New Investigations into the Visual Representation of Objects and Scenes

Alex Burmester
Supervisor: Dr. Guy Wallis

Doctor of Philosophy
University of Queensland
St. Lucia, Queensland
Australia 4072
2005
Abstract

This thesis deals with human visual perception of scenes, where ‘scene’ is a term used to refer to a complex visual stimulus consisting of a large number of different elements/objects and subtending most, if not all, of the visual field. This thesis deals less with initial, low-level perceptual processing of such stimuli and more with the high-level processing that follows. Specifically, the thesis deals with processes in vision that act on information about the real-world objects in the visual scene. In visual terms, an object can be defined as an entity that is coherent in space and time, is restricted to a localised portion of the scene and also has some meaning to the observer viewing it (i.e., they can use and/or identify it). In keeping with this object-oriented approach, areas such as the attentional selection of objects in a visual scene, the recognition of those objects and the maintenance of information about objects in memory will all be dealt with by the thesis; whereas areas such as the perception of basic visual qualities (e.g., orientation, colour, luminance and contrast) will not. The thesis begins by reviewing literature on scene perception, an area which deals with how humans explore a complicated visual environment (the scene) and select objects from it to assist in their current behavioural goals. Generally, it is thought that observers use covert visual attention and overt eye-movements to orient the high-resolution parts of the retinas (the foveas) towards salient parts of their visual environment. The conventional view is that processing related to recognition of objects then ‘takes over’ in a serial manner after this exploration. Despite the apparent separation of these attentional and recognition processes, they are still critically dependent on one another, especially in the context of the perception of a complete scene (this may not be apparent in the perception of isolated visual stimuli). The experiments of this thesis aim to elucidate this interaction by using paradigms from the scene perception and visual attention literature. Although this is a large body of literature, there are still several key questions that it has not answered. The first of these relates to the perception of dynamic (motion-based) scenes. The literature on scene perception has largely concentrated on the perception and representation of static (motionless) scenes, generally presented in photographic format, either in isolation or in a series with other scenes. Because the visual world of humans is filled with motion, studies of scene perception must head towards incorporating
dynamic, motion-based elements into their experimental programs to retain and expand their ecological validity. This is something the current thesis attempts to do. The first few experiments investigate the perception of dynamic scenes by using a basic visual search paradigm. The purpose of these initial experiments is to precisely quantify the relative contribution different static and dynamic stimulus attributes have in guiding and focusing visual attention. The second unanswered question in the scene perception literature relates to the interaction between the immediate selection and recognition of objects and their maintenance in memory for performance of a later task. There is much physiological evidence, coming from studies using single cell recordings, that looks at connections between temporal and frontal cortical areas thought to sub-serve the maintenance of visual information. There is also brain imaging evidence that looks at specific loci in the brain where working visual representations may be stored and/or supported. However, there is little work using EEG technology, which can give a picture of gross physiological function at small time scales. Because of the immense speed of temporo-frontal interactions, and the wide spatial extent of the areas involved, EEG is an excellent candidate for examining such activity. Therefore, this thesis attempts to examine the neural underpinnings of working visual representations in humans using EEG. The first EEG experiment of this thesis uses the same, simple visual stimuli as the behavioural experiments in order attain maximum control over the difficulty of the attentional and recognition tasks under consideration. The remaining EEG experiments use more complex stimuli, in order to further generalise the findings of the initial EEG experiment and the previous behavioural experiments. The data from the behavioural and EEG experiments will most likely complement each other, as the behavioural experiments aim to quantify the extent to which various visual features contribute to the formation of working visual representations while the EEG experiments seek to elucidate the neurophysiology behind the formation of such representations. Overall then, the behavioural experiments are looking primarily at how such representations are formed whereas the EEG experiments are looking primarily at where they are formed.
Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(Alex Burmester
Supervisor: Dr. Guy Wallis)
# Table of Contents

## Chapter 1 Introduction

1.1 Experiments of the Thesis .................................. 5

## Chapter 2 The Perception and Analysis of Visual Scenes

2.1 What is a ‘Visual Scene?’ ................................. 8
2.2 Viewing Visual Scenes .................................... 9
2.3 What Occurs During a Single Fixation? ................. 12
2.4 What Occurs Across Multiple Fixations? ............... 16
2.5 Change Blindness ........................................ 18

## Chapter 3 Visual Attention

3.1 What is Attention? ....................................... 22
3.2 Theories of Attention .................................... 23
3.3 Visual Attention in Search Tasks ....................... 26
3.3.1 Similarity Theory ................................. 26
3.4 Features and Conjunctions ............................. 27
3.4.1 Feature Integration Theory ....................... 27
3.5 Attention as a Spotlight or a Zoom Lens? ........... 31
3.6 Object Based Visual Attention ......................... 33
3.7 Coherence Theory ..................................... 35

## Chapter 4 Object Recognition

4.1 Objects and Visual Attention ............................ 39
4.2 Why Object Recognition is a Problem in Visual Perception .... 40
4.3 Approaches to the Study of Object Recognition .......... 41
4.3.1 Structural Descriptions .......................... 41
4.3.2 View-Based Approaches .......................... 42
4.3.3 Invariant Properties and Feature-Based Approaches .... 43
4.4 Neurophysiology of Object Recognition ................. 45
4.5 The Timecourse of Object Recognition Processes ...... 49
Chapter 1

Introduction

Visual perception can be described as the process of an observer receiving information\(^1\) conveyed by the structured light coming from the environment around them. However, perception is not simply a passive process of taking in information. Instead, it involves the filtering and use of this information by neurological systems, in order to facilitate the actions and behaviour\(^2\) of the organism (Gregory, 1997). A central question in visual perception research is how humans perceive such a continuous, stable and extensive three-dimensional world given that they receive visual information through two relatively small and flat sheets of neural tissue (the retinas), connected to the brain via the optic nerve - an apparatus that initially seems quite primitive for such a purpose (O’Regan, 1992). This thesis, like much visual perception research, aims to make a contribution to the answering of this question (henceforth referred to as the ‘broad question’ of visual perception).

