Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 097

A Brief Introduction to Cortical Representations of Objects

Guy Wallis[1] & Heinrich Bülthoff[2]

February 2002

[1] Department of Human Movement Studies; University of Queensland, St. Lucia 4072; QLD Australia, E–mail: gwallis@hms.uq.edu.au
[2] Department Bülthoff, E–mail: heinrich.buelthoff@tuebingen.mpg.de

This report is available in PDF–format via anonymous ftp at ftp://ftp.kyb.tuebingen.mpg.de/pub/mpi-memos/pdf/TR-097.pdf. The complete series of Technical Reports is documented at: http://www.kyb.tuebingen.mpg.de/bu/techr/

A Brief Introduction to Cortical Representations of Objects

Guy Wallis & Heinrich Bülthoff

Abstract. To understand how objects are recognized and represented in the human brain is still one of the ultimate goals of cognitive science. In this article, we will collect evidence from mainly neurophysiological studies which suggest that object recognition is achieved by hierarchical processing in the brain and that the representation of objects is distributed and view-based. Furthermore, these studies propose that the temporal coherence of the visual input plays a fundamental role in the learning of object representations.

1 Introduction

As viewing distance, viewing angle or lighting conditions change, so too does the image of an object which we see. Despite the seemingly endless variety of images that objects can project, the human visual system remains able to rapidly and reliably identify them across huge changes in appearance. Understanding how humans achieve this feat of recognition has long been a source of debate. Despite a concerted effort, researchers are still undecided even about the most fundamental questions of how objects are represented in cortex. This article gives a brief overview of some theoretical approaches in the context of mainly neurophysiological evidence. It also considers the related question of objects within a physical context, that is the analysis of visual scenes. Scene analysis is relevant to the question of object recognition because scenes are initially recognised at a holistic, object-like level, providing a context or 'gist' which itself influences the speed and accuracy of recognition of the constituent objects (Rensink, 2000). A precise characterisation of gist remains elusive, but it may well include information such as global color patterns, spatial frequency content, correlational structure, anything which is useful for categorising or recognising the scene.

To provide an anatomical framework it is instructive to review the major functional divisions of visual cortex. Visual processing begins in the hinter most part of neocortex, in the occipital lobe. From there, information flows down into the temporal lobe, forming the ventral stream; and up into the parietal lobe, forming the dorsal stream - see figure 1. On the basis of the neuropsychological and single cell recording data, theorists proposed a functional division between these streams. The dorsal stream was likened to the task of deciding 'where' an object is, and the ventral stream 'what' an object is (Ungerleider & Haxby, 1994). In this review we mainly focus on the 'what' stream, since it is seen as the centre of object recognition, but as later comments will reveal, an integrated model of scene perception will almost certainly require a wider reaching approach encompassing all four lobes.

2 The ventral stream

The path from primary visual cortex to the inferior temporal lobe (IT) passes through as many as ten neural areas before reaching the last wholly visual areas - see figure 1. Early recordings in the temporal lobe reported neurons selective for faces, and later recordings were able to verify that these cells could not be excited by simple visual stimuli, nor as the result of an emotional response to seeing a particular face - see (Logothetis & Sheinberg, 1996; Rolls, 1992).

One striking feature of the reponse properties of neurons in IT that the further down the ventral stream one looks, the more specialised and selective the neurons become. Of especial interest to the field of object recognition was the discovery that along with increasing selectivity, many neurons became tolerant to shifts in stimulus position, changes in viewing angle, size/depth or illumination, or the spatial frequencies present in the image - see (Rolls, 1992).

