and test set were a subset of the dataset described above. A few details were modified for each task in the S2 and C2 layers to make sure the model tested the hypothesis we set forward in that experiment and to avoid having S2 templates sampled from the test set. For the same reasons, we used different set of videos for each experiment. Here we describe these slight modifications.

For the action recognition task [Figure 4-7], templates were sampled from videos of four of the five actors performing all five actions and at all five views. In the experimental model, all S2 cells of the same size, with templates sampled from videos of the same actor-action pair (regardless of viewpoint) were wired to the same C2 cell yielding a C2 layer composed of 60 complex cells. In the control model we scrambled the association between templates and videos of origin (after sampling). The training set for this experiment was composed of 600 videos of four of the five actors performing all five actions at either the frontal or side viewpoints. The test set was composed of 150 videos of the fifth actor performing all five actions at either the frontal or side viewpoint. We only used either one of the viewpoints to train or test so as to verify the ability of the model to recognize actions within the same view and to generalize across views. This train/test split was repeated five times, using each actor for testing once and re-sampling the S2 templates each time.

For the actor recognition experiment [Figure 4-9], templates were sampled from videos of three of the five actors performing all five actions at all five views. In the experimental model, all S2 cells of the same size, with templates from the videos of the same actor-viewpoint pair (regardless of action), were wired to the same C2 cell yielding a C2 layer composed of 45 complex cells. The training set for this experiment was composed of 600 videos of the two held out actors performing four of the five actions at all viewpoints. The test set was composed of 150 videos of the two left out actors performing the fifth action at all five viewpoints. The experiment was repeated five times changing the two actors that were left out for identification and the action used for testing, the S2 templates were re-sampled each time.

The form only classification experiment [Figure 4-11] was conducted using the method described above for the action recognition experiment with the only difference that the test set was composed of videos that only featured one frame repeated for the entire duration of the clip. The motion only classification experiment was also conducted using the method described above for the action recognition experiment with the only differences being that only 100 form depleted videos of the held out actor were used for testing and that the experiment was not repeated using different actors for the test phase due to the prohibitive cost of acquiring human annotation for joint location in each frame (see 4.4.1).

4.5 Supplemental figures

4.5.1 C1 model performance on action recognition dataset

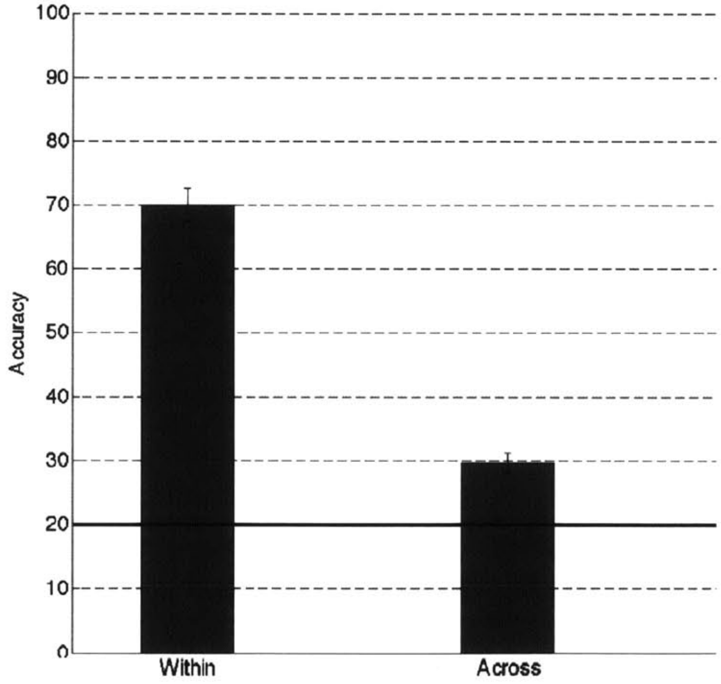


Figure 4-14: The output of the C1 layer of the model, similar to the motion energy model [2, 158], can classify action within view (no invariance) with relatively high accuracy, but cannot classify action invariant to viewpoint.

