Comparing computational models of vision to human behaviour

Thomas Colvin, MSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy, September 2017

Abstract

Biological vision and computational models of vision can be split into three independent components (*image description, decision process,* and *image set*). The thesis presented here aimed to investigate the influence of each of these core components on computational model's similarity to human behaviour. Chapter 3 investigated the similarity of different computational image descriptors to their biological counterparts, using an image matching task. The results showed that several of the computational models could explain a significant amount of the variance in human performance on individual images. The deep supervised convolutional neural net explained the most variance, followed by GIST, HMAX and then PHOW. Chapter 4 investigated which computational decision process best explained observers' behaviour on an image categorization task. The results showed that Decision Bound theory produced behaviour the closest to that of observers. This was followed by Exemplar theory and Prototype theory. Chapter 5 examined whether the naturally differing image set between computational models and observers could partially account for the difference in their behaviour. The results showed that, indeed, the naturally differing image set between computational models and observers was affecting the similarity of their behaviour. This gap did not alter which image descriptor best fit observers' behaviour and could be reduced by training observers on the image set the computational models were using. Chapter 6 investigated, using computational models of vision, the impact

of the neighbouring (masking) images on the target images in a RSVP task. This was done by combining the neighbouring images with the target image for the computational models' simulation for each trial. The results showed that models behaviour became closer to that of the human observers when the neighbouring mask images were included in the computational simulations, as would be expected given an integration period for neural mechanisms.

This thesis has shown that computational models can show quite similar behaviours to human observers, even at the level of how they perform with individual images. While this shows the potential utility in computational models as a tool to study visual processing, It has also shown the need to take into account many aspects of the overall model of the visual process and task; not only the image description, but the task requirements, the decision processes, the images being used as stimuli and even the sequence in which they are presented.

List of Contents

List of Tables

List of Illustrations

Acknowledgements

I would like to say a big thanks to my supervisor Jon Peirce for his constant support and guidance.

I would like to show my appreciation to my secondary supervisor Alain Pitiot for his ideas and advice.

I would also like to thank my office mates for their advice and help throughout the PhD; Richard James, Mike Long and Zack Ellerby.

I am grateful to my parents, John Colvin and Gabriella Gibson, for all of their encouragement and support.

A special thanks has to go to Byron Katie for keeping me sane, as well as Klaudia Kania for all her love and acceptance.

Chapter 1 - Literature Review

1.1. Using computational models to understand human vision.

Early research into human vision made promising strides forward in understanding the mechanisms employed by the mammalian physiology (Hubel & Wiesel, 1962, 1968). The sheer complexity of the human visual system soon revealed itself and it became apparent that new approaches needed to be made. Marr (1982), in his book *Vision,* highlighted the need to incorporate computer models and explicit algorithms which could be tested against human biology in order to gauge our understanding of the visual system. Marr (1982) reasons that the ideal process of using computational models to understand the visual system would follow three cyclical stages. First a theory needed to be formed based on observations of the biology, next computational algorithms implementing said theory need to be created, finally the fit of these algorithms to what is biologically implemented needed to be ascertained.

Since Marr first proposed this theory, vision research has evolved, and a number of highly sophisticated computational models have been produced. In recent years, these models have been rapidly approaching human performance for certain, reasonably constrained, natural image-based tasks (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010; Russakovsky et al., 2015). The current models being produced differ enormously in their implementations and comparisons to knowledge of human biology in their attempt to reach

human performance. Some of these models have been based heavily upon theory and understanding of human biology (Jarrett, Kavukcuoglu, Ranzato, Lecun, & Ieee, 2009; Krizhevsky, Sutskever, & Hinton, 2012; Serre, Oliva, & Poggio, 2007). On the other hand, a large number of models of image recognition have not limited themselves to theory and biology and have instead created models based on optimized mathematical algorithms which only loosely follow, or completely ignore, knowledge about mammalian biology (Lazebnik, Schmid, & Ponce, 2006; Lowe, 2004; Pass & Zabih, 1999). The current models provide a rich base of different ideas which can be compared to human observers. The fact that some of these models are based on human ingenuity rather than anything known about the biology provides new, out-of-the-box avenues for investigating human biology.

Following Marr's approach to understanding the visual system, research is needed to assess the similarity of these different computational models to human observers. Recent research has focused on two main areas of comparisons; comparing computer models to human neural activity and comparing computer models to human behavior.

This rest of this chapter will focus on; (1) outline a basic model of image processing (2) describing different image descriptors (3) outline a brief history of computational models of vision in the form of image database competitions, (3) outlining the literature comparing computer models of vision to human

behavior, (4) outlining the literature comparing computer models of vision to neurology in fMRI studies.

1.2. Model of Visual Processing

In order to understand the nuances of the comparisons being made between computational model and human observers a general framework of visual processing is presented here. While previous works have often referred to "computational models of vision", they do not necessarily include all the steps that are needed for a visual task to be performed. For instance, many studies focus on the *image descriptor* (whereby input images are encoded and stored) with relatively little consideration to the necessary *decision process* that must be performed on that image descriptor for a task to be performed. We therefore present a general framework for considering the overall process in Figure 1.1. This framework will be referred to throughout the thesis and can be considered the backbone on which the various experiments presented in this thesis are based.

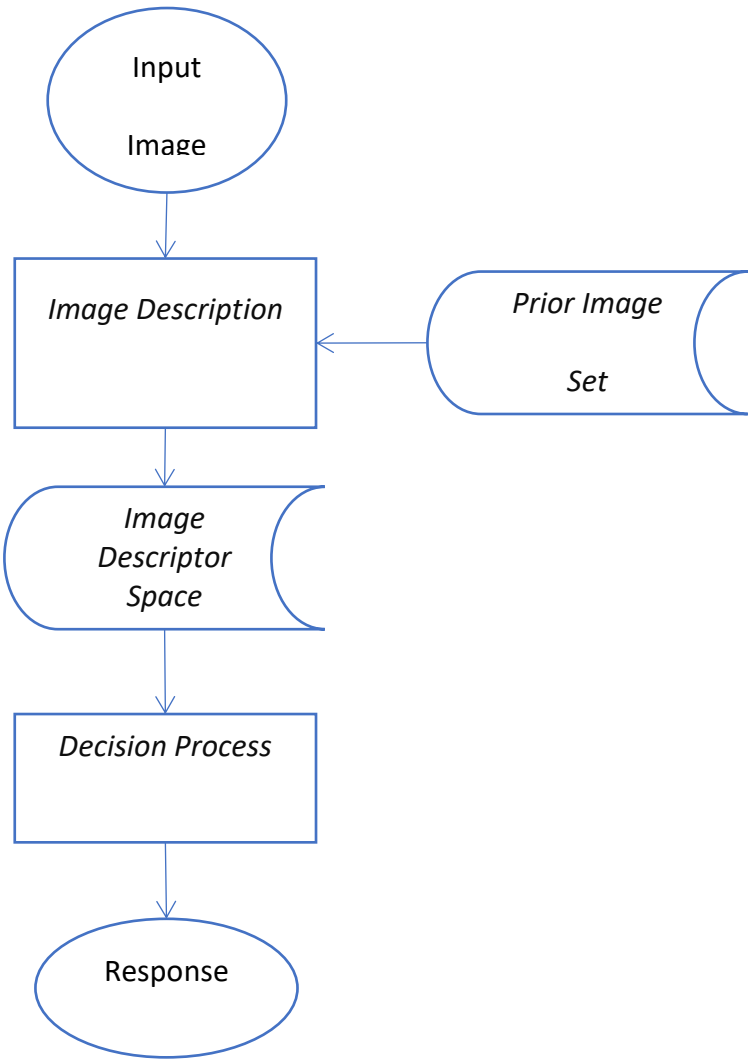


Figure 1.1. A basic model of visual processing based on how humans and computer models are likely to handle a visual processing task. Square boxes represent processes being applied. Curved arrow boxes represent sets of data. In this model, both the input image and the pevious set of images are passed through an image discriptor to describe the images. The output image descriptions and are then applied to the image descriptor space. A decision is then made based on the input image and the previous image set and is task dependent.

The *model of visual processing* presented here introduces new key words, which are written in *italics*. The *model of visual processing* is made up of various components. These components are shared between both human observers and computational models and thus it can be seen as a general model describing the processes that happen in both.

The *model of visual processing* starts with the initial input of an image. For both the computational model and human observer this is in the form of light intensity values. Light intensity values are a notoriously hard format of information (Ghodrati, Farzmahdi, Rajaei, Ebrahimpour, & Khaligh-Razavi, 2014) to base decisions upon and in order to produce meaningful behaviors the format of information needs to be changed to something more useful. In the *model of visual processing* these light intensity values get passed through an *image descriptor* to form an *image description*. In computational models of vision the *image descriptor* takes the form of a mathematical algorithm which extracts information or finds features within the image. The *image description* created by the *image descriptor* generally takes the form of a vector representing the multiple dimensions of which the image was assessed. In the human brain the *image descriptor* takes the form of the processes which happen to information through the visual cortex. The *image description* used in the brain is likely localized in different visual areas; for objects the final *image description* is likely held in Area IT (Bell, Hadj-Bouziane, Frihauf, Tootell, & Ungerleider, 2009; Hung, Kreiman, Poggio, & DiCarlo, 2005; Kiani, Esteky, Mirpour, & Tanaka, 2007), while scene information is likely held in a number of

different areas such as the parahippocampal place area (Aguirre & Desposito, 1997; Epstein & Kanwisher, 1998), the retrosplenial cortex (Maguire, 2001; Vann, Aggleton, & Maguire, 2009), and the transverse occipital sulcus (Nasr & Tootell, 2012). Once this *image description* has been made it is added to the *image descriptor space*. The *image descriptor space* also stores the *image descriptions* of all the images in the *image set*. Here the *image set* refers to all of the images that the computational model or biological system has had previous exposure to and has access to when making a decision. In the *model of visual processing* presented here, the last step is calculating a *decision process* by which an output is produced. This *decision process* is task dependent, a categorization task is going require different computations than an image similarity task, and so the *decision process* may take various forms.

The *model of visual processing* presents the view that a computational model must be viewed as a complete process, from input image to behavioral output. Often in the previous literature a computational model of vision can be an ambiguous term, it can sometimes refer to an *image descriptor* without an explicit decision process attached to it. Good examples of this comes from the *image descriptors* GIST and SIFT (Lowe, 2004; Oliva & Torralba, 2001) that on their own they only produce an *image description* and lack a specified *decision process*. To clarify, in the terminology used throughout this thesis, any reference to a computational model refers to the general process (image descriptor, image set, decision process etc.). While each individual process (like

GIST) that calculates an *image description* will be referred to as an *image descriptor.*

When comparing a computational model of vision to human observers it is important to consider each of these three components, as varying any one might affect the computational model's similarity to human observers. A prime example of this comes from if the *image set* is not fully considered. A computational model could be identical to human observers in *image description* and *decision process,* but if the computational model uses an *image set* that differs dramatically from images that human observers have encountered then the output behavior is likely to differ. Thus this *model of visual processing* encourages the similarity between computer model and human to be examined from a more complete perspective of *image descriptor,* prior *image set* and also the *decision process.*

In the previous literature, computational models have been compared to humans in terms of neural activity and behavior. At first glance it may be tempting to assume that these metrics measure the same thing, namely similarity of computer model to human vision. On closer inspection, these different types of comparisons are subtly different when put in the perspective of the *model of visual processing*. When neural activity is being compared to computational models it is assessing the similarity of the *image descriptions* being produced (e.g. if two images are close in patterns of neural activity are they similarly close in the computer model's *image description*). Again, in the

more general model of this thesis, it is not that neural activity is being compared to computational models. Rather, it is being compared to computational *image descriptors,* ignoring the element of a *decision process.* On the other hand comparing computer models to observers' behavior studies the whole process; similarity of the output of an *image description* paired with a *decision process*.

Comparing computational models to human behavior has the additional complexity of type of task. It is likely that the *image description* is task invariant, while the *decision process* is highly task variant. The level of uncertainty of the *decision process* employed in humans on a task can vary. In very simple tasks, such as image recognition task the *decision process* is almost guaranteed to be based upon distance of images' description in the *descriptor space* (Attneave, 1957; Shepard, 1962a, 1962b, 1987). In a complex task such as image categorization there is a high degree of uncertainty of the *decision process* employed by humans (Ashby & Maddox, 2005, 2011). In tasks where the *decision process* is already relatively known comparing computer models to human behavior can be used to assess the similarity of *image descriptors.* On more complex tasks where there is a large degree of uncertainty of the *decision process*, computational models can be compared to human observers to assess the similarity of various *decision processes.*

Comparing computational models to behavior and neural activity can work synergistically. They both offer an assessment of similarity between

computational *image descriptions* and neural *image descriptions*. If both produce a similar result for each computational *image description* then it is likely the true similarity is somewhere in that region.

It is important when making these comparisons to consider exactly what these similarity measurements mean. In the case of comparing computational models to neural activity and behavior the similarity measures produced are with respect to the output of the process. These measures of similarity are blind to the underlying algorithmic calculations which calculated the output. It is possible to conceive of multiple methods that all produce the same output and would thus score the same on these measures of similarity. Therefore, it would be incorrect to assert that if a computational model is similar to human observers on these metrics that they are performing calculations in a similar manner. In order to assess if computational models are carrying out the same algorithmic calculations as human biology then cell recordings and other methods are more appropriate. Instead, counter intuitively, the metrics of similarity described here are more relevant at pulling out the differences between computational models and human observers. As a model becomes more different it is easier to assert that the computational model is processing information in a different way to human observers. This could either be due to differences in algorithm or perhaps missing components altogether. These metrics of similarity described here can therefore still provide a loose general assessment on the extent of understanding of the algorithms used in the human visual system in general.

1.3. State of the art computational models of vision

This section focuses on providing a brief overview to common computational models of vision mentioned in the literature as well as those specifically used in this thesis. The list of models presented here are by no means exhaustive, but instead has been designed to provide a sweeping overview of the many different types of models out there. For a more complete list of models in greater detail there are a number of reviews which may be helpful (Andreopoulos & Tsotsos, 2013; Khaligh-Razavi, 2014; Mikolajczyk & Schmid, 2005).

The definition outlined in the previous section of a computational model was an *image descriptor* paired with a *decision process*. It is common in the literature to refer to an *image descriptor,* without an explicit *decision process* attached, as a "computational model". Keeping with the definition used in the literature, this section can be thought of as a list of *image descriptors* and when the authors have paired the *image descriptor* with a specific *decision process* then that too shall also be described. Equally when a specific *decision process* is described it is usually because it has been considered to work well when paired with the *image descriptor,* but it doesn't mean that other *decision processes* could not be applied to the *image descriptor.*

Color histograms. A color histogram is an *image descriptor* which describes the image based on the number of pixels of a given color in the image. Color histogram algorithms come in many forms (Hsu, Chua, & Pung, 1995; Pass &

Zabih, 1999; Stricker & Dimai, 1996). In general color histogram algorithms scale the image so that it contains a standard number of pixels. The algorithm then takes the images pixels and convert them into a color space with a reduced palette of discrete colors. A histogram is then formed of the different colors of the image and the image is described as a vector containing the histogram values. Standard color histograms are unable to capture the spatial layout of the color information, but some have worked around this by dividing the image spatially into subsections and creating multiple histograms (Hsu et al., 1995). These regions can also be made to have slight invariance by creating overlapping regions (Stricker & Dimai, 1996). A popular example is the Joint Histogram (Pass & Zabih, 1999) which is multidimensional and takes advantage of the fact that other image properties can be constructed into a histogram similarly to color. Four additional image properties are formed into histograms, edge density surrounding a pixel, a measure of texture, gradient magnitude and also the pixels rank within the light intensity values of its closest neighbors. By using additional image properties in the same way as color they have provided a method which in their tests Pass and Zabih, (1999) was shown to be superior to using color alone. Color histogram techniques are largely used in image data base retrieval as well as assessing image similarity.

Geometric Blur. The Geometric Blur descriptor is designed with the concept in mind that object recognition is a problem solvable by deformable shape matching. Geometric blur is based on the observation that objects of the same class or category often take very similar shapes. Any variations when matching

objects could be solved by geometric transformations that can deform the object's shape into alignment. This approach is particularly useful when objects are viewed from different angels or distances, such that the object's shape can easily be distorted to fit the original image. Geometric blur is calculated by selecting points of interest of an image. This can either be uniformly done (Khaligh-Razavi & Kriegeskorte, 2014) or by selecting points of interest through the use of line detectors (Belongie, Malik, & Puzicha, 2002; Berg, Berg, & Malik, 2005; Berg & Malik, 2001). Spatial blurring is then applied around each point of interest, with increasing blur for pixels further from the interest point. This is done with the intention of aiding point matching, as spatial blurring allows for detailed information to be taken directly around the interest point, while also allowing for some coarse context from the surrounding region. Matching then occurs on these points of interest between the original image and the one in question. Geometric blur descriptor was primarily designed with the purpose of image matching but has been applied to the task of image categorization (Zhang, Berg, Maire, & Malik, 2006).

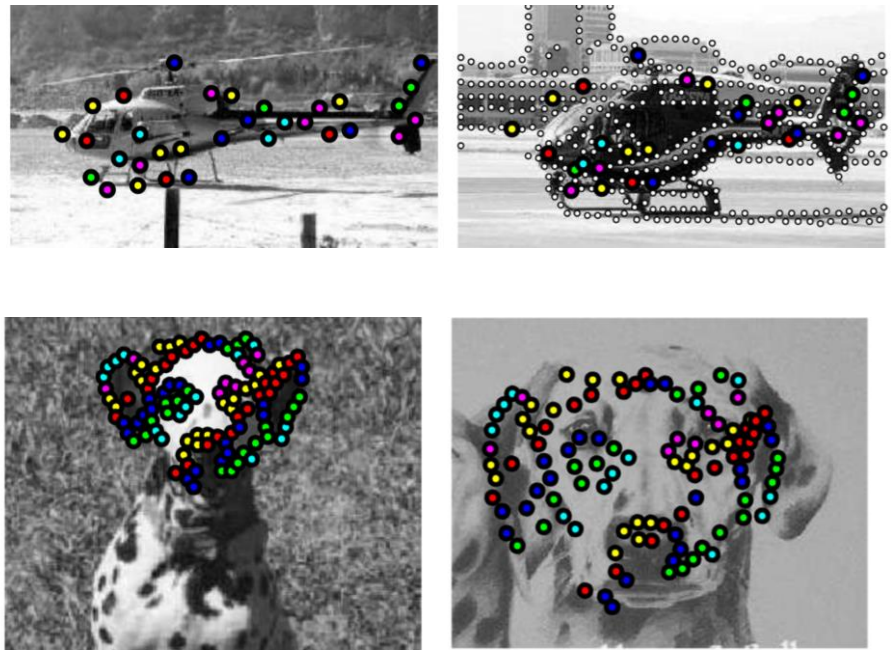


Figure 1.2. Examples of the geometric blur feature points extracted from sample images (Helicopter and a Dog) and applied to a new image of the same image category. Illustrations taken from Berg et al., (2005).

Local Binary Patterns. The Local binary patterns descriptor is an image descriptor designed using texture as the key component of forming an image description. It runs on the idea that different objects have different texture patterns and images belonging to the same category have more similar textures than objects belonging to a different category. Local binary patterns are not computationally costly and have been shown to be relatively robust to illumination changes (Ojala, Pietikainen, & Harwood, 1994, 1996). This descriptor has also been shown to be a good image descriptor for tasks that are not regarded as primarily a texture problem, such as face detection and

motion analysis (Pietikainen, Hadid, Zhao, & Ahonen, 2011). A local binary pattern is calculated by dividing an image into X by X windows; usually 12 x 12. Each pixel in the image is compared to the pixels directly surrounding it. These 8 pixel neighbors are classified as either a 0 or a 1 depending on if their luminance value is greater or smaller. This gives an 8-digit binary number describing each pixel in the image. A histogram for each window is created counting the frequency of the values assigned to each pixel. These histograms are concatenated to produce a vector used to describe the image.

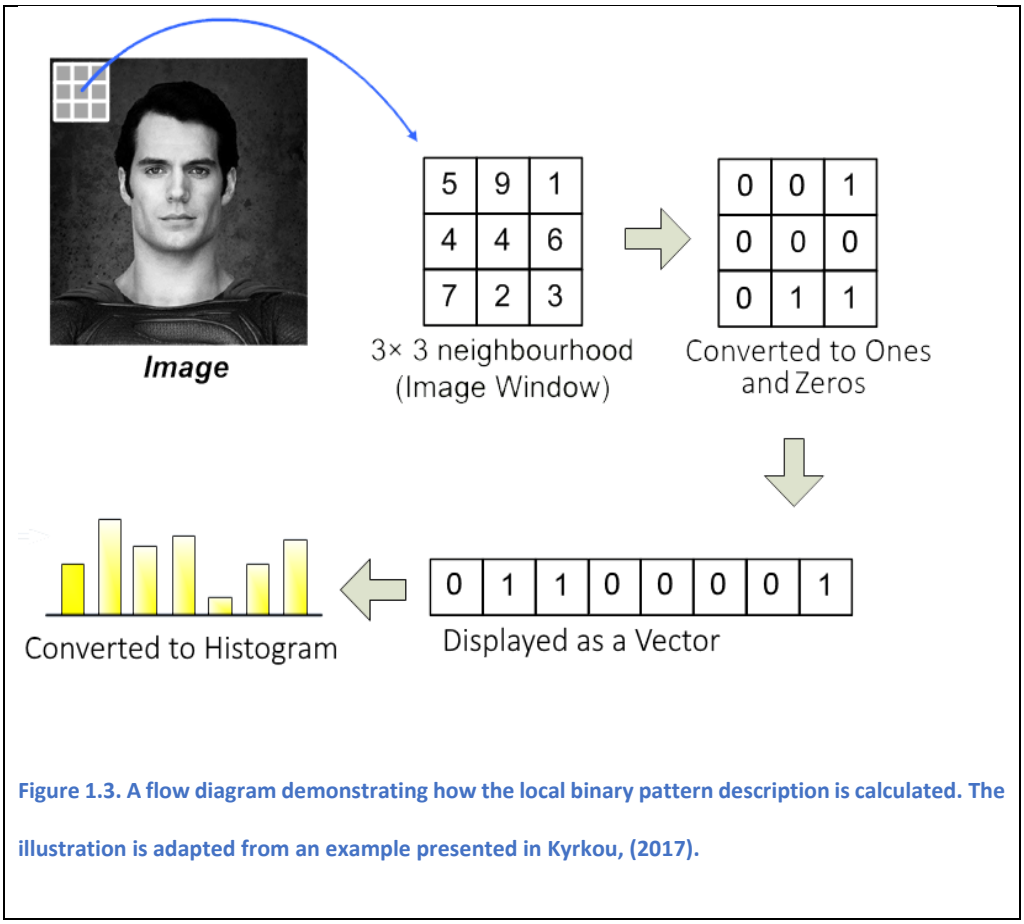


Figure 1.3. A flow diagram demonstrating how the local binary pattern description is calculated. The illustration is adapted from an example presented in Kyrkou, (2017).

GIST. The GIST image descriptor is an algorithm which bases its image description upon the spatial frequency information immediately available in

the image. GIST is modeled on the filtering transformations known to be performed in early visual cortex (Hubel & Wiesel, 1962, 1968), where spatial frequency information is extracted through Gabor-like filters. GIST is often referred to as a model of rapid, purely feedforward processing in humans when time is limited. GIST has been shown to excel as an image descriptor of scenes (Oliva & Torralba, 2001), categorizing scenes in a similar manner to humans on scales of naturalness, openness and roughness. Figure 1.4 shows how the spatial frequency information in an image is informative about the category it belongs to.

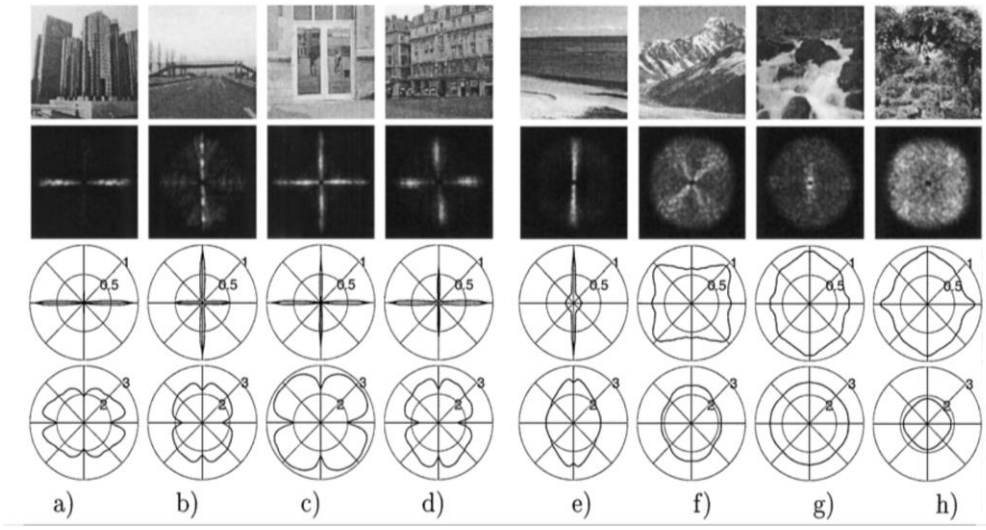


Figure 1.4. An example of how examining the spatial frequency information of an image is informative as to which category the image belong. The top row displays the image examined. From left to right (a to h) the image categories of tall buildings, highway, urban close-up, city centre, coast, mountain, natural close-up views and forests are displayed. Beneath each image is the spectrogram of the image (energy spectra have been multiplied to enhance visibility of high spatial frequency information). Taken from Oliva & Torralba, (2001).

Scale Invariant Feature Transform (SIFT). SIFT is a popular image descriptor that in its original form is mainly a method of image matching (Lowe, 2004). One of the major advantages of SIFT is it is scale invariant and is relatively robust to many common image transforms as well as image rotation and affine distortion. It is designed to approach the task of image recognition from the viewpoint that an object is defined by its features and so the task of object recognition should focus on finding features of interest in an image and then describing them. These features can also be selected in another image and if found to be similar enough then established as the same object. SIFT can be thought of as a model with two sections, feature detection and feature description. Feature detection is usually done by detecting rapid changes in luminance of the image (edges). The description is then created by calculating a histogram of the weighted gradients of the pixels around the feature. Due to SIFTs popularity there have been a number of extensions to the model. Dense SIFT is a modified version of the SIFT descriptor which samples uniformly across the image for features to describe and has been shown to have the same or even better performances than using interest points (Yap, Chen, Li, & Wu, 2010). Due to SIFT's popularity and success as an image descriptor, Muralidharan & Vasconcelos (2010) proposed a biologically plausible variant called BioSIFT. An example of SIFT's output is displayed in

Figure 1.5.

Figure 1.5. Two example images in which SIFT descriptors being found on the images on the left and relocated on the images on the right. Images were taken from Yu & Morel, (2011).

PHOW. PHOW is a specific extension of the SIFT model which is specialized for scene and object categorization. PHOW employs a bag-of-words model which originates from text classification but has since been applied to vision. A bag-of-words model uses the frequency of words of a document to describe the

document. In PHOW each SIFT feature is treated as a word describing the image and the frequency of each SIFT descriptor is used as the image description. Those images containing similar features and feature frequencies are more likely to belong to the same image category. The dictionary of sift features is created through a training set and therefore is specific to the dataset It is created for. The authors paired this image description with a support vector (Cortes & Vapnik, 1995) machine in order to perform a categorization task but, since the image description created by PHOW is flexible, it can also be applied to the task of image similarity.

*HMAX***.** The HMAX model is inspired by simple and complex cells found in the early visual cortex, initially described by Hubel & Wiesel (1962, 1968). Simple cells respond to orientated edges and bar gratings, while complex cells receive input from several simple cells and so respond to complex stimuli. HMAX extends the concept behind simple and complex cells to create a model of visual processing which alternates between layers of simple (S) and complex (C) cells. The original HMAX model followed the structure of layers S1, C1, S2 and C2 (Riesenhuber & Poggio, 1999; Serre, Wolf, & Poggio, 2005). A diagram of the structure of the original HMAX model is seen in Figure 1.6. The cells in layer S1 convolve a set of Gabor filters (varying in phase, receptive field size and orientation) over the image to extract the initial features. Next, each cell in C1 pools over S1 cells of similar orientation and position, in the form of a max function. This creates position and size invariance of features in layer C1. Layer S2 performed a weighted summation over the cells in layer C2, this can

be thought of as combing features. The final layer C2 performs a max pooling operation on the cells of layer S2 which have extracted similar features but at different positions. By alternating between the S and C layers the model's output becomes invariant to small shifts in scale or position. The original model proved to be very popular in image processing and several additional extensions to the original model have been proposed. Serre et al., (2007) extended the model to a total of 9 layers (sticking to alternating S and C layers), and adding extra pathways by which information could bypass layers. This information bypass pathway was inspired by the visual cortex in which information from low level visual areas can bypass the intermediate visual areas and feed directly into higher areas (Nakamura, Gattass, Desimone, & Ungerleider, 1993). Support vector machines (Cortes & Vapnik, 1995) are usually paired with the HMAX description to perform the task of image categorization. HMAX offers a flexible descriptor that can also be easily applied to the task of image similarity.

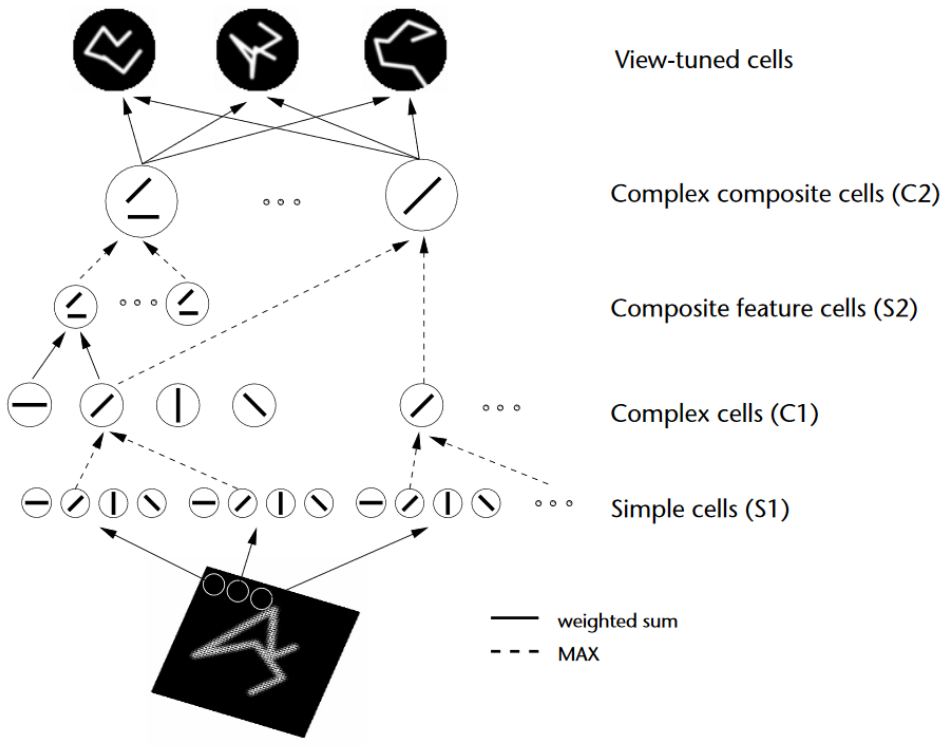


Figure 1.6. A visual diagram of the stages of the HMAX model. Solid lines between layers represented a weighted sum (Layer S1's convolution can also be thought of as a weighted sum). Dotted lines between layers represent a max operation. Figure taken from Riesenhuber & Poggio, (1999).

Combination models. So far, each image descriptor has been outlined separately to one another. While it is useful to think of them separately, as they calculate their image description in different ways, image descriptors can also be combined in order to produce more efficient image descriptions. Recent research, instead of creating novel image descriptors, has discovered the power of combining different image descriptors together to achieve superior performance. These combination image descriptors are usually paired with

novel decision making mechanisms such as advanced support vector machines. There are some notable models from this category of computational models, such as one of the recent winners (Everingham et al., 2015) of the pascal visual object challenge (Everingham et al., 2010). The model was run under the name NUS_SCM (Q. Chen et al., 2015) used histogram of gradients (Dalal & Triggs, 2005), local binary pattern (Ojala et al., 1996) and a color invariant model of SIFT (van de Sande, Gevers, & Snoek, 2010) to create its image description. This was paired with a context support vector machine decision process that was sub-class aware. Other notable models to be mentioned briefly in this thesis are Semantic scene attributes model (Patterson & Hays, 2012) and Never Ending Image learner (X. L. Chen, Shrivastava, & Gupta, 2013).

Convolutional Neural Networks. Convolutional neural networks were inspired by the hierarchical structure of human and primate vision. They are a family of hierarchical models with several stages of feature extraction formed by convolutional complemented by operations such as max pooling and output normalization. Convolutional networks consist of many layers and as these layers increase so too does the complexity of features that they extract. A network trained on faces may have, by its second layer, features resembling eyes and noses and, by its fourth layer, whole face representations might be seen. The main advantage of neural networks is that they typically learn from experience which features are informative or not. This is in contrast to other image descriptors that for which the informative features to be detected are predetermined by the creator of the model. Convolutional models showed

early success (Fukushima, 1980; Lecun, Bottou, Bengio, & Haffner, 1998) and have been applied to other tasks such as auditory or text analysis.

Deep Supervised Convolutional Neural Net. Deep supervised convolutional neural nets belong to the family of convolutional neural nets. These networks have two key properties that make them stand out from traditional convolutional networks; the fact that they consist of many layers (deep) and that the employ supervised learning. Up until recently convolutional neural nets have been largely limited in size due to computing power. In recent years, with advancements in GPU power, as well as the algorithms that implement them upon multiple GPUs, convolutional networks have been able to reach unprecedented sizes with upwards of 19 layers (Simonyan & Zisserman, 2014). Convolutional networks with more than around eight layers have been labeled as 'deep' to highlight their size. Convolutional neural nets have had a number of different learning algorithms proposed for them over the years. Supervised learning algorithms have been shown to perform particularly well (LeCun, Bengio, & Hinton, 2015). Supervised learning is used when a set of training images labeled with the correct category terms are passed through the convolutional neural net, this allows learning to occur which allows the convolutional neural net to optimize itself to the features which best explain the differences in object categories. This is opposed to unsupervised learning which allows the convolutional network to decide which features are relevant without any categorical knowledge. Deep convolutional supervised networks have dominated image competitions far surpassing other methods of image

classification (Russakovsky et al., 2015). Deep supervised convolutional networks were originally described in the ImageNet competition (Krizhevsky et al., 2012). This neural network contained 60 million parameters with 650,000 neurons and was comprised of eight layers; 5 convolutional layers, followed by 3 fully connected layers. The neural network was trained on 1.2 million high-resolution images from the ImageNet LSVRC-2010 contest (Russakovsky et al., 2015). The original model used the output of the eighth layer to make a decision of an image's category. It did this by applying a 1,000-way soft max on the output of layer 7. Layer 8 is, therefore, thought of as the decision process, with layer 7 as the primary *image description*. At the time of creation this model set a new bench mark for convolutional models' performance.

1.4. Brief History of Computational Model Competitions

Computational model competitions have originated due to a need for standardized testing in the performance of different computational models. Originally when a researcher was determining a new model's effectiveness they would run the model on a number of tests. Unfortunately, different research labs were using different data sets to create these tests, and thus it was difficult to compare models across research groups. Therefore, a need grew for standardized tests and data sets in which many researchers could publish the results of their best-performing models and performances across models could be compared. The result has been annual computer model competitions with published image data sets (Everingham et al., 2010; Fei-Fei,

Fergus, & Perona, 2007; Russakovsky et al., 2015; Torralba, Fergus, & Freeman, 2008; Xiao et al., 2010). Originally these image sets contained just one type of category image such as scenes, places or objects, but as the image sets have become more advanced all three have been incorporated. Computational models in these competitions are examined on a range of tasks, such as object and scene categorization as well as object segmentation and detection. These computational model competitions provide a history of how computational models have evolved in their design and performance over the years.