Early vision research did little to solve the ‘broad question’ but instead made contributions regarding the responses of humans and animals to relatively simple, isolated visual stimuli presented on a flat screen of some sort. For instance, the pioneering work of Hubel and Wiesel (1968) showed how responses of different groups of neurons in the macaque monkey’s visual cortex were sensitive to different blobs, lines and colours presented on a projection screen. The contribution of this type of work is that it has elucidated information processing strategies em-

---

\(^1\)Formally, information is defined as a message communicated from a sender (e.g., the visual environment) to a receiver (e.g., the observer) and can be contrasted with noise, which inhibits the communication of this message (Shannon, 1948). Information is measured by the extent to which it reduces the uncertainty of the receiver about some variable (e.g., the visual environment). In other words, more information means less uncertainty.

\(^2\)The terms ‘action’ and ‘behaviour’ are not used interchangeably here. Rather, action is defined as a unitary, single movement of an organism on a short time scale while behaviour is defined as the sequencing of a number of such actions into a series which helps the organism attain some goal established prior to the initiation of that sequence (Fuster, 1997b).
ployed by the brain which enable it to filter information from the retinal image and break it down into elementary features. However, such work can do little to demonstrate what happens to the information once it has been filtered. The current thesis deals with these more complicated processes, which are best examined with reference to the actions and behaviour of an organism.

As was mentioned previously, perception proceeds ultimately to sub-serve the actions and behaviour of the organism (Gregory, 1997). Because there is a constant link in the natural environment between what an organism sees and what it does, an organism is usually looking at something because it provides stimulus for a prospective movement. Understanding this perception-action link is critical for understanding what happens to visual information after it is processed in the neural substrate by early (striate) visual areas. The reason for this is that the extra-striate flow of information is often mediated by cognitive processes related to the behavioural goals of the observer (Goodale & Humphrey, 1998). One of these behaviourally-relevant processes dealt with by this thesis is attention, a process that serves to select stimuli of interest\(^3\) to the organism, so that they may enter conscious awareness and be processed further (Hatfield, 1998). Selection of a stimulus usually also involves the recognition of that stimulus by the observer, so that they may identify\(^4\) how it can facilitate their actions and behaviour (Logothetis, 1996). Once an observer has recognised and attended to a visual stimulus, action can be directed to that stimulus. This action can be immediate or delayed. When action towards a stimulus is delayed, information about that stimulus must be maintained across the period of the delay. This maintenance of information is typically referred to as memory (Fuster, 1997b). The short-term maintenance of information sub-serving behavioural goals is termed working memory (Baddeley, 2003).

Discussions of memory and the maintenance of sensory information inevitably involve the concept of representation, and so this concept will be briefly discussed here. A representation can be defined as a coherent piece of information linked to a previous sensory event or stimulus, otherwise known as a ‘functioning isomorphism’ of that event or stimulus (Gallistel, 2001). The reason a representation is linked to a sensory event or stimulus is because it is typically a transformation of information coming from that event or stimulus. The term transformation, as used here, means that the information from that stimulus is retained, but that it

---

\(^3\)The term ‘interest’ can reflect both the basic visual salience of the stimulus and/or its relevance to current behaviour.

\(^4\)This identification need not occur inside conscious awareness - it can be implicit or automatic for highly familiar stimuli.
The thesis aims to contribute to the broad question of visual perception by examining interactions between the three behaviourally-relevant processes of object recognition, attention and working memory. The thesis does this by attempting to identify the neural substrate which supports the short-term temporal continuity of visual experience and attempting also to identify how and where in the brain a working neural representation of visual input is maintained. These issues cannot be examined without reference to three processes just identified and their interaction with one another.

Because the topics just discussed are very broad in scope, the thesis focuses on aspects of them that have received relatively little attention in the literature. One aspect of the attention-recognition-memory interaction that has received little attention is the role of dynamic stimulus attributes in this interaction. Conventionally, these topics have been examined using visual stimuli of a static nature. However, because the human visual environment is filled with dynamic stimuli, the influence of such stimuli must be examined to enable experiments in this area to expand and retain their ecological validity.

1.1 Experiments of the Thesis

The first experiments of the thesis relate to how attention is allocated across the visual field in different circumstances. Experiments 1, 2 and 3 use two-dimensional search arrays and investigate the effect dynamic stimulus attributes (e.g., motion, changes in location) have on an observer’s visual search efficiency and their ability to notice change in the search field. In all three experiments, the effect that changes of different stimulus properties have on change detection performance is analysed. Experiment 1, however, consists of a more basic setup than the other two as it is designed to establish psychophysical thresholds for the detection of various types of stimulus change. In Experiment 2, these thresholds are used to calculate various manipulations of the types of stimulus change and their magnitudes. This experiment involves conditions where, like Experiment 1, observers are required to detect change in a single search element but also includes conditions where they are required to detect multiple changes in a single search element. The aim of this experiment is to use the thresholds computed from Experiment 1 to establish how multiple changes occurring simultaneously affect the dependent measures of accuracy and reaction time. More broadly, this experiment aims to determine the extent to which various types of change are equivalent (or not) in
their ability to attract an observer’s attention and be detected. Experiment 3 extends this paradigm and this aim by requiring an observer to detect multiple changes occurring in different search elements simultaneously.

Experiment 4 uses stimuli from the previous three experiments but measures EEG activity rather than behavioural performance. Because of the sensitivity of the EEG measure to artifacts such as eye movements, the stimuli in this experiment are presented under more controlled conditions (e.g., all stimuli are stationary). This experiment, like experiments 1-3, involves manipulations to the set size, type of change and the magnitude of change of search elements. These manipulations are structured in such a way to systematically change the difficulty of the change detection task and this difficulty will be the primary independent variable under consideration in the analysis of EEG data from this experiment. Because this experiment is using EEG rather than behavioural measures, it shifts the focus from how detectable change in various stimulus attributes is to an examination of the neural correlates of change detection under various stimulus conditions. The remainder of experiments in this thesis also use the EEG measure.