4.5.2 Effect of eye movement on action decoding

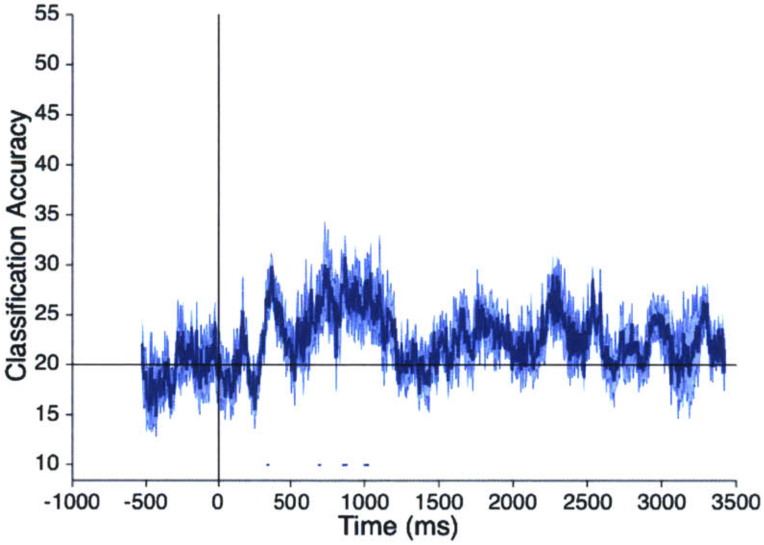


Figure 4-15: We train a classifier on the output of eyetracking data for five subjects as they view five actors perform five actions from five views. Results are from the average of five subjects. Lines at bottom of plot indicate significance with p<0.01 permutation test. Both decoding performance and behavioral performance are lower than for full videos. We cannot decoding significantly above chance for more than two consecutive 5ms time bins, suggesting that the subjects eye movements cannot account for the above decoding performance.

Chapter 5

Conclusions

5.1 Summary of findings

Humans have the remarkable ability to understand a visual scene in a fraction of a second and can learn new visual categories from few labeled examples. What are the computational steps underlying this ability and how can they be implemented in computer systems? This thesis examined these questions by analyzing how object and action recognition develop in the human brain, and tested the mechanisms underlying these findings with a computational model of visual cortex. In particular, this research led to the following findings.

5.1.1 Temporal association training is a robust method for learning invariance in development

In Chapter 2, we showed that the HMAX model can employ a temporal association training rule [49, 189] for learning position-invariance, rather than having this invariant architecture hardwired. The learned system works as well as a traditional hard-wired instantiation of HMAX on position-invariant recognition. We next found that we could replicate recent physiology experiments showing that a cell's preferred stimulus could be switched in a given position through "altered training", which disrupts position invariance [107]. Despite single unit disruption, however, the readout from all model units (analogous to an entire population of IT cells) did not confuse different stimuli, even with extensive amounts of altered training. These results suggest that, despite errors that may occur during training due to visual disruptions such as changes in lighting, occlusions, or noise in the biological mechanism, temporal association learning is a robust method to learn invariant recognition in development.

5.1.2 Size- and position-invariant object recognition have fast dynamics that match feedforward models of visual cortex

We next investigated the mechanisms behind size- and position-invariant object recognition in the brain. By developing and applying new methods for MEG decoding, we showed that object identity can be read out as early as 60 ms after stimulus onset. Size- and position invariant visual signals were present in the MEG data later, between 100-150 ms after stimulus onset. In addition we found that invariance to smaller transformations arose earlier than invariance to larger transformations (i.e. we could decode across to a two-degree position shift before a four-degree position shift). This is consistent with a feedforward hierarchical model, such as the HMAX model, where the early layers employ complex cell pooling over a set of simple cells to build invariance to a small, local region, and the final layer achieves global size and position invariance by pooling over all sizes and spatial locations. Finally, we performed source localization to see where the neural signals originated. The signals were localized primarily to visual areas and moved along the temporal lobe with time, consistent with past studies of size and position invariance in the ventral stream [126, 155, 172, 80, 152]. However, MEG's spatial resolution was not high enough to pick up differences in information between nearby visual regions. These results demonstrate how the temporal resolution of MEG can be used to test and confirm computational models of sensory processing in the brain.

5.1.3 Actor- and viewpoint-invariant action representations arise quickly and match a feedforward model of visual cortex

In order to study how people recognize the actions of others across changes in actor and viewpoint, we filmed a new, naturalistic, yet controlled dataset of different actors performing different actions at different views. We found that action can be read out of MEG data after only 200 ms (after only 6 frames of video), and this early representation is invariant to both actor and view. We next encoded these insights into a hierarchical model, based on an extension of the HMAX model in which filters and pooling regions have both spatial and temporal components to respond to dynamic stimuli. To this existing base model, we added complex cell pooling over different viewpoints of the same action. To the best of our knowledge, this the first time a video model has included pooling over such non-affine transformations. Based on the MEG timing results, we added these new pooling operations to the same model layer that pools over affine transformations, size and position.

Like the human MEG data, the model could also perform actor and view invariant action recognition, providing a computational explanation of the MEG results. Using both MEG and model data we saw that, in addition to decoding action invariant to actor, we could also decode actor identity invariant to action. Finally, we investigated the distinct roles of form and motion in action recognition. Humans can recognize actions in impoverished stimuli, in which either form or motion information is removed, though with a small but significant drop in performance. These results suggest that both form and motion information are important for invariant action recognition and demonstrate the importance of examining natural video stimuli to understand action recognition.

5.1.4 Common themes

This work focuses on invariant recognition in the brain. The timing results yielded by MEG decoding confirm previous findings that object recognition occurs very quickly

[138, 171], and yield new insights about the intermediate timing and computational steps leading to object and action recognition in humans. Specifically for objects, size and position invariance develops in stages (100-150ms) after initial low-level, non-invariant recognition (60 ms), while for actions, invariant and non-invariant representations occur at the same time (200ms). Interestingly, the same class of feedforward hierarchical models, can explain all of these timing results. Combined with the recent empirical successes of these models [99, 101, 93, 164, 71] , which now claim to exceed human performance on large-scale recognition tasks [71], and new analyses of their computational properties [5, 6], the results in this thesis suggest that this model class may be a viable solution to the problem of feedforward object recognition. A wide range of visual problems outside the domain of object recognition, however, remain to be solved.