A great deal of this work originally had to do with neurons selective for faces, but although face cells account for as much as 20% of neurons in some regions of IT and STS, they only account for around 5% of all cells present in inferior temporal cortex. In the early 1990s, Tanaka and his colleagues (Tanaka, Saito, Fukada, & Moriya, 1991) showed that many of the remaining neurons are selective for complex combinations of features, including a basic shape with bounded light, shaded or colored bounded regions, and that these neurons also demonstrate useful invariance properties. This work has served to dispell the idea of a

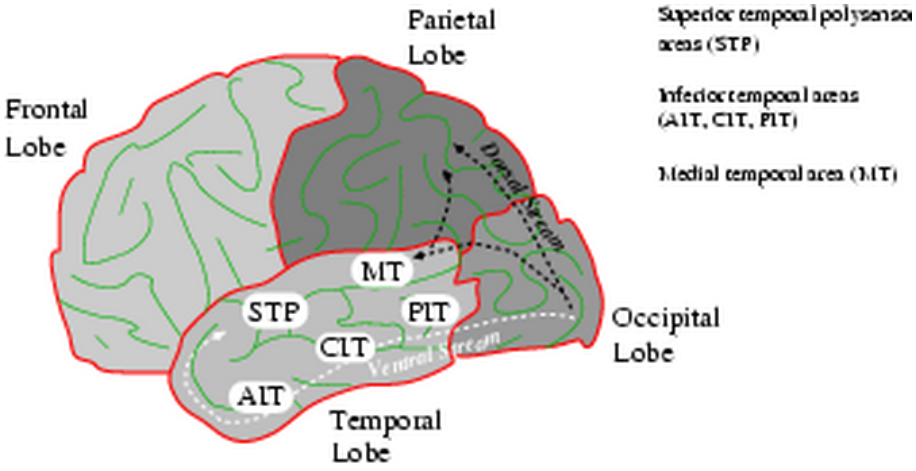

Figure 1: Principle divisions of neocortex, including the main areas of the temporal lobe. The dark arrows indicate information flow along the dorsal stream. The light arrow indicates flow along the ventral stream.

special stream designed specifically for face recognition.

Recent work has focused on the issue of how the cellular response properties of temporal lobe neurons change over time. Several studies have shown that repeated exposure to a particular object class results in changes in the number of neurons selective for that stimulus - e.g. (Logothetis & Sheinberg, 1996; Miyashita, 1993; Rolls, 1992). In humans, we should not be surprised if a car enthusiast has neurons tuned to the appearance of a yellow VW Beetle, or that a lepidopterist has ones tuned to an Orange Tip butterfly.

3 The dorsal stream

Abstracting an object's form from its precise location, size, or orientation is clearly important for tasks such as recognition and categorisation. However, there are plenty of situations in which an object's location and orientation are important, not least when we want to interact with that object by picking it up or wielding it appropriately. Processing of location and orientation appear to be the major concern of neurons in the parietal lobe. These neurons form part of the dorsal stream. In humans, damage to the parietal lobe severely affects the localisation of objects within a scene, leading to disorders such as visual neglect, and it appears that the dorsal stream is intrinsically linked to the control of visual attention and eye-movements (Ungerleider & Haxby, 1994).

Of course many tasks require the interaction of the two types of information, both where and what an object is. Our lepidoptorist would like to be able to net a Tortoiseshell fluttering amongst Red Admirals. This raises an as yet unanswered question of how these types of information interact and where various representations are held. It turns out that there are plenty of routes which information could take between the temporal and parietal lobes - including directly, via the occipital lobe, or via the frontal lobe. It has been shown, for example, that later stages of IT (AIT/CIT) connect to the frontal lobe, whereas earlier ones (CIT/PIT) connect to the parietal lobe (Webster, Bachevalier, & Ungerleider, 1994). One aim of any modelling work must be to investigate the possible significance of these connections.

4 A processing hierarchy

One of the striking features of the ventral stream is its hierarchical structure. Neurons in the latter regions of the temporal lobe can be thought of as sitting on the top of a processing pyramid - see figure 2. Receptive field size grows steadily larger the further up this pyramid one looks, and the response times of neurons also rise systematically (Rolls, 1992).