Experiment 5 involves participants being presented with recognisable objects and fractal stimuli in rapid steams and tests their recognition performance for these stimuli. Rather than the comparatively simple process of detecting change in simple stimuli, then, this experiment requires the retention of more complex stimulus information. The main independent variable under consideration in Experiment 5 is the timing of rapid presentation and how this affects the maintenance of object information as revealed by fronto-temporal EEG activity. This experiment is based on the Rapid Serial Visual Presentation (RSVP) and Repetition Blindness (RB) paradigms discussed in Chapter 3 and is designed to establish a timecourse of the neural interactions between the temporal and frontal lobes in maintaining visual object information.

Participants in Experiments 6 and 7 are involved in tasks where their performance depends on maintaining visual information for up to 30 seconds, under various stimulus conditions. Observers are required to recognise the stimulus presented before the delay amongst a number of distractor stimuli presented after the delay. As well as being presented in temporal (but not spatial) proximity to distractor stimuli, the target stimulus is presented under different spatial and temporal conditions (e.g., rotated in space, presented for a shorter time). In both experiments, a control group has their behavioural responses analysed while an experimental group has their behavioural and EEG responses analysed. The EEG
conditions are aimed at examining the neural substrate that enables the maintenance of visual information they require. This examination will be made not only with reference to the EEG data, but also with reference to the behavioural data from the whole thesis. By doing this, this thesis aims to elucidate the interaction between recognition, memory and attention and examine the neural substrate sub-serving that interaction. However, as was mentioned before, the examination of this interaction occurs primarily in the context of dynamic visual stimulus attributes.
Chapter 2

The Perception and Analysis of Visual Scenes

2.1 What is a ‘Visual Scene?’

The history of research in visual perception has included several different theoretical approaches to fundamental questions such as ‘what does it mean to see?’ and ‘what is the nature of visual input?’ For instance, there are many researchers who consider that visual input should be treated as an image or series of images (see Marr, 1982) and others who think it should be treated as a dynamic array (see Gibson, 1979). In fact, both of these approaches may be more or less relevant depending on the type of perceptual task under consideration (Goodale & Humphrey, 1998). However, because these different approaches exist, terms as broad as ‘visual scene’ may be interpreted in different ways depending on the theoretical leaning one has. Therefore, at least for the initial sections of this thesis, the term ‘visual scene’ will be restricted by the following explicit definition:

A relatively static input taking up most of the visual field, that is presented to an observer long enough for them to explore it visually with movements of their eyes. This ‘static input’ could be a photograph presented to a visual observer (in, say, an experimental situation) or it could be the natural visual environment which they are currently observing.

Later in this thesis, the meaning of a ‘dynamic visual scene’ will be considered, but for now the discussion will be restricted to static scenes. Presenting a static picture to observers is a way of attempting to mimic, inside a laboratory, the complex visual stimulation that bombards humans in the real world. As with any research in visual perception and visual cognition, the area of scene analysis is concerned with how observers respond to visual stimuli in the real world and
not just in the laboratory. However, the study and discussion of this process is often limited to relatively static or unchanging scenes because these are the easiest to use while still maintaining experimental control. In current research, scenes are often presented to participants simply as photographs on a computer screen, to which they are required to make a forced-choice response on a computer input device. Experimental control can also be enhanced by creating artificial visual scenes via computer (see Figure 2.1) - creating such scenes can allow the experimenter to control high-level features of the scene. Also, low level properties of natural scenes can be controlled somewhat by applying statistical algorithms to the pictures that equalize them along dimensions such as Fourier power\(^1\) (see Johnson and Olshausen (2003) for an example of this).

This chapter of the thesis deals with the initial process of an observer exploring a scene and examining its global features and properties. Later chapters look at visual processes that involve more localised segments of visual space (e.g., object recognition) and also deal with how visual processes interface with more complex cognitive processes such as working memory and how visual information is used in these processes.

Figure 2.1: An example of computer-generated scenes.

2.2 Viewing Visual Scenes

An observer can passively or actively scan or view their visual environment. Here, what is meant by ‘passive viewing’ is that the observer scans the visual scene

\(^1\)Fourier Analysis involves extracting component sinusoidal functions from any complex periodic function. A two-dimensional image can be considered a complex periodic function, and so has component sinusoidal functions (visualised as gratings of various frequencies and orientations). Equalising a series of images on Fourier Power involves modifying the images so that all low-frequency Fourier components are equal and only the high-frequency ones are different.
with no particular goal in mind. Active viewing, on the other hand, means that the observer is exploring the visual scene with some purpose - usually this means the observer is looking for a particular visual ‘target’ within that scene. If an observer is actively looking for a target in a visual scene, this target will be relevant to the task in which the observer is currently engaged. The term ‘task’ here is used to refer to the actions of the observer on a relatively short time scale. This task could be an behavioural/physical task involving motor movements or a psychological task involving internal planning and thought. With these considerations taken into account, the process of ‘scene analysis’ is defined as an observer scanning their eyes over a visual scene and directing them to the most interesting parts of that scene (Intraub, 1997), where what is ‘interesting’ could be related simply to visual distinctiveness or, alternatively, to what the observer is looking for a priori, based on their cognitive state and goals. The first experiments conducted to determine what an observer considers ‘interesting’ in a visual scene were conducted by Yarbus (1967).

Yarbus, regarded as one of the pioneers in the study of eye movements and scene analysis, recorded the eye movements of participants while they were allowed to visually explore pictures presented to them for a fixed period of time (see Fig 2.2). Yarbus (1967) demonstrated that humans do not scan visual scenes with their eyes in a regular fashion, but instead make fast, jumpy eye movements (called saccades) that place the foveas over ‘regions of interest’ and alternate between several of these regions within the one scene. Initially, the ‘regions of interest’ in a scene appeared to be a relatively objective property - that is, different participants would look at the same parts of the same scene for around the same periods of time. This was true as long as participants were allowed to fixate freely on the scene. However, if participants were given instructions to look for specific things or they were asked to remember particular portions of the scene, fixation behaviour changed depending on the instructions. Therefore, the ‘regions of interest’ appear to depend on the scene being looked at and/or the cognitive task being performed. Regardless of task demands however, Yarbus found that observers made up to three separate eye movements per second, meaning fixation durations lasted from around 300 milliseconds to 1 second.