5.2 Future directions

5.2.1 Moving towards more naturalistic vision experiments

Traditionally, both computer vision and visual neuroscience have focused on object classification in static images. While this paradigm has led to some notable successes, it deemphasizes both the wide array of tasks humans tackle with vision and the dynamic nature of human visual input.

Humans are able to quickly infer rich concepts, beyond basic categorization, from images and videos. Recently work has been done to address new vision tasks such as image captioning in deep neural network systems [181, 92]. There still exist a wide range of other visual tasks that have not been explored with these systems and likely require new neural insights, such as the incorporation of top down information. These visual tasks include fine-grain recognition [23, 40], tasks involving attention [30, 86, 91, 143], and of particular interest, visual social perception (e.g. understanding the intentions and emotions of people from visual input) [4], which is still largely understudied in both neuroscience and AI.

In addition to considering new visual tasks, moving toward studies using natural movie stimuli has many advantages. Natural videos contain rich visual information that closely mimic humans' daily visual experience. Recent work has shown that natural video stimuli lead to better data-driven parcellation of visual cortex across subjects [25, 81]. In addition, while a great deal of social and high-level visual information can be inferred from images, movies contain rich social narratives that provide information at many different levels, from action categories (e.g. walking versus running invariant to actor) to higher level social dimensions (e.g. is an interaction between two people friendly or hostile?). Understanding these aspects of visual perception is an important next frontier for visual neuroscience and AI.

5.2.2 Combining temporal and spatial information in the human brain

The research in this thesis demonstrates the importance of temporal information in constraining neural computations. Fully understanding the neural mechanisms and network structures, however, will require both high resolution spatial and temporal information. This work indeed raises several questions that should be addressed with better spatial resolution to more fully understand the underlying mechanisms. For example, our MEG data suggests that neural representations are highly dynamic, while fMRI studies show clear spatial localization for object and action representations. Are the neural dynamics due to activity in a single (or sparse group of) brain region(s), or is the changing neural representation due to information moving to different brain regions? High spatiotemporal resolution data is also important for disentangling the roles of feedforward versus feedback processing in the brain.

MEG source localization results produced by standard algorithms and constrained by subjects' anatomy were not enough to provide high resolution decoding in different visual regions (Chapter 3, appendix). New source localization methods have improved resolution, but are computationally expensive and difficult to implement [128, 100]. These methods are also notoriously difficult to evaluate as there is no ground truth.

This is usually overcome with experiments on phantom sources and other simulations, but there is no good way to evaluate them on real data.

Performing fMRI and/or physiology experiments using the same stimuli as MEG experiments and combining results across imaging techniques provides a promising way to probe the human brain with both high spatial and temporal resolution. For example, recent high temporal resolution ECoG studies have shown the role of visual feedback in recognizing occluded objects [169]. In addition, recent work [24] comparing MEG and fMRI using representational similarity analysis (RSA) [98] revealed dynamics in object recognition were due both to movement along the ventral stream and feedback to V1. Other recent work used the combination of fMRI, DTI and MEG to reveal a new region and functional mechanism for attentional control [12]. Applying the combination of fMRI and MEG to problems beyond object recognition (see section 5.2.1) has great promise for furthering our understanding of the human brain.

5.3 Conclusion

The work in this thesis focused on the development of invariant object and action recognition in the human brain. Using MEG decoding we could examine human visual processing with high temporal resolution, making it possible to see how invariant information evolves over time and break down the computational steps required to go from low-level features to invariant representations for objects and actions. By capturing the new insights from MEG decoding in computational models, we could more concretely test theories inspired by this timing data.

We found that object recognition occurs very quickly in the human brain and that size- and position-invariance develop in stages. In addition, that action recognition also occurs on a similarly fast time scale and the very first representations that support action decoding area already invariant to changes in actor and view. Finally, we found that a single class of hierarchical feedforward models provides a computational explanation for both the object and action recognition timing results. The methods and results in this thesis provide a basis for studying the timing and order of the neural computations underlying a wide range of visual and cognitive tasks. In the future, using MEG decoding data in conjunction with higher spatial resolution data on more realistic tasks can further progress our understanding of human visual recognition.

Bibliography

[1] David J Acunzo, Graham Mackenzie, and Mark C W van Rossum. Systematic biases in early ERP and ERF components as a result of high-pass filtering. *Journal of neuroscience methods*, 209(1):212–8, July 2012.

[2] E H Adelson and J R Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A, Optics and image science*, 2(2):284–99, February 1985.

[3] T. D. Albright. Cortical processing of visual motion., 1993.

[4] Truett Allison, Aina Puce, and Gregory McCarthy. Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, 4(7):267–278, July 2000.