One of the earliest standarsised data sets computational models were tested on was the Caltech 101 data set (Fei-Fei et al., 2007). This data set contained 101 image categories. Each image category contained 40-800 images, with the average of around 50 images per category. Computational models were commonly trained on around 15-30 images per class. The competition ran from 2003 until 2006. A number of different computational models were classified as the top performers, with very little difference in the top published results at the end of 2006, the top two having very close correct categorisation rates. The top performer at the end of 2006 employed a geometric blur (Berg & Malik, 2001) image descriptor, paired with a combination of support vector machine and nearest neighbour algorithm (Zhang et al., 2006) as its decision process. This computational model performed at an accuracy rate of 66% with 30 training examples. The second highest performance was by the PHOW descriptor, scoring 64% accuracy with 30 training images (Lazebnik et al., 2006). The majority of entries to Caltech 101 focused on creating novel image

descriptors which had high performances and nearly all were paired with support vector machines.

The Pascal visual object classes challenge (Everingham et al., 2010) contained 20 object classes, with 22,591 images in total. The competition ran from 2005-2012 and was charactersised by models that combined multiple image descriptors in a single decision process. For example, the best performing model for object categorisation (Everingham et al., 2015), NUS_SCM (Q. Chen et al., 2015), used a number of different image discriptors in order to generate its image discription; histogram of gradients (Dalal & Triggs, 2005), local binary pattern (Ojala et al., 1996) and a color invariant model of SIFT (van de Sande et al., 2010). They attributed their success to a context support vector machine which created sub-class aware object detection and classification.

More recent competitions have used much larger scale image databases. The ImageNet large scale visual recognition challenge (ImageNet) (Russakovsky et al., 2015) contains 1.2 million images spread across 1000 image categories; each image category contains 700-1300 images. The competition has run from 2010 – present and, during this time, the most successful computational models have changed quite dramatically. In the first two years the winning computational models (Lin et al., 2011; van de Sande, Uijlings, Gevers, & Smeulders, 2011) consisted of variations of the SIFT descriptor (Lowe, 2004; Perronnin & Dance, 2007) mixed with other image descriptors (Ahonen, Hadid, & Pietikainen, 2006) combined with variations of support vector machines. The

design of these models were popular in the earlier image competitions of Caltech 101 and Pascal visual object class challenge. In 2012 a deep convolutional neural net took first prize by a considerable margin (Krizhevsky et al., 2012) in the field of image categorization, and in 2013 almost every entry used large-scale convolutional neural networks. By 2014 deep convolutional networks won on all three of the tasks the competition offered; image classification, single-object localization, and object detection. Innovations in convolutional neural nets came largely from the availability of such a large training set offered by the ImageNet competition, along with the design of efficient algorithmic implementation and massive computing resources offered by new GPUs.

Image competitions compare computational models to a set of idealised responses pre-determined by the researcher. This means that computational models are being compared to an idealised set of human responses; human behavior free of any restriction. Although these competitions do not compare computational models to real human behavior they are still able to provide a general assessment. If a model has near human performance then it would suggest that this model would be intresting to examine further. Alternatively, if a model does not have near human performance then it is unlikely to be performing calculations in a similar manner to human observers.

1.5. Comparing computational models to human behaviour

Studies which compare computational models of vision to human observers' behavior come in two different flavors. Some studies compare a single computational model to human behavior, providing an in-depth analysis of a single model, akin to a "case study". We will consider some examples of such studies first. Other studies, which we will move on to second, compare multiple models with behavior to try and determine which performs best.

Serre, Oliva, & Poggio (2007) performed a case study for their HMAX model. HMAX was designed on known principles of the human visual cortex and had not been optimized to match human *behavioral* characteristics. The researchers were interested in examining if it was able to predict human behavior in terms of error rates and reaction times. They showed that the HMAX model was able to detect whether images contained objects such as a body in the distance, a body nearby etc., and that the profile of performance for each category was similar to that of human observers in a speeded perceptual task. This study was the first to demonstrate that a state of the art computational model of vision could predict human error rates in an image categorization task.

Oliva & Torralba (2001) developed the GIST image descriptor and demonstrated that human observers' ratings of scene properties such as naturalness, openness and roughness could be captured by the low-level properties of the image described by their GIST descriptor. Additionally, they

showed that GIST was able to retrieve images that human observers would rate as similar on these property ratings. This would suggest that GIST is able to produce image descriptions which organize images in a similar way to human observers.

Several studies have investigated the relationship between scene recognition and object recognition. This usually takes the form of object or scene recognition tasks when objects and scenes are either consistent or inconsistent (Davenport & Potter, 2004; Henderson & Hollingworth, 1999; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Palmer, 1975). These studies show that objects are more easily recognized in scenes they are likely to be found in. A proposed mechanism for this phenomenon of scene-object interaction is a dual-system account (Davenport, 2007; Davenport & Potter, 2004) in which scene recognition and object recognition interact and can have a facilitation or inhibitory effect. Mack & Palmeri (2010) proposed an alternative explanation to the dual-system interaction theory. Using the GIST image descriptor paired with a linear discriminant analysis they showed that images which had objects consistent with the scene were further from the decision boundary than images which had objects inconsistent to the scene. In a linear discriminant analysis, the further a point is from the decision boundary the easier it is to categories as belonging to that category. They went on to show that an image's distance from the decision boundary using the model was able to account for the behavioral advantage in humans in consistent vs inconsistent object-scene images.

A single scene or object can be categorised on a number of different dimensions. For example, a beach scene can be categorised as open or navigable (attribute level), it can be categorised as a beach (basic level) and also as outdoors (superordinate level). A number of studies have shown that categorising an image on these different levels reveals different reaction times; with attribute level having the slowest and superordinate having the fastest reaction times (Joubert et al., 2007; Kadar & Ben-Shahar, 2012; Loschky & Larson, 2010). The effect has been labeled the superordinate advantage and has been used to argue for a heirarchical approach to human image categorisation (Joubert et al., 2007; Kadar & Ben-Shahar, 2012; Loschky & Larson, 2010). Sofer, Crouzet, & Serre (2015) proposed an alternative explanation to this effect. Using GIST, paired with a linear discriminant analysis, they showed that the differences in reaction times to categorizing different images could be attributed to the distribution of images around a decision boundary. If images were closer to the boundary, this had the effect of making the images harder to categorize, and also resulted in slower reaction times. To examine this theory in greater detail, the researchers went on to select pools of images that were either close to the decision boundary or further away and demonstrated the effect could be reversed by artificially selecting the pools of images used in the categorization task. Sofer, Crouzet, & Serre (2015) is the first study listed so far to examine categorization rates on a per-image bases irrespective of image category.

When determining which computational models best fit human behavioral data, multiple computational models need to be compared to the same behavioral data set. These studies focus on comparing computational models to invariant object recognition; the ability to recognize the same object from different angles.

Ghodrati, et. al. (2014) measured the performance of human observers and a number of computational models in an object invariant recognition task. Stimuli were generated using a program that used 3D models. These allowed objects to be varied in rotation as well as by the background they were placed. In this study six different models were compared to human behavior. The computational models assessed were a V1-like model, HMAX (Serre et al., 2007), GMAX (Ghodrati, Khaligh-Razavi, Ebrahimpour, Rajaei, & Pooyan, 2012), Stable (Rajaei, Khaligh-Razavi, Ghodrati, Ebrahimpour, & Abadi, 2012), SLF (J. Mutch & Lowe, 2008) and a deep convolutional neural network (Krizhevsky et al., 2012). All of these models, with the exception of the deep convolutional neural network, are variants of the HMAX model (Serre et al., 2005). As a control, they entered the stimulus' raw pixel values into a support vector machine. They compared human behavior to the different computational models on the overall percentage correct and the percent correct for each image using a modified version of representational similarity analysis (Nili et al., 2014) to fit the behavioral task. They found that under small image variations, such as small image rotations or background shifts, the models were able to perform nearly as well as human observers, in overall percent correct.

They also showed that at the individual image level observers and computer models behavior matched closely. As the size of the image variations increased human observers still performed remarkably well, but the computational models suffered greatly, in terms of overall percent correct. As would be expected, at the individual image level as image variations increased the computational models performed less similarly to the human observers. The results suggest that computational models perform similarly to humans in optimal conditions, but as the task of image categorization becomes harder differences in performance and behavior become apparent.

Kheradpisheh, et. al. (2016) extended the study by Ghodrati et al., (2014), by examining a number of deep convolutional neural nets. They examined a total of eight deep convolutional models and used the HMAX model (Serre et al., 2007) as a benchmark. All of the deep convolutional models are variants of Krizhevsky et al., (2012) deep convolutional network, with the exception of the very deep model (Simonyan & Zisserman, 2014) which consisted of 19 layers. All of the neural networks were trained on the ImageNet database (Russakovsky et al., 2015), with the exception of one model (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014) which included scene images extracted from search engines and the SUN database (Xiao et al., 2010). Similarly to Ghodrati et al., (2014) they evaluated each computational model on similarity to human behavior on the overall percentage correct as well as the percent correct for each image. The result found showed that some of the deep convolutional neural nets were able to reach human performance even at large object

rotations, on overall percent correct. Generally the deeper a convolutional neural net was the more human-like it behaved at the individual image level even at large object rotations. Surprisingly some of the deep convolutional neural nets had a profile of correct behaviour for each individual image which were indiscernible from human observers at high image variations. Overall they showed that deep supervised convolutional neural nets provided a good fit for human object invariant recognition at the individual image level.

Research investigating the similarity between computational models of vision and human behaviour is still a relatively new field and is thus missing a full scope of investigative studies. There are two main issues with this field. The first is that the majority of research has focused on assessing a single model's fit to behavioural data, with only a few studies examining multiple models fit to human behaviour. This has made making comparisons between different models difficult. Secondly, the majority of research in this area has compared computational models to human behaviour at the category level, rather than at the individual image level. By making the comparisons more specific a greater amount of detail about how a computational model fits human behaviour can be obtained. In order to remedy this investigative studies need to focus on comparing multiple computational models to a single behavioural data set which matches computational models' behaviour to human observers on a single image basis. The research, on the whole, demonstrates the story that computational models which base their image description on low level visual properties (e.g. GIST) are able to explain a significant proportion of

variance in observers' behaviour when paired with a linear decision bound. However, deep supervised convolutional models out perform these models and provide the closest account of any type of computational model at explaining observers' behaviour.

1.6. Comparing computational models to neural activity

1.6.1. Representational Similarity Analysis

Several fMRI studies have been conducted comparing *image descriptors* to human neural activity. The main method of comparing image descriptors to neural activation is through representational (dis-)similarity analysis (RDA/RSA) (Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014). Representational similarity analysis (RSA) is an alternative application of multivariate pattern analysis (Haxby et al., 2001) which allows patterns of neural activity in response to different stimuli to be compared to the structure of computational image descriptions.

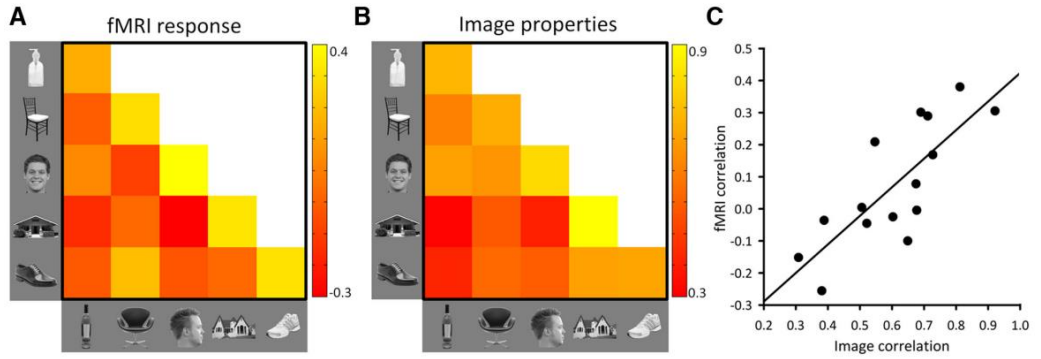


Figure 1.7. The example (Rice, Watson, Hartley, & Andrews, 2014) presented here illustrates the process of representational similarity analysis. Two similarity matrices are shown, one from neurological data, as seen in A, and one from a computational model, as seen in B. In C the cells of the similarity matrices are plotted against each other and a correlation run. The result of this correlation reveals the similarity of how the different image categories are structure between biological vision and the image descriptor being examined.

Performing an RSA is a 2-step process; we investigate the representation of a particular system (e.g. a particular image descriptor or an fMRI dataset) and then we investigate the similarity of those representations between systems.

In the first step, to investigate the representation *within* a system, a similarity matrix is created across images (e.g. Khaligh-Razavi & Kriegeskorte, (2014)) or across categories (e.g. Watson, Hartley, & Andrews, (2014)), depending on the resolution of the analysis. This is essentially a correlation matrix, whereby each cell is the correlation between the representations in the system of two images (for example, the correlation between a pair of GIST vectors, or the correlation between a pair of fMRI response vectors). The matrix tells us, essentially, which

images the system considers to be “similar” and this will differ according to the system (is the system sensitive to scene semantics for instance, or just to low-level properties).

The second step is to compare how similar these representations are *between* systems. This can be measured with a single correlation of the similarity matrices for any pair of systems.

An example of this approach is shown in Figure 1.7, in which a similarity matrix of 5 different image *categories* has been calculated for an image descriptor as well as for the neural activity observed in human participants. The similarity matrix for the computational image descriptor was obtaining by comparing the similarity of each image description in each image category against each other and the average similarity taken. The similarity matrix for the neural responses was calculated by comparing the similarity of each image’s fMRI response vector in each image category against each other and the average similarity taken. Once the similarity matrices had been constructed the values in the cells of each matrix are plotted against each other as each represents the same pairwise comparison between two images. A correlation is then run to establish if there is a relationship between the two matrices. Correlations with high variance explained show that the pattern in which images are represented are very similar. Correlations with a small or no variance explained would suggest that the structure of image descriptions are differing largely from each other.

In summary, RSA does not directly compare image descriptors to human neural activity, but instead estimates the fit of the pattern of responses of the computational image descriptor to the observed pattern of responses in the neurological activity.

1.6.2. Which image properties best explain neural activation in the visual cortex

The concept that the properties of a stimulus are key in determining evoked neural activation (O'Toole, Jiang, Abdi, & Haxby, 2005) has spurred research investigating which properties of an image best explain the observed neural activation. Two camps have developed, one stating that neural activation is largely in response to the low level visual properties of an image (Andrews, Watson, Rice, & Hartley, 2015). While, another camp states that along with the low level properties of an image, knowledge about categories is a requirement to explain neural activation (Khaligh-Razavi & Kriegeskorte, 2014).

Watson et al., in a series of studies investigated the GIST *image descriptor's* ability to predict neuronal activity using RSA. GIST is an image descriptor which solely uses the low level visual properties of an image in order to create its image description. These studies have shown that the structure of image descriptions produced by GIST correlates with structure of evoked neural activity from a variety of scene images (city, indoor, coast, forest, mountain) (Watson, Hymers, Hartley, & Andrews, 2016), supporting the current literature that scene perception is mediated though the low level properties of an image

(Oliva & Torralba, 2001). They have also shown that neural activity in response to images of objects (bottles, chairs, houses and shoes) as well as faces is predicted by GIST's image descriptions (Rice et al., 2014).

A number of studies have been conducted in which the low level image properties of an image have been varied to determine the extent that the evoked neural activation will vary based upon this. These have shown that the neurological activation varied with these low level visual property changes, even though the semantic category of the image had not changed (Coggan, Baker, & Andrews, 2016; Coggan, Liu, Baker, & Andrews, 2016; Watson, Young, & Andrews, 2016). Watson, Young, et al., (2016) later showed that alternative versions of the GIST descriptor (spectral and spatial) correlate with changes in neural activity when these low level properties were changed.

Previous research had always viewed that the clustering of neural activity to the same category of objects was evidence for a 'categorical/modular' response (Kanwisher, McDermott, & Chun, 1997; Kanwisher & Yovel, 2006). A number of studies have been produced questioning this assumption and examined if grouping based on low level visual properties could explain this clustering effect. The cells of a RSA matrix can be broken down into two groups, those classified as within-category (correlations of the same category of image, e.g. faces against faces) and between-category (correlations of images not of the same category, e.g. shoes against faces) (Haxby et al., 2001; Kriegeskorte & Kievit, 2013). Rice et al., (2014) demonstrated that if all the within-category

points from the RSA correlation were removed that a significant correlation still remained. This shows that the similarity of neural activity between two categories is predicted by the similarity of the low level visual properties of the images. For example, the similarity of a face is to a shoe in neural activity is predicted by the similarity of their low level visual properties, demonstrating that category knowledge may not be necessary for clustering to occur. In a second paper investigating this clustering effect Watson, Hartley, & Andrews, (2017) employed a cluster analysis to organize images based on their low level visual properties using the GIST image descriptor. These clusters did not correspond directly to semantic categories, yet the clusters of images showed 'grouping' of neural activity in observers. This would suggest that 'grouping' of neural activity of images of the same image category is due to shared low level image statistics, rather than the visual system being made up of modules dedicated to specific processing of categories.

All of this research has led to the hypothesis that, similar to low level visual areas, high level visual areas are organized by low level visual properties, albeit in a more complex manner (Andrews et al., 2015). However, some research has shown that this may not be the total story. There is a whole class of computational models that are classified as supervised models. These models still obtain their image description from low level visual properties, but modulate their response based on known categorical principles. These models often have greater levels of performance in general visual perception tasks than their unsupervised counterparts (image descriptions based on low level

visual properties alone) (Ghodrati et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Kheradpisheh et al., 2016; Krizhevsky et al., 2012).

A number of studies have examined the pattern of neurological activity in response to objects and scenes using a range of different computational models. These computational models vary from ones based purely on low level visual properties to those supervised by categorical knowledge. Aminoff et al., (2015) compared different computational models to the visual areas known to be responsible for scene perception; the Parahippocampal Place Area (PPA), Retrosplenial Cortex (RSC) and Occipital Place Area (OPA). The results found showed that the models Never Ending Image learner (NEIL) (X. L. Chen et al., 2013) and Semantic scene attributes (SUN) (Patterson & Hays, 2012) explained the most activation of any of the models examined in the PPA and the TOS, while GIST best accounted for activity in the RSC. NEIL and SUN are both computational models which create their image description based on learned categorical knowledge. The results of this study suggest that although GIST is able to predict activation to scene images, certain areas responsible for scene understanding are better explained by low level visual properties supervised by categorical knowledge.

A number of studies have examined which computational models best explain neural activation in inferior temporal (IT) cortex in response to images of objects. IT is considered to contain the final representation of objects used by

the visual cortex (Bell et al., 2009; Kiani et al., 2007; Kriegeskorte, Mur, Ruff, et al., 2008).

Yamins et al., (2014) compared a number of computational models to neural activation to objects; SIFT (Lowe, 2004), V1-like model, V2-like model, HMAX (Serre et al., 2007) and a four layer deep supervised convolutional model. They compared these computational models using RSA to neurological activation in area V4 and area IT. Area V4 is thought of as the precursor to area IT. The results showed that the deep supervised convolutional neural network described the neural activation better than any other model. Comparing the output of each layer of the deep convolutional model to area IT showed that with each layer of the deep convolutional model better explained the amount of variance in the neural activation. It was also shown that the penultimate layer image description of the deep supervised convolutional neural network was the most similar to the activation seen in area V4, mimicking biological findings. Cadieu et al., (2014) performed a follow up study, comparing a number of deep supervised convolutional networks ability to explain neural activation patterns in area IT to objects. They used the deep convolutional networks of Yamins et al., (2014), Zeiler & Fergus, (2014) and Krizhevsky et al., (2012), as well as the models previously mentioned in Yamins et al., (2014). They again showed that the deep convolutional models performance was superior to models that derive their image description from the low level visual properties of an image. These study suggests that deep supervised convolutional models provide a better fit to neural activation found in the

visual cortex over image descriptors which solely employ low level visual properties.

Khaligh-Razavi & Kriegeskorte (2014) compared a mammoth number of computational image descriptors to neural activity in area IT using RSA. A total of 37 computational image descriptors were compared in total, an almost exhaustive list of the state of the art image descriptors. Additionally, they used a bootstrap method of RSA so that they could estimate the noise celling; the maximum variance explained possible given the noise in the data. They found that computational models based on low-level visual properties, such as GIST, indeed did explain some of the variance in neural patterns of activation, but this was a long way off fully explaining the pattern of activity found in area IT. They showed that a deep supervised convolutional network (Krizhevsky et al., 2012), that was linearly reweighted to fit the categories being tested, fully explained the structure of neural activation found in area IT to object stimuli, given a noise ceiling. This study shows that image descriptors that base their image description on the low level properties of an image do explain some of the variance in patterns of neural activation found, but in order to fully explain the activation categorical knowledge needs to be employed.

Research comparing computational image descriptors to neurological activity is diverse in the different computational models which have been examined. The majority of these computational image descriptors have found to correlate in some way to with evoked neural activation. This is surprising as these image

descriptors have never been optimized to predict neural activity and yet the majority still predicted neural activity. This would suggest that, irrespective of their implementations, the majority of image descriptors organize objects in a similar manner to each other and also to biological vision. The fact that image descriptors based on low level visual properties readily correlated with neural activation can be taken as a sign that high level image representations are based on low level visual properties. Alternatively, this may not be the whole story as it has been shown that deep supervised convolutional neural nets are closest, out of all the models examined, to neurological patterns of activation. Deep supervised neural nets base their image description on low level visual properties, but then modulate their response with respect to categorical knowledge. Neural nets are based on human biology and therefore could provide the closest approximation to a computational model of human visual perception. Together from this body of research it would suggest that high level scene and object representations in biological vision are constructed from low-level properties, but are then adjusted to fit categorical knowledge. Interestingly this body of research finds little difference between the structural representation of scenes and objects; categorically supervised over low level property image descriptor models best fit neural representations. This is contrary to previous literature which has viewed them as using entirely different mechanisms in order to create their image property representations (Barrow & Tenenbaum, 1978; Biederman, 1987; Marr, 1982; Potter, 1975).

Applying computer vision models to neural data may allow us to better understand how scene and object information is encoded in neural systems.

1.7. Overview of thesis

In this thesis, we investigate the similarity of different computational models to human observers' behavior. Specifically, the aim is to investigate not only the *image descriptors* in the models, but the contributions of several other components of the model, such as the contribution of the *decision process* and the *image set*.

Chapter 2 outlines and explains the various methods used in this thesis; the core components used to create the computational models, as well as the methods of comparison between observers' behavioral data and the computational models.

Chapter 3 investigates the similarity of different computational *image descriptors* to their counterpart employed by human observers. This is done through comparing the behavior of computational image descriptors to humans on an image recognition task.

Chapter 4 investigates the similarity of different *decision processes* to human observers in an image categorization task. An image categorization task was chosen as it provides a task with which the decision process in human observers is hotly debated (Ashby & Maddox, 2005, 2011).

Chapter 5 investigates the effect of altering *image set* statistics of humans and computational models. This chapter aims to see if human observers can be made to respond closer to computational models of vision by over training them on the image set used by the computational model.

Chapter 6 investigates the mechanism by which observers are producing image descriptions when image duration is small.

Chapter 7 overviews the key findings in the thesis. Advantages of the methods used within this thesis are discussed as well as the direction future research would benefit.

Chapter 2 - General Methods

2.1. Introduction

As outlined in Chapter 1 - Model of Visual Processing, computational models can be comprised of three main components; image descriptor, decision processes and the image set. In the studies presented in this thesis multiple computational models comprised of different variants of these components have been constructed. This section provides a list of the components used in the computational models as well as materials used in the experiments (e.g. Image set).

2.2. Image Descriptors

Four different *image descriptors* were examined; GIST (Oliva & Torralba, 2001), PHOW (Lazebnik et al., 2006), HMAX (Serre et al., 2007) and a deep supervised convolutional neural network (Krizhevsky et al., 2012). How each of these formed its image description is summarized here.

GIST derives its image description by dividing the image into a 4x4 grid, giving 16 non-overlapping windows. Oriented Gabor filters in 8 orientations and 4 different spatial scales convolve with each of these 16 windows. The mean filter response intensity in each window is then measured. This generates a vector of 512 (32 x 16) values. This results in an output that represents the image in terms of spatial frequencies and orientations present at different positions across the image. The code used to calculate the GIST image

description is freely available at http://people.csail.mit.edu/torralba/code/spatialenvelope/ (Oliva & Torralba, 2001).

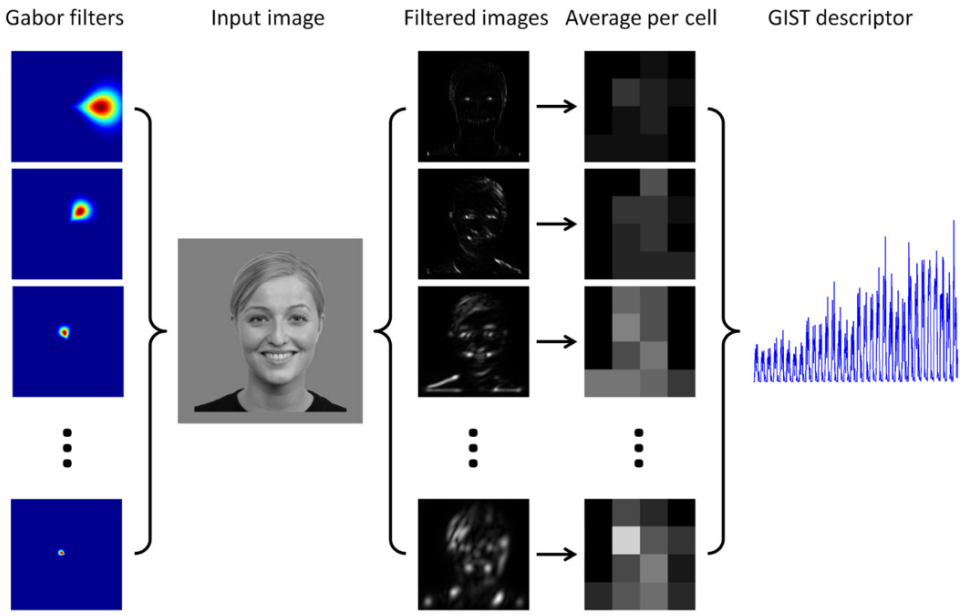


Figure 2.1. An illustration of the GIST descriptor on an example image. Different Gabor filters are convolved with the input image generating the filtered images. These filtered images are then split into a 4 x 4 grid and the mean intensities taken in each window to form the GIST description. Taken from Rice et al., (2014).

The PHOW image descriptor represents an image based on the number of SIFT features from a learned dictionary found in an image. As PHOW extends the dense SIFT image descriptor, for the use of image classification, the dense SIFT image descriptor will first be described and then the extension added by PHOW shall be explained.

The dense SIFT descriptor is a variant of the original SIFT descriptor. In the original SIFT descriptor feature points were located and then described. In comparison, the dense SIFT descriptor uniformly samples the image and uses these samples to create the SIFT features. The utility of this idea can be seen in the example of scene recognition, in which information about the whole image is useful rather than just information about specific points. This sampling of the image is done by splitting the image into 16 by 16 pixel patches with a spacing of 8 pixels to create overlap between patches. Each patch of 16 by 16 neighborhood of pixels creates its own SIFT feature. This is done by dividing the patch down further into 4 by 4 blocks, and an eight bin histogram of the weighted orientation gradient is calculated for each block. The process by which a histogram of weighted orientation gradients is explained in Figure 2.2. This produces a 128-dimensional vector using these concatenated histogram of each block as the SIFT feature for that patch. The dense SIFT description is then the list of the SIFT features uniformly sampled through the image.

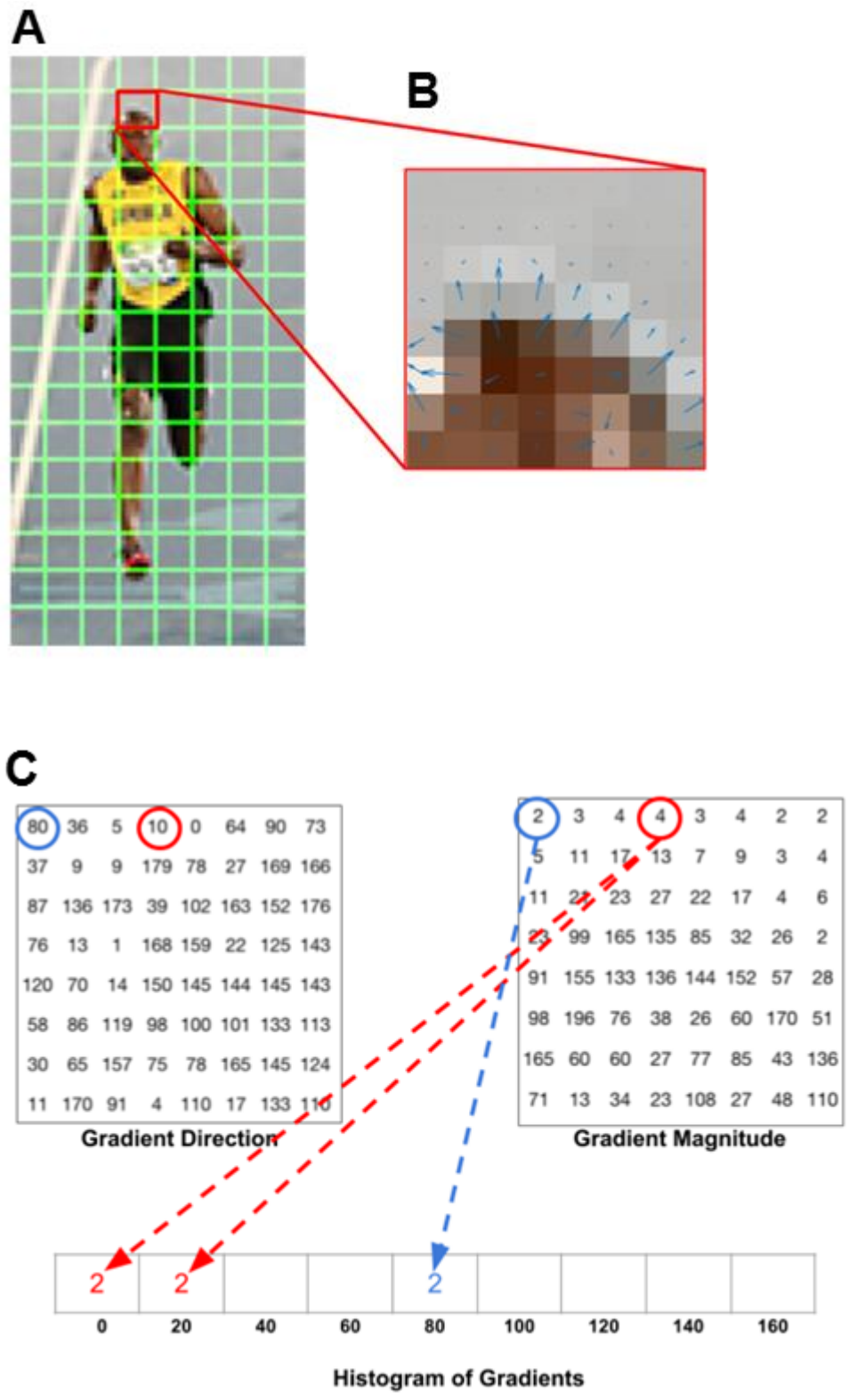


Figure 2.2. An example of how a weighted gradient histogram is constructed. A) displays an image of a runner split into non-overlapping 8 by 8 pixel windows (a simplified version of a dense SIFT 16

by 16 overlapping windows). B) demonstrates a close-up of one window. Blue arrows display the gradient orientation scaled by the magnitude. C) shows two matrices, the left most matrix represents the gradient direction at each pixel and the right most matrix represents the gradient magnitude. Underneath is the histogram of gradients which is constructed from these two matrices. The bins of the histogram refer to the gradient direction, while the values which get added to the bins are the gradient magnitudes. The example drawn in red demonstrates a typical example, while the example drawn in blue demonstrates how the pixel is calculated if the pixel falls between bins. Images obtained from Mallick, (2016).

PHOW initially trains a dictionary of the different SIFT features found in the image set. This is done by calculating dense SIFT descriptors on a training sample of images. Here 30 images were randomly chosen for each image category from the mask image set. The resultant SIFT features were then quantized using k-means clustering to a dictionary size of 200. For each image being described by PHOW a spatial pyramid of three levels is then created and the histogram of the dictionary SIFT features was calculated for each bin. This is illustrated in Figure 2.3. The concatenated version of the histograms was used as the PHOW representation of the image. The code is available at http://slazebni.cs.illinois.edu/research/SpatialPyramid.zip (Lazebnik et al., 2006).

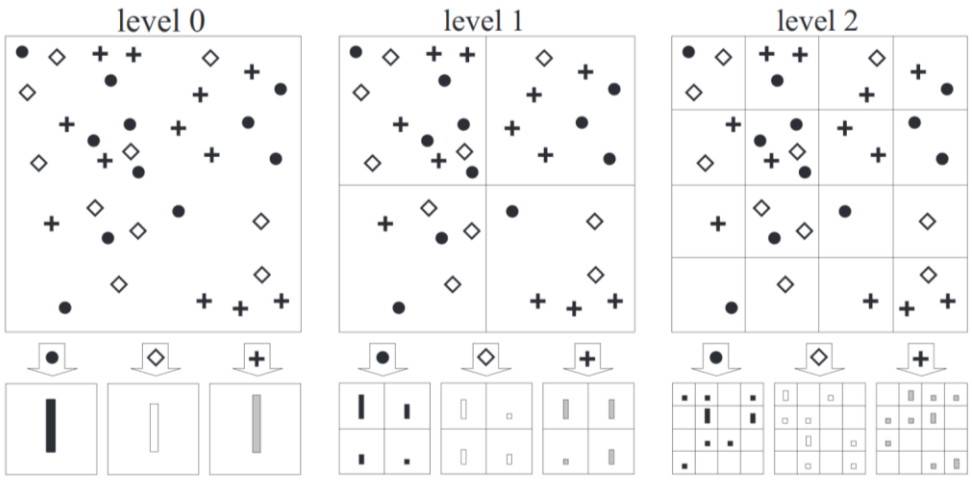


Figure 2.3. A visual depiction of spatial pyramiding of an image. In the above example a dictionary size of three SIFT features is shown. Level 0 shows the location of each feature in the image and below the histogram bins for each feature. Level 1 spits the image into 4 windows and calculates the histogram bins for each SIFT feature in each window. Level 2 splits the image into 16 windows and calculates the histogram bin for each feature. The concatenated version of the histogram for each spatial level acts as the PHOW image description. Image taken from Lazebnik et al., (2006).

The HMAX model used here is the standard model from Serre et al., (2007). The model is inspired by the work of Hubel & Wiesel (1962, 1968) on simple and complex cells. The model's architecture will first be described and then a detailed analysis of the processes occurring within the layers will be described. The model here possesses two pathways (a main pathway and a bypass pathway), both consisting of alternating simple (S) and complex (C) layers. The main pathway starts from layer S1 and ends on layer S4 (S1, C1, S2, C2, S3, C3, S4). The second pathway implements a bypass pathway, in which information from lower layers can influence higher layers by introducing an additional S2b, C2b layer which feeds of layer C1 and feeds directly to S4. The bypass pathway

is designed to mimic biological vision, in which information from low level visual areas can bypass intermediate areas and feed directly into higher visual areas (Nakamura et al., 1993). Figure 2.4. demonstrates the architecture of the model as well as the area it corresponds most closely with in the visual cortex. Figure 2.5. shows a visual representation of the processes which are occurring in the first two S layers and the first C layer. In general S layers perform a summation operation on their inputs, while C layers perform a max operation on their inputs. By alternating between S and C layers the model's output becomes invariant to shifts in scale or position. The S1 layer's cells convolve a bank of Gabor filters across the image (can be thought of as a summation of pixel intensities using the weights of a Gabor filter). This filter bank consists of 96 different filters (two different phases, four orientations and 17 receptive field sizes). Each of the cells in the C1 layer receives the output of a group of S1 cells with the same preferred orientation, but at slightly different positions and sizes. The pooling over cells in S1 causes the cells in C1 to be invariant to small changes in size and position. Layer S2 pools the activity of a local neighborhood of C1 cells, as a result the complexity and size of their preferred stimuli is increased. Layer C2 pools over Layer S2 units that are tuned to prefer the same stimuli, but at different locations and scales. Layers S3 and C3 perform the same process as S2 and C2 only iterated one more time to increase in feature size, invariance and complexity. Layers S2b and C2b corresponding to the bypass layers which perform the same operations as their S2 and C2 counterparts, yet pool two to three times as many cells from the layers before

them. This causes them to represent more elaborate features, but with less tolerant to image changes. The final layer S4 sums from all C layers to form complex whole image representations. The HMAX model described here performs unsupervised learning to decide the weights used in S layer summation (from layers S2 and onwards). This is done by passing training examples through the model, the weights of S cells are then altered according to the activity they perceive in their receptive field. This has the effect that patterns of activity which regularly occur within the model become enhanced. This learning adapts the model to the image statistics of the natural environment and its units become tuned to common image features. Here the model was trained on 30 images from each image category from the mask image training set. The code used is freely available at http://cbcl.mit.edu/software-datasets/pnas07/index.html (Serre et al., 2007).