It is thought that there are many different reasons that part of a visual scene could be a ‘region of interest’ to an observer. Some of these relate to visual

---

2Saccades were first demonstrated in late nineteenth-century studies on reading (Coltheart, 1999).
3The fovea is the central portion of the retina that contains the greatest density of light- and colour-sensitive cones and, because of this, receives the highest resolution visual input.
properties of the scene itself and others relate to the internal psychological state of the observer, or the task they are performing. These are categorized as ‘bottom-up’ and ‘top-down’ processes respectively. In general, a bottom-up process is one initiated by simple processes which in turn activate more complex processes and a top-down process is one in which the reverse occurs. An example of a bottom-up process would be if part of a visual scene is a distinctive colour relative to the rest, an observer might be drawn to look at it because of factors such as colour contrast. After looking at this stimulus, the processes involved in detecting colour contrast would occur before higher level processes concerned with identifying the stimulus’ semantic properties. An illustration of a top-down process would be a situation where an observer was looking for their friend’s face in a crowd. In this situation, the observer would probably be drawn to look at peoples’ faces rather than the part of the scene containing the greatest colour contrast (e.g., the person wearing the brightest and most colourful jumper), because the observer is concentrating on finding their friend (i.e., they are conducting a goal-based search which is a top-down process). In visual terms, bottom-up processes generally reflect a search for low-level features in the input while top-down processes generally reflect a search for higher-level features. Also, top-down processes are dependent to a much

---

4The bottom-up/top-down dichotomy is used extensively in psychological and neuroscience literature and not just in areas relating to visual perception. Neurologically, a bottom-up process is considered as one that proceeds from earlier, more basic brain areas to later, more complex areas and is usually governed by information external to the observer. A top-down process, by contrast, is initiated in more complex brain areas and proceeds to more basic areas (Fuster, 1997a). This interpretation is neurological, but the concept is not always interpreted this way - it can be explained in the context of a hierarchy of psychological functions as well.
greater extent on the volition of the observer than are bottom-up processes.

Since saccades occur so frequently during viewing of a visual scene, they are obviously an important part of any theoretical analysis of scene perception. However, the saccades themselves are of enormously short duration and, during them, the retinal input is blurred (see Figure 2.3). Furthermore, many researchers suggest a central mechanism of ‘saccadic suppression’ acts to further suppress retinal input during saccades (Wurtz, 1969). Regardless of the cause, visual perception is virtually non-existent during the saccade. These results have led many researchers to posit that the pattern of fixation points and the actual events occurring during a fixation are of far more importance than what occurs during a saccade (Intraub, 1999; O’Regan, 1992). Taking this into account, we will first look at studies dealing with what occurs during a single fixation and then look at studies dealing with patterns of fixation behaviour, and what occurs internally during scene exploration with multiple fixations.

![Figure 2.3: A depiction of what happens to retinal input at different time points during a saccade - fixation, saccade start and when the saccade reaches high velocity.](image)

2.3 What Occurs During a Single Fixation?

The development of the tachistoscope in the late nineteenth century made the presentation of brief visual stimuli possible, and this piece of equipment was used for much of the twentieth century’s program of vision research, up until the advent of computers with significant visual display capabilities. The rationale for using brief, stationary stimuli is that the information intake of the visual system occurs on a fixation by fixation basis, and that each fixation lasts only a brief period of time. Stimuli presented for 100 msec or less are thought to be the best for probing what happens in a single fixation, as the latency of an eye movement is always greater than 100 msec after a fixation is started (Coltheart, 1999). A paradigm was developed by Sperling (1960) for studying the nature of visual processes occurring within a single fixation. Sperling (1960) presented subjects with a 3 x 3 or 3 x 4 array of letters for 200 msec and they were told to remember as many as possible. He found that subjects reported only around 3-4 letters.
However, after considering that this might reflect a limit on the number of letters subjects could recall, rather than the number of letters that may have been stored in a temporary sensory register, he modified the experiment.

The modified experiment is a prototype for what is now known as the ‘partial report’ paradigm. It involves observers being presented with an auditory tone after the presentation of the visual display, where the pitch of this tone identifies the row of letters the observer must report (observers had previously been trained with what tone matched what row of letters). Therefore, in each trial the observer is asked only to report a single row of three letters. The results using this methodology were that observers successfully reported whatever row they were cued to report. However, the report was only complete (i.e., 3 out of 3 letters correct) if the tone cue was given 100 msec or less after presentation of the letters. It became less complete if the delay was longer (see Figure 2.4). This result indicates that a ‘free recall’ method is not the most appropriate when studying the mechanisms of early visual memory - as visual memory has a higher capacity than the memory involved in free recall and there appears to be a bottleneck between the two. The visual memory apparently revealed by his experiments was termed ‘iconic memory’ by Sperling and belongs to the general class of ‘sensory memory’. Sensory memory is thought to code sensory input in a way that is relatively raw and untransformed - for this reason, it requires a high capacity and because it has a high capacity, it is fast-decaying (for the sake of information processing efficiency) (Humphreys & Bruce, 1989). Figure 2.5 shows the hypothesised role of such a memory store relative to memory that uses more abstract codes (short- and long-term memory).

Another significant result of Sperling’s studies is that what was presented visually before or after the array of letters affected the response of the observers - if anything other than a black field was presented, the partial report superiority over the full report only lasted 0.5 sec, whereas it lasted around 5 sec if only blank fields were presented. Subsequently, this effect was termed ‘visual masking’. A stimulus functions as a mask if it reduces the detectability of the target stimulus (Turvey, 1973; Coltheart, 1999). A mask presented before the target stimulus is termed a ‘forward mask’ and one presented after is termed a ‘backward mask’, indicating the position they have relative to the target stimulus and the direction (in temporal terms) of their influence upon sensory memory.

Figure 2.6 shows the effect of a masking stimulus on the recall of one out of 5.