[5] Fabio Anselmi, Joel Z Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations with low sample complexity: the magic of sensory cortex or a new framework for machine learning? (001), 03/2014 2014.

[6] Fabio Anselmi and Tomaso Poggio. Representation learning in sensory cortex: a theory. (026), 11/2014 2014.

[7] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On Invariance and Selectivity in Representation Learning. March 2015.

[8] Akiyuki Anzai, Xinmiao Peng, and David C Van Essen. Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, September 2007.

[9] Michael J Arcaro, Stephanie A McMains, Benjamin D Singer, and Sabine Kastner. Retinotopic organization of human ventral visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(34):10638–52, August 2009.

[10] Elizabeth Ashbridge and David Perrett. Generalizing across object orientation and size. In *Perceptual Constancy: Why Things Look as They Do*, page 560. Cambridge University Press, 1998.

[11] Sylvain Baillet. Encyclopedia of Computational Neuroscience. chapter Forward an. Springer New York, New York, NY, 2013.

[12] Daniel Baldauf and Robert Desimone. Neural mechanisms of object-based attention. *Science (New York, N.Y.)*, 344(6182):424–7, April 2014.

[13] Evgeniy Bart and Shimon Ullman. Class-based feature matching across unrestricted transformations. *IEEE transactions on pattern analysis and machine intelligence*, 30(9):1618–31, September 2008.

[14] Michael S Beauchamp, Kathryn E Lee, James V Haxby, and Alex Martin. FMRI responses to video and point-light displays of moving humans and manipulable objects. *Journal of cognitive neuroscience*, 15(7):991–1001, October 2003.

[15] Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological Studies of Face Perception in Humans. *Journal of cognitive neuroscience*, 8(6):551–565, November 1996.

[16] M C Booth and E T Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 8(6):510–23, September 1998.

[17] C Busettini, G S Masson, and F A Miles. Radial optic flow induces vergence eye movements with ultra-short latencies. *Nature*, 390(6659):512–5, December 1997.

[18] Charles F Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS computational biology*, 10(12):e1003963, December 2014.

[19] R Campbell, C A Heywood, A Cowey, M Regard, and T Landis. Sensitivity to eye gaze in prosopagnosic patients and monkeys with superior temporal sulcus ablation. *Neuropsychologia*, 28(11):1123–42, January 1990.

[20] Susan Carey and Elsa Bartlett. Acquiring a Single New Word. July 1978.

[21] Thomas Carlson, Hinze Hogendoorn, Hubert Fonteijn, and Frans A J Verstraten. Spatial coding and invariance in object-selective cortex. *Cortex; a journal devoted to the study of the nervous system and behavior*, 47(1):14–22, January 2011.

[22] Thomas Carlson, David A Tovar, Arjen Alink, and Nikolaus Kriegeskorte. Representational dynamics of object vision: the first 1000 ms. *Journal of vision*, 13(10):1–, January 2013.

[23] Patrick Cavanagh. Representations of Vision: Trends and Tacit Assumptions in Vision Research. chapter What's up, page 349. Cambridge University Press, 1991.

[24] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–62, March 2014.

[25] Bryan R Conroy, Benjamin D Singer, J Swaroop Guntupalli, Peter J Ramadge, and James V Haxby. Inter-subject alignment of human cortical anatomy using functional connectivity. *NeuroImage*, 81:400–11, November 2013.

[26] David D Cox, Philip Meier, Nadja Oertelt, and James J DiCarlo. 'Breaking' position-invariant object recognition. *Nature neuroscience*, 8(9):1145–7, September 2005.

[27] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical Surface-Based Analysis. *NeuroImage*, 9(2):179–194, 1999.

[28] R Desimone. Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience*, 3(1):1–8, January 1991.

[29] R Desimone, TD Albright, CG Gross, and C Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.*, 4(8):2051–2062, August 1984.

[30] R Desimone and J Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18:193–222, January 1995.

[31] R Desimone and L G Ungerleider. Multiple visual areas in the caudal superior temporal sulcus of the macaque. *The Journal of comparative neurology*, 248(2):164–89, June 1986.

[32] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–41, August 2007.

[33] James J DiCarlo and John H R Maunsell. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of neurophysiology*, 89(6):3264–78, June 2003.

[34] John P. Donoghue. Connecting cortex to machines: recent advances in brain interfaces. October 2002.

[35] P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science (New York, N.Y.)*, 293(5539):2470–3, September 2001.

[36] R. Dubner and S.M. Zeki. Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Research*, 35(2):528–532, December 1971.

[37] Wolfgang Einhäuser, Christoph Kayser, Peter König, and Konrad P Körding. Learning the invariance properties of complex cells from their responses to natural stimuli. *The European journal of neuroscience*, 15(3):475–86, February 2002.

[38] Yasmine El-Shamayleh and J Anthony Movshon. Neuronal responses to texture-defined form in macaque visual area V2. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(23):8543–55, June 2011.

[39] S. Engel. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, 7(2):181–192, March 1997.