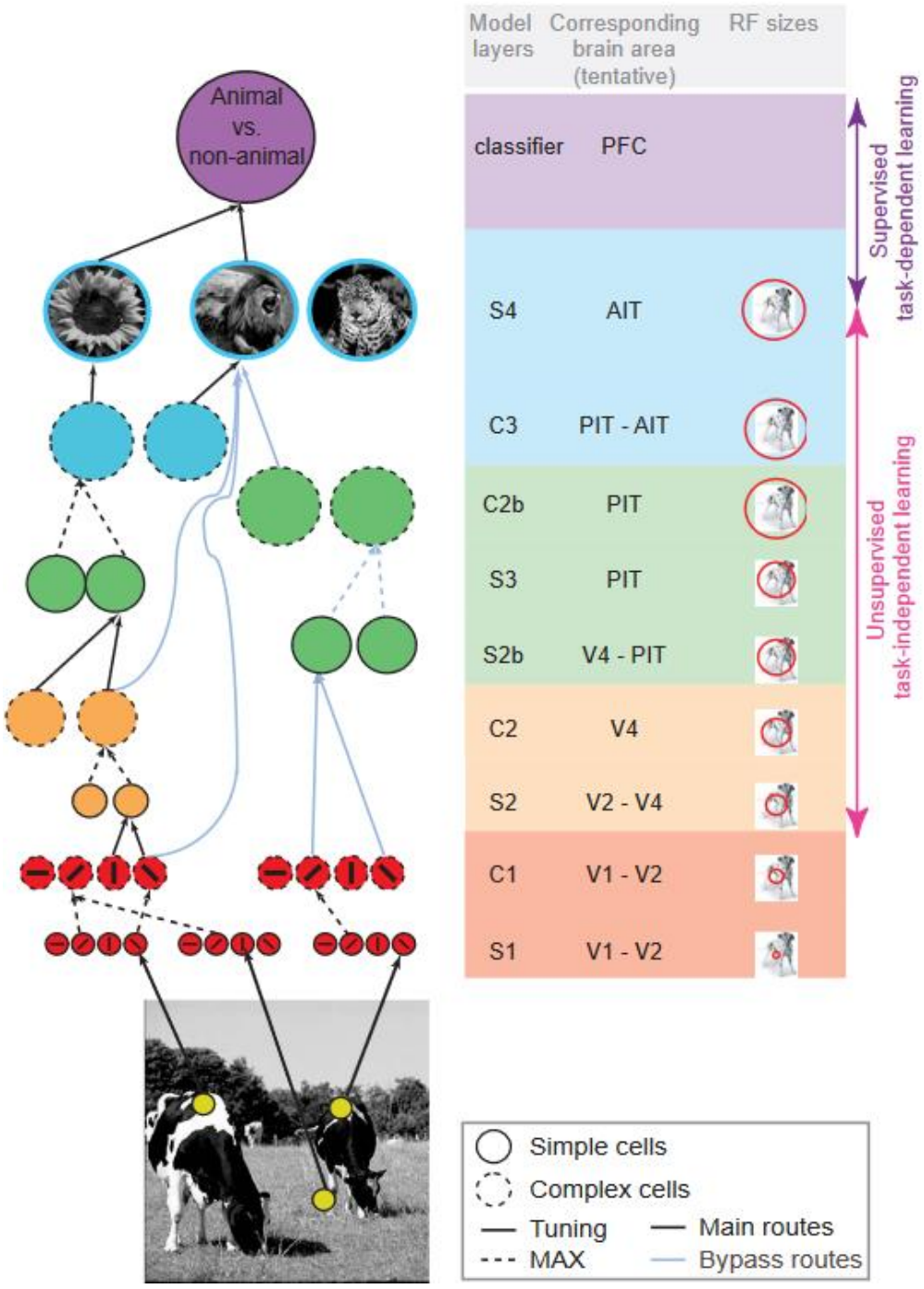


Figure 2.4. A diagram of the architecture of the HMAX model used. Image modified from Serre et al., (2007).

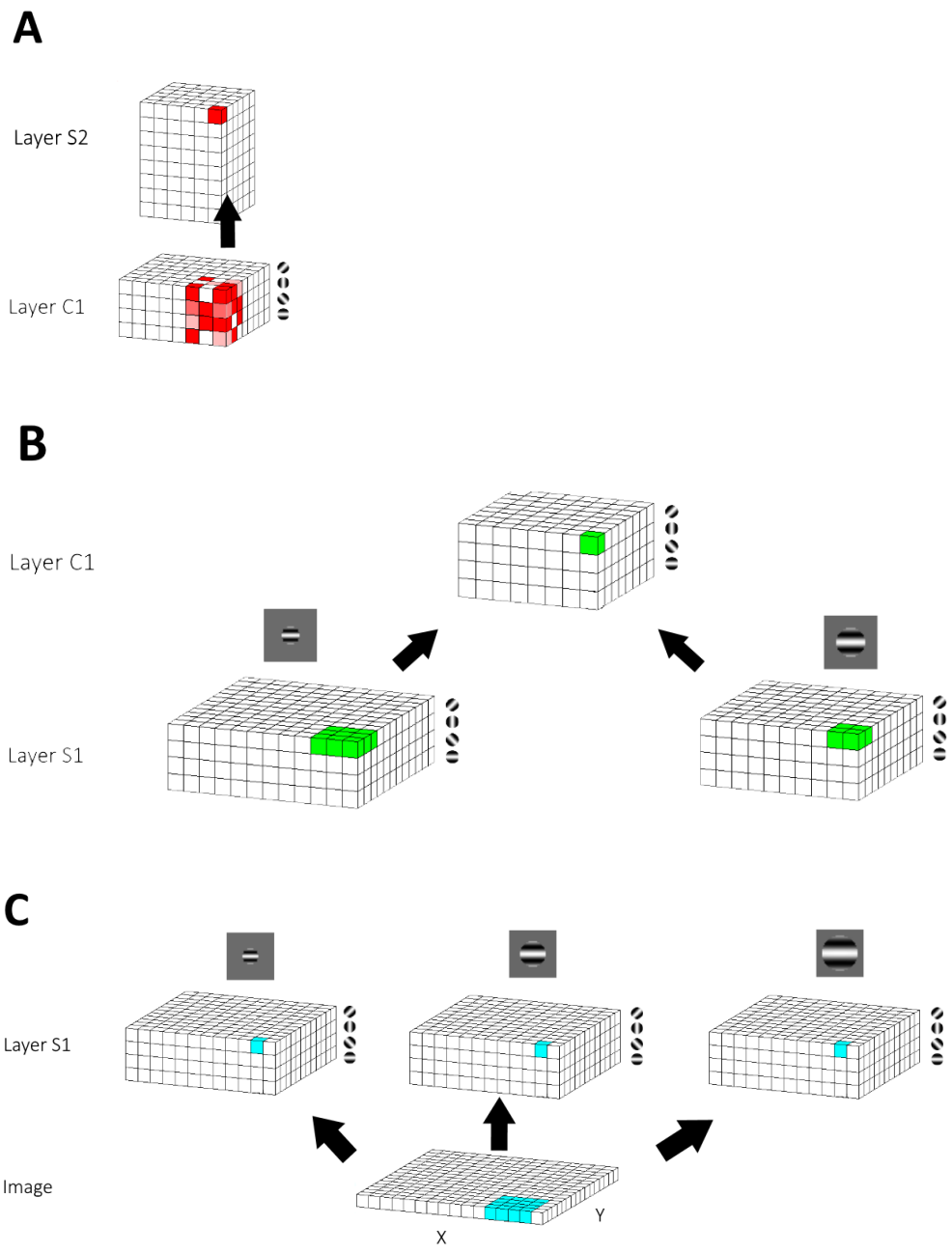


Figure 2.5. A visual depiction of the processes occurring in the first two S layers and the first C layer in the HMAX model used. C) shows an image being processed by layer S1. The area of the image being examined is highlighted in blue. Three banks of cells are shown and above them with different receptive field sizes being convolved to the image. Each layer of the bank of cells shown in layer S1 corresponds to an orientation of the Gabor filter convolved to the image at different X and Y locations. B) shows the processes occurring from Layer S1 to Layer C1. A max operation is performed

on cells of S1 with similar spatial location and same preferred orientation, highlighted in green. A) displays the process which occurs between layer C1 and S2. A weighted summation, seen in red highlight, of C1 cells which is across orientation and similar spatial location is occurring. The weights used for the summation are learned through unsupervised training, in which the weights mimic the activity seen in the receptive fields of the cells during training. Modified from Mutch, (2010; 2008).

The deep supervised convolutional network examined here is the winner of the ImageNet 2012 competition (Krizhevsky et al., 2012). The neural network is comprised of two kinds of layers, convolutional layers and fully connected layers. A general explanation of the processes that happen in each of these two types of layers will be described, as well as the basics behind supervised learning in convolutional neural nets. After this explanation the basic architecture of the supervised convolutional network shall be described.

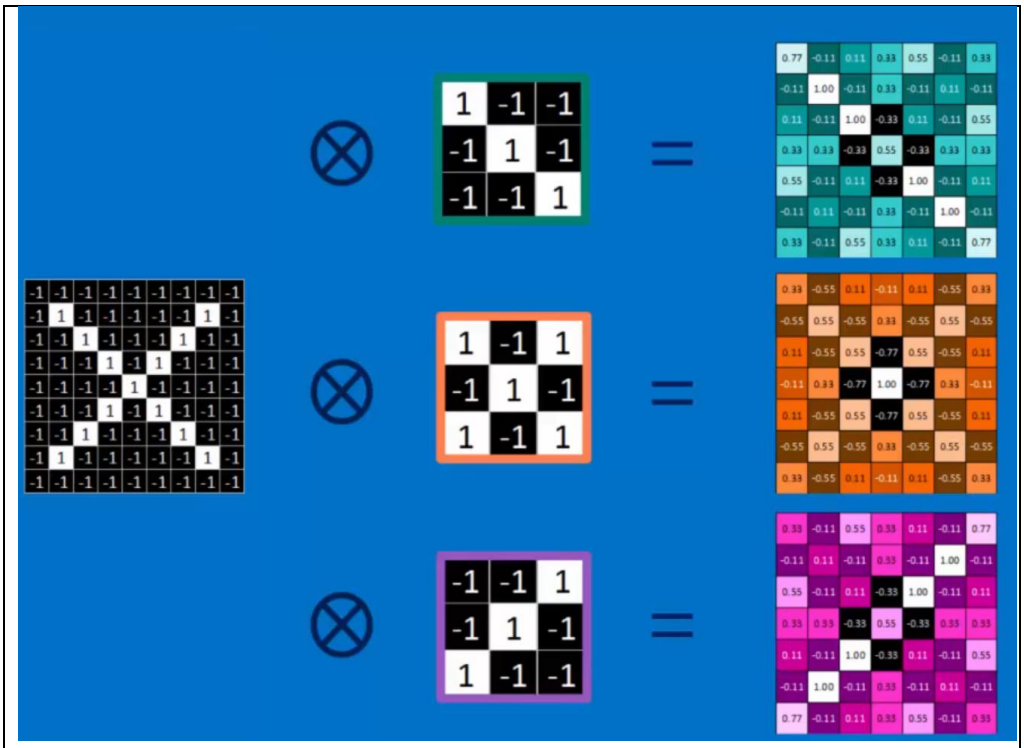


Figure 2.6. Convolution of a set of three filters onto an image.

The first and third filter show an output which demonstrates that the filter best fits along the diagonal from left to right and right to left respectively. The middle filter shows an output which demonstrates that it best fits the centre of the image. Taken from Rohrer, (2016).

Convolutional layers, as the name suggests, perform convolution of a bank of filters on their inputs creating a stack of filtered images; one filtered image for each filter. The output images show where the filters best match the image. This can be seen in Figure 2.6 which displays an input image of an 'X' being convolved with 3 different filters. There are two additional operations which can also occur in a convolutional layer; the rectified linear unit (ReLU) and max pooling.

ReLU is a function that turns any negative values in a stack of images into a 0. An example of this can be seen below in Figure 2.7. The ReLU function adds a nonlinearity to the system which lets it represent more complex features than convolution would alone. The ReLU function is used over other nonlinear functions, such as tanh and sigmoid, due to it allowing the network to train faster.

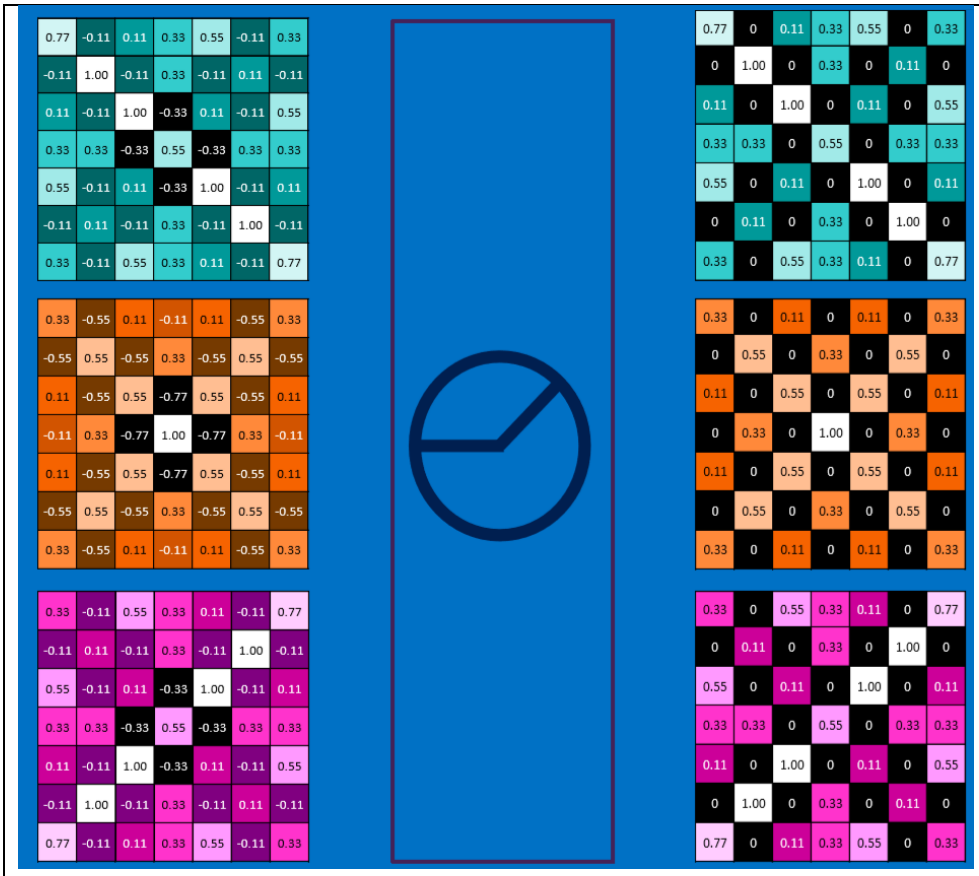


Figure 2.7. A demonstration of a ReLU function which turns any negative value in a stack of images to 0. Taken from Rohrer, (2016).

Max pooling is a process by which the image stack is shrunk. A window is passed over the image and the maximum value is taken. Shrinking through max pooling

is convenient to reduce the size of the stack. Max pooling creates outputs which care less about where the feature was located in the image and so invariance to feature position is created. A visual demonstration of max pooling is seen in Figure 2.8.

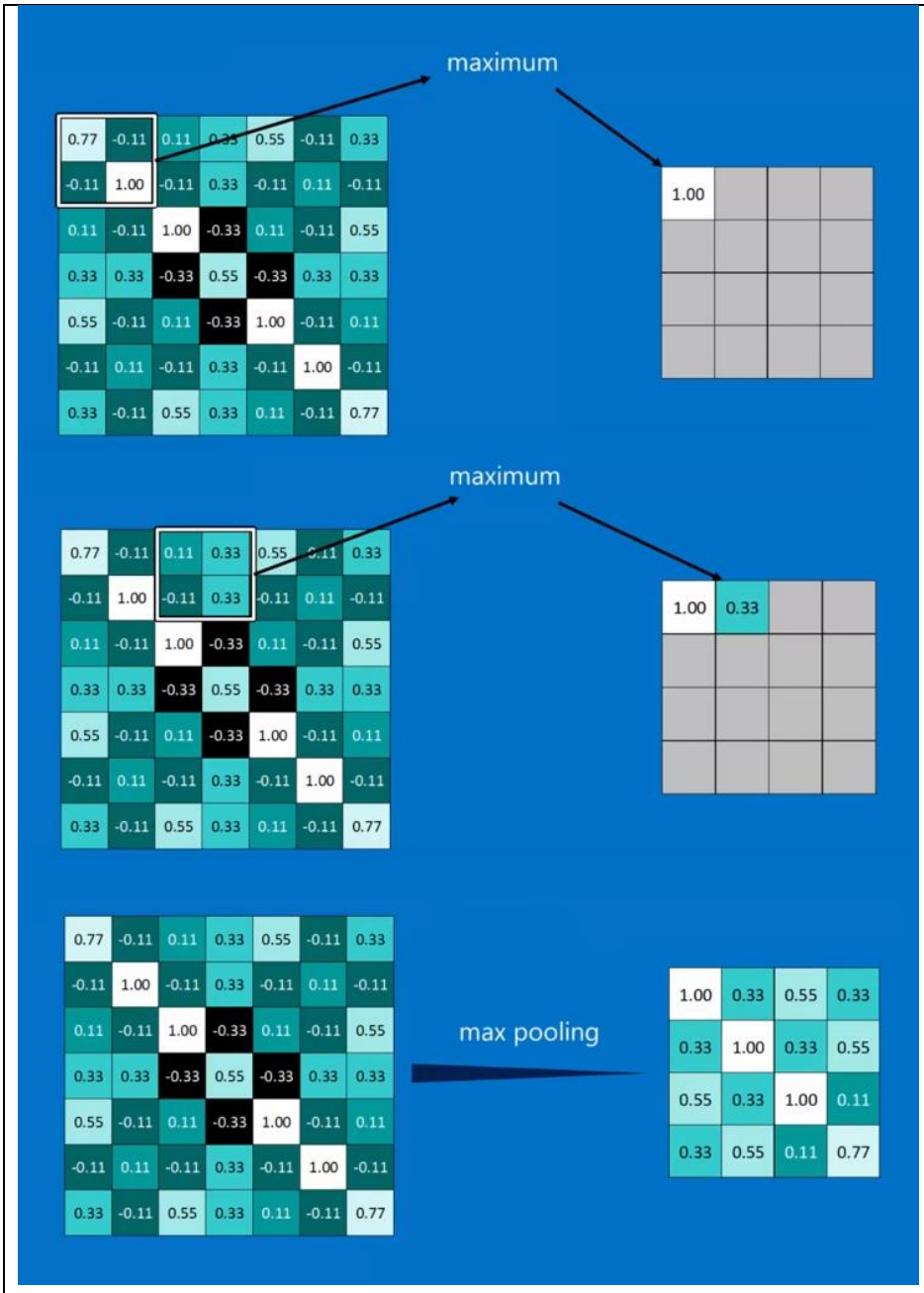


Figure 2.8. An example of max pooling for the first feature. A window size of 2 by 2 pixels is moved across the image using a stride of 2 pixels. Taken from Rohrer, (2016).

The second kind of layer is the fully connected layer. These layers possess a number of neurons which perform a weighted sum of their inputs. The weighing by which each neuron sums its inputs are different and so complex

combinations of features are able to be represented very quickly by stacking multiple fully connected layers. A diagram of a fully connected layer can be seen in Figure 2.9. The output of a fully connected layer normally has a non-linear operation added to it, for example the ReLU. Adding this nonlinearity to the output of these layers has the same purpose as in the convolutional layers, to an increase in the complexity of functions the network can produce.

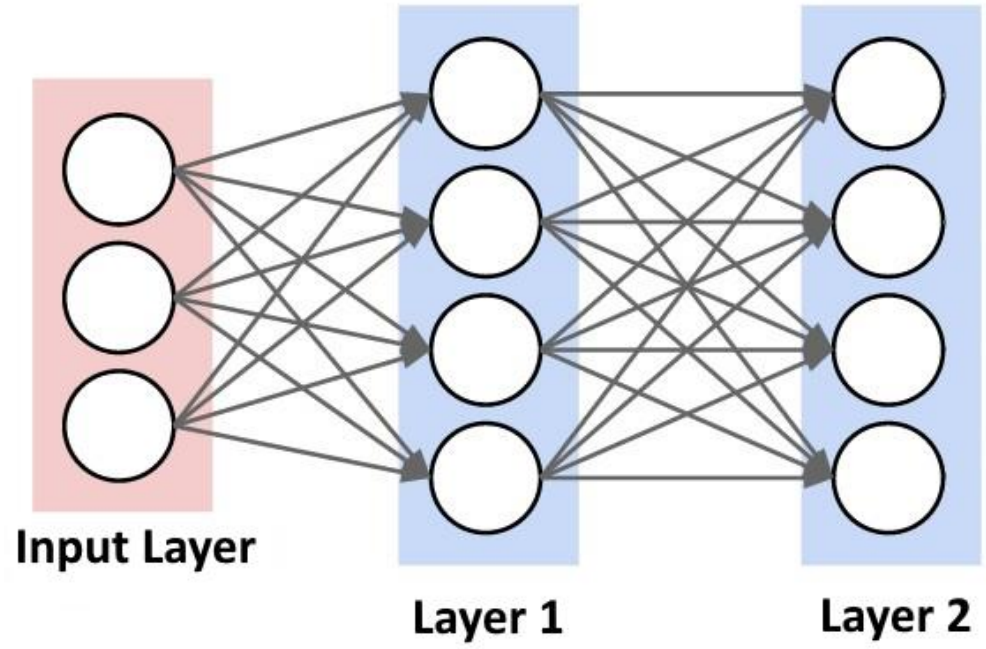


Figure 2.9. A diagram of two fully connected layers stacked on top of each other. In layer 1 each neuron performs a weighted sum of its 3 inputs. In layer 2 each neuron performs a weighted sum on each of its 4 inputs. Taken from Rohrer, (2016).

Convolutional neural networks have an incredible ability to learn which features in a set of images are relevant to the task of image categorization. This learning is done through trying to minimize the output of the cost function of

the neural network. The cost function calculates how well a neural network performed on a set of images, in essence the difference between the output the model gave and the output that we wanted it to have. If the cost function returns a value that is small or 0 then the network is performing extremely well or optimally. The filters in the convolutional layers and the weights in the fully connected layers are all a set as variables which can be altered. It is possible to alter these variables so that the output of the cost function changes. Back propagation with gradient descent is the process by which the filters and the weights in the neural network are altered to reduce the output of the cost function and make the network learn about the training set of images. Gradient descent for a single weight in an example convolutional network is shown in Figure 2.10. Learning in a convolutional neural network is a slow process that requires many training iterations and examples. Many training examples are needed to stop the neural network learning rules about images which do not generalize from a training set of images to the test set. While, many iterations of training are needed as the weights and features are only altered a small amount at a time so they do not overshoot or miss their local minimum contribution to the cost function.

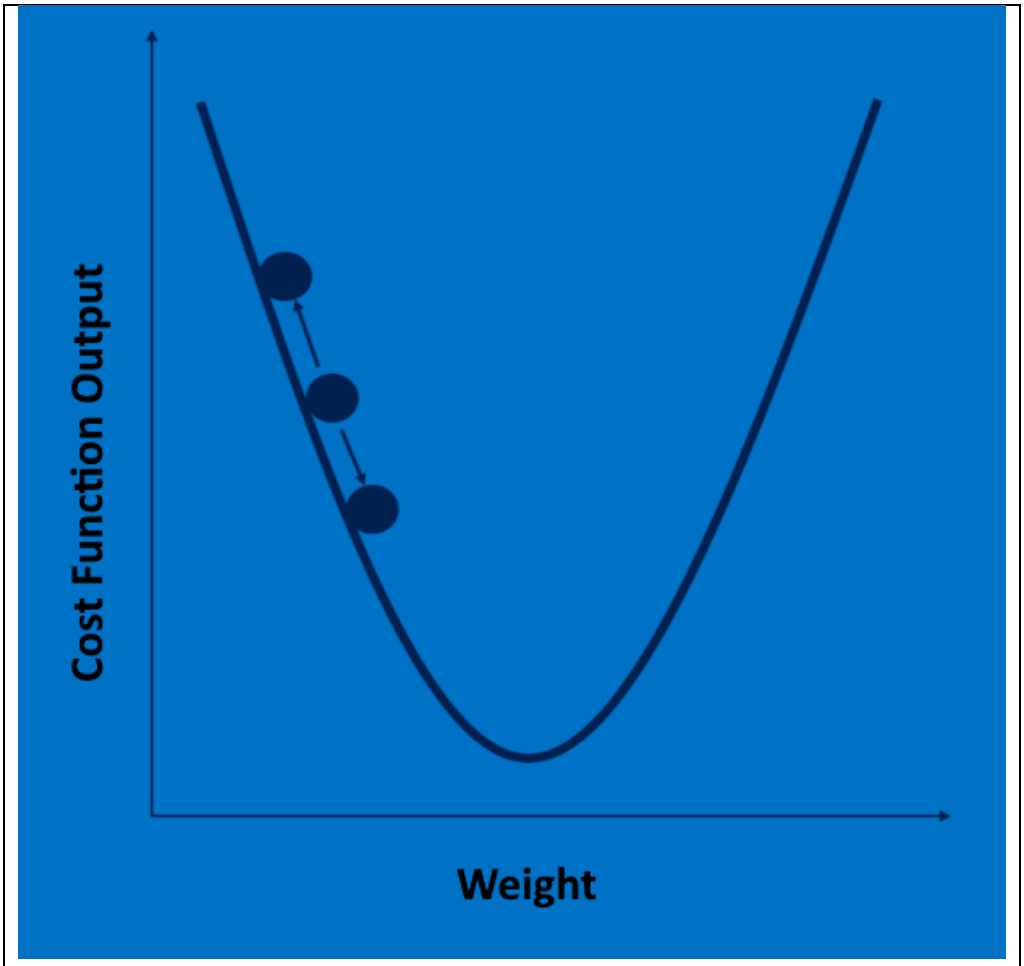


Figure 2.10. A demonstration of how changing a weight in a convolutional neural network can affect its cost function. By following the gradient, the weight can be altered to find its minimum contribution to the cost function. Taken from Rohrer, (2016).

The neural network employed here came as the pre-trained winner of the ImageNet 2012 competition (Krizhevsky et al., 2012). The model contained 60 million parameters with 650,000 neurons and was comprised of eight layers; 5 convolutional layers, followed by 3 fully connected layers. Max pooling occurred after the first, second and fifth convolutional layer. The ReLU non-linearity was applied to the output of every convolutional and fully connected

layer. An input image to the network always took the form of a 224 x 224 x3 matrix. The first convolutional layer convolved 96 filters of size 11 x 11x 3. The second layer took the pooled output from the first and convolved 256 filters of size 5 x 5 x 48. The third layer took the pooled output of the second layer and convolved 384 filters of size 3 x 3 x 256. The fourth and fifth layer convolved the output of the layer before it and had 384 filters of size 3 x 3 x 192 and 256 filters of size 3 x 3 x 192 respectively. The fully connected layers had 4092 neurons each. The architecture of the model is seen in Figure 2.11.

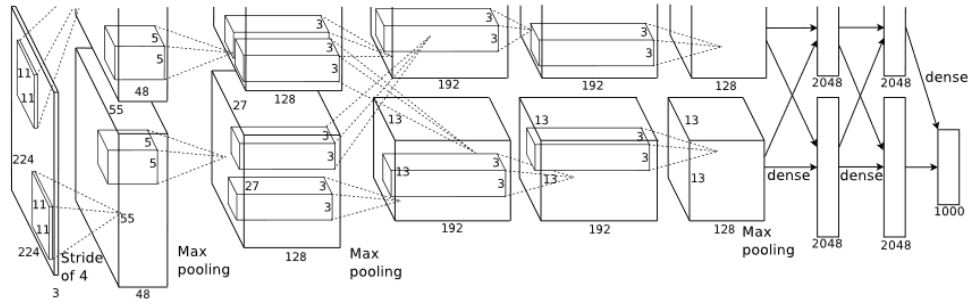


Figure 2.11. An illustration of the architecture of the network used here. Filter sizes are shown here as well as the output size of each layer. Taken from Krizhevsky et al., (2012)

The neural network was trained on 1.2 million high-resolution images from the ImageNet LSVRC-2010 contest (Russakovsky et al., 2015). The original model used the output of the eighth layer to make a decision of an image's category. It did this by applying a 1,000-way soft max (a method of turning the output of a network into probabilities) on the output of layer 8. This final layer (Layer 8) is, therefore, thought of as the decision process, with layer 7 as the primary image description. For our purposes, layer 7 was used as the image description

for the deep supervised convolutional model. Implementation of this model can be found at http://caffe.berkeleyvision.org/ (Jia et al., 2014).

2.3. Decision Processes

Two types of decision processes are used in this thesis, those used in the image *recognition* tasks and those used in the image *categorization* tasks. The decision processes used in the image *recognition* tasks, in which participants are asked to identify whether a stimulus was present using a match-to-sample procedure, are based on Euclidean distance between the images' image descriptions. The image *categorization* task ("Was there a photograph of a mountain?") explored decision processes from prototype theory, exemplar theory and decision boundary theory.

A single decision process was used in the recognition tasks. This process produced a value that we call a *Disc* score (see below, Chapter 2 - Standardization of computational models' outputs) which is a measure of the evidence for correctly discriminating the target image from all other images (either a single distractor image or multiple in a rapid serial visual presentation procedure). *Disc* scores were standardized to vary from 0 to 1.

Standard Image Recognition. The decision process used for the standard image recognition tasks calculated the ease with which the computational model could tell two images apart. This was done by taking the Euclidean distance between the *target* image and the *distractor* image to produce the *Disc* score.

As the *Disc* score (distance) gets larger the easier it is to tell the two images apart and so is also the evidence for a correct recognition.

Each of the decision processes used in image *categorization* tasks calculated the strength of the evidence for an image to be placed in a particular category, which was called the image's *Cat* score (see below, Chapter 2 - Standardization of computational models' outputs). If the decision process correctly categorizes the image then the *Cat* score is positive and increases with confidence. Otherwise, the *Cat* score is negative or extremely small. Calculation of each decision process uses a leave-one-out cross-validation method for each image in each computational model. *Cat* scores were standardized to vary from -1 to 1. Further details of the calculation for each decision theory are described below.

Prototype Theory. For each image being categorized, the prototype to which it was being compared was calculated by obtaining the mean image description of all of the images in its category excluding the image being categorized. The mean was chosen as the prototype, as each dimension of the image descriptor spaces is continuous. The distance between the image's description and the description of the primed category prototype was calculated. For this decision process, *larger* distances suggest *less* evidence for a correct categorization. Therefore, in the normalization step to generate the *Cat* score for this decision process we also subtract the distance score for the image from the maximum

distance measured for that computational model, thereby flipping the magnitudes of the scores.

Exemplar theory. Here we use 9-nearest neighbors in order to classify the new image. The choice of 9-nearest neighbors is arbitrary, but falls between numbers successfully employed in literature for image categorization (Kim, Kim, & Savarese, 2012; Zhang et al., 2006). As the question asked to observers was 'Does the image belong to category X?' the version of exemplar theory used here categorized images based on a one-versus-all rule. If one of the 9 closest images was of the category primed then it was calculated as a +1, if it was not of that category a -1 was assigned. This value was then weighted by the image's distance to the image being categorized. A weighted distance was used as it has been shown to aid exemplar models in predicting human performance (Getty, Swets, & Swets, 1980; Nosofsky, 1986).

Decision Bound Theory. Here a Linear Discriminant Analysis (LDA) was employed to generate the decision bound of a category using the one-versus-all rule; the linear boundary optimally separates the images of one category from all other images. The *signed* distance of the image from the decision bound was used as the evidence for categorization; if the image falls on the correct side of the decision bound then a positive *Cat* score is used, whereas a negative value is used for images falling on the wrong side of the bound (that are miss-categorized).

2.4. Image Set

The image set described here was used in the experiments as stimuli as well as the computational models as training and testing images. All images were taken from the LabelMe scene database (http://cvcl.mit.edu/database.htm). This database was picked, because it consisted of images which were all the same size and there was little overlap between image categories (Oliva & Torralba, 2001; Watson et al., 2014). All images were converted to greyscale and had a resolution of 256 $\times$ 256 pixels. The luminance profiles of the images were normalized using the luminance histogram function of the SHINE toolbox (Willenbockel et al., 2010), such that very simple image differences, such as mean luminance, did not provide cues as to the image category. The computational models and the observers only ever saw the normalized images.

Images consisted of four categories; Buildings, Mountain, Ocean and Trees. These image categories were chosen based on previous work (Watson et al., 2014). The categories of Mountain, Ocean and Trees are designed to capture natural scenes, while the category Cities is designed to reflect man made scenes. There were 830 images that are referred to as *mask* images, 120 images that are referred to as *targets* and another 120 images referred to as *distractors*. These names denote the way in which the images were used in the various experiments presented in this thesis. Example images can be seen in Figure 2.12.

Figure 2.12. Example scene images taken from the set of images used in the experiment that have been grey scaled and histogram luminance corrected.

2.5. Impact of binning data during comparisons

Representational similarity analysis (RSA) is the main method of comparison between computer models and human observers, in both the areas of human behavior (Ghodrati et al., 2014; Kheradpisheh et al., 2016) and neurological activation (Khaligh-Razavi & Kriegeskorte, 2014; Watson et al., 2014).

There are two main reasons why RSA was not used in this thesis. The first is that RSA is limited in the number of images it can examine, as each image needs

to be compared to each other image. This means that, as the number of images examined increases, the number of trials needed to perform that analysis increases exponentially. As a result every study that has employed RSA has been limited to 100-150 images in total. Here the number of images examined varied from 240 to 36,000 images in the case of the temporal blurring analysis. The second reason is that often each image (*categorization* task) or image pairing (*recognition* task) varied in the number of trials presented to observers (this could be as low as 1 trial per image pairing in the image *recognition* task). This prevented binning for each image or image pairing.

While RSA bins data based on either images or image categories, here we chose to bin data based on a set proportion of the number of trials, pooling across participants. A fixed-effects approach rather than a random/mixed-effects approach is taken since here the primary interest is in the average response of all participants. While it would be interesting to examine the data on the participant level, looking how individual participants responses differ from one another and the computational models, it is outside the scope of this thesis.

In both image categorization and image recognition experiments binning of trials followed the same method. For each trial the computer model calculated an unstandardized *Disc* or *Cat* score, the estimated difficulty that the observer would have in producing a correct response. Trials were then ordered based on their *Disc* or *Cat* score and allocated to bins. For each bin the average *Disc* or C*at* score was calculated and plotted against the mean of observers'

behavior in that bin (e.g. reaction time and accuracy). This allowed a regression to be run to assess the variance in observers' behavioral data which was explained by the computational model in the same manner as RSA.

The robustness of the proposed binning method was examined in Chapter 3 - Experiment 1, where a number of different bin sizes were explored to examine the effect of bin size on the data. The results found that as the number of bins increased the computational models which were found to produce a significant fit to observers behavior did not change. The order in which the computational models best explained observers' behavior also did not change suggesting that bin size had little effect on the overall trend of results.