One possible explanation for the presence of such a hierarchy is that the visual system is gradually building representations of ever increasing complexity to produce neurons which respond to combinations of inputs themselves forming the effective stimuli for later neurons. By responding to local combinations of neurons co-active in the previous layer, arbitrary spatial arrangements of the same features should fail to activate the same neuron. This should then reduce the chance of finding the trigger features supporting recognition in random arrangements of the features, an issue often referred to as the 'feature binding problem'. Some of the most selective and view-invariant responses belong to cells in the superior temporal areas. These neurons appear to pool the outputs of view selective AIT cells. One unsupervised explanation as to how the STPa neu-

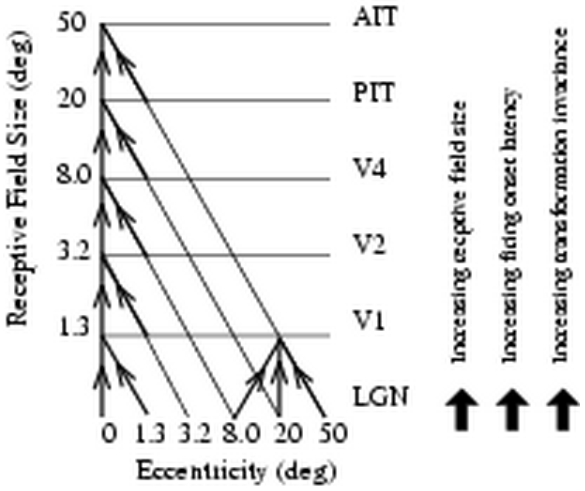

Figure 2: Schematic of convergence in the ventral processing stream. The steady growth in receptive field size suggests that neurons in one layer of the hierarchy receive input from a select group of neurons in the preceding layer. The time taken for the effects of seeing a new visual stimulus increases systematically through the hierarchy, supporting the notion of a strictly layered structure.

rons know which AIT neurons to group together is discussed in the next section.

Neuroanatomists tell us that there are at least as many connections running back as there are forward in the ventral stream, and this is important when one comes to devise models. The precise use of these connections remains unclear. Some theorists have argued that they are used in recall, and it is true that the act of remembering visual events causes activity to spread into primary visual areas. They may also control visual attention. Certainly attending to specific regions of our visual environment has been shown to facilitate the processing of signals in topographically matched regions of visual cortex, which may well be due to selectively raising activity (or lowering activation thresholds) of neurons along the processing hierarchy. One important role which the connections almost certainly do play is in relaying 'top-down' influences on recognition, due to expectations or selective attention, perhaps prompted by the gist of a scene. Such influences include contextual priors, which in this case are functions that govern the likelihood of seeing a particular object in a particular context. Our lepidoptorist will implicitly change these priors with habitat, improving the chances of correctly distinguishing a Caper White in the rocky bush of South Australia from a Cabbage White on the Meadows of Southern England.

Apart from its role in relaying attentional mechanisms, some have argued that the backward projecting connections play an integral role in normal visual processing. Some have gone as far as to suggest that each neural region forms a recurrent attractor network, each connected through the cortex up to and including the temporal lobe. Whilst such models may be required to deal with confusing or low quality images, there is good evidence that timing constraints prohibit such a model from acting during the rapid recognition of everyday, familiar objects (Thorpe, Fize, & Marlot, 1996).

5 Encoding objects in the temporal lobe

Despite the apparent selectivity of temporal lobe neurons it is important to realise that they are not all-singing, all-dancing 'super cells', selective for a single entity in the manner proposed in early theories of object representation. On the contrary, many of the cells reported in the literature responded to several examples of objects within their particular object category. Evidence is emerging that object encoding is achieved via small ensembles of firing cells which both efficiently and robustly code for individual objects.

Under a distributed scheme, many hundreds or thousands of neurons - each selective for its specific feature - would act together to represent an object. Although many of these features represent only small regions of an object, others appear to represent an object's outline, or some other global but general property. In addition, the neural representation of these features is more sophisticated than a simple template, since they may exhibit invariance to scale and size, something typical of temporal lobe neurons.