[40] Boris Epshtein, Ita Lifshitz, and Shimon Ullman. Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14298–303, September 2008.

[41] R Epstein and N Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, April 1998.

[42] Michèle Fabre-Thorpe. The characteristics and limits of rapid visual categorization. *Frontiers in psychology*, 2:243, January 2011.

[43] L Fei-Fei. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, 2010.

[44] D. J. Felleman and D. C. Van Essen. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, 1(1):1–47, January 1991.

[45] B. Fischer and E. Ramsperger. Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, 57(1), 1984.

[46] B Fischl, A Liu, and A M Dale. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE transactions on medical imaging*, 20(1):70–80, January 2001.

[47] Bruce Fischl, Martin I. Sereno, and Anders M. Dale. Cortical Surface-Based Analysis. *NeuroImage*, 9(2):195–207, 1999.

[48] Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H Salat, Evelina Busa, Larry J Seidman, Jill Goldstein, David Kennedy, Verne Caviness, Nikos Makris, Bruce Rosen, and Anders M Dale. Automatically parcellating the human cerebral cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 14(1):11–22, January 2004.

[49] Peter Földiák. Learning Invariance from Transformation Sequences, March 1991.

[50] Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–81, July 2013.

[51] Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science (New York, N.Y.)*, 330(6005):845–51, November 2010.

[52] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, January 1988.

[53] Helen L. Gallagher and Christopher D. Frith. Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, 7(2):77–83, February 2003.

[54] J L Gallant, J Braun, and D C Van Essen. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science (New York, N.Y.)*, 259(5091):100–3, January 1993.

[55] J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J Neurophysiol*, 76(4):2718–2739, October 1996.

[56] Martin A Giese and Tomaso Poggio. Neural mechanisms for the recognition of biological movements. *Nature reviews. Neuroscience*, 4(3):179–92, March 2003.

[57] Melvyn A. Goodale and A.David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, January 1992.

[58] K Grill-Spector, T Kushnir, T Hendler, S Edelman, Y Itzchak, and R Malach. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human brain mapping*, 6(4):316–28, January 1998.

[59] Kalanit Grill-Spector, Tammar Kushnir, Shimon Edelman, Galia Avidan, Yacov Itzchak, and Rafael Malach. Differential Processing of Objects under Various Viewing Conditions in the Human Lateral Occipital Complex. *Neuron*, 24(1):187–203, September 1999.

[60] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature reviews. Neuroscience*, 15(8):536–548, June 2014.

[61] C. G. Gross, C. E. Rocha-Miranda, and D. B. Bender. Visual properties of neurons in inferotemporal cortex of the Macaque. *J Neurophysiol*, 35(1):96–111, January 1972.

[62] E. Grossman, M. Donnelly, R. Price, D. Pickens, V. Morgan, G. Neighbor, and R. Blake. Brain Areas Involved in Perception of Biological Motion. *Journal of Cognitive Neuroscience*, 12(5):711–720, September 2000.

[63] Emily D Grossman and Randolph Blake. Brain Areas Active during Visual Perception of Biological Motion. *Neuron*, 35(6):1167–75, September 2002.

[64] Emily D Grossman, Nicole L Jardine, and John A Pyles. fMR-Adaptation Reveals Invariant Coding of Biological Motion on the Human STS. *Frontiers in human neuroscience*, 4:15, January 2010.

[65] Marcos Perreau Guimaraes, Dik Kin Wong, E Timothy Uy, Logan Grosenick, and Patrick Suppes. Single-trial classification of MEG recordings. *IEEE transactions on bio-medical engineering*, 54(3):436–43, March 2007.

[66] Matti Hämäläinen, Riitta Hari, Risto J. Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. Magnetoencephalography:theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, April 1993.

[67] Matti S. Hämäläinen, Fa-Hsuan Lin, and John C. Mosher. Anatomically and Functionally Constrained Minimum-Norm Estimates : MEG: An Introduction to Methods. In *MEG: An Introduction to Methods*. 2010.

[68] J V Haxby, M I Gobbini, M L Furey, A Ishai, J L Schouten, and P Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293(5539):2425–30, September 2001.

[69] JV Haxby, EA Hoffman, and MI Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, June 2000.

[70] John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature reviews. Neuroscience*, 7(7):523–34, July 2006.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. February 2015.

[72] D J Heeger. Linking visual perception with human brain activity. *Current opinion in neurobiology*, 9(4):474–9, August 1999.

[73] J Hegdé and D C Van Essen. Selectivity for complex shapes in primate visual area V2. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(5):RC61, March 2000.

[74] J. Hegde and D. C. Van Essen. A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4. *Cerebral Cortex*, 17(5):1100–1116, June 2006.

[75] C J Holmes, R Hoge, L Collins, R Woods, A W Toga, and A C Evans. Enhancement of MR images using registration for signal averaging. *Journal of computer assisted tomography*, 22(2):324–33, 1998.

[76] D H Hubel and T N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160:106–54, January 1962.

[77] D H Hubel and T N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–43, March 1968.