2.6. Standardization of computational models' outputs

As *Cat* and *Disc* scores (for *categorization* and *recognition* experiments respectively) are based upon image distances and as such they can vary greatly in magnitude on different image descriptors as well as decision processes. Standardization of *Disc* and *Cat* scores happened at the level of the bins. It did not matter whether standardization occurred pre- or post-binning, as no transforms occurred that would affect the results. Standardization of *Disc* and *Cat* scores followed a slightly different process.

Disc scores were standardized for each computational model to range from 0 to 1. The zero point of each computational model was the lowest value bin the model produced. This was achieved by taking away the lowest bin *Disc* score

away from all of the bin *Disc* scores the computational model produced. Next in order to standardize the highest point *Disc* scores were divided by the highest bin *Disc* score the computational model produced. Thus, the binned *Disc* scores varied from zero to one, with one being the value which indicates the group of images which were the easiest to tell apart by the computational image descriptor and zero being the hardest. This is summarized by the Equation 1 below.

$$\text{Disc bin}_{\text{standardized}} = (\text{Disc bin}_{\text{raw}} - \text{Disc bin}_{\text{Min}})/\text{Disc bin}_{\text{Max}} \qquad \text{(Equation 1)}$$

Cat scores were standardized for each computational model to vary from -1 to 1. This was done by dividing by the greatest absolute value of either the highest or lowest bin the model produced.

Chapter 3 -Comparing computational Image descriptors to human behaviour.

3.1. Introduction

The first aspect of the Model of Visual Processing that we sought to investigate was the image descriptor component. The main aim was to understand which computational image descriptors structure their image descriptions in a similar manner to biological vision.

In order to map out the structural organization of observer's image descriptions, an image *recognition* task (a match-to-sample task) was used. This is a task which requires observers to pick out a target image from two images; one target and one distractor. This task becomes more difficult as the two images become closer in perceived similarity. It is therefore possible, by examining observers' correct responses and reaction times, to measure how similar two images are in the observer's descriptor space (Shepard, 1958, 1962a, 1962b, 1987; Torgerson, 1952). This can be repeated over a number of trials for a number of images to map out the structure of observers' image descriptions.

By using an image recognition task the structure of image descriptions of different computational image descriptors can be mapped out in much the same way. Computational models, each comprising various image descriptors, can be used to simulate behavior on each trial. Here these computational

models employed a decision process that calculates the Euclidean distance of the target image and distractor image in descriptor space. The distance between the two images in descriptor space is the difficulty in which the computational model is able to tell apart the two images. Comparing the observers' behavioural responses with the responses from the computational model makes it is possible to gauge the similarity the of image descriptions between a computational image descriptor and biological vision.

There have already been a number of studies investigating the similarity of computational models to human observers in behaviour (Ghodrati et al., 2014; Kheradpisheh et al., 2016). These studies, however, focus on categorization tasks; a complex task where there is a high degree of uncertainty of the *decision process* employed by humans (Ashby & Maddox, 2005, 2011). Due to the degree of uncertainty of the decision process it is difficult to determine the similarity of image descriptor separate from that of the decision process. By comparison, an image recognition task is a simple task where the *decision process* is almost guaranteed to be based upon distance of images' description in the *descriptor space* (Attneave, 1957; Shepard, 1962a, 1962b, 1987). The experiments presented in this chapter aim to fill the gap in the literature and use a behavioural approach to examine more closely the similarity between computational image descriptions and biological vision.

Previous studies comparing computational models to human behaviour have examined the results at a categorical level and with a limited image set size

(around 50-100 images) (Mack & Palmeri, 2010; Serre et al., 2007). Here we use image set size of 240 images and compared computational models to human observers at the image level. By expanding the image set size and comparing the data at a much finer level more information can be brought out of the data on exactly how similar these computational models are to human behaviour.

Two experiments are presented here. Both are image recognition studies but differ in their design. The first was an image recognition task that employed a Yes/No, delayed match-to-sample procedure. Examining the results of the first experiment indicated that the design may have influenced the observed similarity of computational models and human observers; computational models were unable to predict observers' reaction times and observers' correct responses were close to ceiling. A second experiment, using a 2AFC, match-to-sample procedure, was used to confirm the results of the first experiment. This experimental design allowed observers to respond as soon as they made a decision and also removed the subjective nature of a Yes/No response by forcing observers to pick an image. Both experiments are added as they highlight design issues for image recognition tasks.

Here we constructed computational models with a range of different image descriptors from the previous literature, GIST, HMAX, PHOW and deep convolutional network. Each image descriptor was chosen specifically due to its history in the literature.

3.2. Experiment 1

3.2.1. Methods

Observers

41 Nottingham University students took part (25 female; range 19 – 41 years; mean age 23.2). All participants had normal or corrected to normal vision. Observers were given the option of compensation in the form of an inconvenience allowance or course credits. Written consent was obtained for all the observers, with the study being approved by the University of Nottingham Ethics Committee.

Apparatus

The experiment was programmed in PsychoPy (Peirce, 2007), and was run on a Lenovo desktop with 3.7 GHz, Intel Xeon E5-1620 v2 processor and NVIDIA NVS 310 graphics card. The viewing distance was held constant with a chin rest at 57cm from the monitor screen. The monitor was a Iiyama ProLite GB2488HSU set to a 1920 x 1080 resolution, and with a 144Hz refresh rate. To ensure good timing of image presentation, images for each trial were loaded onto the graphics card during the inter-trial interval. Timing of all briefly-presented stimuli in the rapid serial visual presentation (RSVP) was controlled by presenting the stimuli for a fixed number of screen refreshes. We verified that the system reliably presented these stimuli within an RSVP sequence without dropping any frames.

Design and Procedure

The experiment consisted of a block of practice trials, followed by two blocks of main trials, the whole experiment took around 40 minutes to complete. Observers were offered a break between the two blocks of main trials, this was done to avoid fatigue of the observers. Before the trials started a set of instructions were given to the observer. These instructions told the participant that their task was an image recognition task. They would first be presented with an image (the *target*), after which a stream of images would be displayed and they would have to respond, by pressing a key, if the image was in the stream. Participants were also told they should respond as quickly and as accurately as possible once the stream of images had ended. In order to familiarise observers with the experiment they were given a practice block of trials before the main block of trials. This practice period followed the same task design, but the target images were taken from a different set of images (actually the *mask* image set; all the pools of images used in the practice and main trials are described in Chapter 2 -General Methods). The practice period lasted no longer than 20 trials and could be self-terminated by the observer once they were happy they understood the task.

Each block of main trials consisted of 240 trials. Each trial started with a fixation dot which lasted 500 ms. An image prime from the target image set was then displayed for 1000 ms. This image prime then disappeared and was followed by a fixation point lasting 1000 ms. After the fixation dot disappeared an RSVP stream was presented. This RSVP stream consisted of a sequence of 6 pictures

presented for either 6, 12, 18, 24 screen refreshes (42, 83, 125, or 167 ms respectively with our 144 Hz monitor) per image. Image presentation times were varied to examine if different computational models better explained observers' behavior at different presentation times. This analysis was however never conducted. After this stream of images had finished observers had to then report whether the target primed image appeared in the RSVP stream. On 50% of trials the *target* image was present, on the other 50% of trials a different image but of the same category was present (a *distractor*). Primed *target* images or their *distractor* counterpart could appear in the serial positions of 2, 3, 4 or 5 in the RSVP stream, the first or last position was not used to ensure that the target or distractor image was forward and backwards masked. Target and distractor image positions were balanced over the trials. The other images used to make the RSVP stream consisted of images that were of a different image category to the target image primed. Observers responded with the left arrow key if they thought the target image matching the prime had been displayed, while they used the right arrow key if they thought that the target image was absent. A visual explanation example of a single trial is displayed in Figure 3.1.

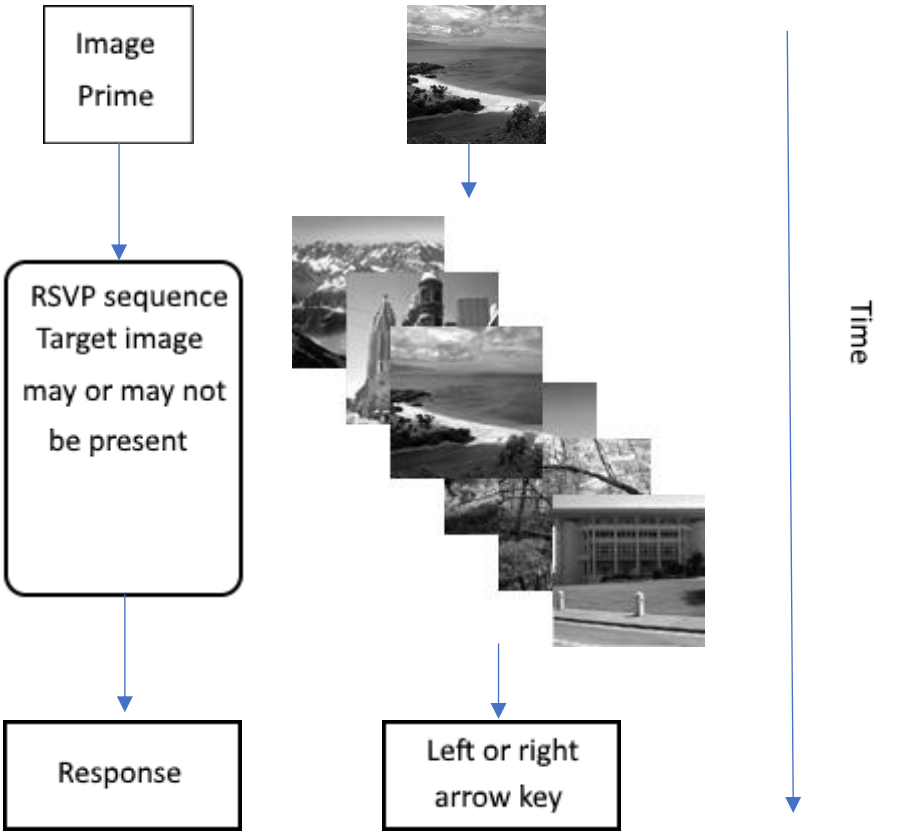


Figure 3.1. A diagram showing the flow of the experiment. On the left a verbal description is given, while the right a visual representation is shown. The images of the RSVP sequence are displayed at the same location, here in the visual representation they are displayed at an angle to show the different images being displayed. The target image has also been made to 'pop' out in the visual representation so that is more clearly visible in the diagram of the RSVP stream.

3.2.2. Results

Trials in which the observer took longer than two seconds to respond were excluded from the analysis (2.8% of trials). This criterion for exclusion was chosen in order to limit observers' responses to rapid feedforward response rapid feedforward response based on instinct rather than cognitive reasoning. The descriptive statistics of observer's performance in the experiment are first

presented, before being compared to the *image descriptors* paired with a Euclidean distance *decision process*. The descriptive statistics can be seen in Figure 3.2.

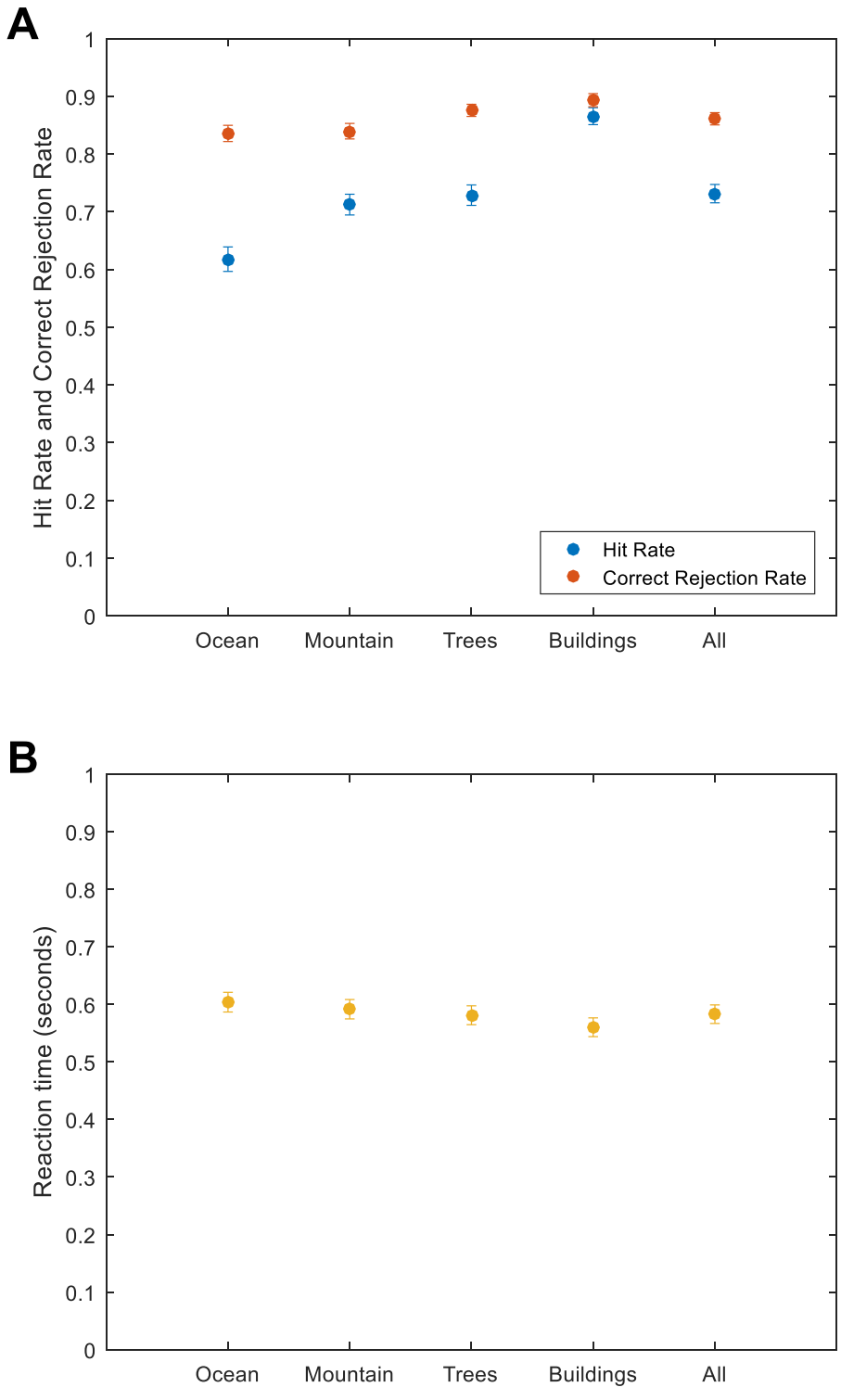


Figure 3.2. Observers performance in the four different image categories as well as when they are all pooled together. A) Hit rate, correct rejection rate are plotted. B) Observer's reaction time (in seconds) are all plotted. Error bars shown are the standard error of the mean.

Three separate 1x4 repeated measures ANOVAs were run for the dependent measures of hit rate, correct rejection rate and reaction times respectively. This was done to see if the dependent variables varied across image category. Observers' hit rate was shown to vary significantly across image category ($F(3,120) = 81.216$, $p < .001$, $\eta_p^2 = .81$). Observers' correct rejection rate was also seen to significantly vary across image categories ($F(3,120) = 11.222$, $p < .001$, $\eta_p^2 = .22$). Observers' reaction times were also seen to significantly vary across image category ($F(3,120) = 16.046$, $p < .001$, $\eta_p^2 = .29$).

On "distractor" trials (target absent), the ability of the four image descriptors (GIST, HMAX, PHOW, and deep convolutional neural net) to distinguish targets from distractors (the *Disc* score, as described in Section 2.6) was calculated and compared with observers' actual responses. Only "distractor" trials were analysed as there was a distance between the *target* image observers were primed with and the *distractor* image observers perceived in the RSVP. Human observers' accuracy on target absent trials (correct rejection rate) was compared to computational models. A full explanation of how this comparison was made can be found in Chapter 2 - General Methods.

For each bin the average *Disc* score, accuracy, and reaction time was calculated. *Disc* score was then plotted against each behavioural measure, accuracy, and reaction times. A number of different bin sizes were explored to

examine the effect of bin size on the results. Correlation coefficients for each image descriptor and bin size were calculated and are summarised in Table 3.1. Figure 3.3 plots the results of the four different image descriptors when the number of bins is 120.

Table 3.1. The results of correlating different image descriptors' *Disc* scores against observers' accuracy and reaction times. Significance values are uncorrected for multiple comparisons. Different bin sizes are investigated. Regressions in the direction predicted have a positive r value and are indicated by green shading.

	Accuracy				Reaction times			
	GIST	HMAX	PHOW	DCN	GIST	HMAX	PHOW	DCN
Number of bins 10 (955-956 trials per bin)	*p*= .009 *r*= .77	*p*= .086 *r*= .57	*p*= .110 *r*= -.54	*p*< .001 *r*= .89	*p*= .132 *r*= -.51	*p*= .169 *r*= -.47	*p*= .956 *r*= .02	*p*= .132 *r*= -.51
Number of bins 30 (318-319 trials per bin)	*p*< .001 *r*= .59	*p*= .076 *r*= .31	*p*= .207 *r*= -.24	*p*< .001 *r*= .68	*p*= .095 *r*= -.31	*p*= .101 *r*= -.31	*p*= .942 *r*= .01	*p*= .027 *r*= -.40
Number of bins 60 (159-160 trials per bin)	*p*< .001 *r*= .50	*p*= .11 *r*= .21	*p*= .200 *r*= -.17	*p*< .001 *r*= .61	*p*= .066 *r*= -.24	*p*= .122 *r*= -.20	*p*= .864 *r*= -.02	*p*= .023 *r*= -.29
Number of bins 120 (79-80 trials per bin)	*p*< .001 *r*= .37	*p*= .071 *r*= .17	*p*= .208 *r*= -.12	*p*< .001 *r*= .49	*p*= .079 *r*= -.16	*p*= .153 *r*= -.13	*p*= .861 *r*= -.02	*p*= .016 *r*= -.22

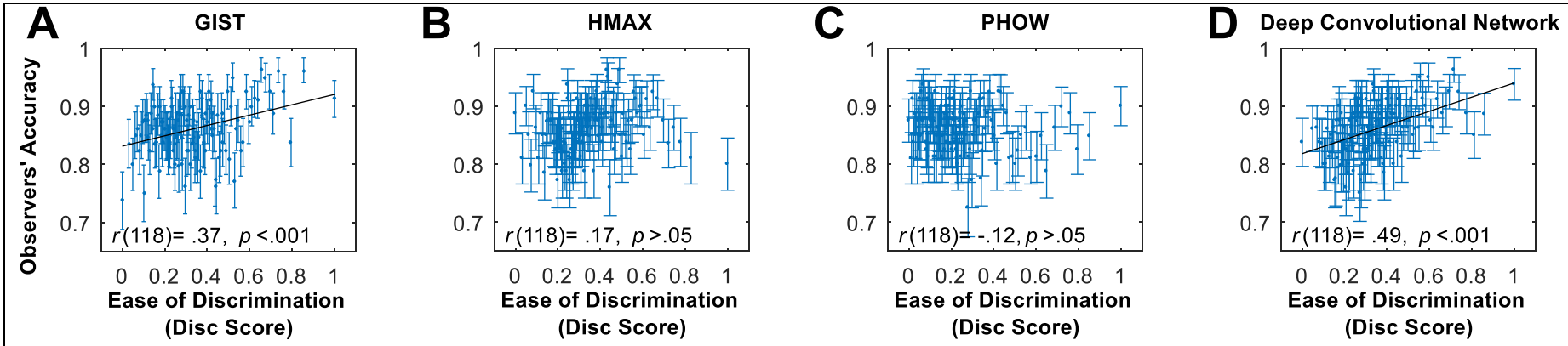


Figure 3.3. Plotting observers' accuracy data against different models' *Disc* scores. This was done for (A) GIST, (B) HMAX (C) PHOW (D) Deep convolutional Network. Standard error of the mean has been plotted. The bin size used for each graph is 120. A score of .6 means 60% of distractor images were correctly rejected.

The results table (Table 3.1) indicates that bin size had little effect on which image descriptors were found to significantly fit human behaviour. Two models were found to have a good fit to human behaviour. The image descriptors GIST and the deep supervised convolutional neural net were largely able to predict human observer's behaviour on a single trial basis in the image recognition task. These models are able to predict correct responses, but largely are unable to predict reaction times. This is probably due to experimental design, observers were unable to respond during the RSVP procedure and instead were told to make their response after. This means that observers could have already decided upon a response before they were able to execute it, indicating that the measured reaction times may not reflect true reaction times.

3.2.3. Discussion

The aim of this study was to try and determine the ability of each image descriptor to explain human observers' behaviour. Out of the four computational image descriptors examined three produced significant fits to human behaviour, in terms of accuracy. The deep supervised convolutional neural net was the only model able to explain a significant amount of variance in observers' reaction times. The deep supervised convolutional neural net (Krizhevsky et al., 2012) was shown to have closest fit to observers' data, being able to fit observers' accuracy and reaction times, this was then followed by GIST (Oliva & Torralba, 2001). The image descriptors HMAX (Thomas et al., 2007) and PHOW (Lazebnik et al., 2006) failed to explain human performance

on any aspect of the task. The results support the growing evidence image descriptors which create their image description from the low level visual properties of the stimuli can account for a significant proportion of variance of the structure of image description in humans (Leeds, Seibert, Pyles, & Tarr, 2013; Rice et al., 2014; Watson et al., 2014; Watson, Young, et al., 2016). Additionally, it has been shown that deep supervised convolutional networks provide the best known account the structure of human image descriptions (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

In the current study the main metric of interest was accuracy on target absent trials (correct rejection rate). Observers scored higher than expected, with their accuracy being around 85%. At this level behavioural responses were potentially being affected by the ceiling effect. This would be expected to reduce the overall variance of the behavioural data as the top end of the behavioural variance was excluded. As the computational models were not designed to take this into account, the true variance explained of the computational models may be higher than the results reported here. This is investigated further in Experiment 2 below.

The analysis also examined the impact of binning trial data to turn discrete human responses into a more probabilistic metric. A number of different bin sizes of the trial data were investigated, and the data confirmed that bin size had no effect on which image descriptors best fit human behaviour. The ranking of fit of each image descriptor was preserved irrespective of bin size.

3.3. Experiment 2

3.3.1. Introduction

The main aim of this chapter was to assess the ability of different computational *image descriptors* to explain the structure of human observers' image descriptions through a behavioural experiment. We wanted to replicate this with an additional study to check the potential impact of the ceiling effect and look for effects of reaction time more explicitly.

In Experiment 1 the metric of human behaviour used to assess observers' image descriptions was that of accuracy on target absent trials (correct rejection rate). Observers' accuracy was much higher than expected, with observers scoring around 85% in Experiment 1. At this level, observers' accuracy was almost at ceiling. This ceiling effect could have distorted the observed structure of image descriptions by eliminating the top end of variance in the data set. This would have affected the computational image descriptor's ability to explain human behaviour as it was unable to take into account this ceiling effect. In Experiment 1 a Yes/No task was used ("Was the target present in the stream?") whereas in this experiment we used a 2AFC task ("Which of the two images was present?"). The former depends on both sensitivity and participant's internal thresholds (to say "yes") whereas the latter does not. This internal threshold could partially explain the ceiling effect of the previous study, if participants were conservative and only responded "yes" when they were absolutely sure the *target* image had been seen.

A number of different studies have shown that reaction times can also be used as a measure of the structure of human image descriptions (Ashby, Boynton, & Lee, 1994; Mack & Palmeri, 2010; Sofer et al., 2015). Experiment 1 did not find this, probably due to the experimental design which restricted observers to respond at the end of the RSVP procedure and not when they had gathered enough evidence to make the decision. Experiment 2 examines the structure of human image descriptions based on reaction times by allowing observers to respond as soon as they feel appropriate. This change in experimental design allows reaction times to reflect visual processing demands and so letting reaction times reflect the structure of observers' image descriptions.

A difference in observers' behavioural metrics across image categories was found in Experiment 1. The data from Experiment 2 allows us to further investigate why this may be the case. This aspect of the data is considered more extensively in Chapter 6 - Investigating temporal blurring.

3.3.2. Methods

Observers

Seventy Nottingham University students (55 female, 15 male; age 18-31 years) took part in the experiment. All were volunteers who were either paid for participation or given course credits. All signed a consent form and all procedures were approved by the University of Nottingham Psychology ethics committee.

Apparatus

The experiment was programmed in PsychoPy (Peirce, 2007), and was run on a Dell desktop with 4 GHz, Intel Core 2 Duo processor. The CRT monitor was set to a 1024 x 768 resolution, with an 85Hz refresh rate. The room was normally illuminated. Images were loaded before the sequence was run and were presented precisely on a specific number of frames. The viewing distance was held constant with a chin rest at 57cm.

Design and Procedure

The experiment consisted of a block of 20 practice trials, followed by a main block of 240 trials. Each trial consisted of a categorization task which lead onto an image recognition task. This was done so that a single experiment could provide two different data sets. In the current chapter we are mainly concerned with the results from the image recognition section of the task, while Chapter 4 - Experiment 1 is concerned with the results from the image categorization task.

Each trial began with a fixation cross which lasted 500ms. The fixation was then followed by a text prime which consisted of the name of an image category (Ocean, Mountain, Trees, Buildings). After this prime disappeared a RSVP procedure was conducted. This consisted of six images presented rapidly (All the images used in the practice and main trials are described in Chapter 2 - Image Set). For main trials, images were presented for either 2, 4, 6 or 8 screen refreshes (24, 47, 70, or 94 ms) per image. Image presentation times were

varied so that it could be examined if different models better explained observers' behavior at different presentation times. This analysis was however never conducted. On practice trials image presentation times were kept constant at 94 ms, to make them easier. On 50% of trials a *target* image matching the category prime could appear in any of the image positions in the RSVP procedure, except from the first or the last image position. Target images were only ever presented once (and never in the practice period). Images used as target images in the practice period did not come from the *target* pool of images (they came from the *mask* pool of images). This was done to limit the exposure observers had to images from the target pool of image. Once the RSVP procedure had ended observers then were required to respond indicating if they had perceived an image matching the text prime (*categorization task*).

On target-present trials the experiment moved onto the image *recognition* task. The image recognition task was displayed irrespective of whether the observer had responded correctly. The image recognition part of each trial started with a centralised fixation cross which lasted 500 ms. After the fixation, the target image and a distractor image were presented at the same time equidistant apart from the fixation cross. The distractor image was of the same image category as the target, but was never presented in the trials as either a target or a mask. The target and distractor image stayed on the screen until observers had made a 2-alternate-forced-choice decision about which of the two images had been presented in the RSVP sequence. Practice trials followed the same format as the main trials except the distractor image was replaced

with a mask image matching the image category primed. On both practice and main trials observers had to respond by pressing the left or right arrow key to pick out the target image they had seen in the RSVP procedure. A visual representation of this experiment can be seen in Figure 3.4.

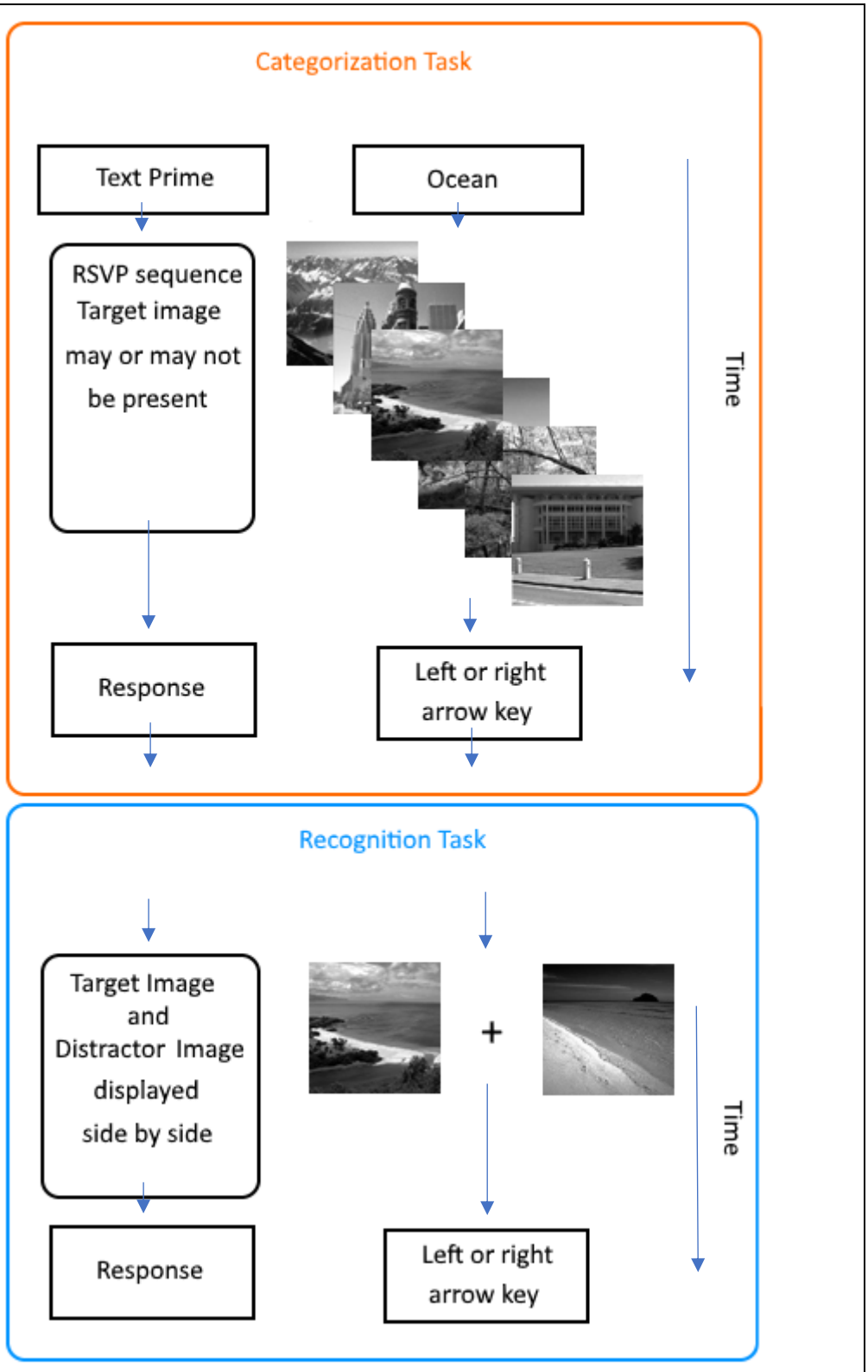


Figure 3.4. A visual representation of Experiment 2. The flow of each experiment moves from top to bottom. The left side gives a verbal description, while the right displays a single trial.

3.3.3. Results

Only trials in which the observer responded correctly in the categorization task were analysed. This was to make sure that observers had seen the target image or else they would be guessing for the image recognition task. Trials in which the observer took longer than two seconds to respond were also excluded from the analysis (2.4% of trials). This criterion for exclusion was chosen in order to limit observers' responses to rapid feedforward response. The descriptive statistics of observer's performance across the image categories can be seen in Figure 3.5.

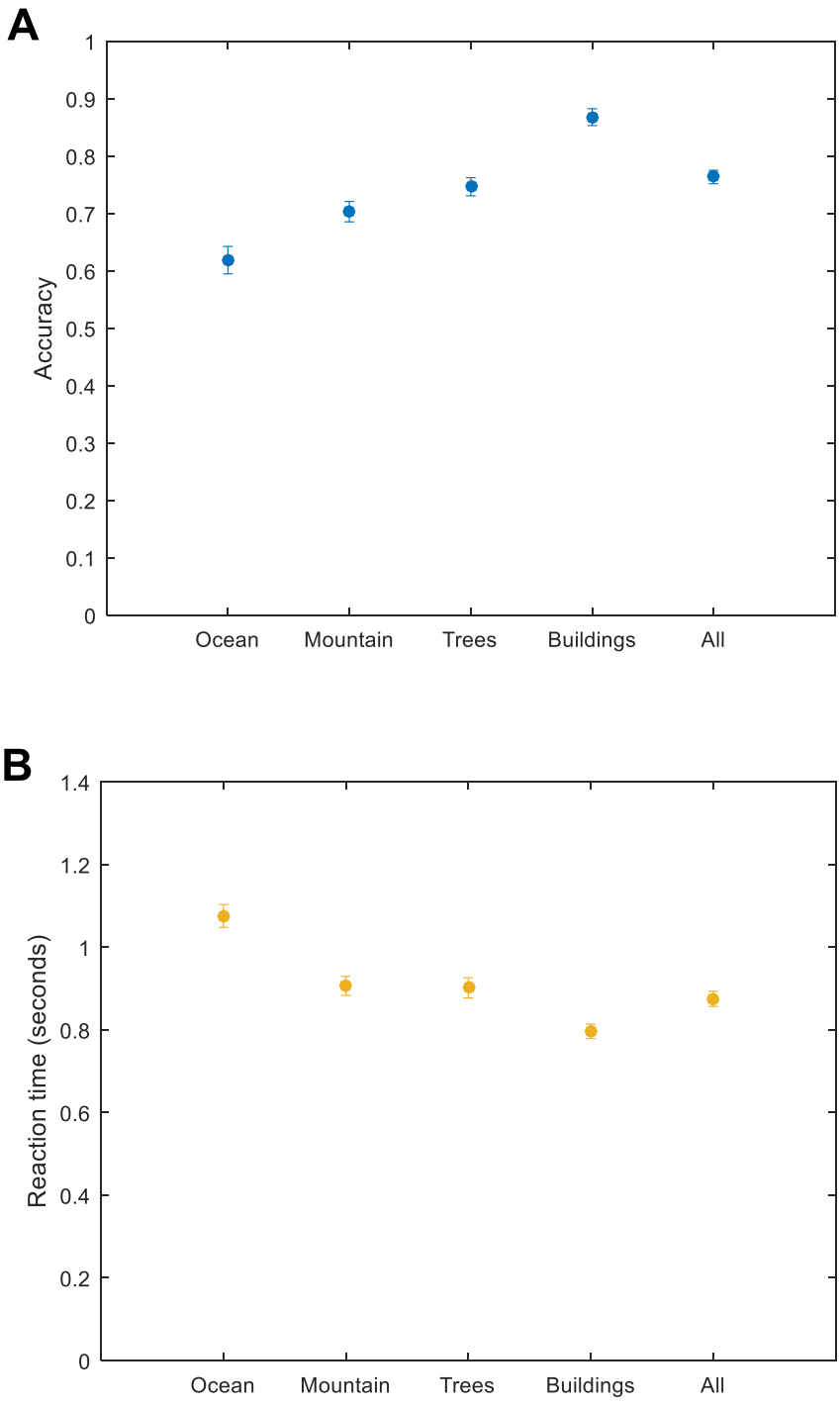


Figure 3.5. Observers performance in the four different image categories as well as when they are all pooled together. A) plots observers' Accuracy (a score of .6 means 60% of targets were detected**), while B) plots observer's reaction time (in seconds). Error bars shown are the standard error of the mean.**

Two separate 1x4 repeated measures ANOVAs were run for the dependent measures of hit rate and reaction times respectively. This was done to see if the dependent variables varied across image category. Examining if observers' hit rate varied across image category, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 33.49$, $p < .001$, ($\varepsilon = .74$)). The results show that observers' hit rate was shown to vary significantly across image category ($F(2.22,153.33) = 41.115$, $p < .001$, $\eta_p^2 = .37$, Greenhouse-Geisser corrected). Examining if observers' reaction times vary across image category, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 69.12$, $p < .001$, $\varepsilon = .60$). Observers' reaction times were seen to significantly vary across image category ($F(1.8,124.19) = 28.429$, $p < .001$, $\eta_p^2 = .29$, Greenhouse-Geisser corrected).

Computational models were constructed by pairing each image descriptor with the standard image recognition decision process (the Euclidean distance between the target image and the distractor image). For further details see Chapter 2 - General Methods.

Due to the fact that there was a varying number of trial for each target distractor pairing trials were divided into 120 bins (36-37 trials per bin). For each bin the average *Disc* score, accuracy (hit rate as target was always present) and reaction time was calculated and plotted against each other. *Disc* score was then plotted against each behavioural measure and a correlation run. Correlation coefficients are summarised below in Table 3.2.