Additionally, experiments can use ‘lateral masks’ which are spatially separate from the target stimulus, but not temporally separate - masking appears to act both spatially and temporally under the right conditions (Humphreys & Bruce, 1989).
Figure 2.4: The results of Sperling’s partial report method. From Sperling (1960).

an array of letters. Also shown in the figure is the hypothesised ‘iconic image’
that results from the integration of the array of letters with the mask stimulus.
The results of a study conducted by Averbach and Coriell (1961) are shown
in Figure 2.7. In this study, a masking effect was only found as long as the
inter-stimulus interval (ISI) was around 100 msec. It is now thought, because of
experiments like this, that iconic memory fades a little after 100 msec following
stimulus onset, so stimuli must be presented within this window for them to be
integrated with the initial stimulus (Coltheart, 1999). From these considerations,
it is clear that visual masking offers a way to study the capacity iconic memory
has for integrating visual information within a single fixation. However, the main
bearing that studies of visual masking have on scene perception is that the visual
input occurring at one fixation may act as a backward mask on the input of the
previous fixation and/or may act as a forward mask on the input of the next
fixation (Irwin, Brown, & Sun, 1988).

A later paradigm developed to investigate further the properties of visual
memory is rapid serial visual presentation (RSVP) (Raymond, Shapiro, & Arnell,
1992). This paradigm involves presenting the observer with a series of pictures,
one after the other, for a brief period of time each (usually around 100 msec).
The purpose of this short presentation duration is, as with Sperling’s experi-
ments, to limit the influence of eye movements and investigate what information
is processed from single fixations. As further insurance against eye movements, observers are usually instructed to fixate on a central cross on the screen during these experiments. Typically, results from RSVP studies show that recall memory for objects and words presented is very bad and recognition memory\(^6\) is only slightly better (Intraub, 1999). A phenomenon of interest that occurs in some RSVP experiments is termed ‘repetition blindness’ (RB) - the finding that observers are unable to report or realise when prompted that an object or word has been repeated in the RSVP sequence within a second or so (Bavelier, 1999). Furthermore, if an object is presented once and then presented with a change in orientation, size or location within the picture, RB still occurs for that object (Kanwisher, Yin, & Wojciulik, 1999) - therefore, RB is invariant to certain changes in low-level visual input as long as high-level features (i.e., the same object) remain.

Another important finding from these experiments is that when there is an inter stimulus interval (ISI) of around a second or more, memory for presented pictures is very good and RB does not occur. Also, the bad memory for stimuli presented sequentially was found not to be due to low-level masking effects (Kanwisher et al., 1999). Therefore, these effects have been attributed to an interruption of higher-level encoding of the stimulus. Whereas masking is thought to interrupt encoding into sensory memory, RSVP style presentation is thought to interrupt encoding into a short-term memory store, where the code is more abstract (Intraub, 1999). Because this coding is more abstract, it is thought to take a longer time to occur after stimulus offset (greater transformation of the information has to occur, compared to the relatively raw and untransformed na-

\[^6\]Recognition memory is what is tested when observers are asked to answer yes/no when presented with stimuli that were and were not presented as part of the original RSVP presentation.
ture of coding in sensory memory). This would explain why memory is bad for stimuli presented in these paradigms unless the ISI is extended to a second or so (Kanwisher et al., 1999). The concepts of a high-capacity sensory store that encodes in visual terms coupled with the concept of a lower capacity short-term store that codes in more abstract terms are the most important for understanding what takes place in the visual system during a single fixation. They provide a necessary foundation for understanding empirical data related to the viewing of whole scenes, to which we will now turn.

2.4 What Occurs Across Multiple Fixations?

Theories of scene perception have always asked how the comprehension of a complete, expansive visual environment is achieved from individual fixations on regions of interest. Initial theories suggested that there was an internal, high-
capacity visual buffer in the brain to which each view was added in a piecemeal fashion (see Irwin et al. (1988); Irwin (1991) for a review of these theories). This buffer was supposed to be spatiotopic\textsuperscript{7} in its representation of visual space. This approach, however, has been severely criticised in more recent times on both theoretical and empirical grounds. In theoretical terms, the approach is reminiscent of the infinite regress problem that has been discussed in the context of the philosophy of the mind. Having an internal screen simply moves the locus of visual input inward from the retina and does little to solve the problems of visual perception, which centre on feature extraction from the image/array in order to facilitate the organism's actions (see Ryle (1949) and Dennet (1991) for philosophy regarding the infinite regress and its relationship to Descartes' mind-body dualism).

\textsuperscript{7}The term 'spatiotopic' is used when a system, process or physical substrate represents spatial location by a two-dimensional co-ordinate system representing the whole possible sensory field of an observer (e.g., the whole visual field) - this is contrasted to the term 'retinotopic', which refers to co-ordinates representing physical location on the retina. The key difference is that the retinotopic co-ordinate of a visual stimulus will change dramatically with eye and head movements, whereas the spatiotopic co-ordinate will not. In other words, spatiotopic co-ordinates are robust to eye and head movements, retinotopic ones are not. Because of this, spatiotopic representation occurs higher in the cortical/psychological hierarchy than does retinotopic representation, as it necessitates extra computation.
A different objection, raised by O’Regan (1992) suggests that the ‘internal screen’ approach overly complicates the apparatus required for visual representation - a high-capacity buffer would require massive computational power early in the visual hierarchy, and it is unclear how the different views from different fixations would be integrated into a single spatiotopic representation. This integration would have to be complicated, because each fixation is not equal in terms of factors like retinal resolution - but creating an extended spatiotopic representation would have to take this into account (Rensink, 2000a). O’Regan (1992) also suggests that there is no need for such a system because the eyes are exploratory sense organs - the information about the external environment is available quickly and easily because the eyes can move rapidly to foveate any ‘regions of interest’. Studies aimed at exploring these theoretical issues generally fall into the category of trans-saccadic memory - studies which aim to discover what information observers are able to maintain from one fixation to the next (across saccadic eye movements) (Coltheart, 1999). Note that the use of the term ‘memory’ here does not necessarily refer to the ability to explicitly recall visual detail in a verbal report - it can also refer to the sensory memory previously discussed. This is an important distinction because it can affect the way in which the memory is measured experimentally.