[78] D. H. Hubel and T. N. Wiesel. Ferrier Lecture: Functional Architecture of Macaque Monkey Visual Cortex. *Proceedings of the Royal Society B: Biological Sciences*, 198(1130):1–59, May 1977.

[79] Alexander C. Huk, Robert F. Dougherty, and David J. Heeger. Retinotopy and Functional Subdivision of Human Areas MT and MST. *J. Neurosci.*, 22(16):7195–7205, August 2002.

[80] Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science (New York, N.Y.)*, 310(5749):863–6, November 2005.

[81] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–24, December 2012.

[82] Leyla Isik, Joel Z Leibo, and Tomaso Poggio. Learning and disrupting invariance in visual recognition with a temporal association rule. *Frontiers in computational neuroscience*, 6:37, January 2012.

[83] Leyla Isik, Ethan M Meyers, Joel Z Leibo, and Tomaso Poggio. The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology*, 111(1):91–102, January 2014.

[84] Elias B Issa, Alex M Papanastassiou, and James J DiCarlo. Large-scale, high-resolution neurophysiological maps underlying FMRI of macaque temporal lobe. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(38):15207–19, September 2013.

[85] M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol*, 73(1):218–226, January 1995.

[86] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[87] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A Biologically Inspired System for Action Recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[88] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, June 1973.

[89] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–85, May 2005.

[90] N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(11):4302–11, June 1997.

[91] N Kanwisher and E Wojciulik. Visual attention: insights from brain imaging. *Nature reviews. Neuroscience*, 1(2):91–100, November 2000.

[92] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. December 2014.

[93] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. IEEE, June 2014.

[94] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–5, March 2008.

[95] Holle Kirchner and Simon J Thorpe. Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision research*, 46(11):1762–76, May 2006.

[96] E Kobatake and K Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3):856–67, March 1994.

[97] Zoe Kourtzi and Nancy Kanwisher. Activation in Human MT/MST by Static Images with Implied Motion. *Journal of Cognitive Neuroscience*, 12(1):48–55, January 2000.

[98] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–41, December 2008.

[99] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[100] Camilo Lamus, Matti S Hämäläinen, Simona Temereanca, Emery N Brown, and Patrick L Purdon. A spatiotemporal dynamic distributed solution to the MEG inverse problem. *NeuroImage*, 63(2):894–909, November 2012.

[101] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011*, pages 3361–3368. IEEE, June 2011.

[102] R.M Leahy, J.C Mosher, M.E Spencer, M.X Huang, and J.D Lewine. A study of dipole localization accuracy for MEG and EEG using a human skull phantom. *Electroencephalography and Clinical Neurophysiology*, 107(2):159–173, April 1998.

[103] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989.

[104] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[105] Joel Z. Leibo, James Mutch, and Tomaso Poggio. Why The Brain Separates Face Recognition From Object Recognition. In *Advances in Neural Information Processing Systems (NIPS)*, Granada, Spain, 2011.

[106] Joel Z Leibo, Jim Mutch, Lorenzo Rosasco, Shimon Ullman, and Tomaso Poggio. Learning Generic Invariances in Object Recognition: Translation and Scale. December 2010.

[107] Nuo Li and James J DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science (New York, N.Y.)*, 321(5895):1502–7, September 2008.

[108] Nuo Li and James J DiCarlo. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron*, 67(6):1062–75, September 2010.

[109] Qianli Liao, Joel Z Leibo, and Tomaso Poggio. Unsupervised learning of clutter-resistant visual representations from natural videos. *arXiv preprint arxiv:1409.3879*, 2014.

[110] Angelika Lingnau and Paul E Downing. The lateral occipitotemporal cortex in action. *Trends in cognitive sciences*, April 2015.

[111] Hesheng Liu, Yigal Agam, Joseph R Madsen, and Gabriel Kreiman. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2):281–90, April 2009.

[112] N K Logothetis and D L Sheinberg. Visual object recognition. *Annual review of neuroscience*, 19:577–621, January 1996.

[113] Nikos K. Logothetis, Jon Pauls, and Tomaso Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563, May 1995.

[114] Ellen M. Markman. *Categorization and Naming in Children: Problems of Induction.* MIT Press, 1991.

[115] Timothee Masquelier, Thomas Serre, Simon Thorpe, and Tomaso Poggio. Learning Complex Cell Invariance from Natural Videos: A Plausibility Proof. December 2007.

[116] W H Merigan and J H Maunsell. How parallel are the primate visual pathways? *Annual review of neuroscience*, 16:369–402, January 1993.

[117] Ethan M. Meyers. The neural decoding toolbox. *Frontiers in Neuroinformatics*, 7, May 2013.

[118] Ethan M. Meyers. The neural decoding toolbox. *Frontiers in Neuroinformatics*, 7, May 2013.

[119] Ethan M Meyers, David J Freedman, Gabriel Kreiman, Earl K Miller, and Tomaso Poggio. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology*, 100(3):1407–19, September 2008.

[120] Lars MICHELS, Markus LAPPE, and Lucia Maria VAINA. Visual areas involved in the perception of human movement from dynamic form analysis. *Neuroreport*, 16(10):1037–1041.