Table 3.2. The results of correlating different image descriptors *Disc* scores against observers' accuracy and reaction times in both Experiment 2 and Experiment 1. These values are uncorrected for multiple comparisons. Regressions in the direction predicted have a positive r value and are indicated by Green shading. Number of bins is 120.

	Accuracy				*Reaction times*			
	GIST	HMAX	PHOW	DCN	GIST	HMAX	PHOW	DCN
Experiment 1 (79-80 trials per bin)	p< .001 r= .37	p= .071 r= .17	p= .208 r= -.12	p< .001 r= .49	p= .079 r= -.16	p= .153 r= -.13	p= .861 r= -.02	p= .016 r= -.22
Experiment 2 (36-37 trials per bin)	*p*< .001 *r*= .63	*p*= .002 *r*= .28	*p*=.906 *r*= -.01	*p*< .001 *r*= .62	*p*< .001 *r*= -.63	*p*< .001 *r*= -.33	*p*= .284 *r*= -.10	*p*< .001 *r*= -.79

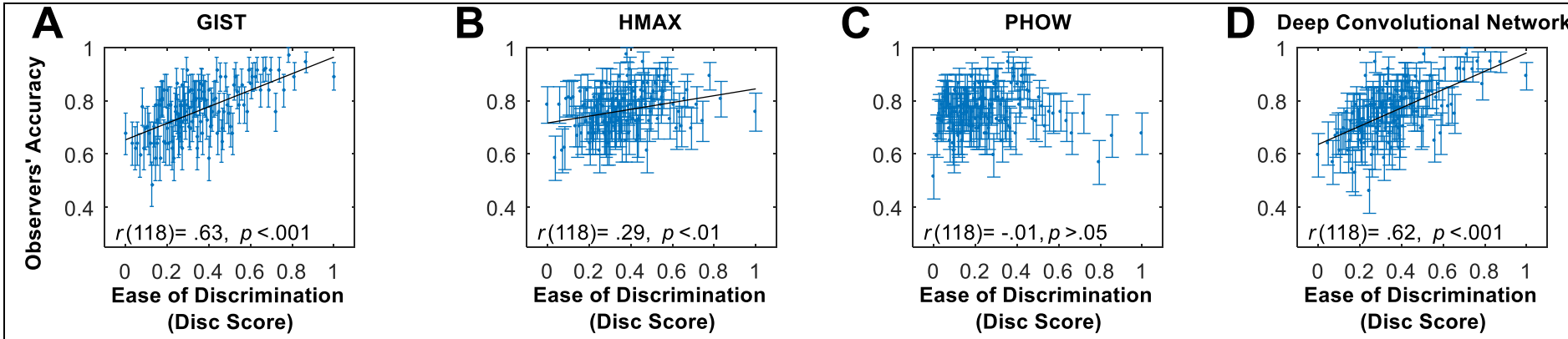


Figure 3.6. Plotting observers' accuracy data against different models' *Disc* scores in Experiment 2. This was done for (A) GIST, (B) HMAX (C) PHOW (D) Deep convolutional Network. Error bars represent +/- 1 SEM. The number of bins used for each graph is 120. A score of .6 means 60% of targets were detected.

Table 3.2 shows the correlations of the models against the human observers for the models, including a reiteration of the data from Experiment 1, for easy comparison. The image descriptors GIST, HMAX and the deep supervised convolutional neural net were largely able to predict human observer's behaviour on a single trial basis in the image recognition task. This confirms the results found from Experiment 1. Additionally, Experiment 2 found significant correlations in the reaction time data, showing that image descriptors are able to predict observers' reaction times when observers are able to respond without any procedural delay.

3.3.4. Discussion

The results of Experiment 2 replicated and extended the results from Experiment 1. The deep supervised convolutional neural net was shown to have the best performance at explaining observers' accuracy, this was followed by GIST in their ability to explain the explaining observers' accuracy. Additionally, Experiment 2 found that HMAX was able to explain observers accuracy and reaction time data. In a similar manner to Experiment 1, Experiment 2 failed to find evidence that the image descriptor PHOW is predictive of human behaviour.

In Experiment 2 the main metric of interest (observers accuracy) was below ceiling; around 75% for the four image categories. Ensuring that observers' accuracy were below ceiling offered an undistorted view of the structure of observers' image descriptions as the top end of participants' variance was not

eliminated. Since the ceiling effect was avoided it would be expected that each image descriptor would explain a greater amount of variance in Experiment 2 than in Experiment 1, which potentially suffered from the ceiling effect. This appears to be supported by the data; the correlations between model and behaviour in Experiment 2 appear consistently higher than Experiment 1, although the pattern of results across models appears unchanged.

Experiment 1 found little indication that the computational image descriptors examined could explain observer's reaction times. This may have been because the experimental design limited observers to respond once the RSVP procedure had finished and thus stopped reaction times from reflecting the structure of observers' image descriptions. In order to examine this further, Experiment 2 employed a design which allowed observers to respond as soon as they wanted to. When observers were allowed to make a response when they were ready, reaction times were indeed predicted by the computational image descriptors. The results from the reaction time data show the same pattern of model performance as the results from observers' accuracy. The deep convolutional neural net was found to explain the greatest amount of variance in the reaction time data, closely followed by the image descriptor GIST and HMAX respectively. PHOW was unable to explain a significant amount of the variance in observes' reaction times.

3.4. General Discussion

The main aim of this chapter was to examine the similarity of different computational image descriptors to biological vision in terms of how they structurally organise an image set. Two studies were presented with this purpose in mind. Both studies employed an image recognition task. The first study used observers' accuracy on target absent trials and reaction times to estimate the structure of observers' image descriptions. This encountered the problem that observers' accuracy were close to ceiling and so the measurement of observers' image descriptions could have been distorted. Additionally, computational models were unable to predict observers' reaction times, indicating that reaction times did not reflect observers' image processing requirements. A second Experiment was conducted with observers' accuracy below ceiling. This Experiment found both observers' accuracy and reaction time data could be explained by the computational models.

Both the studies produced consistent results with each other. Out of the four image descriptors examined three produced significant fits to human behaviour. The image descriptor which provided the best fit to human observers' behaviour was the deep supervised convolutional neural net (Krizhevsky et al., 2012). This is in line with both the neuroscientific literature (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014) and the human image categorization behaviour literature (Ghodrati et al., 2014; Kheradpisheh et al., 2016). From the results presented here and the

trend in the literature, deep supervised convolutional neural nets provide the closest account for biological image descriptions from the computational models presently examined.

GIST produced the second closest fit to human behavior after the deep supervised convolutional model in the two Experiments. This is somewhat surprising as GIST is rather a simple image descriptor. GIST employs no learning and forms its image description based on the low level visual properties of the image. This finding is, however, in line with the previous literature, in which, GIST has been shown to predict human image descriptions (Rice et al., 2014; Watson et al., 2014; Watson, Young, et al., 2016), as well as being useful as an image descriptor for explaining human behavior in image categorization tasks (Mack & Palmeri, 2010; Sofer et al., 2015).

HMAX was also found to have a significant fit to human behavior. HMAX was, however, found to explain less of the variance in observers' behavior than GIST and the deep convolutional network model. The finding that HMAX produces an image description which has a significant fit to human image descriptions is in line with previous findings from the neurological literature (Khaligh-Razavi & Kriegeskorte, 2014; O'Toole et al., 2005) as well as behavioral studies examining image categorization rate (Ghodrati et al., 2014; Kheradpisheh et al., 2016; Serre et al., 2007). Interestingly in studies where more than one image descriptor was examined, other models are usually found to produce a better fit to human data than HMAX.

PHOW was the only image descriptor which did not find a significant fit to human behaviour, although previous research has found it to have some explanatory power (Khaligh-Razavi & Kriegeskorte, 2014; Leeds et al., 2013). Studies in which PHOW, along with other image descriptors have been examined, have shown that PHOW's variance explained is considerably lower than other image descriptors that produce a significant fit (Khaligh-Razavi & Kriegeskorte, 2014). Here, perhaps due to the experimental design, or the more stringent criteria of asking the model to explain human behaviour on a per trial basis, caused a lack of a significant finding. PHOW is based on a 'bag of features' model, where the number of different SIFT features in an image is used as the image description. While a 'bag of features' model can produce high level of correct categorizations (Lazebnik et al., 2006), it is possible that the 'bag of features' model is dissimilar to image description processes employed by the human visual system.

In both the experiments presented here, the image descriptor space was represented as a simple Euclidean space. This form of measuring distances was chosen due to its simplicity and also its popularity in computational vision (Pass & Zabih, 1999) and psychophysics studies (Shepard, 1958, 1962a, 1962b, 1987; Torgerson, 1952). The results found show that this decision process works remarkable well when paired with computational image descriptors at explaining human behaviour, in image recognition tasks. Euclidean distance is not the only one way in which distance between two points can be calculated and there are a number of different ways distance measures can be taken.

There have been a number of studies examining image retrieval from large image databases. Some studies have found that different measures of similarity, other than Euclidian distance, have the best performance in retrieving similar images from the database (Malik & Baharudin, 2013; Sharma & Batra, 2014). Additionally, some psychophysics experiments have found that other ways of measuring distance better match observers behaviour, such as a weighted Euclidean distance (Getty et al., 1980; Nosofsky, 1986). Furthermore, some studies have employed descriptor space transforms, such as principle component analysis, to better represent human observers' behaviour (Mack & Palmeri, 2010). While the current research shows that Euclidean distance works as a way of measuring distance in descriptor space, further research could use the experimental paradigm presented here to investigate into the many different ways descriptor space could be represented in biological vision.

Both Experiment 1 and Experiment 2 found that observers' behavior, in terms of hit rate, correct rejection and reaction times, changed based on image category. There are a number of different possible explanations for this. A possible explanation is that due to experimental design different image categories had different masks. This category dependent masking means that some image categories had a harsher masking that other categories, which could have led to this effect. Another possible explanation for this is that the distribution of images in descriptor space varied across categories, some image categories may be more spread out than others in descriptor space. The changes in how tightly, or loosely, packed images are together would affect

their difficulty to tell them apart from one another, causing the categorical effect. A similar explanation to this has been used by Sofer et al. (2015) to explain how hit rates and reaction times can vary depending on the category individuals are being asked to categorize an image to (Greene & Oliva, 2009; Joubert et al., 2007; Kadar & Ben-Shahar, 2012; Loschky & Larson, 2010). While this is almost certainly a factor in the categorical change in behavior seen in observers, it is unlikely that these reasons are the sole reasons, as effect sizes were quite large (η_p^2 = .81 in Experiment 1 for hit rate). Another explanation for this categorical change in behavior is that due to the RSVP procedure observers were obtaining temporally blurred image descriptions of the target image. This temporal blurring was having a differing impact on different categories and therefore adding to this effect. This hypothesis was investigated using Experiment 2 presented here in Chapter 6. In this reanalysis, we added a temporal blur component whereby the two neighboring mask images were added (in a variety of weights) to the target image, prior to forming the image description.

Chapter 4 - Investigating the decision processes in an image categorization task.

4.1. Introduction

The main aim of this chapter was to investigate the decision process observers were using in order to categorise images. Early research on category learning investigated a number of different possible mechanism by which humans could be categorizing images. This research found that, surprisingly, in certain circumstances, each theory could be supported and that no single mechanism could explain all of the observed data. The main contenders were prototype theory, exemplar theory and decision boundary theory (Ashby & Maddox, 2005).

This research led to the idea that observers were not using a single rule to categories images, but were instead using multiple rules depending on the circumstance (Ashby & Maddox, 1994; Ashby & Townsend, 1986; Lockhead, 1966; Shepard, 1964). Studies employing fMRI methods looked to see if different brain networks active when observers are using different strategies (Konishi et al., 1999; Lombardi et al., 1999; Rao et al., 1997). The results of these studies suggest different brain areas become active depending on the strategy observers are employing to perform the categorization task (Ashby & Maddox, 2011).

One problem with these previous studies is that they have typically used *image descriptions* based on observable characteristics of the images, such as shape and light contrast (Lamberts, 2000), distortion in line patterns (Homa, Sterling, & Trepel, 1981), distance in dot patterns (Posner & Keele, 1968, 1970) and even distortion of faces (Reed, 1972). While these image descriptions seem intuitive and are easy to report verbally, they are unlikely to reflect neural image descriptions, given what we know from single-unit recordings (Hubel & Wiesel, 1962, 1968). Potentially, the use of inappropriate image descriptions is the reason that the question of optimal decision process has not been resolved.

Here we compare the ability of different categorization decision processes to explain human behavior when combined with a range of recent biologically-motivated image descriptors (GIST, HMAX, PHOW, and deep supervised convolutional neural net). Three different decision processes are examined and are as follows.

Prototype theory was one of the earliest theories of image categorization. Prototype theory proposes that category learning is driven by individuals creating a single "prototypical" representation of a category. New items are accepted as a member of the category if they are similar enough to the prototype (Homa et al. 1981; Posner and Keele 1968, 1970; Reed 1972; Rosch 1973, 1975; Smith & Minda 1998). Prototype theory has the general prediction that as an image gets closer in similarity to the prototype, the easier it is to classify.

Exemplar theory proposes that category learning is driven by the exemplars of a category. Category decisions are based on comparing the new stimulus to the closest neighborhood of images to it. The stimulus is then assigned to the category for which it has the closest relatives (Brooks 1978; Estes 1986; Hintzman 1986; Lamberts 2000; Medin and Schaffer 1978; Nosofsky 1986). Exemplar theory would predict that the category of images closest to that new image would predict categorization.

Decision bound theory proposes that subjects create a decision boundary in the descriptor space that splits the space into category regions. When the observer is presented with an unfamiliar stimulus the side of the decision boundary the image falls on determines the assigned category (Ashby and Gott 1988, Ashby and Townsend 1986, Maddox and Ashby 1993; Dongjian et al., 2010; Sofer et al., 2015). This theory makes the prediction that as an image gets closer to this decision line, the harder it is to categorize.

4.2. Experiment 1

4.2.1. Methods

To study the impact of different decision processes we used data collected in the *categorisation* task in Chapter 3 - Experiment 2 (two tasks were conducted simultaneously but the analyses in that chapter focused on the *recognition* task).

Briefly, observers were presented with a written prime of a category (e.g. “ocean”). A RSVP sequence of images was then presented, and observers were probed as to whether or not any image in the RSVP sequence was of that category. They were also probed as to which image was seen, as analysed in the previous Chapter. For a full description of the Observers, Apparatus as well as the design and procedure please see Chapter 3 - Experiment 2 - Methods.

4.2.2. Results

Trials in which the observer took longer than two seconds to respond were excluded from the analysis (4.3% of trials). This criteria for exclusion was chosen in order to limit observers’ responses to rapid feedforward response based on instinct rather than cognitive reasoning. The descriptive statistics of observer’s performance in the categorization task are first presented, before being compared to the computer models. The descriptive statistics can be seen in Figure 4.1.

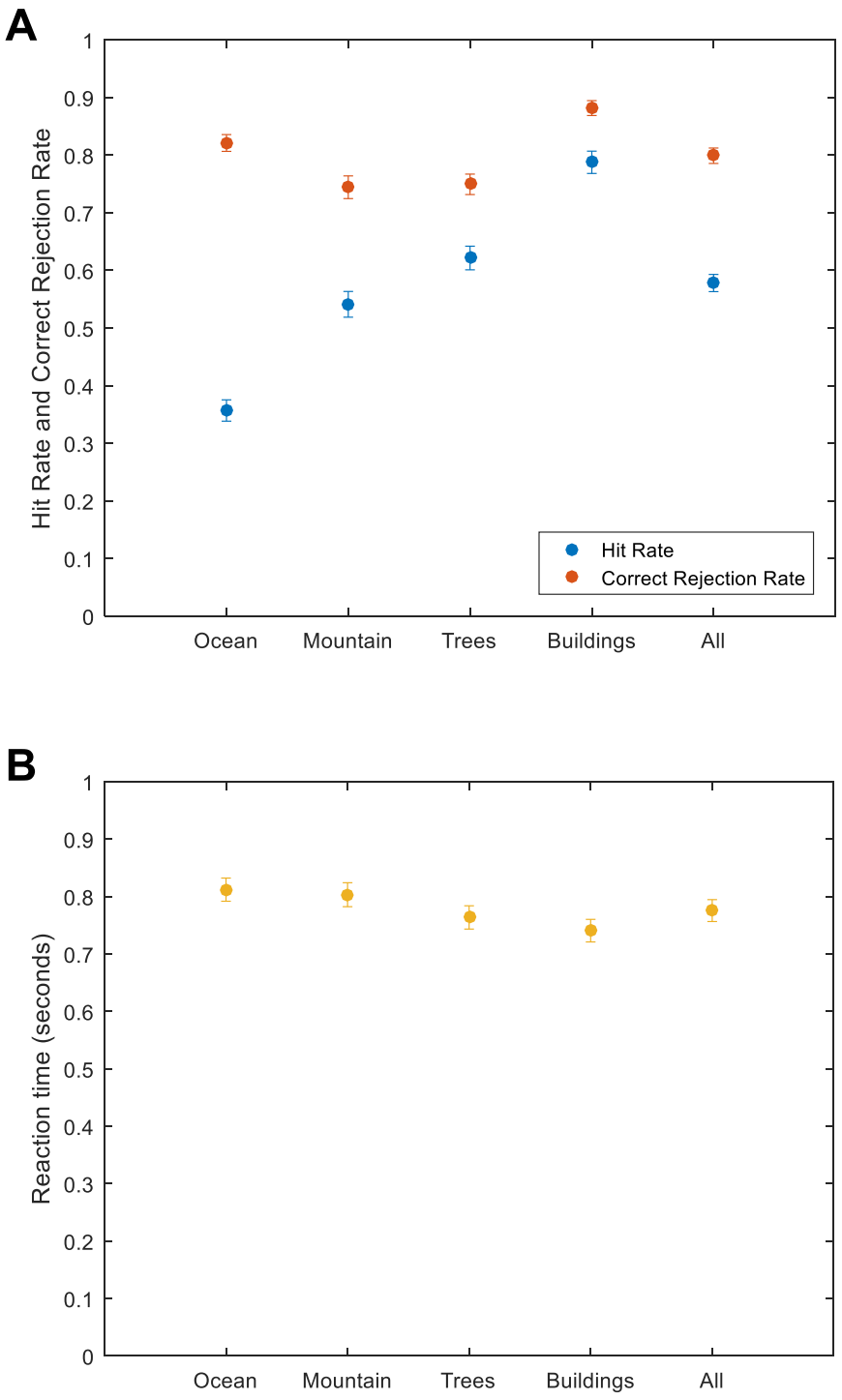


Figure 4.1. Observers performance in the four different image categories as well as when they are all pooled together. A) observers' hit rate, correct rejection rate are plotted, B) observers' reaction time (in seconds) are all plotted. Error bars shown are the standard error of the mean.

Three separate 1x4 repeated measures ANOVAs were run for the dependent measures of hit rate, correct rejection rate and reaction times respectively. This was done to see if the dependent variables varied across image category. Observers' hit rate was shown to vary significantly across image category ($F(3,207) = 128.822$, $p < .001$, $\eta_p^2 = .65$). With respect to observers correct rejection rate, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 15.12$, $p = .010$, $\varepsilon = .89$). Observers' correct rejection rate was also seen to significantly vary across image categories ($F(2.67,183.998) = 35.387$, $p < .001$, $\eta_p^2 = .34$, Greenhouse-Geisser corrected). With respect to observers reaction times, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 26.09$, $p < .001$, $\varepsilon = .37$). Observers' reaction times were also seen to significantly vary across image category ($F(2.49,171.67) = 19.764$, $p < .001$, $\eta_p^2 = .22$, Greenhouse-Geisser corrected).

Computational models were constructed for all the different variations of image descriptors (GIST, HMAX, PHOW, deep convolutional neural net) and decision processes (prototype theory, exemplar theory, and decision bound theory). The output of the computational models' *Cat scores* were compared to observers accuracy on target present trials (hit rate). Target absent trials could not be compared as it was uncertain as to which image the observers were responding to. A full explanation of the image descriptors, decision

process, image set, standardization and binning of computational model outputs are explained in Chapter 2 - General Methods.

Table 4.1. The results of correlating different image descriptors Cat scores against observers' accuracy and reaction times in Experiment 1. These values are uncorrected for multiple comparisons. Regressions in the direction predicted have a positive r value. Green shading indicates significant correlations in the direction expected, while red shading indicates significant correlations in the opposite direction expected. Number of bins is 120 with 67-68 trials in each bin.

	Accuracy				Reaction times			
	GIST	HMAX	PHOW	DCN	GIST	HMAX	PHOW	DCN
Prototype Theory	p = .004 r = -.26	p = .003 r = -.27	p = .195 r = .12	p <.001 r = -.42	p = .32 r = .09	p = .017 r =.22	p = .277 r = -.10	p = .130 r =.14
Exemplar Theory	p = .296 r = .10	p = .126 r = -.13	p <.001 r = .37	p = .004 r = .26	p = .027 r = -.20	p = .678 r = -.04	p <.001 r = -.31	p <.001 r = -.14
Decision Bound Theory	p <.001 r = .33	p = .436 r = -.05	p = .010 r = .24	p <.001 r = .55	p <.001 r = -.33	p = .926 r = -.01	p < .001 r = -.34	p < .001 r = -.47

The results table (Table 4.1) is best examined first from the perspective of the decision process and then from the perspective of image descriptor. The effects of each decision theory were largely consistent across the different image descriptors in the direction of the significant effects found. Prototype theory only produced significant results in the reverse direction to that expected; images furthest from the prototype of that category were the easiest to be categorized. Both exemplar and decision bound theory produced significant positive correlations. From the perspective of the image descriptors three out of the four image descriptors (GIST, PHOW and deep convolutional neural net) were able to explain a significant proportion of human observers' behavior, in terms of accuracy on target present trials and reaction times, when paired with either exemplar theory or decision bound theory. HMAX failed to find any significant correlations in the direction that was expected and this was even reversed in the case of combining it with prototype theory.

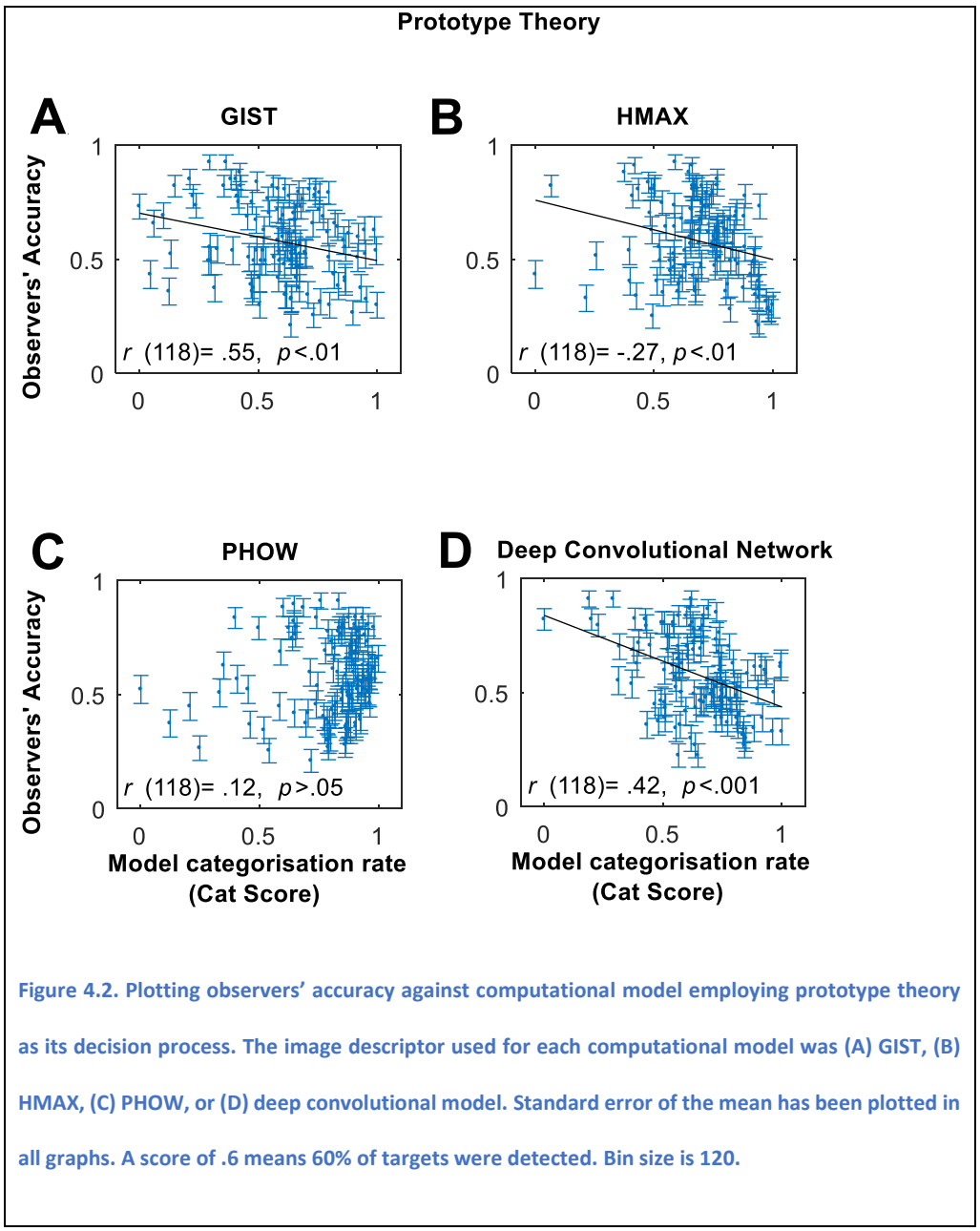


Figure 4.2. Plotting observers' accuracy against computational model employing prototype theory as its decision process. The image descriptor used for each computational model was (A) GIST, (B) HMAX, (C) PHOW, or (D) deep convolutional model. Standard error of the mean has been plotted in all graphs. A score of .6 means 60% of targets were detected. Bin size is 120.

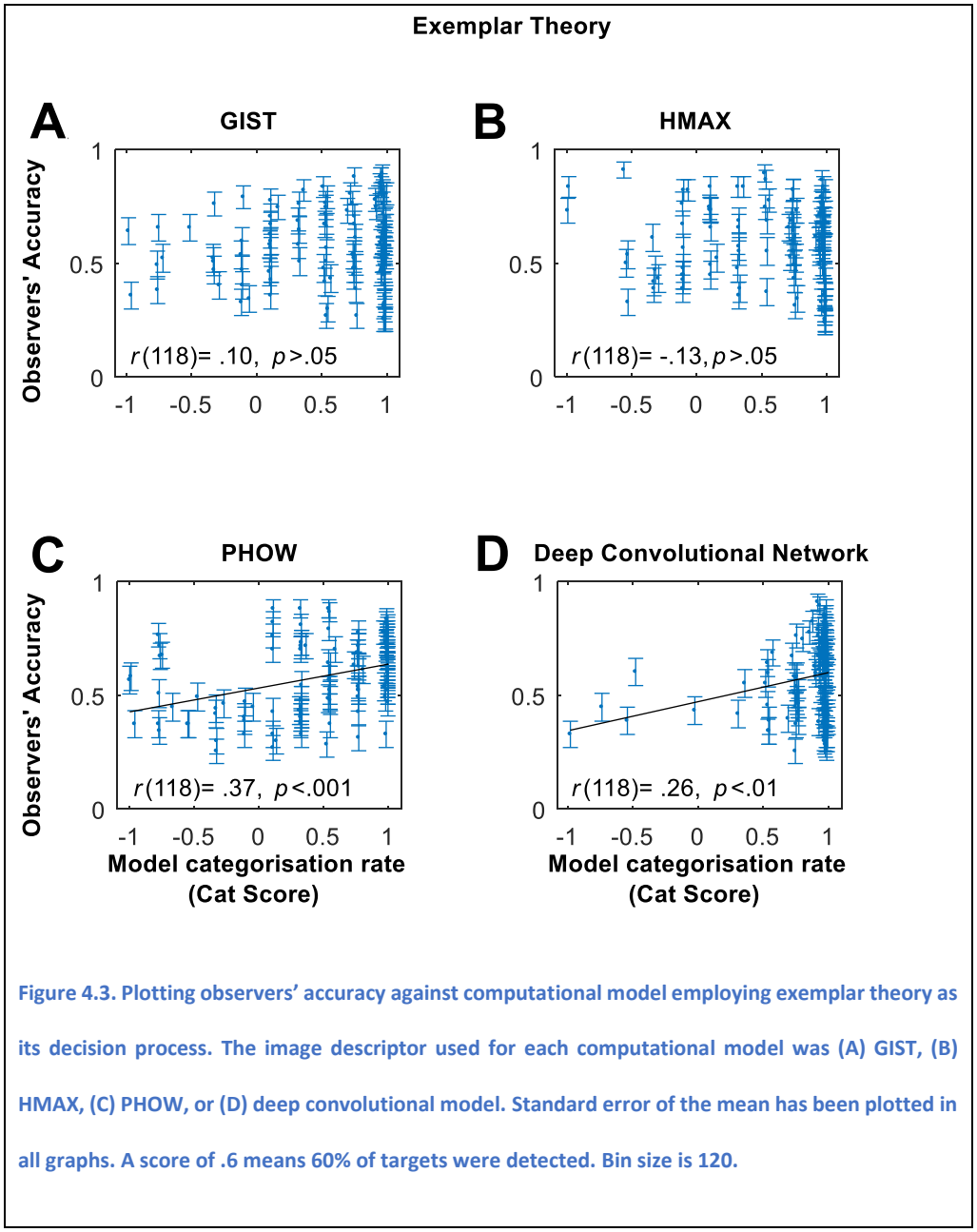


Figure 4.3. Plotting observers' accuracy against computational model employing exemplar theory as its decision process. The image descriptor used for each computational model was (A) GIST, (B) HMAX, (C) PHOW, or (D) deep convolutional model. Standard error of the mean has been plotted in all graphs. A score of .6 means 60% of targets were detected. Bin size is 120.

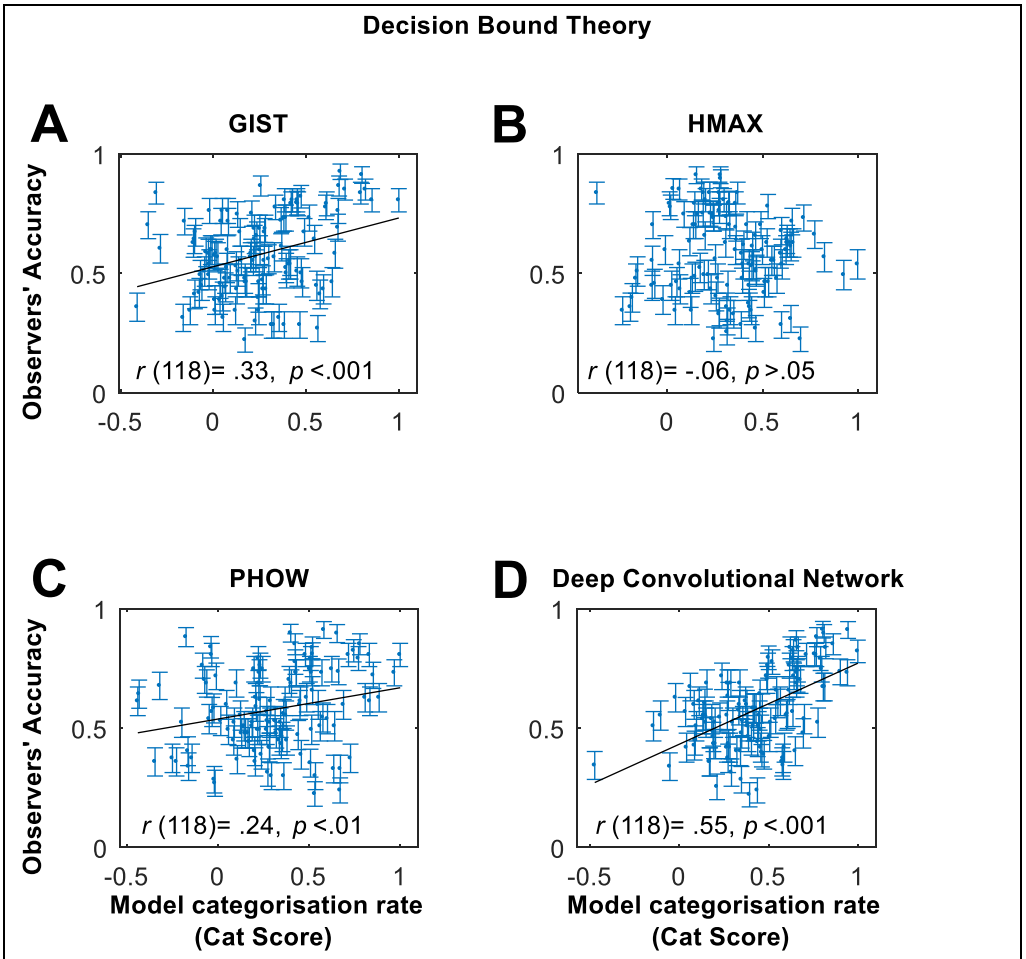


Figure 4.4. Plotting observers' accuracy against computational model employing decision bound theory as its decision process. The image descriptor used for each computational model was (A) GIST, (B) HMAX, (C) PHOW, or (D) deep convolutional model. Standard error of the mean has been plotted in all graphs. A score of .6 means 60% of targets were detected. Bin size is 120.

4.3. General Discussion

The main aim of this chapter was to examine the similarity of different computational decision processes to the decision process employed by biological vision in an image categorization task. Previous research in this area has traditionally applied computational decision process onto image

descriptions created on human interpretations (Lamberts, 2000; Posner & Keele, 1968, 1970; Reed, 1972). These do not reflect the known first stages of biological visual processing (Hubel & Wiesel, 1962, 1968). Here we test computational decision processes with image descriptions which are designed to resembling biological image descriptions, in order to determine if one decision process out performs the others (Khaligh-Razavi & Kriegeskorte, 2014; Rice et al., 2014; Watson et al., 2014).

Three different decision processes were examined; prototype, exemplar and decision bound theories. These decision processes were combined with four computational image descriptors, GIST, HMAX, PHOW, and a deep supervised convolutional network.

Each of the decision processes investigated here produced consistent trends across each of the image descriptors it was paired with. Prototype theory consistently demonstrated it was able to explain a significant proportion of the variance in human observer's behavior in terms of accuracy on target present trials (hit rate) and reaction times. This, however, was not in the direction that would be expected; images that were *further* from the category prototype were *easier* it is to identify as belonging to that category. This finding could be likened to the idea that atypical images stand out and so are easier to correctly categorize. Additionally, this result could be because all of the images tested were reasonably close to the prototype; images were selected on the basis of clearly belonging to one category or the other. In the experiment presented we

examined the ease with which an image could be categorized into its own category. If images outside of the primed image category could be included in the analysis then the effect might disappear. It was not possible to test this hypothesis as an RSVP procedure was used and it wouldn't be possible to determine which image observers were responding to when the target image wasn't present.

Exemplar theory was able to explain a significant proportion of observers' behavior when paired with the majority of image descriptors. Surprisingly, exemplar theory performed especially well when paired with PHOW, a computational image descriptor which previously in this thesis had shown little evidence to match biological vision. Additionally, when exemplar theory was paired with GIST, a computational image descriptor known to match biological image description processes (Rice et al., 2014; Watson et al., 2014), it failed to explain observers' accuracy. As previously mentioned, comparisons between computational models and human observers are unable to distinguish between a computational model which is performing the same algorithms as biological vision, and a model which has a reasonable performance, and so correlates with human behavior, but is ultimately performing calculations in a different way. There are a number of reasons here why the latter is the case. Chapter 3 examined which image descriptors best fit biological vision. The order the image descriptors fit biological vision, from best to worse was, deep convolutional network, GIST, HMAX, PHOW. However, exemplar theory found a different trend, PHOW, deep convolutional network, GIST, HMAX, suggesting

that its calculations may be differing from biological visions. Examining Figure 4.3 also demonstrates that a lot of the bins are clustered at the top end of the computational models' *Cat* score, showing that it has a high correct categorization rate. Because of this relatively few bins are spread out to the lower end of the *Cat* scores showing that correlations are being driven by relatively few bins. This is especially highlighted in the case of the deep supervised convolutional neural net. Here Exemplar theory is summarized as a decision process which works, but it probably differing to the one employed by biological vision.