Implementing representation in a distributed code brings with it several advantages. First, the representations are robust to cell damage: Since hundreds or thousands of neurons react to the presence of a single object, the death of one neuron within the ensemble will not adversely affect recognition accuracy or speed. Second, a distributed representation provides immediate recognition generalisation to novel stimuli: A new object can be represented distinctly from all other stimuli by using a unique combination of the many, well established feature selective neurons already present. In so doing, each neuron brings knowledge of how its feature changes in appearance with changes in viewpoint. The numerous beneficial, emergent properties of a distributed representation have long been realized by neural network theorists - see (Deneve, Latham, & Pouget, 2001).

In addition to the general encoding and topological organisation of IT cortex, work has also been carried out to establish what functional organisation might be present. Some researchers have made moves to describe the functional organisation of IT. Cells were tested and their key stimulating features characterised,

revealing a columnar structure in which groups of neurons appear to respond to similar, though subtly different collections of features. Neighbouring columns seem to bear less in common. This in part reflects the findings of other researchers who have described localised 'clusters' or 'patches' of face cells - see (Rolls, 1992) - and is taken by some as evidence for local excitatory and more diverse inhibitory connections within the processing layers akin to those used in competitive networks - see (Wallis & Bülthoff, 1999; Riesenhuber & Poggio, 2000).

6 Models of object representation and recognition

There have been a huge number of systems for object recognition proposed over the years. Some largely inspired by the desire to build intelligent machines, and some by the desire to describe human recognition processes. This section summarises some of the popular models and their relevance, or otherwise, to human object recognition (see also review articles on the topic: (Wallis & Bülthoff, 1999; Riesenhuber & Poggio, 2000)).

One family of models, which owes its heritage to AI research in the 1970s, sees the need to extract cues to 3D structure. Using texture gradients, linear perspective, structure from motion, etc. it seeks to transform the retinal image into a fully fledged internal 3D model, capable of rotation, scaling, translation and therefore matching to a store of known objects. Various means for achieving this reconstruction have been proposed, though perhaps the most preeminent is the geon theory of Biedermann, and its associated network model called JIM (Hummel & Biederman, 1992). Unfortunately, whilst there are plenty of neurons sensitive to cues such as terminated edges or complex forms of motion, neurophysiologists have yet to find evidence for large quantities of the types of neural analysers which these types of models would predict, and even less evidence for the set of 36 3D volumetric building blocks which Biederman's theory claims are combined to represent all objects. What is more, there is only limited evidence for the neural synchronisation mechanism which it uses to bind elements of activated geons, and there no evidence of neurons purely selective to spatial relationships of parts such as 'left-of', 'above' etc. Nonetheless, some form of structural representation must surely exist, particularly in defining object categories for distinguishing a quadruped from a biped, or telephone from an elephant. The JIM model is one of the very few models focused on human object recognition which provides a principled means of extracting and representing structure.

As an alternative to this type of bottom-up object reconstruction, a number of approaches to object recognition have looked at the possibility of matching the incoming image to a large collection of 2D images or whole 3D objects. This process takes a number of different forms. In some models the image of the object is normalised for size and location and then simply matched pixel by pixel to a stored set of images. Of course, simple 3D transformations such as depth rotation lead to non-trivial changes in the 2D projected image. To compensate for this, some models have employed local distortions of the incoming image in the matching process. Others have presupposed an ability to extract 3D anchor points in the image which allow stored 3D representations to be rotated and scaled in 3D before before the matching process begins. In practice, most of these models work well on predefined sets of objects and small changes in appearance, but are prone to errors if the incoming image changes considerably. Understandably, models which employ local distortion or rotation algorithms are more robust, but this comes with a cost. The models are slow and become slower the more objects are stored in the internal library. The simplest form of 2D template matching is at least fast, and if the process proceeds in parallel, it can scale extremely well as the number of objects increases. However, where all of these models fall down is in explaining our ability to categorise and generalise recognition of new objects to changes in viewing direction.