2.5 Change Blindness

An important phenomenon that has emerged in studies of trans-saccadic memory and visual memory in general is termed ‘change blindness’. This phenomenon is now typically studied using the paradigm depicted in Figure 2.8. The experiment involves a picture (A) alternating with a blank field. A modified version of the picture (A’), in which a significant portion of it has been changed is presented after the blank field, and the sequence (A, blank, A’) repeats until the observer indicates they have noticed the change. Although the changes occur to significant portions of the visual scene, observers perform very badly on this task. Changes are often not noticed until about a minute of the two pictures alternating has passed. During the viewing time, observers are free to explore the scene with their eyes, but do not seem to pick up the change unless they are attending to the portion that is changing, as it changes (Rensink, O’Regan, & Clark, 1997). The purpose of the intervening blank field is to mask the appearance of a flicker in the portion of the picture that is changing. This flicker would attract the observer’s attention, not because it represents a change in the high-level properties of the
scene, but because it is a low-level visual transient (specifically, an apparent motion signal) that acts to capture the observer’s attention automatically (Simons & Levin, 1997).

Studies of change blindness are most centrally interested in how observers build up a representation of their visual environment from multiple fixations. The phenomenon can be described as the inability of observers to notice changes to a visual scene unless those changes produce a visual transient (i.e., a quick change in colour and/or luminance) at a specific position on the retina (Simons & Levin, 1997). Findings from change blindness studies seem to indicate that although humans perceive a relatively detailed, rich visual world at any instant, we actually do not store a detailed visual representation which grows as we change from one view to the next or move our eyes from one position to another (Rensink, 2000a; Simons, 2000; Intraub, 1997).

Figure 2.8: A typical change blindness experiment.

Change blindness studies using the flicker paradigm grew out of earlier studies dealing with visual change. Changes in these experiments were sometimes made during the saccades of the observer (McConkie & Currie, 1996), sometimes in film-like editing cuts (Levin & Simons, 1997) and sometimes with no intervening period. As an example, McConkie (1979) found that observers do not notice
when words presented in alternating case\textsuperscript{8} change as they make eye movements - the cases of the letters swapped each time the observer made an eye movement. Also, Grimes (1996) instructed observers they were to study a photograph for an upcoming recognition test and that it would change periodically. The photograph changed during some saccades made by the observer, but observers did not notice most of the changes that occurred. This poor performance occurred despite the fact that changes were quite visually significant, such as the hats of two men being swapped. These observations have only recently been subjected to a more rigorous theoretical analysis.

As part of the theoretical analysis of change blindness, Simons (2000) listed the following as possible causes of the phenomenon:

- Overwriting of Representations
- First Impressions Last (Initial Representations Persist)
- Nothing is Stored
- Nothing is Compared
- Features are Combined across Representations

‘Overwriting’ refers to the initial representation being overwritten by the changed one. ‘First Impressions’ appears to be a reversal of this where the initial representation overwrites the changed one. ‘Nothing stored’ is similar to the ‘world as an external memory’ idea of O’Regan (1992), in that it is not necessary or economical for detailed representations to be stored. The causes given by Simons are dealing with representations based on several fixations and so are not the same as the ‘overwriting’ of visual masking or the fast-decaying storage (comparable to the ‘nothing stored’ idea) of sensory memory. They are instead referring to higher level representations, possibly located in a short-term store (see Figure 2.5). Certainly, change blindness studies seem to indicate an absence of any lasting detailed representations of a scene, in contrast to the ‘internal screen’ approach (Rensink et al., 1997). Therefore, the question remains: is a more abstract representation of scene structure created and, if so, how?

The question of what kind of representation is built up in scene analysis has been studied by Rensink and colleagues using various modifications of the change blindness paradigm. Rensink et al. (1997) state that change blindness is probably the result of a failure to allocate visual attention properly across the visual scene.

\textsuperscript{8}Each letter in these words is the opposite case to the previous letter - ThIs Is An ExAmPlE.
In the natural environment, changes are accompanied by something moving, and motion produces a visual transient in the visual field that will act to capture attention, even if it occurs in the periphery (see Rensink et al., 1997; Simons & Levin, 1997; Pashler, 1998). Therefore, changes capture attention in natural circumstances and so attention is properly allocated as a result. In the artificial context of the laboratory, the changes are not accompanied by these transients and so attention is guided to regions of interest not defined by them, but defined by static visual properties such as contour density, colour contrast, visual detail etc. Because searching among static visual properties is a process that is less automatic than attentional capture by a transient (see Posner, Snyder, & Davidson, 1980), the process of searching for high-level changes could be considered a task where allocating attention properly is difficult and, in most cases, inefficient, perhaps because we are not used to it (see Rensink (2000a) for more detail on this concept). The theoretical position of Rensink and colleagues will be considered in more detail in the next chapter. However, before reviewing this position, it is necessary to explore the concept of visual attention in more general terms and this is done at the start of the next chapter.
Chapter 3

Visual Attention

3.1 What is Attention?

Typically, a visual scene in our natural environment contains a large amount and a wide variety of information at any one time. However, only a small portion of this information will be relevant to the observer’s task at hand - it is the selection of this task-relevant information against a background of task-irrelevant information in a visual scene that defines the process of visual attention (Desimone & Duncan, 1995). Because visual attention studies grew historically out of studies relating to attention in general (whether in the visual or auditory modalities, or across modalities), the basic theoretical ideas to come out of research on attention in general will be reviewed before turning to visual attention specifically.

The general populace has a concept of attention suggesting that it is some force or quantity that can be devoted (‘paid’) to various stimuli or withheld from them (i.e., ‘ignoring’). This ‘folk psychology’ treatment of attention suggests that it relies on the volition or will of the observer, but it is important to note that attention can be attracted involuntarily as well by, for example, loud noises in the auditory field or fast movements in the visual field (see Pashler (1998) for a more thorough discussion of the folk psychology of attention). Attention was originally defined in an academic context by William James as follows:

“It is the taking possession in the mind, in clear and vivid form, of one out of several simultaneous possible objects or trains of thought.”

(James, 1890/1950).

This definition is useful even today, because it identifies the phenomenon of selectivity - something that is universally accepted as being integral in defining what is known by the term ‘attention’ (Hatfield, 1998). Furthermore, James’ statement also indirectly identifies the idea of capacity limitations in attentional processing. This indirect identification occurs because the constructs of selectivity and

22
capacity limitation are logically inseparable, as one necessarily implies the other (Pashler, 1998). For instance, if an observer needs to select a particular stimulus to attend to, it must be because they do not have the capacity to attend to all available stimuli simultaneously. In other words, because there are capacity limitations in the amount of attention that can be devoted to stimuli, it is necessary for the attentional process to be selective.