Decision bound theory explained the most variance in the behavioral data, with a significant correlation with three out of the four image descriptors. Performance across the image descriptors form best to worst was deep supervised convolutional model, GIST, PHOW and HMAX. This followed largely the same trend as in Chapter 3 in which the image descriptors, rather than decision processes, were the focus of the study. Interestingly, decision bound theory produced significant correlations when paired with PHOW, which performed poorly in previous experiments. An explanation of this finding is that PHOW is an image descriptor which was originally created for the purpose of image classification. PHOW therefore works a lot better for the purpose of image classification than image recognition.

There have been a number of studies over recent years that decision bound theory provides a good approximation to the mechanism biological vision is

using to categorize images. Two studies have used decision bound theory paired with GIST in order to explain observer's behavior in image categorization tasks (Mack & Palmeri, 2010; Sofer et al., 2015). A number of studies have also tried to predict human behavior based upon measured brain activity. These studies have shown that decision bound theory works extremely well at predicting observers image categorization behavior based on MEG (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Ritchie, Tovar, & Carlson, 2015) and fMRI (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014). All of this research supports the notion that when image descriptions approximate those used by biological vision decision bound theory provides a good account for image categorization in human observers (Ritchie & Carlson, 2016).

While decision bound theory works well, there may be other, more complex, decision processes which outperform it. Zhang, Berg, Maire, & Malik, (2006) showed that a decision process which utilized both exemplar theory and decision bound theory principles had a higher categorization rate than either of the two theories alone. It could be possible that biological vision is performing a process similar to decision bound theory, but the exact nature of the decision process may be slightly different.

It is also important to note that the categorization literature has shown that the mechanism observers are using to categories images is likely to be task dependent (Konishi et al., 1999; Lombardi et al., 1999; Rao et al., 1997). For example, this is particularly pronounced in task which use categories which can

be separated by verbal rules versus categories where no verbal rules exist to separate them. The majority of studies in the recent decade have used natural images and so almost all of the studies fall under the category of tasks which have no clear cut rules for categorization (Mack & Palmeri, 2010; Sofer et al., 2015). It could be that decision bound theory works particularly well for these kinds of studies, but much simpler mechanisms are being employed in rule based tasks. It would be interesting to apply the methods here, image descriptions used which reflect biology, to rule based tasks in order to determine if decision bound is still the optimal strategy.

In this study observers were restricted, until the RSVP procedure had finished, before they were able to respond. In previous studies when this was the case computational models failed to be able to explain observers' reaction times. Here, however, computational models are able to explain a significant proportion of observers' reaction times. The results of the reaction time data follow closely with the accuracy data. This would suggest that in some circumstances, even when observers are delayed in their responding, reaction times can reveal the inner processing of observers.

Previous studies examining computational models similarity to observers have compared at a general level; overall accuracy for a category, examining a single decision process or a single image descriptor (Mack & Palmeri, 2010; Serre et al., 2007; Sofer et al., 2015). In this study a rigorous comparison between computational models and human observers was made. Multiple decision

processes and image descriptors were examined. Comparisons were made at the fine detailed level of each trial. Reaction times as well as observers' accuracy were compared to the computational models. Additionally, all comparisons were made on a single data set, allowing comparisons between computational models to be straightforward. All of this had the advantage of painting a broad picture which reveals results which would not have been shown by individual studies examining small elements of the whole picture (e.g. although exemplar theory predicted a significant amount of observers' behavior, it is unlikely to be the algorithm biological vision is using to categorizing images in this experiment). While studies examining single elements can reveal important information, studies examining multiple elements are crucial to understanding the puzzle that is the human visual system.

Chapter 5 -Investigating the effect of image set

5.1. Introduction

Computational models and human observers naturally differ in the image sets on which they were trained. In the current chapter, we investigated the extent to which this natural difference could account for differences in their behavior.

Computational models are trained on a finite (albeit increasingly large) image set, which usually ranges from between 1,000 to 1.2 million images (Fei-Fei et al., 2007; Russakovsky et al., 2015). These are generally images that are photogenic; long shots including the whole object or scene. Human observers, on the other hand, have had a life time to accumulate their image set and so have access to a much larger range of images (Gibson, 1969; McGraw, Webb, & Moore, 2009), and environmental conditions (e.g. fog).

The difference in image sets used by the computational models and human observers is likely to cause them to respond differently; they are making decisions based on different information. This poses a problem when trying to determine the similarity between a computational model and human observers, this is illustrated by the following example. Imagine a computational model which is identical to human observers in *image description* and *decision process,* but is using a vastly different image set. Even though the computational model is performing the same algorithms as the biological system the output behavior would be different.

In order to obtain a better measure of the similarity between computational models and human observers, the gap between the two image sets needs to be closed. This could be done by increasing the variety and number of images that computational models are trained on. This is naturally happening over time as image sets become larger (Russakovsky et al., 2015), but that doesn't solve the problem that images taken by photographers are likely not to be the same as the scenes naturally encountered by the eye. An alternative approach could be to train the *observer* on the image set used by the c*omputational model*. This method would employ observers' natural ability for perceptual learning (Goldstone & Hendrickson, 2010; Werker & Tees, 2002), in which human observers' perceptual system naturally adapts to better discriminate stimuli categories with which it is presented. Training observers on the image set used by the computational model ensures that the human observers have had access, and a chance to optimize, to the same images the computational model is using. By training observers on the image set the computational model is using, the observers' image statistics are likely to be steered to be more like the computational model.

Here we aim to investigate if human observers can be made to respond closer to the computational model by training them on the image set used by the computational models. Additionally, by using this method it is possible to gauge the extent to which the intrinsic differences in image sets, between computational models and human observers, influence the similarity of their behaviour.

An experiment of three phases is presented here, an initial testing session (pre-training), followed by 8 training sessions in which the participants were repeatedly exposed to the image set, and then a final testing session (post-training). Pre-training sessions give an approximation of the initial similarity between the computational models and humans. Post-training sessions measure the similarity of the computational models to humans, after training. The results of the pre- and post-training sessions can then be compared against each other to determine the effect that training had on their similarity. We used separate experimental designs for testing sessions and training sessions so that any improvement in models' ability to predict human behavior can be attributed to a closing of the gap between the two image sets, rather than familiarity with the task.

5.2. Methods

5.2.1. Observers

Twelve Nottingham University students (seven female; age 18-24 years) took part in the experiment. All were volunteers who were given an inconvenience allowance. All signed a consent form and all procedures were approved by the University of Nottingham Psychology ethics committee.

5.2.2. Apparatus

The experiment was programmed in PsychoPy (Peirce, 2007), and was run on a Lenovo desktop with 3.7 GHz, Intel Xeon E5-1620 v2 processor and NVIDIA

NVS 310 graphics card. The viewing distance was held constant with a chin rest at 57cm from the monitor screen. The monitor was a Iiyama ProLite GB2488HSU set to a 1920 x 1080 resolution with an 144Hz refresh rate. Images for each trial were loaded onto the graphics card during the inter-trial interval. Timing of all briefly-presented stimuli (e.g. in the rapid serial visual presentation (RSVP) task) was controlled by presenting the stimuli for a fixed number of screen refreshes. We verified that the system reliably presented these stimuli within an RSVP sequence without dropping any frames.

5.2.3. Design and Procedure

The experiment consisted of ten sessions, each taking place on separate days and all of them taking place within three weeks (allowing participants a certain level of flexibility, while ensuring a degree of consistency in gaps between sessions). The ten sessions were split into three phases; Phase 1: one initial testing session (referred to as pre-training), Phase 2: eight training sessions and Phase 3: one final testing session (post-training). The task in the testing sessions followed a Yes/No RSVP procedure and, the task for the training sessions was a simple Yes/No task with single images rather than an RSVP presentation. This allowed us to compare the model with performance on two different tasks that placed quite different temporal constraints on the observers. The different pools of images used as the image set in this experiment are described in Chapter 2 - Image Set.

Testing Sessions (pre- and *post-training).* Each session consisted of 480 main trials. Each trial began with a fixation cross which lasted 500 ms. This was then followed by a text prime, which was the name of one of the four image categories (Buildings, Mountain, Ocean or Trees). The text prime was presented on the screen for one second. Participants then viewed an RSVP sequence of 6 pictures presented for 10 screen refreshes (69.4 ms at our 144 Hz refresh rate) per image. Participants had to report whether any of the images were of the category that was primed. There was a 50% chance on any trial that a target image matching the prime was present. Target images could not appear as the first or last image in the RSVP sequences, but could appear in the serial positions 2, 3, 4, or 5. Target position was balanced over the trials. The target image came from the *target* pool of images, while the other images in the RSVP sequence were from the *mask* pool of images. After the RSVP sequence participants were presented with text, reminding them of the prime and asking if they saw the corresponding image. This text remained on the monitor until the participants responded, by pressing arrow keys on the keyboard to indicate if they had seen an image matching the primed image category. If the participant responded slower than 500 ms text was displayed on the screen requesting a faster response. During the first testing session participants were given a practice block of 20 trials, which consisted of images from only the *mask* pool of images. This initial practice session was to make sure that observers understood the task. A diagram illustrating the trial structure of testing sessions is shown in Figure 5.1.

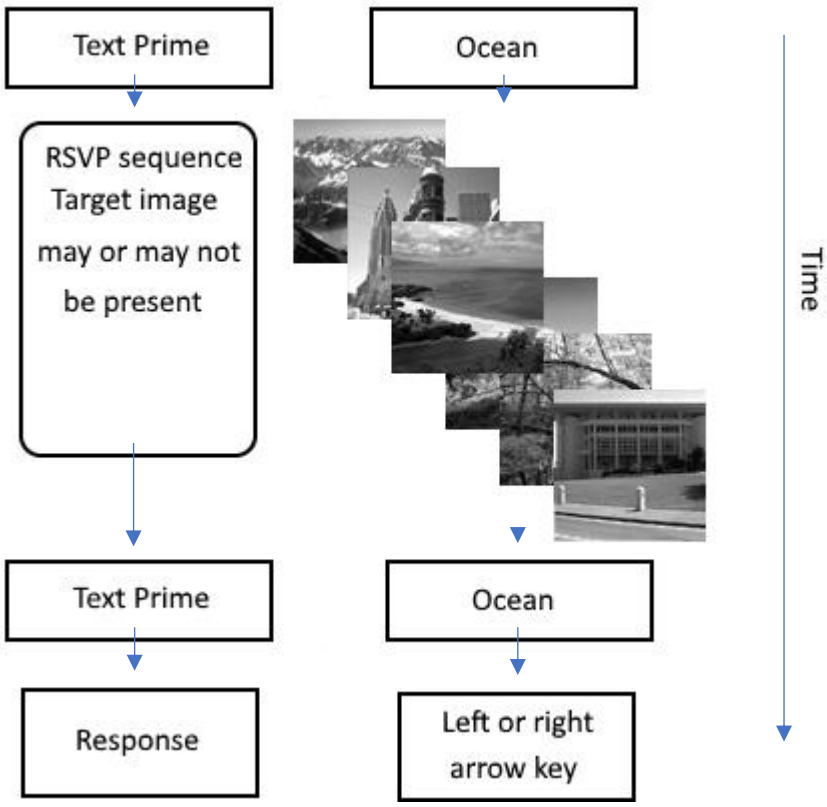


Figure 5.1. A diagram showing the trial structure of the testing session. On the left a verbal description is given, while on the right a visual example is shown.

Training Sessions. The training sessions were deliberately set up to use a different task, so that any effect of training would be caused by changes in visual perception of the images rather than procedural learning where the participants had improved their ability to attend to or process the RSVP sequences. The key difference was that there was no RSVP presentation; just a single image presented for a single frame (6.95 ms). Each trial began with a fixation cross lasting 500 ms. This was then followed by a text prime, which was the name of one of the four image categories (Buildings, Mountain, Ocean or

Trees). The prime lasted on the screen for one second. A single image was presented for 1 frame (6.94 ms). The screen was then left blank until the participant responded. If the participant responses were greater than 500 ms then text asking the participant to respond faster was displayed. Every 100 trials participants received a screen showing the number of trials they responded to correctly and their reaction time, both of which stayed on the screen until they pressed a key to move on. This form of feedback was given to encourage performance and as a way of allowing participants optional breaks. Images used in this session came from the *mask* pool of images; the pool of images the computational models were using to base their decisions. Images from the *target* pool of images was not used as we wanted to limit exposure the observers had to these images to keep them novel; ensuring observers had to calculate their response rather than using memory. Each session consisted of 425 main trials. A diagram showing the trial structure for training sessions is shown in Figure 5.2.

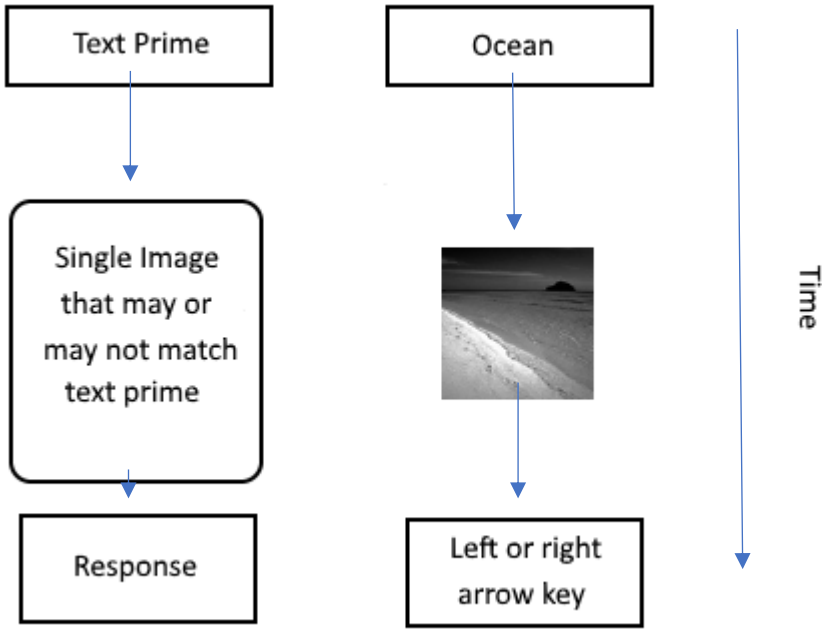


Figure 5.2. A diagram showing the trial structure of training sessions. On the left a verbal description is given, while on the right a visual example is shown.

5.3. Results

Any trials in which observers responded slower than 500 ms were excluded from the analysis (4.2% of trials) and participants were warned on such trials that they should respond faster. This exclusion criterion was used to encourage participants to respond very rapidly. The results of the 3 different phases of the experiment (*pre, training* and *post)* are presented separately, due to the difference in their methodologies. The descriptive statistics of the observers' performance in testing and training phases of the experiment are presented first, then the results are examined to see if observers' performance increased

through the sessions as a result from training, finally observers' performance in the different phases are compared to the output of the computational models.

Figure 5.3 presents the descriptive statistics of the observers' behavior from both testing sessions, while Figure 5.4 presents the descriptive statistics of the observers' behavior in all the training sessions.

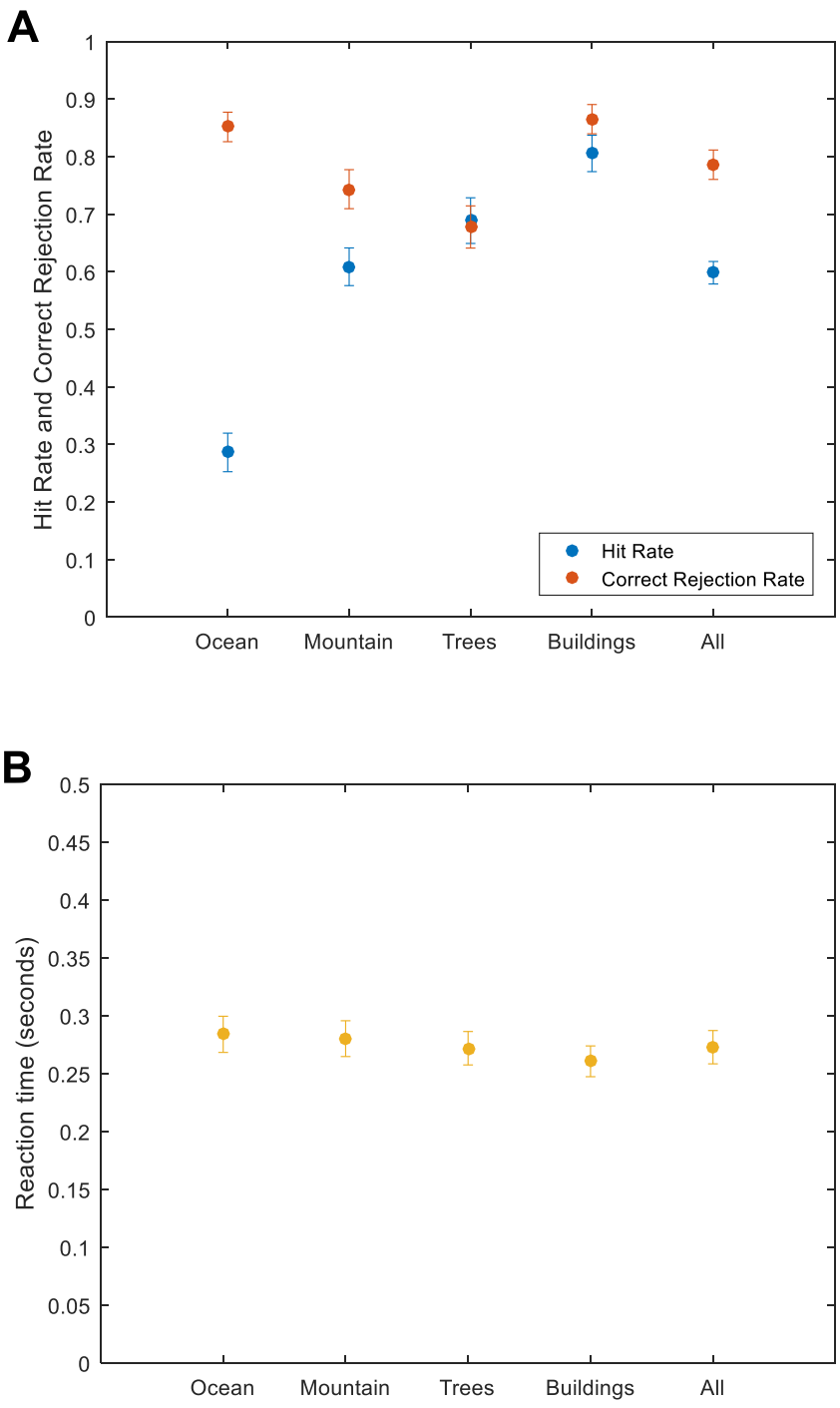


Figure 5.3. Observers performance in both the testing sessions combined in the four different image categories as well as when they are all pooled together. A) plots hit rate and correct rejection rate, while B) plots observers' reaction time (in seconds). Error bars shown are the standard error of the mean.

Three separate 1x4 repeated measures ANOVAs were run for the dependent measures of hit rate, correct rejection rate and reaction times respectively in the pre- and post-training sessions combined. This was done to see if the dependent variables varied across image category. Observers' hit rate, correct rejection rate and reaction times was shown to vary significantly across image category ($F(3,33) = 45.017$, $p < .001$, $\eta_p^2 = .80$, $F(3,33) = 20.068$, $p < .001$, $\eta_p^2 = .65$, $F(3,33) = 14.283$, $p < .001$, $\eta_p^2 = .57$, respectively).

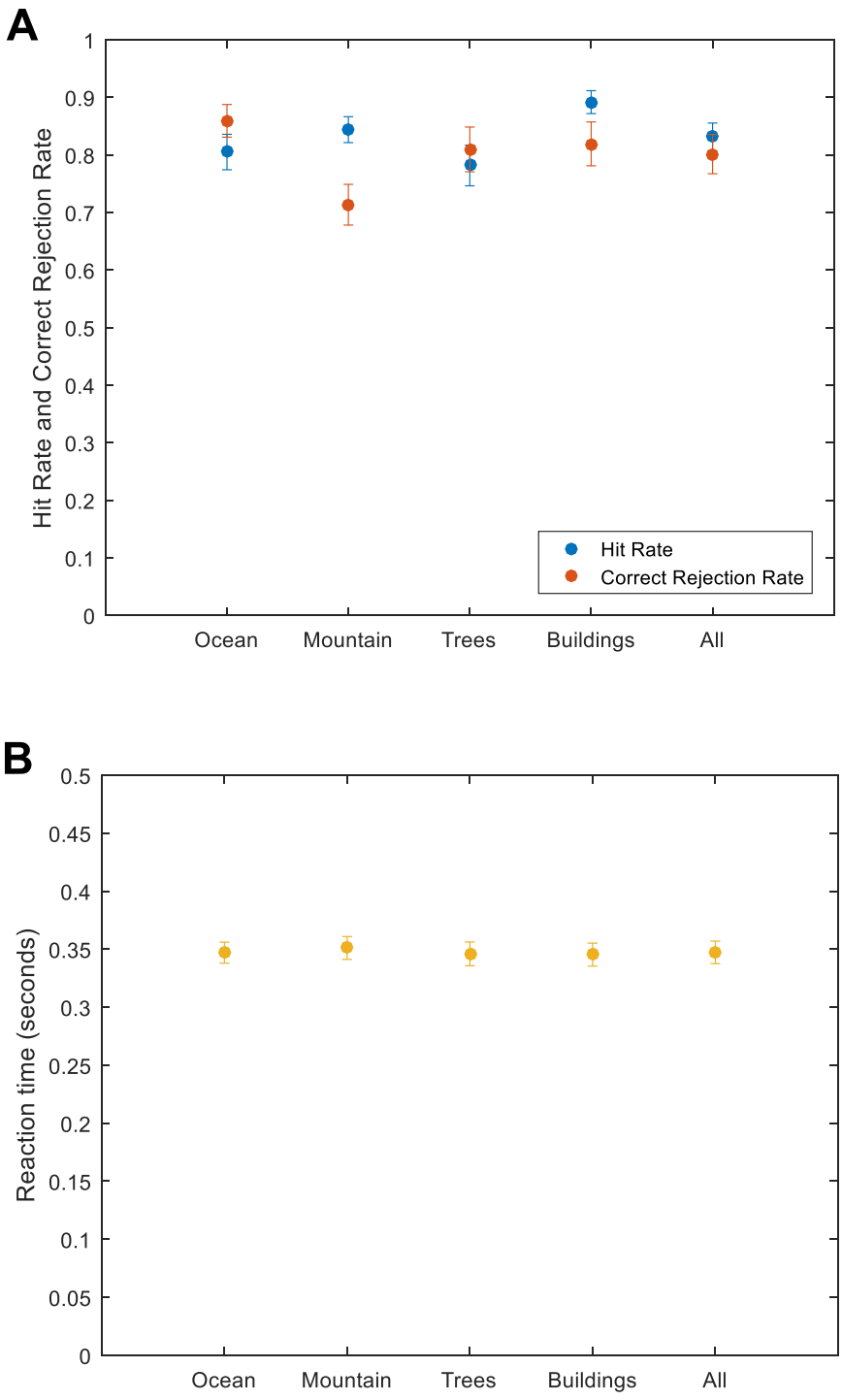


Figure 5.4. Observers performance in all the training sessions combined in the four different image categories as well as when they are all pooled together. A) plots observes' hit rate and correct rejection rate, while B) plots observers' reaction time (in seconds) are plotted. Error bars shown are the standard error of the mean.

Three separate 1x4 repeated measures ANOVAs were run for the dependent measures of hit rate, correct rejection rate and reaction times respectively in the training sessions combined. This was done to see if the dependent variables varied across image category. Observers' hit rate, correct rejection rate and reaction times was shown to vary significantly across image category ($F(3,33) = 9.188$, $p < .001$, $\eta_p^2 = .46$, $F(3,33) = 29.982$, $p < .001$, $\eta_p^2 = .73$, $F(3,33) = 3.296$, $p = .032$, $\eta_p^2 = .23$, respectively).

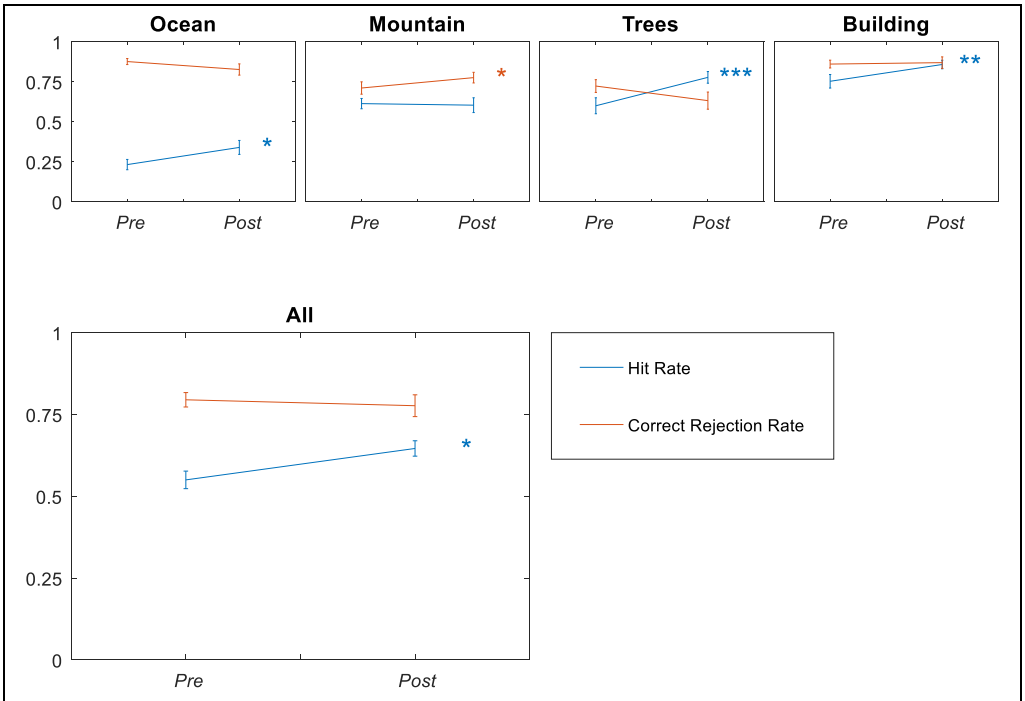


Figure 5.5. Observers' performance in the pre-training and post-training testing sessions. Each image category is shown separately and also when they are all grouped together. Hit rate and correct rejection are plotted together on the same y-axis. Significance markers are colour coded to the lines and plotted next to any significant differences (p<0.05: *, p<0.01:, p<0.001:***). Error bars shown are the standard error of the mean.**

To determine whether training had any impact on performance we conducted paired- sample T-tests comparing hit rate, correct rejection rate and reaction times in the *pre*- and *post*-training session. On average, observers had significantly better hit rates in the *post*-training session (M= 64.61, SE= 2.35) than in the *pre*-training session (M= 55.01, SE= 2.68), when all the image categories were pooled together ($t(11)$= -3.02, p <.05, d= 1.10). Breaking this down into categories reveals, on average, significantly better hit rates in the *post*-training session in the category of ocean (M= 33.96, SE= 4.36), trees (M= 77.74, SE= 3.64) and buildings (M= 85.78, SE= 2.66) when compared to their

pre-training session counterparts, ocean (*M*= 23.23, *SE*= 3.22), trees (*M*= 59.99, *SE*= 4.99) and buildings (*M*= 75.30, *SE*= 4.22) (ocean, $t(11)= -2.88$, $p< .05$, $d= .81$; trees, $t(11)= -4.84$, $p< .001$, $d= 1.17$; building, $t(11)= -3.33$, $p< .01$, $d= .86$). On average, observers had significantly better correct rejection rates in the *post*-training session in the category of mountain (*M*= 77.58, *SE*= 3.33) when compared to the *pre*-training session (*M*= 71.01, *SE*= 3.86), $t(11)= -2.64$, $p<.05$, $d= .52$). No significant differences were observed in the reaction time data (which is therefore not plotted in Figure 5.5 or Figure 5.6).

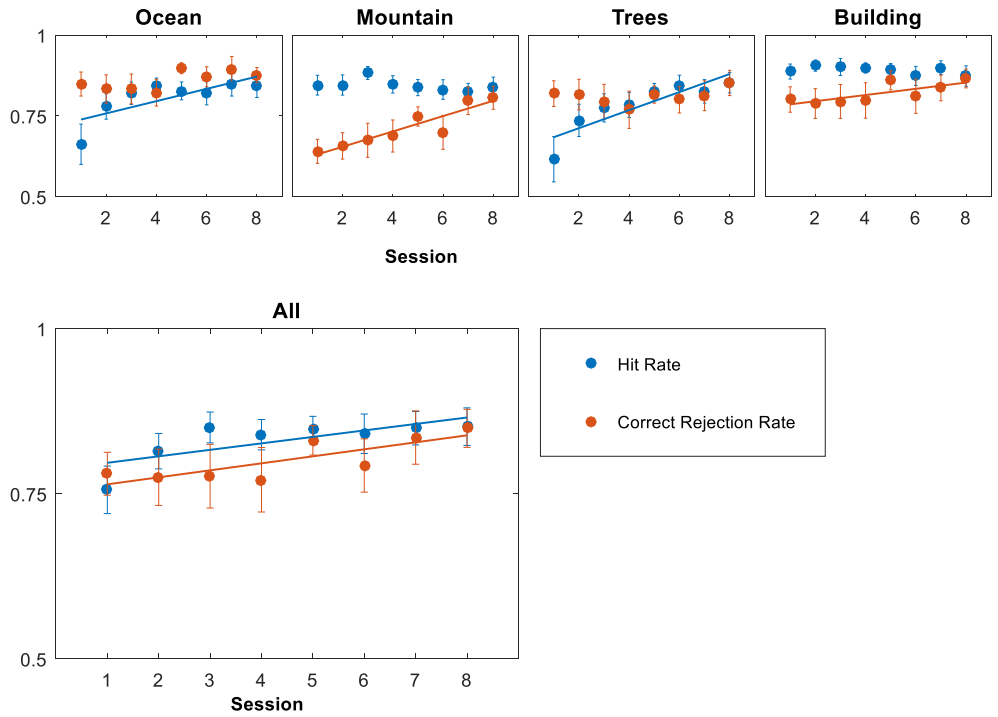


Figure 5.6. Observer's performance in the training sessions. Each image category is shown separately and also when they are all grouped together. Hit rate and correct rejection rate are shown as significant changes are observed. Reaction times are not displayed as no significant changes were observed. Significant regressions are plotted as trend lines. Error bars shown are the standard error of the mean.

Linear regressions were calculated to examine if any change in hit rate, correct rejection rate or reaction time occurred across the training sessions. This can be seen in Table 5.1. The results show that when all the image categories are pooled together that observers hit rate and correct rejection rate increase over the sessions. When this effect is dissected by image category it seems that each image category had its own pattern of increased performance over the training sessions. In some cases (trees and oceans) observers learnt to better detect

images that belonged to those categories. For other categories (mountains and buildings) observers learnt to better identify images which were not of that image category. In all image categories performance increased over the sessions.

Table 5.1. The linear regressions conducted on the human observers' data in the training sessions. Positive regressions are highlighted in green.

		R^2	slope	p Value	Intercept
Ocean	Hit rate	.57	.019	.031	.72
	Correct Rejection Rate	.47	.0081	.061	.82
	Reaction Time	.39	-.0016	.10	.35
Mountain	Hit rate	.27	-.0038	.18	.86
	Correct Rejection Rate	.86	.024	.0010	.61
	Reaction Time	.011	0.00031	.81	.35
Trees	Hit rate	.78	0.028	.0036	.66
	Correct Rejection Rate	.12	0.0033	.40	.79
	Reaction Time	.083	-.00065	.49	.35
Buildings	Hit rate	.26	-.0026	.20	.90
	Correct Rejection Rate	.59	.0096	.026	.78
	Reaction Time	.091	-.00076	.47	.35
All	Hit rate	.54	.0098	.038	.79
	Correct Rejection Rate	.67	.011	.013	.75
	Reaction Time	.090	-.00073	.47	.35

One aspect to note is that the current study used two different tasks for the testing (pre-and post-training) and the training sessions. In the former an RSVP task was used in which the rapid presentation of stimuli result in substantial masking form one to the next. In the training sessions stimuli were presented for brief periods but not masked. This influenced hit rate, which was worse in both testing sessions than in the training period and varied by image category.

Computational models were constructed for four image descriptors (GIST, HMAX, PHOW, deep convolutional neural net) all paired with decision bound theory, keeping with the findings from Chapter 4. For further details on how comparisons between computational models and human observers were made see Chapter 2 - General Methods.

The results are presented in Table 5.2. Correlations highlighted in green are in the direction predicted by the model.

Table 5.2. The results of the various different computational models when Observers' accuracy (on target present trials) or reaction times is regressed against them. Significance markers are presented next to any significant differences (all uncorrected for multiple comparisons). Green shading indicates significant correlations in the direction the computational model predicted.

	Pre-training (22-23 trials in each bin)				Training (174-175 trials in each bin)				*Post*-training (23-24 trials in each bin)			
	GIST	HMAX	PHOW	DCN	GIST	HMAX	PHOW	DCN	GIST	HMAX	PHOW	DCN
Decision Bound Theory (Accuracy)	*p*= .026 *r*= .20	*p*= .637 *r*= .04	*p*= .215 *r*= .11	*p*< .001 *r*= .34	*p*= .035 *r*= .19	*p*= .038 *r*= .19	*p*= .503 *r*= .06	*p*< .001 *r*= .43	*p*= .043 *r*= .19	*p*= .740 *r*= .00	*p*= .081 *r*= .16	*p*< .001 *r*= .35
Decision Bound Theory (Reaction time)	*p*= .29 *r*= -.10	*p*= .783 *r*= .00	*p*= .12 *r*= -.14	*p*= .013 *r*= -.23	*p*< .001 *r*= -.33	*p*= .062 *r*= -.17	*p*= .206 *r*= -.12	*p*< .001 *r*= -.47	*p*< .001 *r*= -.34	*p*= .959 *r*= .00	*p*= .026 *r*= -.21	*p*< .001 *r*= -.48

The data in Table 5.2 demonstrates a number of different computational models were able to significantly predict observers' behavior within each of the 3 phases of the experiment; pre-training, training and, post-training.

In Figure 5.7 we examine which of the image descriptors best approximates human observers' behavior. These graphs show behavioral data taken from the observers' training sessions and plotted against each of the image descriptors paired with decision bound theory. Three of the four image descriptors were significantly correlated with human performance. The strongest relationship was seen for the deep convolutional network model descriptor.