[121] Mortimer Mishkin and Leslie G. Ungerleider. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, 6(1):57–77, September 1982.

[122] J Movshon, E Adelson, M Gizzi, and W Newsome. The analysis of moving visual patterns. In C Chagas, R Gattas, and C Gross, editors, *Pattern Recognition Mechanisms*, pages 117–151. Experimental Brain Research, 1997.

[123] Jim Mutch, Ulf Knoblich, and Tomaso Poggio. CNS: a GPU-based framework for simulating cortically-organized networks. *MIT-CSAIL-TR-2010-013*, 2010.

[124] WT Newsome, MS Gizzi, and JA Movshon. Spatial and temporal properties of neurons in macaque mt. *Investigative Ophthalmology and Visual Science*, 24(106):365–375, 1983.

[125] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology : CB*, 21(19):1641–6, October 2011.

[126] LG Nowak and J Bullier. The Timing of information transfer in the visual system. In K Rockland, J Kaas, and A Peters, editors, *Cerebral Cortex: Extrastriate Cortex in Primates*, page 870. Plenum Publishing Corporation, 1997.

[127] M. W. Oram and D. I. Perrett. Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *J Neurophysiol*, 76(1):109–129, July 1996.

[128] Wanmei Ou, Matti S. Hämäläinen, and Polina Golland. A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946, 2009.

[129] A Pasupathy and C E Connor. Responses to contour features in macaque area V4. *Journal of neurophysiology*, 82(5):2490–502, November 1999.

[130] Marius V Peelen and Paul E Downing. Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, 93(1):603–8, January 2005.

[131] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):S199–209, March 2009.

[132] D I Perrett, J K Hietanen, M W Oram, and P J Benson. Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 335(1273):23–30, January 1992.

[133] D.I. Perrett, P.A.J. Smith, A.J. Mistlin, A.J. Chitty, A.S. Head, D.D. Potter, R. Broennimann, A.D. Milner, and M.A. Jeeves. Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report. *Behavioural Brain Research*, 16(2-3):153–170, August 1985.

[134] Marios G Philiastides, Roger Ratcliff, and Paul Sajda. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(35):8965–75, August 2006.

[135] Marios G Philiastides and Paul Sajda. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral cortex (New York, N.Y. : 1991)*, 16(4):509–18, April 2006.

[136] M. Piccardi. Background subtraction techniques: a review. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 4, pages 3099–3104. IEEE, 2004.

[137] David C. Plaut and Geoffrey E. Hinton. Learning sets of filters using back-propagation. *Computer Speech & Language*, 2(1):35–61, March 1987.

[138] M C Potter. Short-term conceptual memory for pictures. *Journal of experimental psychology. Human learning and memory*, 2(5):509–22, September 1976.

[139] Mary C Potter. Recognition and memory for briefly presented scenes. *Frontiers in psychology*, 3:32, January 2012.

[140] Mary C Potter, Brad Wyble, Carl Erick Hagmann, and Emily S McCourt. Detecting meaning in RSVP at 13 ms per picture. *Attention, perception & psychophysics*, 76(2):270–9, February 2014.

[141] Aina Puce and David Perrett. Electrophysiology and brain imaging of biological motion. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1431):435–45, March 2003.

[142] Roger Ratcliff, Marios G Philiastides, and Paul Sajda. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6539–44, April 2009.

[143] John H Reynolds and David J Heeger. The normalization model of attention. *Neuron*, 61(2):168–85, January 2009.

[144] M Riesenhuber and T Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–25, November 1999.

[145] Anna W Roe, Leonardo Chelazzi, Charles E Connor, Bevil R Conway, Ichiro Fujita, Jack L Gallant, Haidong Lu, and Wim Vanduffel. Toward a unified theory of visual area V4. *Neuron*, 74(1):12–29, April 2012.

[146] E T Rolls. Learning mechanisms in the temporal lobe visual cortex. *Behavioural brain research*, 66(1-2):177–85, January 1995.

[147] E T Rolls. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–18, August 2000.

[148] S. Romdhani and T. Vetter. Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 986–993. IEEE, 2005.

[149] Guillaume A Rousselet. Does Filtering Preclude Us from Studying ERP Time-Courses? *Frontiers in psychology*, 3:131, January 2012.

[150] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[151] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. page 43, September 2014.

[152] Nicole C Rust and James J Dicarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(39):12978–95, September 2010.

[153] R Saxe and N Kanwisher. People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4):1835–42, August 2003.

[154] P H Schiller and K Lee. The role of the primate extrastriate area V4 in vision. *Science (New York, N.Y.)*, 251(4998):1251–3, March 1991.

[155] Matthew T. Schmolesky, Youngchang Wang, Doug P. Hanes, Kirk G. Thompson, Stefan Leutgeb, Jeffrey D. Schall, and Audie G. Leventhal. Signal Timing Across the Macaque Visual System. *J Neurophysiol*, 79(6):3272–3278, June 1998.