A possible solution to this final problem is based on a further alternative for how objects are represented and recognised. This approach once again suggests that objects are stored as images or multiple views (Bülthoff & Edelman, 1992). However, rather than being stored as a single template, each view is represented as a collection of small picture elements, each tolerant to small view changes (Wallis & Bülthoff, 1999; Riesenhuber & Poggio, 2000). Such a system immediately reaps the benefits of a distributed encoding system described above, in terms of robustness and transformation generalisation for novel objects, it also accords with the types of neural response properties known to exist in the ventral stream.

In practice, many systems base recognition on the combination of pictorial features. Some have simply attempted to look across the entire image for tell tale features irrespective of relative position, as evidence for the presence of one object rather than another. Of course, models which throw away spatial information in this way, run into the problem of 'recognising' random rearrangements of the features triggering recognition. This is not the case for real neurons responsive to faces which often reduce their response to faces in

which the features appear jumbled up (Logothetis & Sheinberg, 1996; Rolls, 1992). Nor is it true for cells responsive to more abstract features (Tanaka et al., 1991), indeed this is an example of the feature binding problem. As described in the hierachy section, one solution to this problem is to combine features gradually over a series of stages, achieving translation invariance step by step. This has inspired many theorists to take this approach in object recognition. One of the first to construct a truely hierachical model was (Fukushima, 1980). His Neocognitron is an elegant example of how piecewise combinations of features can lead to comprehensive translation and scale invariance whilst at the same time retaining object specificity and thereby avoiding one form of the 'feature binding problem'. Fukushima's ideas accord well both with elements of the known neurophysiology of the ventral stream and a view-based scheme of object representation and has inspired a whole series of models - see (Wallis & Rolls, 1997; Riesenhuber & Poggio, 2000). The Riesenhuber et al. paper also describes their development of Fukushima's model and how it predicts the use of a non-linear weighting mechanism on the inputs to neurons of each layer. Likewise Wallis et al. describe their own model, which is once again hierarchical and convergent but simpler in structure. Despite its simplicity it has been shown to be able to learn invariant representations of objects without recourse to non-local learning mechanisms; supervised learning; specialist neural populations; or specific, prescribed connectivity. An important omission from such models is any explicit representation of object structure. As mentioned above, structure may well be important for higher levels of catagorisation, for distinguishing broad categories such as insects from mammals. Image based approaches, on the other hand, are probably of more importance for within category discrimination such as a Peacock butterfly from a Meadow Brown.

One aspect which the hierachical feedforward models lack is an account of the effects of top-down information due to expectation or selective attention. As such they only really deal with recognition within the high acuity centre of the visual field and would require some other mechanism for locating and fixating objects. One hierachical model which does consider this is due to (Olhausen, Anderson, & Essen, 1993). It selects targets by controlling the breadth and number of pathways present in the model's hierachy. Recognition is achieved using a classical object matching algorithm which immediately suffers from the disadvantages described above, but the model does provide insight into a possible mechanism for object selection and with it, an additional solution to the problem of translation and scale invariance.

7 Temporal order

Whilst it is possible to conceive of the ventral stream building features to represent individual views of objects, the question still remains as to how neurons learn to treat their preferred feature as the same, irrespective of size or location. Indeed, ultimately, one would like to understand how neurons learn to recognise objects as they undergo non-trivial transformations perhaps due to changes in viewing direction or lighting.

One solution to this problem is to assume that each neuron receives some external information as to the identity of a particular stimulus. Of course, this simply begs the question of where this information originates in the first place. To describe a potential solution it is worth reflecting on what clues our environment gives us about how to associate the stream of images that we see in everyday life. Recently, several theorists have argued that our natural environment provides a temporal cue to object identity. This cue emerges from the simple fact that we often study objects for extended periods. This then provides us with a simple heuristic for deciding how to associate novel images of objects with stored object representations. Since objects are often seen over extended periods, any unrecognised view coming straight after a recognised one is most probably of the same object. This heuristic will work as long as accidental associations from one object to another are random and associations from one view of an object to another are experienced regularly. There is every reason to suppose that this is actually what will happen under normal viewing conditions, and that by approaching an object, watching it move, or rotating it in our hand we will receive a consistent associative signal capable of bringing all of the views of the object together.