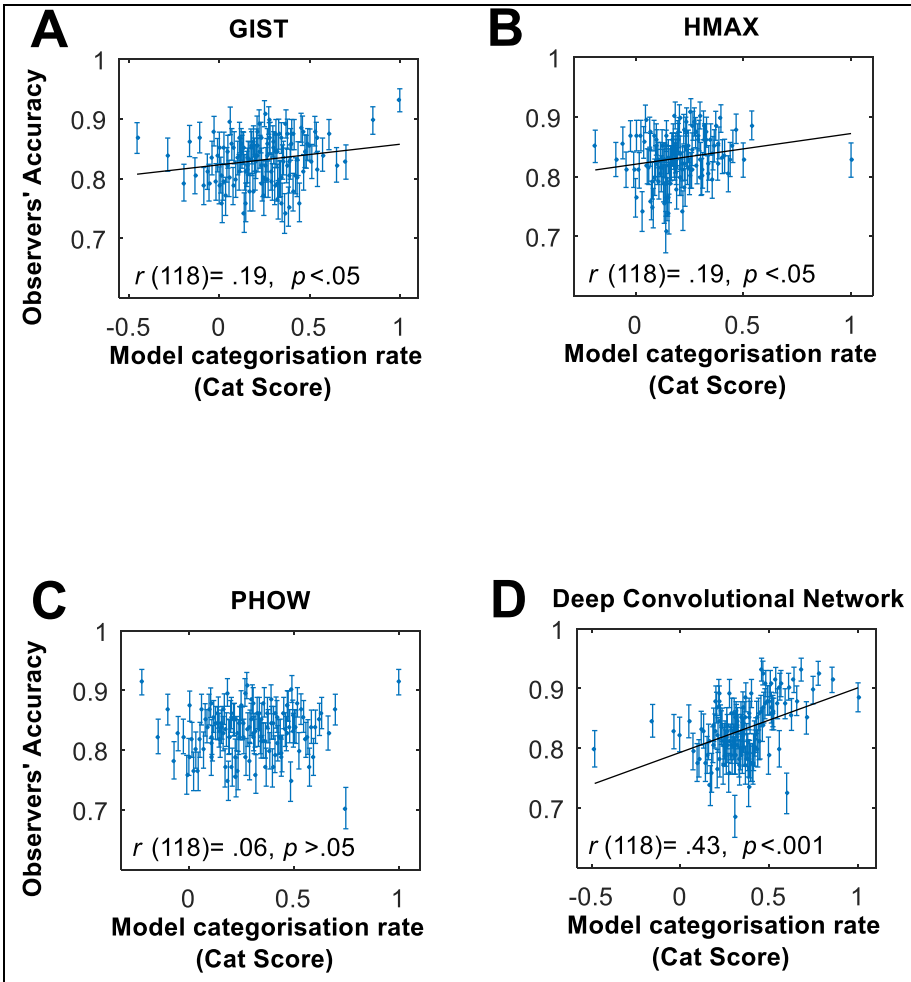


Figure 5.7. Plotting observers' behavioural data in the training sessions against different models' performance comprised of a decision bound paired with different image descriptors; (A) GIST, (B) HMAX, (C) PHOW and (D) Deep convolutional network. Standard error of the mean has been plotted in all four panels. A score of .6 means 60% of targets were detected. Number of bins is 120.

The comparisons which show the clearest signs that training caused the observers to behave closer to computational models is seen in the two best image descriptors (Deep convolutional model and GIST), combined with a decision bound method, before and after training (see Figure 5.8), correlated against observers' reaction times.

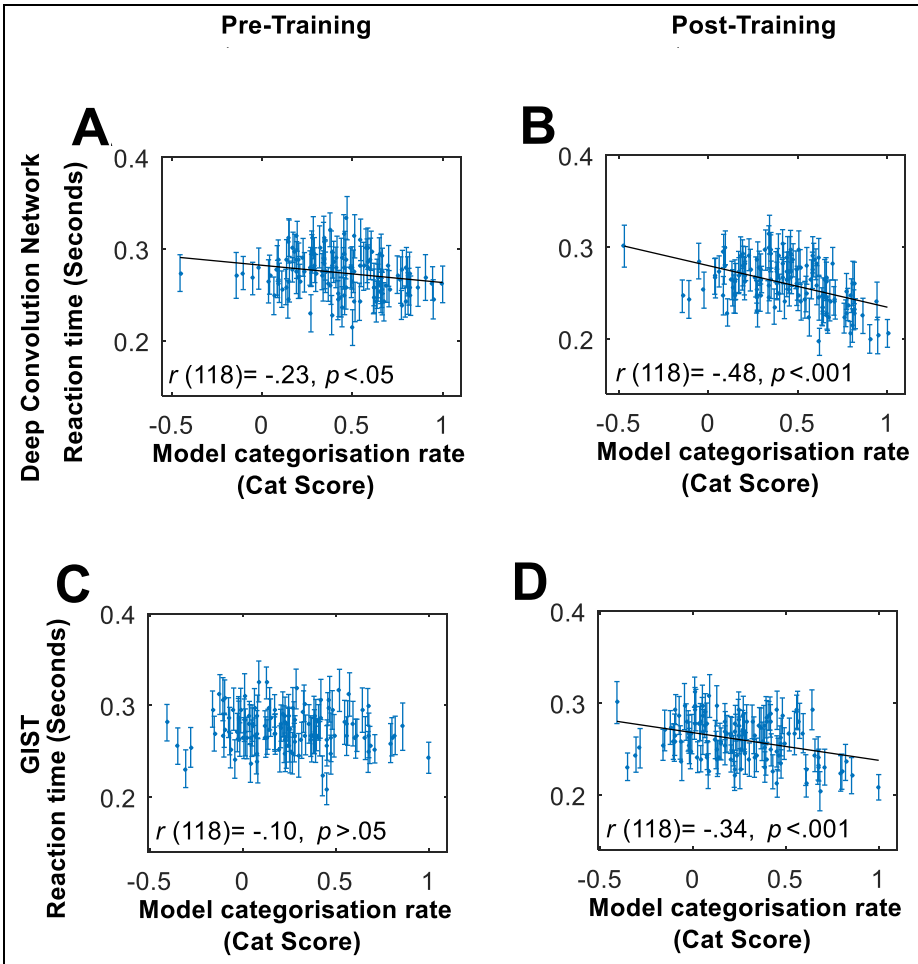


Figure 5.8. Comparing different models' ability to explain observers' behavioural data between pre-training and post-training sessions. Observers' pre-training session behaviour plotted against a models' performance comprised of a decision bound paired with (A) deep convolutional network and (C) GIST. Observers' post-training session behaviour plotted against a model comprised of a decision bound paired with (B) deep convolutional network and (D) GIST. Standard error of the mean has been plotted in all four panels, but the error bars are smaller than the data points in all cases. A score of .6 means 60% of targets were detected.

5.4. General Discussion

Here we investigated whether human observers could be made to respond closer to the computational models by training them on the image set the computational models were using. Additionally, we were interested in attempting to gauge the extent to which the intrinsic differences in image sets, between computational models and human observers, influence the differences of their behaviour.

Observers' performance (not compared to the computational models) across the training sessions and a comparison between pre- and post-training sessions indicates whether they learned during the experiment (perceptual learning). Observers' performance, in terms of hit rate and correct rejection rate, was shown to increase during the training sessions and the comparison between pre- and post-training session revealed that observers' hit rate increased significantly. Reaction times were seen to be stable across the training sessions and when the pre-training session was compared to the post-training session. The increase in observers' performance, in terms of hit rate and correct rejection rate, is a good indicator that observers could learn the image set and increase their performance. The absence of any significant change in reaction times, may have been due to a ceiling effect caused by the experimental design; observers were prompted to respond faster than 500 ms.

Comparisons over the three phases in the ability of the computational models to predict observers' performance, suggests that observers' performance

became closer to the computational models through training. The maximum significant correlation coefficient was smaller in the pre-training session than in training or post-training. This suggests that computational models, on the whole, found it easier to predict observers' performance in the training sessions and the post-testing session than in pre-training. On a closer examination of this result, looking at pre- and post-training sessions specifically, the increase in models' performance to predict observers' behavior is located in the domain of reaction times and not observers' accuracy. The change in computational models' ability to explain observers' reaction times are displayed in Figure 5.8. This change is moderately large, with an increase of around .20 correlation coefficient with the best performing image descriptors (GIST and the deep convolutional neural net). Observers' accuracy, and not reaction times were seen to change due to training. It is therefore surprising that computational models showed a greater ability to explain observers' reaction times and not their accuracy data after training. This result would suggest that although no change in average reaction times occurred due to training, this didn't mean that changes were not happening at a much finer level in observers' reaction times, reflected here by the computational models' ability to better predict them. It is possible that due to the small participant size that no change in models' ability to explain observers' accuracy was found, with an increased participant sample size and a greater number of trials it is possible that computational models may be able to better predict observers' accuracy data after training.

From the study presented here a slightly different message on which computational image descriptors best explain human behavior was observed. The deep supervised convolutional network still provided the closest fit to observers' behavior, in terms of accuracy and reaction times. This was again followed by GIST, as the second-best image descriptor at explaining human behavior. However, HMAX and PHOW were shown to have roughly similar abilities to predict human behavior, both only finding one significant correlation in the accuracy data in training sessions and reaction times in post-training, respectively. A possible explanation for this effect is the large reduction in number of participants. Previous studies presented in this thesis used participant sizes of 40-70, while this experiment used only 12 as it was a longitudinal study.

The deep supervised convolutional neural net's image description process has been trained on 1.2 million images from the ImageNet database (Russakovsky et al., 2015). This training has optimized its image description for performance on the image categorization task used in the 2012 ImageNet competition (Krizhevsky et al., 2012). Here, while the computational model did not have access to the 1.2 million images it did keep the optimization of its image description algorithm. The deep supervised convolutional neural net could explain a significant amount of observers' behavior in the pre-training data set, where the other image descriptors struggled. It could be that due to the image descriptors' prior exposure to a large image set it was still able to perform the task in a similar manner to humans. Khaligh-Razavi & Kriegeskorte, (2014)

showed that image descriptors which had undergone supervised learning in their image description process, compared to those that didn't, better approximated the structure of observers' image descriptions. Supervised learning offers potentially another mechanism by which the image set gap between image descriptors and computational models can be closed.

The results presented here suggest that as human observers become more familiar with the image set the computational model is using, the closer their behavior is to the computational model. At first glance, this would seem to pose a problem to the existing literature as studies often use no training or only a small number of training trials before beginning the main block of trials. However, this may not be the case. The results here show that while observers' behavior altered to become closer to that of the computational model, no change in which image descriptors, or the order in which model best approximates biological vision occurred. This would suggest that while researchers stand a better chance at detecting if a model is similar to human observers no change in the overall pattern of results is likely to occur. The results, and experimental design, presented here indicate that if the computational model has around 1000+ images it is likely to be sufficiently similar enough to humans to poses no major problem. However, researchers should still aim for larger image sets or allow for proper training of observers on the image sets used before conducting the experiment.

Chapter 6 - Investigating temporal blurring

6.1. Introduction

In the current chapter, we used computational models to investigate if, in an RSVP task, observers were experiencing temporal blurring during their image description process.

Several studies have shown that the human visual system accumulates a signal over time to form a single perception (Sweet, 1953; Westheimer & McKee, 1977). Consequently, if a stimulus is flashed on and off fast enough the visual system will perceive it as a single object (Hecht & Smith, 1936).

An image presented by itself needs a duration of around 20 ms to be correctly identified (Thorpe, Fize, & Marlot, 1996). If this image is masked in an RSVP sequence then the duration that each image needs to be presented increases to around 125 ms (Potter, 1975). Even though the minimum duration an image needs to be displayed in a RSVP increases due to masking, it is unclear if, at this duration, the visual system is able to form an image description which is not influenced by the masks either side of the target. It is possible that during a RSVP task images are temporally blurred together when forming a single image description.

The current chapter has two main aims. First, to investigate if computational models better predict observers' behavior when temporal blurring is included in their calculations. Second, if observers are experiencing temporal blurring

then how much more of the variance do the models explain once temporal blurring is included as a variable? If it is found that computational models better approximate observers' behavior when temporal blurring is included then this, as a method, could be used to study the integration window (time course and profile) of the image description process.

Here we re-analyzed behavioral data from the RSVP tasks in previous chapters; Chapter 3 - Experiment 2, which presented an image recognition task, and Chapter 4 - Experiment 1, which presented an image categorization task. In this reanalysis, we added a temporal blur component whereby the two neighboring mask images were added (in a variety of weights) to the target image, prior to forming the image description.

6.2. Experiment 1

6.2.1. Methods

The behavioural data comes from Chapter 3 - Experiment 2 which was a 2AFC image recognition task. For a description of the Observers, Apparatus as well as the design and procedure please see Chapter 3 - Experiment 2 - Methods.

Modeling temporal blurring

Temporal blurring was included into the computational model by presenting the model with a combined image of the *target* image and the *mask* images presented temporally either side of it. Temporal blurring of the target image with the mask images followed a simple function which was defined by extent

of the temporal blurring of each mask image. Calculation of the temporal blurring took a percentage of each mask image (B) and then added to a percentage of the target image (1-2*B), maintaining a total of 100%. This is illustrated by the example of temporal blurring value of 0.1; 10% of the luminance values of the forward and backwards mask are taken and added to 80% the luminance value of the target image.

Decision Process

Temporal Blurred Image Recognition. In the most general case the ability to identify which of two images looks most like a target is given by the differences in the distance between the sample and each of the images; if the difference in distances is great then the decision about which image is the target becomes easy. In the previous analyses the distance between the sample image and the target was zero (they were the same image) and so the measure here reduced to simply the distance between the sample and distractor. With the addition of blurring, which applies only to the sample (the image, as it was presented during the RSVP sequence) not to the target during the decision stage (it was presented for a prolonged duration with no masks at this point). Therefore, in the analysis here the predicted difficulty becomes the difference in distances between the target image with its blurred counterpart and the target image with the distractor image.

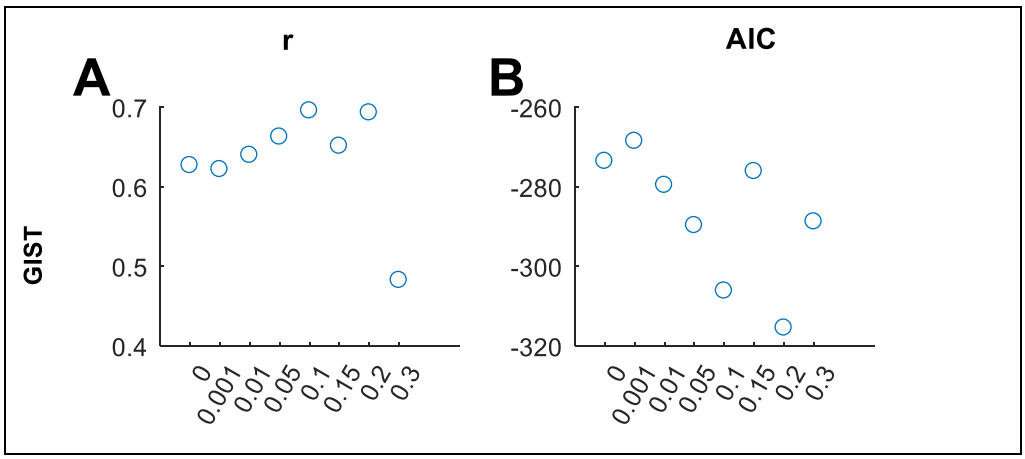
6.2.2. Results

Here the four Image descriptors (GIST, HMAX, PHOW and the deep supervised convolutional neural net) were paired with a decision process which had been adapted to accommodate temporal blurring. Original analysis of this behavioural data, described in Chapter 3 - Experiment 2, found that PHOW did not produce image descriptions which fit human observers. It is still included in this analysis to examine if PHOW can explain human observers' behaviour when temporal blurring is included.

As with Chapter 3 - Experiment 2, only target-present trials in which the observer responded correctly in the categorization task were analysed. This was to make sure that observers had seen the target image or else they would presumably be guessing for the image recognition task. Trials in which the observer took longer than two seconds to respond were also excluded from the analysis (2.4% of trials).

Several different temporal blur values were examined, ranging from 0.001 to 0.3. These values indicate the weight each mask (B) and can be used to determine the weight the target image was given (1-2*B). These values were chosen to cover a large proportion of the different values that were possible, as it was not known exactly which value would best approximate human observers.

As the blurring step introduces a new parameter into the computational model, we present the data in terms of both the correlation between the model's performance and observers' behavioural data and, also, in terms of the Aikaike Information Criterion (AIC) to account for the extra variable. AIC scores are used to examine if the increased number of model parameters is justified by the increase in variance explained (as a score becomes more negative it indicates a greater justification). The number of parameters used in calculation of AIC in the zero blur model is 2; the number of parameters for a linear correlation model. The number of parameters used in calculation of AIC in the models including temporal blur is 3; adding extent of temporal blur as another parameter. The results from the temporal blurring analysis are shown in Figure 6.1 and Figure 6.2 for accuracy and reaction times respectively.



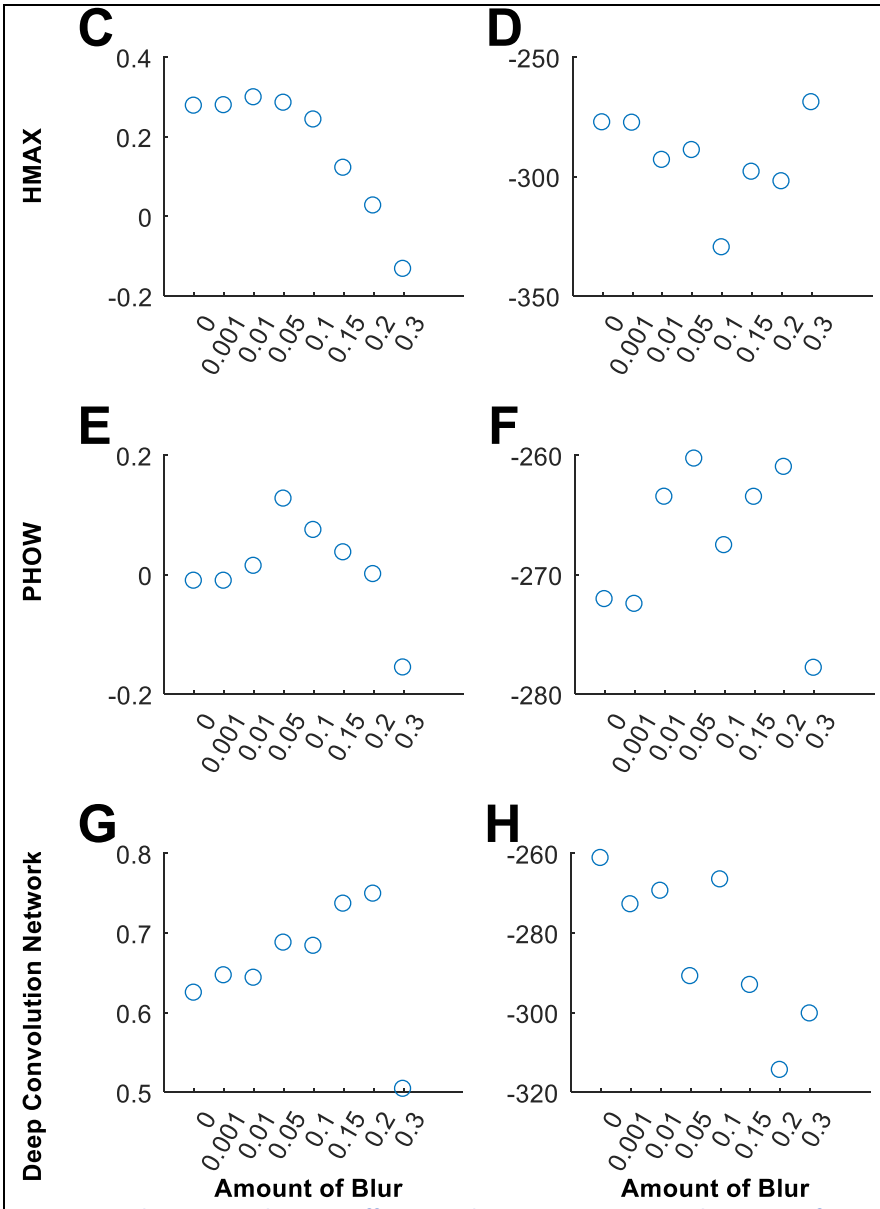
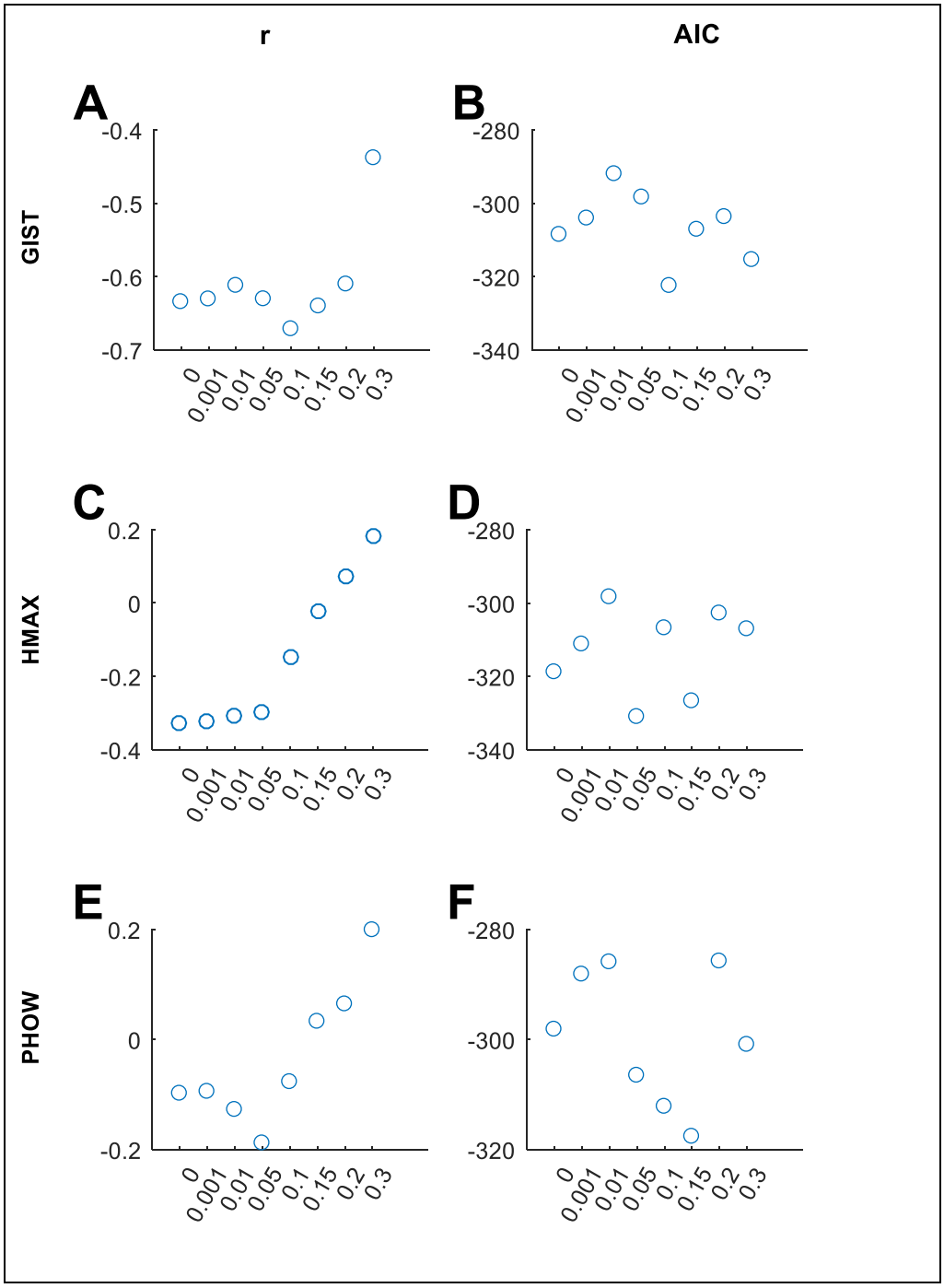


Figure 6.1. Plotting correlation coefficient and AIC criterion against the extent of temporal blurring for each computational model predicting observer's accuracy. Graphs on the left (A, C, E, G) plot variance explained in observers' accuracy data against extent of temporal blurring of the computational model, while graphs on the right (B, D, F, H) plot AIC criterion of observers' accuracy data against extent of temporal blurring. Each row represents a different image descriptor, from top to bottom, GIST (A, B), HMAX (C, D), PHOW (E, F) and deep convolutional neural network (G, H).


r
AIC
A
B
C
D
E
F
GIST
HMAX
PHOW
-0.4
-0.5
-0.6
-0.7
-280
-300
-320
-340
0.2
0
-0.2
-0.4
0
0.001
0.01
0.05
0.1
0.15
0.2
0.3

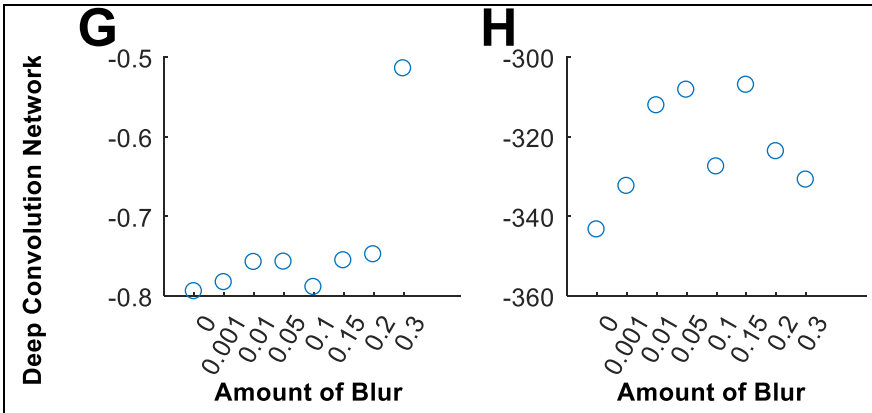


Figure 6.2. Plotting correlation coefficient and AIC criterion against the extent of temporal blurring for each computational model predicting observer's reaction times. Graphs on the left (A, C, E, G) plot variance explained in observers' reaction times data against extent of temporal blurring of the computational model, while graphs on the right (B, D, F, H) plot AIC criterion of observers' reaction times data against extent of temporal blurring. Each row represents a different image descriptor, from top to bottom, GIST (A, B), HMAX (C, D), PHOW (E, F) and deep convolutional neural network (G, H).

The results show that computational models which include temporal blurring tend to better explain human observers' behaviour. This suggests that observers were likely experiencing temporal blurring during the task. Almost all correlation coefficient (r) graphs imply an increased level of variance explained when blurring is considered in both observers' accuracy and reaction times. AIC further demonstrate this and show that the increase in variance explained at the cost of more parameters is justified.

6.2.3. Discussion

The results show that including temporal blurring in the computational models can increase their ability to explain human behaviour, in an image recognition

task. This is largely seen in observers' accuracy data but can still be seen in observers' reaction times. The only cases where the inclusion of temporal blurring did not aid computational models to better predict observers' behaviour was in the case of HMAX and the deep supervised convolutional neural net, in the domain of observers' reaction times. Image descriptors showed peaks in variance explained at different blur levels. These peaks were consistent in each image descriptor for both the accuracy and reaction time data. The deep convolutional model and GIST both showed peaks at around 10-20% temporal blur. PHOW showed a peak at around 5% and HMAX a peak at around 1% blur of each mask onto the target image. PHOW originally showed no evidence in being able to predict human observers' accuracy data, as temporal blur was added it did show a small increase in its ability to predict observers' behaviour. The results would suggest that top performing models benefit from higher levels of temporal blurring compared to the other models.

AIC scores were used to examine if the increase in variance explained is justified by the addition of an extra parameter; temporal blur. AIC results overall followed the trend of the correlation coefficient graphs and largely showed that where a peak formed in the correlation coefficient graph that the model is justified. The results show that the effects of temporal integration windows can be studied with comparisons of computational models.

6.3. Experiment 2

6.3.1. Methods

The behavioural data comes from Chapter 4 - Experiment 1, an image categorization task ("Was an Ocean image present?"). For a description of the Methods please see Chapter 3 - Experiment 2 - Methods, where the experiment was originally described.

Information on how temporal blurring of the stimuli was created are found in Chapter 6 - Experiment 1 - Methods - Modeling temporal blurring.

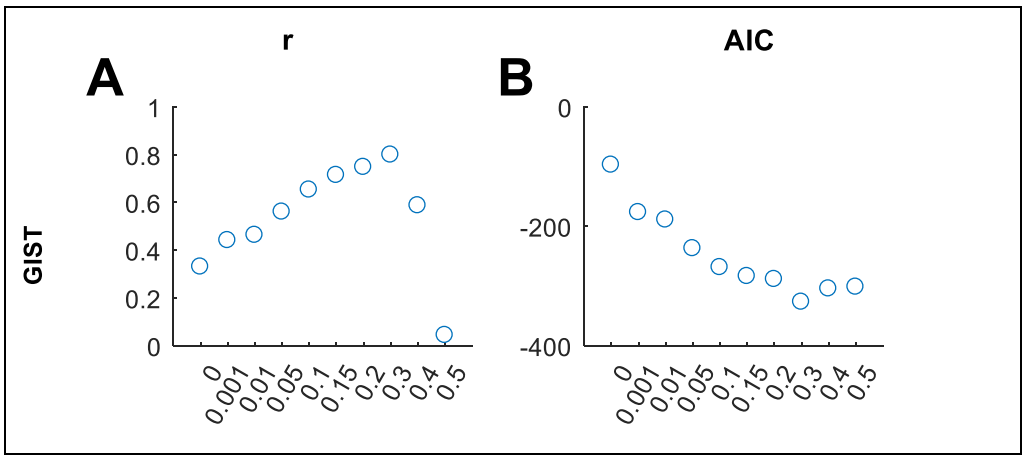
6.3.2. Results

Here the four Image descriptors (GIST, HMAX, PHOW and the deep supervised convolutional neural net) are paired with decision bound theory. Only decision bound theory is considered as it was shown to have the closest similarity to human observers in Chapter 4 - Experiment 1. Decision bound theory needed little altering to cope with the temporally blurred stimuli. It was however altered to leave out the target image and both the mask images when creating the decision bound for each trial.

In a similar manner to Chapter 4 - Experiment 1 trials in which the observer took longer than two seconds to respond were excluded from the analysis (4.3% of trials). This criterion for exclusion was chosen to limit observers' responses to rapid feedforward response based on instinct rather than cognitive reasoning.

Several different temporal blur values were examined, ranging from 0.001 to 0.5. A greater range of temporal blurring values was examined in than in Experiment 1 as it seemed observers were experiencing a greater extent of temporal blurring in this task.

In a similar manner to Experiment 1 the correlation between the model's performance and observers' behavioural data, as well as the model's AIC score are presented. The number of parameters used in calculation of AIC in the zero blur model is 2; the number of parameters for a linear correlation model. The number of parameters used in calculation of AIC in the models including temporal blur is 3; adding extent of temporal blur as another parameter. The results from the temporal blurring analysis are shown in Figure 6.3 and Figure 6.4, for accuracy and reaction time data respectively.



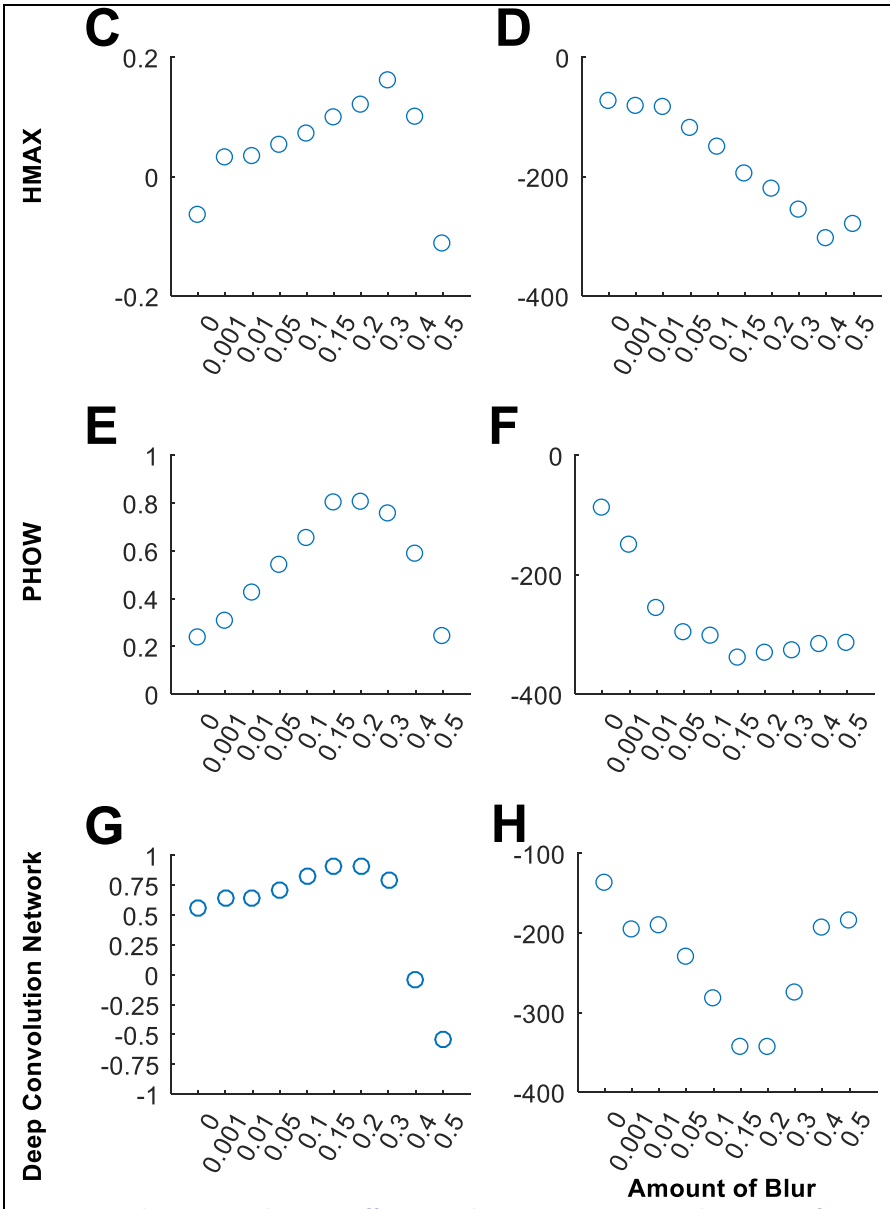
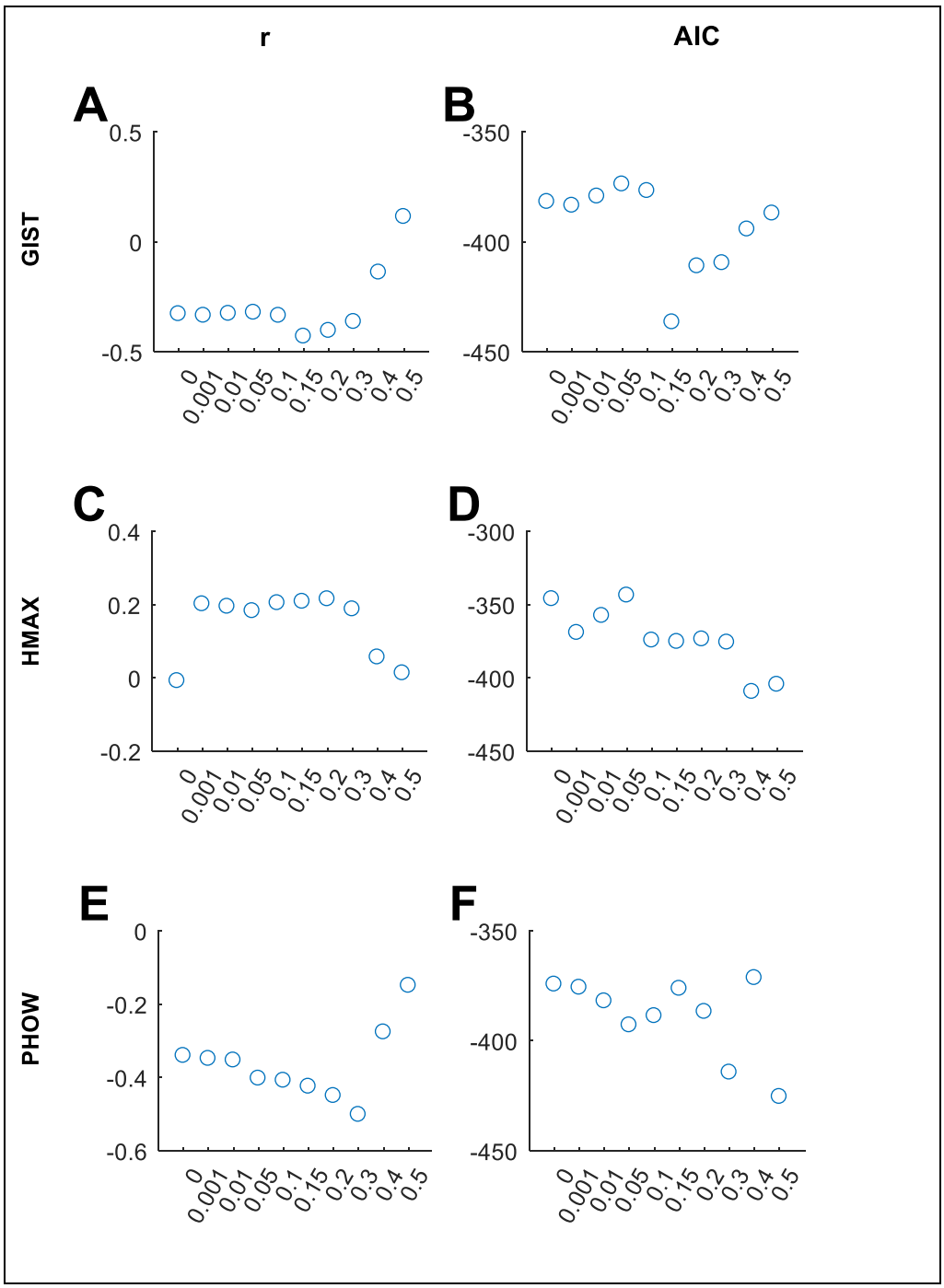


Figure 6.3. Plotting correlation coefficient and AIC criterion against the extent of temporal blurring for each computational model predicting observer's accuracy. Graphs on the left (A, C, E, G) plot variance explained in observers' accuracy data against extent of temporal blurring of the computational model, while graphs on the right (B, D, F, H) plot AIC criterion of observers' accuracy data against extent of temporal blurring. Each row represents a different image descriptor, from top to bottom, GIST (A, B), HMAX (C, D), PHOW (E, F) and deep convolutional neural network (G, H).


r
AIC
A
B
C
D
E
F
GIST
HMAX
PHOW
0.5
0
-0.5
-350
-400
-450
0.4
0.2
-0.2
-300
-0.4
-0.6
0
0.001
0.01
0.05
0.1
0.15
0.2
0.3
0.4
0.5

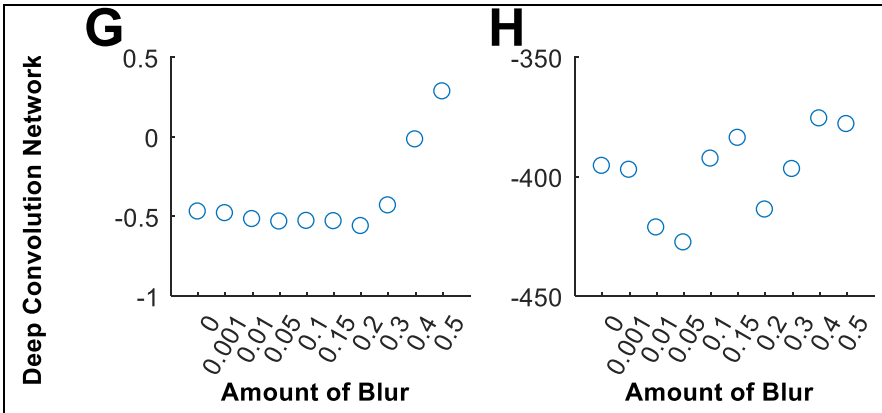


Figure 6.4. Plotting correlation coefficient and AIC criterion against the extent of temporal blurring for each computational model predicting observer's reaction times. Graphs on the left (A, C, E, G) plot variance explained in observers' reaction times data against extent of temporal blurring of the computational model, while graphs on the right (B, D, F, H) plot AIC criterion of observers' reaction times data against extent of temporal blurring. Each row represents a different image descriptor, from top to bottom, GIST (A, B), HMAX (C, D), PHOW (E, F) and deep convolutional neural network (G, H).