[156] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. *MIT-CSAIL-TR-2005-082*, December 2005.

[157] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–9, April 2007.

[158] Eero P. Simoncelli and David J. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761, March 1998.

[159] Jedediah M Singer and David L Sheinberg. Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(8):3133–45, February 2010.

[160] R. Socher. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, June 2009.

[161] Michael W Spratling. Learning viewpoint invariant perceptual representations from cluttered images. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):753–61, May 2005.

[162] S M Stringer, G Perry, E T Rolls, and J H Proske. Learning invariant object recognition in the visual system with continuous transformations. *Biological cybernetics*, 94(2):128–42, February 2006.

[163] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–63, August 2012.

[164] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. September 2014.

[165] Andrea Tacchetti, Pavan K. Mallapragada, Matteo Santoro, and Lorenzo Rosasco. Gurls: A least squares library for supervised learning. *Journal of Machine Learning Research*, 14:3201–3205, 2013.

[166] François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational intelligence and neuroscience*, 2011:879716, January 2011.

[167] K Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19:109–39, January 1996.

[168] K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J Neurophysiol*, 66(1):170–189, July 1991.

[169] Hanlin Tang, Calin Buia, Radhika Madhavan, Nathan E Crone, Joseph R Madsen, William S Anderson, and Gabriel Kreiman. Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron*, 83(3):736–48, August 2014.

[170] C.D. Tesche, M.A. Uusitalo, R.J. Ilmoniemi, M. Huotilainen, M. Kajola, and O. Salonen. Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. *Electroencephalography and Clinical Neurophysiology*, 95(3):189–200, September 1995.

[171] S Thorpe, D Fize, and C Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–2, July 1996.

[172] S. J. Thorpe. NEUROSCIENCE: Seeking Categories in the Brain. *Science*, 291(5502):260–263, January 2001.

[173] Doris Y Tsao, Winrich A Freiwald, Roger B H Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science (New York, N.Y.)*, 311(5761):670–4, March 2006.

[174] S. Ullman and S. Soloviev. Computation of pattern invariance in brain-like structures. *Neural Networks*, 12(7-8):1021–1036, October 1999.

[175] L G Ungerleider and R Desimone. Cortical connections of visual area MT in the macaque. *The Journal of comparative neurology*, 248(2):190–222, June 1986.

[176] K. Uutela, M. Hämäläinen, and E. Somersalo. Visualization of Magnetoencephalographic Data Using Minimum Current Estimates. *NeuroImage*, 10(2):173–180, 1999.

[177] L M Vaina, J Solomon, S Chowdhury, P Sinha, and J W Belliveau. Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11656–61, September 2001.

[178] Joris Vangeneugden, Marius V Peelen, Duje Tadin, and Lorella Battelli. Distinct neural mechanisms for body form and body motion discriminations. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(2):574–85, January 2014.

[179] Joris Vangeneugden, Frank Pollick, and Rufin Vogels. Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. *Cerebral cortex (New York, N.Y. : 1991)*, 19(3):593–611, March 2009.

[180] T Vetter, A Hurlbert, and T Poggio. View-based models of 3D object recognition: invariance to imaging transformations. *Cerebral cortex (New York, N.Y. : 1991)*, 5(3):261–9, January.

[181] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. November 2014.

[182] Stephan Waldert, Hubert Preissl, Evariste Demandt, Christoph Braun, Niels Birbaumer, Ad Aertsen, and Carsten Mehring. Hand movement direction decoded from MEG and EEG. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(4):1000–8, January 2008.

[183] G Wallis and H H Bülthoff. Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4800–4, April 2001.

[184] Guy Wallis and Edmund T. Rolls. INVARIANT FACE AND OBJECT RECOGNITION IN THE VISUAL SYSTEM. *Progress in Neurobiology*, 51(2):167–194, February 1997.

[185] Dirk B Walther, Eamon Caddigan, Li Fei-Fei, and Diane M Beck. Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(34):10573–81, August 2009.

[186] B A Wandell. Computational neuroimaging of human visual cortex. *Annual review of neuroscience*, 22:145–73, January 1999.

[187] KYLIE J. WHEATON, ANDREW PIPINGAS, RICHARD B. SILBERSTEIN, and AINA PUCE. Human neural responses elicited to observing the actions of others. *Visual Neuroscience*, 18(03):401–406, May 2001.

[188] Laurenz Wiskott. How Does Our Visual System Achieve Shift and Size Invariance?, December 2004.

[189] Laurenz Wiskott and Terrence J. Sejnowski. *Neural Computation*, March.

[190] Fei Xu and Joshua B Tenenbaum. Word learning as Bayesian inference. *Psychological review*, 114(2):245–72, April 2007.

[191] Yann Lecun Y-Lan Boureau, Jean Ponce. A Theoretical Analysis of Feature Pooling in Visual Recognition.

[192] Daniel L K Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24, June 2014.

[193] Ying Zhang, Ethan M Meyers, Narcisse P Bichot, Thomas Serre, Tomaso A Poggio, and Robert Desimone. Object decoding with attention in inferior temporal cortex. 2011.