The idea that selectivity and capacity limitations are central to the concept of attention is fairly basic and intuitive, but it has paved the way for later empirical work on attention dealing with its less obvious properties. For instance, the phenomenon of selection in attentional processing allowed it to be later characterised by Broadbent (1958) as a filter that selects some sensory inputs and rejects others\(^1\) (see Figure 3.1). Using this analogy, Broadbent (1958) became a proponent of the ‘early selection’ theory of attentional processing, one of the first empirically testable and formalised theories of this previously nebulous construct.

\[\text{Figure 3.1: A filter-like depiction of Broadbent’s early theory of attention. From Broadbent (1958).}\]

### 3.2 Theories of Attention

Early selection theories of attention, such as that proposed by Broadbent (1958), suggest that all incoming stimuli are processed by perceptual processes, but only

---

\(^1\)Broadbent’s treatment of attention was based largely on the ideas of information theory. Information theory was developed at the start of the computer age, in response to a need to formally and mathematically describe information flow in electronic systems. Later work demonstrated that it could be applied to many physical systems and also to psychological systems, if the structural framework had been sufficiently elaborated. The details of information theory are unnecessary for the current thesis, but more information may be found in Cover and Thomas (1991).
those that are attended to are processed to the extent where they are identified. The distinction between perception and identification should be elaborated here. Perceptual processes capture the physical information from a stimulus - its colour, orientation and luminance for example. Identification is the processing of a stimulus beyond its physical properties. It involves the comparison of the physical information from a stimulus with information stored in memory (Pashler, 1998; Treisman, 1998; Wright & Ward, 1998). Therefore, identification can allow the observer to recognise, semantically tag and even interact with the stimulus. Another important distinction between these two processes is that perception is considered to act in parallel across the sensory field (be it a visual, auditory or tactile field) and identification can only occur serially - this means that multiple stimuli can reach the perceptual stage simultaneously, but only one stimulus at a time can be identified (Pashler, 1998). This reflects a wider and quite old distinction between pre-attentive processes which act in parallel, and attentive processes which act serially (Neisser, 1967). Furthermore, this distinction is similar to that of ‘sensory memory’ from ‘short-term memory’, which was discussed in the previous chapter.

This theory of Broadbent (1958) was termed ‘early selection’ because it proposes that attention acts early in the cortical/psychological hierarchy to select the stimulus required for further processing. It is highly compatible with the filter analogy because it suggests sensory stimuli are processed in multiple channels until they reach the locus of attention (the bottleneck of the filter), where they must enter a single channel. At this point, some stimuli are selected by the attentional filter while others are prevented from entering the attentional processing stage. This filter would be located between the stages of perception and identification. For ‘late selection’ theories, by contrast, the filter is located after both perceptual and identification processes have occurred. It is important to note that most late selection theories do not suggest all available stimuli are processed beyond the perceptual stage. Instead, they suggest that all highly familiar stimuli are - stimuli that the psychological/neurological apparatus has been tuned to deal with (see Deutsch and Deutsch (1963) and MacKay (1973) for examples of late selection theories). Late selection theories can be reconciled with an information processing account using the proposition that familiar stimuli require less information processing resources to be perceived and identified. Late selection theories will not be discussed in greater detail here because, compared to early-selection theories, they have received little empirical support (Pashler, 1998). However, this is not to say they have no merit - indeed, in more recent research on at-
tion, and visual attention specifically, the idea has emerged that both early- and late-selection theories contain ideas of value and the most accurate model of attention may be one that incorporates aspects of both (Treisman, 1998).

The idea that seems to create somewhat of a compromise between, or a union of early- and late-selection theories is that of capacity limitations. The idea of capacity limitations is that during the processing of a higher informational load, more capacity is taken from some bank of cognitive resources than would be during the processing of a low informational load (Palmer, 1998). For instance, Treisman (1998) suggests that the identification processes have a limited capacity, whereas perceptual processes have such a high capacity that it seems unlimited when subjected to many empirical analyses. She suggests further than attention acts to distribute the resources of identification in an optimal manner. As a result of this optimal distribution of resources, when there is an abundance of incoming sensory information, attention acts at an earlier stage (early-selection) to reduce the amount of information allowed through to later (identification) stages in the hierarchy. In contrast, when there is a small amount of incoming sensory information, attention can act at a later stage (late-selection) and allow more stimuli to be identified. By this account, the location of the attentional bottleneck in the hierarchy of processing is variable and will change depending on the demands of the task. A caveat to this idea however, is that stimuli receive more or less processing resources depending on whether they are unfamiliar or familiar (i.e., familiar stimuli require less resources to process than do unfamiliar stimuli) (Treisman, 1998). The more efficient processing of familiar stimuli is a concept that is supported by recent neurophysiological research into highly specialised object recognition areas in the temporal lobe (e.g., see Rolls & Treves, 1990).

In summary, the attentional bottleneck appears to be controlled both by task demands and by the familiarity of the stimulus under consideration. Although this is a more complicated picture of attention, it is one that fits better with empirical data than either early- or late-selection theories in their purest forms. The ideas of capacity limitations, parallel and serial processing and early- and late-selection are of central importance in studies of visual attention, to which we will now move.
3.3 Visual Attention in Search Tasks

In a typical visual search experiment, an observer is presented with an array of small stimuli such as letters, numbers or small shapes. This array will contain a ‘target’ stimulus against a background of ‘distractor’ stimuli. On each trial, an observer is asked to indicate whether or not the target stimulus is present, typically by pressing one of two buttons. The number of distractor stimuli will often be varied, to test the capacity limitations of the attentional processes under study. Also, the similarity between the target and distractors can be varied in order to test how selective the attentional processes are. The dependent measure of such a task can either be the accuracy of the observer’s judgement (how many times they correctly say the target is or is not present when the search field is presented for a fixed time on each trial), or it can be the reaction time (RT) of the observer, in which case the search field is present until the observer presses one of two buttons - ‘target present’ or ‘target absent’ (Palmer, 1998).