It was (Miyashita, 1993) who discovered that many neurons within IT cortex had developed selectivity for small sets of fractal images which he had been using in a short-term memory task. Although this task did not explicitly require the overall test sequence to be remembered, Miyashita noted that these neurons consistently responded well to single images which neighboured one another in the test sequence. For example, one neuron might respond preferentially to images 5, 6 and 7, whereas another neuron would respond to images 37, 38 and 39. The fact that the images were generated randomly, meant that there was no particular reason - on the grounds of spatial similarity - why these images should have become associated together by a single neuron. Instead, the results indicate the importance of temporal order in controlling the learning of neural selectivity. Recent studies of human recognition learning have found evidence for such a mechanism as well (Wallis & Bülthoff, 2001). Taken to-

gether the two sources of evidence provide important preliminary support for the temporal association hypothesis (Wallis & Bülthoff, 1999). Several network models have made successful use of the temporal cue to view association and it forms the core of learning in the model described by (Wallis & Rolls, 1997).

8 Discussion

This article has reviewed much of the current thinking on object recognition. In particular it has proposed the presence of a distributed, view-based representation, in which objects are recognised on the basis of multiple, 2D feature selective neurons. Specialist cells appear to play a role in associating such feature combinations into certain non-trivial image transformations, coding for a certain percentage of all stimuli in a largely view invariant manner. We have also pointed to evidence that a convergent hierachy is used to build invariant representations over several stages, and that at each stage lateral competitive processes are at work between the neurons.

In the final section we argued that temporal association could act as a cue for associating views of objects. It is worth pointing out that if such a mechanism does exist it can only work in the ventral stream since it would *not* be appropriate in the dorsal visual system, in which motion and location are processed (Ungerleider & Haxby, 1994). Indeed, the importance of using temporal association in invariant object recognition, and the importance of not making such assoications in the part of the visual system involved in processing motion and location, might be a fundamental reason for keeping these two processing streams apart.

We also made mention of the analysis of visual scenes both within and beyond the ventral stream. Although much has been said about the roles of the parietal and temporal lobes, relatively little has been said about the frontal lobe. What we do know, is that it acts as a temporary or working memory store and that neurons within the frontal lobe are responsive to combinations of both where and what an object is. It may well turn out that the frontal lobe acts as a running store of objects currently being represented within a scene (Rensink, 2000), and a challenge for models in the future will be to integrate the frontal lobe into the overall picture of scene analysis.

References

Bülthoff, H., & Edelman, S. (1992). Psychophysical support for a two–dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences, USA*(92), 60-64.

Deneve, S., Latham, P., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience., 4*(8), 826-831.

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

Hummel, J., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480-517.

Logothetis, N., & Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience, 19*, 577-621.

Miyashita, Y. (1993). Inferior temporal cortex: Where visual perception meets memory. *Annual Review of Neuroscience, 16*, 245-263.

Olhausen, B., Anderson, C., & Essen, D. V. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13*, 4700-4719.

Rensink, R. (2000). The dynamic representation of scenes. *Visual Cognition, 7*, 17-42.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*, 1199-1204.

Rolls, E. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*(335), 11-21.

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*(66), 170-189.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*(381), 520-522.

Ungerleider, L., & Haxby, J. (1994). 'what' and 'where' in the human brain. *Current Opinion in Neurobiology*, *4*, 157-165.

Wallis, G., & Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences, 3*, 22-31.

Wallis, G., & Bülthoff, H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences*, *98*(8), 4800–4804.

Wallis, G., & Rolls, E. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology*(51), 167-194.

Webster, M., Bachevalier, J., & Ungerleider, L. (1994). Connections of inferior temporal areas TEO and TE with parietal and frontal-cortex in macaque monkeys. *Cerebral Cortex*, *5*(4), 470-483.