6.3.3. Discussion

The results show that including temporal blurring in a computational models' calculations can increased their ability to explain human behaviour, in an image categorization task. Similarly to Experiment 1, this effect is more clearly seen in observers' accuracy data, but can also be seen in observers' reaction time data. The only instance where the inclusion of temporal blurring did not aid computational models in their ability to explain observers' behaviour was for HMAX in the domain of observers' reaction times. An Image descriptors' peaks in variance explained were largely consistent for both the accuracy and reaction time data. The peaks across each image descriptor were more

consistent than in Experiment 1, ranging from 15% to 30%. In Experiment 1 it was shown that top performing image descriptors benefitted from a larger amount of temporal blurring, here all the image descriptors benefit from at least 15%+ temporal blurring.

AIC scores were used to examine if the increase in variance explained is justified by the addition of an extra parameter; temporal blur. AIC results show that for any clear peak in the reaction time or accuracy data that the addition of the extra parameter was justified by the extra variance explained. The results show that indeed observers are likely to be experiencing temporal blurring of the stimuli in the RSVP procedure which is influencing their behaviour.

6.4. General Discussion

Here we aimed to investigate if human observers were experiencing temporal blurring during an RSVP task. The effect of temporal blur was simulated by combining the target image with various weightings of the mask images presented on either side of the target. Computational models which either included, or did not include, temporal blurring in their calculations were created. These computational models were compared to human observers' behavior, in terms of reaction times and accuracy, to examine which computational models best predicted observers' performance in an image recognition and categorization task. The results largely show that models which

included temporal blurring in their calculations were better able to explain human behavior, in terms of accuracy data and reaction time data.

There was a slight difference between the two experiments in the peak temporal blurring value which best predicted observers' behavior. In Experiment 1 the top performing image descriptors (GIST and deep convolutional neural net) required a temporal blur value of around 10-20% to best predict human observers' performance, in terms of accuracy and reaction times. The results of Experiment 2 suggest that, for all image descriptors examined, temporal blur values around 15-30% were required to optimally explain observers' behavior, in terms of accuracy and reaction times. From these results, it appears that the categorization task required greater amounts of temporal blurring to best explain observers' performance. This might be caused by differences in the task. Both the data sets from Experiment 1 and Experiment 2 were collected in the same session. Each trial consisted of three stages; an initial RSVP followed by the categorization task and then the recognition task. The extra delay observers had before they were given the recognition task could have allowed time for higher cognitive processes to occur. These processes could have consolidated the image description, reducing the noise from the mask images and increasing the signal of the target. This consolidation, although quite small, may explain the difference in peak temporal blurring value between the two experiments. Further studies investigating temporal blurring and delayed recall would be needed to answer this question.

It was possible that, once temporal blurring was included in the computational models, the order of which image descriptors best explained human observers' behavior could have changed. This was not the case. Taking the peak variance explained for each image descriptor preserved the same order in which computational models best explained observers' behavior.

The analyses here provided a proof-of-principal that in, a RSVP task, the ability of the models' ability to predict observers' behavior is dependent not only on the target image, but on the mask images that neighbored the target. The fact that the fit of the computational models to behavioral data was sensitive to these temporal effects means that they could be used potentially to study the nature of temporal integration windows in biological systems. There are many ways the analyses could be extended. It should be noted that these analysis are very time-consuming analyses (for instance, computing the HMAX image descriptor for the 10 temporal blur values examined took 1 month of computing time) and, hence, went beyond the scope of the current thesis.

The simulated temporal blur in this analysis simply added some weighted combination of the two neighboring images to the target. This had equal weighting of the forward- and backward-mask with no influence of masks more than one image away from the target. In reality, there are a number of masking profiles that could be studied. A Gaussian profile, in which influence gradually diminished in time, might be the first interesting addition, but also various

forms of non-symmetrical profiles could be used to compare the models with the behavioral performance.

We might also test where in the visual system the “blurring” occurs. The current analysis treated it entirely as being in the image domain, suitable for temporal integration in low-level mechanisms such as in the photoceptors, but it might be that later stages could also be involved. This could be examined by creating weighted combinations in the descriptor space or, indeed, if the image descriptor examined has many layers, then at various layers in the image descriptor (for multi-layer descriptors such as the deep convolutional neural net).

The RSVP procedure in Experiment 1 and Experiment 2 used a range of different image duration values. This was done so that temporal blurring would be more easily detectable; either the peak temporal blur value would have been more spread out or many peaks would have been seen. It appears that here, the result of using multiple image duration values, caused a single peak in the correlation coefficient graphs. It would be interesting to examine the change in peak image blur value needed at different image durations.

The current research demonstrates the strength of computational modeling as a method of revealing the inner mechanisms of biological vision in a non-invasive manner. However, the current research leaves several unanswered questions. Where is temporal blurring taking place within the observers’ visual system? If it is occurring at the early level of the eye or higher up within the

cortex? How does temporal blurring change with image duration? And if spatial blurring could also be occurring with temporal blurring? While these questions are partially answerable from more advanced modeling, it is likely that to fully understand the processes of temporal blurring cell recording studies are needed.

Chapter 7 -General Discussion

This thesis has sought to compare human visual perception with computer models of visual processing with reference to three distinct components (image descriptor, decision process and image set). The thesis presented here aimed to investigate the influence of each of these core components on a computational model's similarity to human behaviour. The primary aims of the thesis were:

- To determine which image descriptors best approximate biological vision through behavioural tasks.
- To investigate the decision processes biological vision is employing in an image categorization task.
- To examine the extent the naturally differing image set between computational models and biological vision can explain the differences in their behaviour.
- To extend current understanding of biological vision and explore whether observers are experiencing temporal blurring when viewing a rapid visual presentation of stimuli.

Each chapter presented in this thesis aimed to answer a different one of these aims.

7.1. Summary of findings

Chapter 3 investigated the similarity of different computational image descriptors to their biological counterpart. In order to focus on the image descriptor, minimizing the influence of the decision process, a 2AFC match-to-sample task was used to map out the structure of observers' image descriptions. The structure of observers' image descriptions was then compared to those produced by different computational image descriptors. The results found that the deep supervised convolutional neural net created image descriptions which were the closest in structure to biological vision. This was followed by the image descriptors GIST, HMAX and then PHOW in their similarity to biological vision.

Chapter 4 investigated the decision process observers were using to conduct an image categorization task. Computational models were constructed by pairing the potential decision processes with the image descriptors from Chapter 3. The computational models were then compared to observers' behaviour. The results suggest that decision bound theory was the optimal decision process at explaining observers' behaviour.

Chapter 5 examined the extent the naturally differing image set between computational models and observers could explain the difference in their behaviour. Observers were trained on the image set the computational models were using. Training was done to 'steer' observers' image statistics in the direction of the computational model. A three-phase experiment was

conducted. An initial testing phase (pre-training), followed by a phase where observers were trained on the image set used by the computational model, and a final testing phase (post-training). The pre- and post-training sessions' performance were then compared to examine if observers' behaviour was closer to the computational models. The results show that indeed human observers can be made to respond closer to the computational models through training.

Chapter 6 investigated the mechanism by which observers were making rapid image descriptions. The data from Chapter 3 and 4 were re-analysed to examine if computational models could better predict observers' behaviour, in RSVP experiments, if temporal blurring was included in their calculations. The results show that if computational models were performing their calculations on temporally blurred stimuli then their behaviour becomes closer to that of the observers. This potentially presents an interesting new method with which to study the integration window in visual processing.

7.2. Advantages of the methods used in this thesis

The current literature examining computational models' similarity to observers consists of a limited number of studies. These studies often examine only a single decision process paired with a single image descriptor (Mack & Palmeri, 2010; Serre et al., 2007; Sofer et al., 2015), with limited studies comparing a number of computational models to a single data set (Ghodrati et al., 2014; Kheradpisheh et al., 2016). This makes it hard from the current literature to

rank different image descriptors or decision processes in order of best fit to human behavior. Additionally, only a few studies have compared computational models to human behavior at the trial, or image level, with the majority focusing on comparing at the level of overall accuracy for a category (Mack & Palmeri, 2010; Serre et al., 2007). The level of detail of these comparisons makes it difficult to distinguish a computational model which has a decent categorization rate, and so fits observers' behavior at the level of category, versus one which can match observers' performance on a per trial or image level.

In this thesis, comparisons between different computational models and human observers were made which goes above and beyond the existing literature. Multiple different tasks were considered (pervious literature only focusing on categorization tasks). Multiple decision processes and image descriptors were examined for each of these tasks. Models were compared to behavior on a trial by trial basis. Reaction times, as well as observers' accuracy, were compared to the computational models. Additionally, many comparisons were made on a single data set, allowing comparisons between different computational models to be straightforward. No previous works had examined the full range of components of human visual perception in combination.

Studying all the components of the model together was useful in identifying interactions. For instance, Chapter 4 showed that the decision process based

on exemplar theory predicted a significant amount of observers' behavior. If this was a single study, based in isolation, one would have drawn the conclusion that exemplar theory provided a good account for observer's categorization process. However, due to several different decision processes being examined, and exemplar theory being paired with a number of different image descriptors, with known similarity to biological vision, it was unlikely that human observers were using this algorithm to categorizing images in this experiment.

If a single model is studied in isolation then the results could be misleading, even if they are positive. HMAX is an image descriptor that is explicitly designed to match neural processing. It has previously been shown to fit observers' behavior, which was considered as evidence for the model's similarity to biological vision (Serre et al., 2007). While the various experiment here also found HMAX to provide a significant fit to human observers' behavior, it actually provides no better account to observers' image description than other image descriptors. PHOW, a computational image descriptor based on mathematical principles of image categorization, has roughly equivalent similarity to human observers' behavior in this thesis. By examining HMAX together with several other image descriptors on multiple data sets, the results suggest there are other image descriptors which, although do not primarily aim to mimic observers' neurology, better approximate the visual systems' image descriptions.

It hasn't previously been determined whether the visual system creates a single image description, which it can use for all tasks (task independent), or if the image description it creates is based on task (task dependent). Here, two different tasks are presented in this thesis. The results show that the computational image descriptors used here had roughly equivalent performance across different tasks. This is shown by the top two image descriptors, deep convolutional network and GIST, being the same for both the categorization task and the image recognition task. Additionally, when temporal blurring was examined, although there was some difference in the extent of blur which best explained that data, temporal blurring was experienced in both tasks, suggesting they were using similar image descriptions. From this data, there is no reason to suggest that the visual system employs distinct image descriptors for different purposes.

While studies examining single elements can reveal important information, this Thesis highlight the fact that examining multiple elements at once are crucial to understanding the puzzle that is the human visual system.

7.3. Future research

From the research presented in this thesis there are several different areas which would benefit from further investigation.

Two different tasks were employed in this Thesis, image categorization and image recognition. On the spectrum of tasks that, computational models of

vision are being designed to perform, these tasks are relatively simple. Several more complex tasks have received special attention in the data science literature such as object detection (recognising an object irrespective of the scene), segmentation (being able to separate out different objects within a scene), and human action classification (predicting an action being performed in a scene, e.g. playing football, shopping, etc.). There exists a number of computational algorithms for these tasks. Little to no research has investigated the similarity of these computational algorithms to human observers' performance on an image or trial by trial basis. Comparing these algorithms to human performance at a trial by trial level might help us to understand which models may be performing the task in a similar manner to biological visual processing. Future research would therefore benefit from focus on a greater breadth of tasks. Additionally, by examining many different tasks the question of whether observers are calculating a task dependent or independent image description can be more fully answered.

The current deep supervised convolutional neural net (Krizhevsky et al., 2012) used in this Thesis came into popularity in 2012 when it came top of the image classification task in the ImageNet competition (Russakovsky et al., 2015). Since then several different deep supervised convolutional neural nets have been created which surpass, in performance, the deep supervised network used here. To name but a few, Simonyan & Zisserman, (2014) created a 'very' deep supervised convolutional neural net which consisted of 19 layers. Zeiler & Fergus, (2014) introduced a convolutional model visualising technique, this

allowed them to fine tune some of the errors in the deep convolutional neural net used in Krizhevsky et al., (2012) which led to an improved performance. As deep supervised convolutional neural nets have been shown to produce behaviour the closest to human observers, future research should focus on this class of models. Some interesting questions which could be asked are, does the size and number of layers effect their similarity to human observers? From a neuroimaging perspective do the image descriptions at specific layers in deep convolutional networks have a high correspondance to the image decriptions at different layers in human biology?

The studies presented here focused on a single image set which used four different categories of scene images. This is a relatively simple image set and much larger and complicated image sets exist. Future research should aim at comparing these models on a variety of different image sests which uses scene as well as object image categories. While recently the gold standard has been to compare computational models to human observers on real life image sets, it may also be useful to examine observers and computational models performance on artificial stimuli, for which neither has been trained extensively.

The final experimental chapter of this thesis examined if observers were experiencing temporal blurring in a RSVP task. Here we showed that observers were indeed experiencing temporal blurring during the task. There are however a number of unanswered questions which relate to the specifics of

temporal blurring. If observers are experiencing temporal blurring, then where is this taking place within the observers' visual system, if it is occurring at the level of the eye or higher up within the cortex? How does temporal blurring change with image duration?

7.4. Conclusion

The main aim of making comparisons between computational models and human behavior is to reveal new information about the inner workings of biological vision. As such, it is important to consider, when making these similarity measurements, exactly what they mean and how they are useful. The similarity measures used here represent the similarity of the output of the computational models and human observers. Ideally it would be good if these similarity measurements reflected how similar the algorithmic calculations were that generated the output. This cannot truly be achieved as it is possible to conceive of multiple different algorithms that produce the same output and would thus score the same on these measures of similarity.

The comparisons described here are perhaps more relevant when viewed from the perspective of the *differences* between computational models and human observers. As a model's behavior becomes more different to that of observers it is easier to assert that the computational model is processing information in a different way, and so varying in their algorithms. While the various experiments and studies described here cannot directly measure the algorithmic similarity of the models and observers, it is possible to determine

the extent of their differences. The research here therefore provides a general assessment on the extent of differences between computational models of vision and the human visual system.

Chapter 8 - References

Aguirre, G. K., & Desposito, M. (1997). Environmental knowledge is subserved by separable dorsal/ventral neural areas. *Journal of Neuroscience, 17*(7), 2512-2518.

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 28*(12), 2037-2041. doi:10.1109/tpami.2006.244

Aminoff, E. M., Toneva, M., Shrivastava, A., Chen, X. L., Misra, I., Gupta, A., & Tarr, M. J. (2015). Applying artificial vision models to human scene understanding. *Frontiers in Computational Neuroscience, 9*. doi:10.3389/fncom.2015.00008

Andreopoulos, A., & Tsotsos, J. K. (2013). 50 Years of object recognition: Directions forward. *Computer Vision and Image Understanding, 117*(8), 827-891. doi:10.1016/j.cviu.2013.04.005

Andrews, T. J., Watson, D. M., Rice, G. E., & Hartley, T. (2015). Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *Journal of Vision, 15*(7). doi:10.1167/15.7.3

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response-time with multidimensional stimuli. *Perception & Psychophysics, 55*(1), 11-27. doi:10.3758/bf03206876

Ashby, F. G., & Maddox, W. T. (1994). A response-time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology, 38*(4), 423-466. doi:10.1006/jmps.1994.1032

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual review of psychology, 56*, 149-178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Year in Cognitive Neuroscience, 1224*, 147-161. doi:10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*(2), 154-179. doi:10.1037//0033-295x.93.2.154

Attneave, F. (1957). Transfer of experience with a class-schema to identification-learning of patterns and shapes. *Journal of Experimental Psychology, 54*(2), 81-88. doi:10.1037/h0041231

Barrow, H., & Tenenbaum, J. (1978). *Recovering intrinsec scene characteristics from images. In Computer Vision Systems.*: Academic Press: New York.

Bell, A. H., Hadj-Bouziane, F., Frihauf, J. B., Tootell, R. B. H., & Ungerleider, L. G. (2009). Object Representations in the Temporal Cortex of Monkeys and Humans as Revealed by Functional Magnetic Resonance Imaging.

Journal of Neurophysiology, 101(2), 688-700. doi:10.1152/jn.90657.2008

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 24*(4), 509-522. doi:10.1109/34.993558

Berg, A. C., Berg, T. L., & Malik, J. (2005, Jun 20-25). *Shape matching and object recognition using low distortion correspondences.* Paper presented at the Conference on Computer Vision and Pattern Recognition, San Diego, CA.

Berg, A. C., & Malik, J. (2001, Dec 08-14). *Geometric blur for template matching.* Paper presented at the Conference on Computer Vision and Pattern Recognition, Kauai, Hi.

Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. In (Vol. 94): Psychological Review.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., . . . DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *Plos Computational Biology, 10*(12). doi:10.1371/journal.pcbi.1003963

Carlson, T., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. S. (2014). Reaction Time for Object Categorization Is Predicted by Representational Distance. *Journal of Cognitive Neuroscience, 26*(1), 132-142. doi:10.1162/jocn_a_00476

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision, 13*(10). doi:10.1167/13.10.1

Chen, Q., Song, Z., Dong, J., Huang, Z. Y., Hua, Y., & Yan, S. C. (2015). Contextualizing Object Detection and Classification. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 37*(1), 13-27. doi:10.1109/tpami.2014.2343217

Chen, X. L., Shrivastava, A., & Gupta, A. (2013, Dec 01-08). *NEIL: Extracting Visual Knowledge from Web Data.* Paper presented at the IEEE International Conference on Computer Vision (ICCV), Sydney, AUSTRALIA.

Coggan, D. D., Baker, D. H., & Andrews, T. J. (2016). The Role of Visual and Semantic Properties in the Emergence of Category-Specific Patterns of Neural Response in the Human Brain. *Eneuro, 3*(4). doi:10.1523/eneuro.0158-16.2016

Coggan, D. D., Liu, W. L., Baker, D. H., & Andrews, T. J. (2016). Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. *Neuroimage, 135*, 107-114. doi:10.1016/j.neuroimage.2016.04.060

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273-297. doi:10.1007/bf00994018

Dalal, N., & Triggs, B. (2005, Jun 20-25). *Histograms of oriented gradients for human detection.* Paper presented at the Conference on Computer Vision and Pattern Recognition, San Diego, CA.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35*(3), 393-401. doi:10.3758/bf03193280

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559-564. doi:10.1111/j.0956-7976.2004.00719.x

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601. doi:10.1038/33402

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2015). The PASCAL Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision, 111*(1), 98-136. doi:10.1007/s11263-014-0733-5

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision, 88*(2), 303-338. doi:10.1007/s11263-009-0275-4

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding, 106*(1), 59-70. doi:10.1016/j.cviu.2005.09.012

Fukushima, K. (1980). Neocognitron - a self-organizing neural network model for a mechanism of pattern-recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193-202. doi:10.1007/bf00344251

Getty, D. J., Swets, J. B., & Swets, J. A. (1980). The observer's use of perceptual dimensions in signal identification. In *Attention and performance VIII* (pp. 361-381): Lawrence Erlbaum Associates Hillsdale, New Jersey.

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., & Khaligh-Razavi, S. M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience, 8*. doi:10.3389/fncom.2014.00074

Ghodrati, M., Khaligh-Razavi, S. M., Ebrahimpour, R., Rajaei, K., & Pooyan, M. (2012). How Can Selection of Biologically Inspired Features Improve the Performance of a Robust Object Recognition Model? *Plos One, 7*(2). doi:10.1371/journal.pone.0032357

Gibson, E. J. (1969). Principles of perceptual learning and development.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews-Cognitive Science, 1*(1), 69-78. doi:10.1002/wcs.26

Greene, M. R., & Oliva, A. (2009). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science, 20*(4), 464-472. doi:10.1111/j.1467-9280.2009.02316.x

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430. doi:10.1126/science.1063736

Hecht, S., & Smith, E. L. (1936). Intermittent stimulation by light VI. Area and the relation between critical frequency and intensity. *Journal of General Physiology, 19*(6), 979-989. doi:10.1085/jgp.19.6.979

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243-271. doi:10.1146/annurev.psych.50.1.243

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology-Human Learning and Memory, 7*(6), 418-439. doi:10.1037//0278-7393.7.6.418

Hsu, W., Chua, S., & Pung, H. (1995). *An integrated color-spatial approach to content-based image retrieval.* Paper presented at the Proceedings of the third ACM international conference on Multimedia.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in cats visual cortex. *Journal of Physiology-London, 160*(1), 106-&.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology-London, 195*(1), 215-&.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*(5749), 863-866. doi:10.1126/science.1117593

Jarrett, K., Kavukcuoglu, K., Ranzato, M., Lecun, Y., & Ieee. (2009, Sep 29-Oct 02). *What is the Best Multi-Stage Architecture for Object Recognition?* Paper presented at the 12th IEEE International Conference on Computer Vision, Kyoto, JAPAN.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrell, T. (2014). Caffe: Convolutional Arhitecture for Fast Embedding. *arXiv preprint arXiv:1408.5093*.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286-3297. doi:10.1016/j.visres.2007.09.013

Kadar, I., & Ben-Shahar, O. (2012). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of Vision, 12*(13). doi:10.1167/12.13.16

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B-Biological Sciences, 361*(1476), 2109-2128. doi:10.1098/rstb.2006.1934

Khaligh-Razavi, S.-M. (2014). What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. *arXiv preprint arXiv:1407.2776*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *Plos Computational Biology, 10*(11). doi:10.1371/journal.pcbi.1003915

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports, 6*. doi:10.1038/srep32672

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97*(6), 4296-4309. doi:10.1152/jn.00024.2007

Kim, J., Kim, B.-S., & Savarese, S. (2012). Comparing image classification methods: K-nearest-neighbor and support-vector-machines. *Ann Arbor, 1001*, 48109-42122.

Konishi, S., Kawazu, M., Uchida, I., Kikyo, H., Asakura, I., & Miyashita, Y. (1999). Contribution of working memory to transient activation in human inferior prefrontal cortex during performance of the Wisconsin Card Sorting Test. *Cerebral Cortex, 9*(7), 745-753. doi:10.1093/cercor/9.7.745

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401-412. doi:10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*, 4-4. doi:10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron, 60*(6), 1126-1141. doi:10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. Paper presented at the NIPS.

Kyrkou, C. (2017). Object Detection Using Local Binary Patterns. Retrieved from https://medium.com/@ckyrkou/object-detection-using-local-binary-patterns-50b165658368

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review, 107*(2), 227-260. doi:10.1037//0033-295x.107.2.227

Lazebnik, S., Schmid, C., & Ponce, J. (2006). *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*. Paper presented at the CVPR.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi:10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the Ieee, 86*(11), 2278-2324. doi:10.1109/5.726791

Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *Journal of Vision, 13*(13). doi:10.1167/13.13.25

Lin, Y. Q., Lv, F. J., Zhu, S. H., Yang, M., Cour, T., Yu, K., . . . Ieee. (2011, Jun 20-25). *Large-scale Image Classification: Fast Feature Extraction and SVM*

Training. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO.

Lockhead, G. R. (1966). Effects of dimensional redundancy on visual discrimination. *Journal of Experimental Psychology, 72*(1), 95-&. doi:10.1037/h0023319

Lombardi, W. J., Andreason, P. J., Sirocco, K. Y., Rio, D. E., Gross, R. E., Umhau, J. C., & Hommer, D. W. (1999). Wisconsin card sorting test performance following head injury: Dorsolateral fronto-striatal circuit activity predicts perseveration. *Journal of Clinical and Experimental Neuropsychology, 21*(1), 2-16. doi:10.1076/jcen.21.1.2.940

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition, 18*(4), 513-536. doi:10.1080/13506280902937606

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110. doi:10.1023/b:visi.0000029664.99615.94

Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision, 10*(3). doi:10.1167/10.3.11

Maguire, E. A. (2001). The retrosplenial contribution to human navigation: A review of lesion and neuroimaging findings. *Scandinavian Journal of Psychology, 42*(3), 225-238. doi:10.1111/1467-9450.00233

Malik, F., & Baharudin, B. (2013). Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain. *Journal of king saud university-computer and information sciences, 25*(2), 207-218.

Mallick, S. (2016). Histogram of Oriented Gradients.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*.

McGraw, P. V., Webb, B. S., & Moore, D. R. (2009). Sensory learning: from neural mechanisms to rehabilitation. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1515), 279-283. doi:10.1098/rstb.2008.0274

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 27*(10), 1615-1630. doi:10.1109/tpami.2005.188

Muralidharan, K., & Vasconcelos, N. (2010). *A biologically plausible network for the computation of orientation dominance.* Paper presented at the Advances in Neural Information Processing Systems.

Mutch, J. (2010). HMAX Models Architecture. Retrieved from http://www.mit.edu/~9.520/spring10/slides/class15-visualneuroscience/class15-hmax.pdf

Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision, 80*(1), 45-57. doi:10.1007/s11263-007-0118-0

Nakamura, H., Gattass, R., Desimone, R., & Ungerleider, L. G. (1993). The modular organization of projections from area-v1 and area-v2 to area-v4 and teo in macaques. *Journal of Neuroscience, 13*(9), 3681-3691.

Nasr, S., & Tootell, R. B. H. (2012). A Cardinal Orientation Bias in Scene-Selective Visual Cortex. *Journal of Neuroscience, 32*(43), 14921-14926. doi:10.1523/jneurosci.2036-12.2012

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *Plos Computational Biology, 10*(4). doi:10.1371/journal.pcbi.1003553

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology-General, 115*(1), 39-57. doi:10.1037/0096-3445.115.1.39

O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience, 17*(4), 580-590. doi:10.1162/0898929053467550

Ojala, T., Pietikainen, M., & Harwood, D. (1994, Oct 09-13). *Performance evaluation of texture measures with classification based on kullback discrimination of distributions.* Paper presented at the Conference A on Computer Vision and Image Processing, at the 12th IAPR International Conference on Pattern Recognition, Jerusalem, Israel.

Ojala, T., Pietikainen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition, 29*(1), 51-59. doi:10.1016/0031-3203(95)00067-4

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145-175. doi:10.1023/a:1011139631724

Palmer, S. E. (1975). Effects of contextual scenes on identification of objects. *Memory & Cognition, 3*(5), 519-526.

Pass, G., & Zabih, R. (1999). Comparing images using joint histograms. *Multimedia Systems, 7*(3), 234-240. doi:10.1007/s005300050125

Patterson, G., & Hays, J. (2012). *Sun attribute database: Discovering, annotating, and recognizing scene attributes.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8-13. doi:10.1016/j.jneumeth.2006.11.017

Perronnin, F., & Dance, C. (2007, Jun 17-22). *Fisher kernels on visual vocabularies for image categorization.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN.

Pietikainen, M., Hadid, A., Zhao, G., & Ahonen, T. (2011). Computer Vision Using Local Binary Patterns. *Computer Vision Using Local Binary Patterns, 40*, 1-207.

Posner, M. I., & Keele, S. W. (1968). On genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3P1), 353-&. doi:10.1037/h0025953

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*(2), 304-&. doi:10.1037/h0028558

Potter, M. C. (1975). MEANING IN VISUAL SEARCH. *Science, 187*(4180), 965-966. doi:10.1126/science.1145183

Rajaei, K., Khaligh-Razavi, S. M., Ghodrati, M., Ebrahimpour, R., & Abadi, M. (2012). A Stable Biologically Motivated Learning Mechanism for Visual Feature Extraction to Handle Facial Categorization. *Plos One, 7*(6). doi:10.1371/journal.pone.0038478

Rao, S. M., Bobholz, J. A., Hammeke, T. A., Rosen, A. C., Woodley, S. J., Cunningham, J. M., . . . Binder, J. R. (1997). Functional MRI evidence for subcortical participation in conceptual reasoning skills. *Neuroreport, 8*(8), 1987-1993. doi:10.1097/00001756-199705260-00038

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*(3), 382-407. doi:10.1016/0010-0285(72)90014-x

Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of the Ventral Visual Pathway. *Journal of Neuroscience, 34*(26), 8837-8844. doi:10.1523/jneurosci.5265-13.2014

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025.

Ritchie, J. B., & Carlson, T. A. (2016). Neural Decoding and "Inner" Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Frontiers in Neuroscience, 10*. doi:10.3389/fnins.2016.00190

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *Plos Computational Biology, 11*(6). doi:10.1371/journal.pcbi.1004316

Rohrer, B. (2016). How do Convolutional Neural Networks work? Retrieved from http://brohrer.github.io/how_convolutional_neural_networks_work.html

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*(3), 211-252. doi:10.1007/s11263-015-0816-y

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104*(15), 6424-6429. doi:10.1073/pnas.0700622104

Serre, T., Wolf, L., & Poggio, T. (2005, Jun 20-25). *Object recognition with features inspired by visual cortex.* Paper presented at the Conference on Computer Vision and Pattern Recognition, San Diego, CA.

Sharma, M., & Batra, A. (2014). Analysis of Distance Measures in Content Based Image Retrieval. *Global Journal of Computer Science and Technology, 14*(2).

Shepard, R. N. (1958). Stimulus and response generalization - tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55*(6), 509-523. doi:10.1037/h0042354

Shepard, R. N. (1962a). The analysis of proximities - multidimensional-scaling with an unknown distance function .1. *Psychometrika, 27*(2), 125-140. doi:10.1007/bf02289630

Shepard, R. N. (1962b). The analysis of proximities - multidimensional-scaling with an unknown distance function .2. *Psychometrika, 27*(3), 219-246. doi:10.1007/bf02289621

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*(1), 54-87. doi:10.1016/0022-2496(64)90017-3

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317-1323. doi:10.1126/science.3629243

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sofer, I., Crouzet, S. M., & Serre, T. (2015). Explaining the Timing of Natural Scene Understanding with a Computational Model of Perceptual Categorization. *Plos Computational Biology, 11*(9). doi:10.1371/journal.pcbi.1004456

Stricker, M., & Dimai, A. (1996, Feb 01-02). *Color indexing with weak spatial constraints.* Paper presented at the Conference on Storage and Retrieval for Still Image and Video Databases IV, San Jose, Ca.

Sweet, A. L. (1953). Temporal discrimination by the human eye. *American Journal of Psychology, 66*(2), 185-198. doi:10.2307/1418725

Thomas, S., Aude, O., & Tomaso, P. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104*(15), 6424-6429. doi:10.1073/pnas.0700622104

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522. doi:10.1038/381520a0

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika, 17*(4), 401-419.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 30*(11), 1958-1970. doi:10.1109/tpami.2008.128

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating Color Descriptors for Object and Scene Recognition. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1582-1596. doi:10.1109/tpami.2009.154

van de Sande, K. E. A., Uijlings, J. R. R., Gevers, T., & Smeulders, A. W. M. (2011). Segmentation as Selective Search for Object Recognition. *2011 Ieee International Conference on Computer Vision (Iccv)*, 1879-1886.

Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience, 10*(11), 792-U750. doi:10.1038/nrn2733

Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage, 99*, 402-410. doi:10.1016/j.neuroimage.2014.05.045

Watson, D. M., Hartley, T., & Andrews, T. J. (2017). A data driven approach to understanding the organization of high-level visual cortex. *Scientific Reports, 7*. doi:10.1038/s41598-017-03974-5

Watson, D. M., Hymers, M., Hartley, T., & Andrews, T. J. (2016). Patterns of neural response in scene-selective regions of the human brain are affected by low-level manipulations of spatial frequency. *Neuroimage, 124*, 107-117. doi:10.1016/j.neuroimage.2015.08.058

Watson, D. M., Young, A. W., & Andrews, T. J. (2016). Spatial properties of objects predict patterns of neural response in the ventral visual pathway. *Neuroimage, 126*, 173-183. doi:10.1016/j.neuroimage.2015.11.043

Werker, J. F., & Tees, R. C. (2002). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 25*(1), 121-133. doi:10.1016/s0163-6383(02)00093-0

Westheimer, G., & McKee, S. P. (1977). Perception of temporal-order in adjacent visual-stimuli. *Vision Research, 17*(8), 887-892. doi:10.1016/0042-6989(77)90062-1

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*(3), 671-684. doi:10.3758/brm.42.3.671

Xiao, J. X., Hays, J., Ehinger, K. A., Oliva, A., Torralba, A., & Ieee. (2010, Jun 13-18). *SUN Database: Large-scale Scene Recognition from Abbey to Zoo.* Paper presented at the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). 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), 8619-8624. doi:10.1073/pnas.1403112111

Yap, K. H., Chen, T., Li, Z., & Wu, K. (2010). A Comparative Study of Mobile-Based Landmark Recognition Techniques. *Ieee Intelligent Systems, 25*(1), 48-57.

Yu, G., & Morel, J. M. (2011). ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Computer Vision - Eccv 2014, Pt I, 8689*, 818-833.

Zhang, H., Berg, A. C., Maire, M., & Malik, J. (2006). *SVM-KNN: Discriminative nearest neighbor classification for visual category recognition.* Paper presented at the Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). *Learning deep features for scene recognition using places database.* Paper presented at the Advances in neural information processing systems.