A result that has been replicated many times in the visual search literature is that RT will increase with the number of distractors and will also increase with an increase in similarity between the target and distractors. Therefore, using RT as an index of difficulty, the easiest visual search tasks are ones with few distractors and a target that is visually very distinctive compared to the distractors.

3.3.1 Similarity Theory

The typical pattern of RT data discussed above forms the basis of a qualitative theory of visual search called ‘similarity theory’, proposed by Quinlan and Humphreys (1987). The idea is simply that search time depends on how easily a target can be distinguished from its background and also on the number of items of information required to identify the target. More formally, the model is described by the following principles (where ‘T’ refers to target and ‘N’ refers to non-target):

- as T-N similarity increases, search efficiency decreases and search time increases
- as N-N similarity decreases, search efficiency decreases and search time increases
- T-N similarity and N-N similarity are related; decreasing N-N similarity has little effect if T-N similarity is low; increasing T-N similarity has little effect if N-N similarity is high.
3.4 Features and Conjunctions

An increase in RT with the number of distractors only occurs when the target stimulus is not easily distinguished from the distractors (Treisman & Gelade, 1980). In other words, when a target ‘pops out’ against its background, it does not matter how large the background is - only when there is no visual ‘pop out’ does the size of the background matter. An example of an easy visual search task that would produce a flat set size vs. RT graph is shown in Figure 3.3. An example of a difficult search task that would produce a steep set size vs. RT graph is shown in Figure 3.4. Both example stimulus displays show a condition where the target is absent (on the left) and a condition where the target is present (on the right). The easy task requires the observer to look for a target that is defined against the background of distractors on one visual dimension (colour: red against a background of blue). The difficult task requires the observer to look for a target defined against the background on two visual dimensions - shape and colour. In this instance, the target is a red square in a background of shapes that can be either circles or squares and red or blue, but none of which are a red square. Therefore, the observer must look for a target defined by a conjunction of the visual features present in the background distractors. The typical results for feature and conjunction searches are shown in Figure 3.2 as the blue and red lines respectively. The issues of searching for features vs. searching for conjunctions of features was first treated in a theoretical context by Treisman and Gelade (1980), with her Feature Integration Theory (FIT).

3.4.1 Feature Integration Theory

Feature Integration Theory (FIT, see Treisman & Gelade, 1980; Treisman, 1998) is schematized in Figure 3.5. The basic idea is that perceptual processing occurs in independent feature maps or feature modules which are spatiotopic maps activated by the relevant feature throughout the visual field. It is proposed that these feature maps exist for basic, physical visual dimensions such as colour, size

---

2 The term ‘feature’ here is used to identify a particular point along a visual dimension or scale. As an example, colour and shape are visual dimensions, circle and square are points on the dimension ‘shape’ while blue and red are points on the dimension ‘colour’.

3 The concept of spatiotopic feature maps is prevalent in literature on visual search and visual attention. There is neurophysiological evidence to suggest such maps exist, but they are undoubtedly higher in the cortical hierarchy than the retino-topic maps previously discussed (Livingstone & Hubel, 1988). The reason for them being higher in the hierarchy is probably because of the extra computation required to create them - they must be robust to movements of the eyes and head, and, because of this may require a ‘discounting’ or ‘correcting’ of these influences using neural computation.
and orientation. FIT suggests that the outputs of the various feature maps are combined in a spatial location ‘master map’ and that this combination requires attention. Therefore, when a stimulus can be differentiated from its background by the output of a single feature map, attention is not required according to FIT (this situation occurs in Figure 3.3). Attention is required to identify stimuli that can only be distinguished from their background by a conjunction of features. In other words, attention is required to identify stimuli that can only be distinguished by the output of several feature maps. This combination of outputs from several feature maps is termed ‘feature integration’ and is supposed to require attention, in contrast with detection based purely on ‘pop-out’ or ‘feature
Visual search tasks that do not require feature integration (i.e., that can be performed on the basis of visual salience of the target) are supposed to proceed in parallel (using pre-attentive mechanisms), whereas those tasks that do require feature integration are supposed to proceed in a serial manner. This is the common explanation given for the difference in reaction time data for both types of tasks (see Figure 3.2). This dichotomy can be linked to earlier theories of attention, such as Broadbent’s - perceptual processes were treated as processing the physical properties of stimuli in parallel, with no capacity limitation; identification, however, was treated as proceeding serially, because it inherently involves limited capacity. Perhaps then, feature searches need only invoke perceptual processes, whilst conjunction searches necessarily invoke identification processes, as well as perceptual processes.

Although there are problems with the original incarnation of FIT (see Pashler, 1998, Chapter 2), the basic concepts of feature maps, feature integration, parallel processing and serial processing have remained and are even used in modern computational models of attention (see Itti and Koch (2000) for an example). An extended and modified version of FIT was created by Wolfe, Cave, and Franzel (1989) and is called ‘Guided Search’. The main difference between Guided Search and FIT is that Guided Search supposes that parallel search processes can guide serial search processes by giving them access to the result of a pre-attentive,

---

4 A salient stimulus can be considered as one which differs from its background on a single visual dimension (colour, orientation etc.) and has a background that is relatively homogeneous in that dimension (Humphreys & Bruce, 1989).
parallel search. For instance, if an observer is searching for a target defined by conjunction of two features, the parallel process that detects one of these features makes available to the serial process the information about which elements in the search field have that feature. This would make the serial process more efficient by restricting it to a subset of the search stimuli.

Theories such as FIT and Guided Search have helped the development of a large body of visual search literature with the aim of determining what ‘features’ can be detected in parallel and which ones must be detected serially. Wolfe and Horowitz (2004) reviewed much of this research and offers the following list of features which, given the amount of work and results relating to them, are undoubtedly detected in parallel by the visual system:\footnote{Wolfe and Horowitz (2004) refers to this list as ‘undoubted guiding attributes’, in keeping with the Guided Search model.}

- colour

- motion
The attributes that Wolfe and Horowitz (2004) lists as ‘probable’ candidates for parallel/guiding processes include:

- luminance onset (flicker)
- luminance polarity
- shape
- curvature

Both FIT and Guided Search suppose that attention is required to integrate maps of these ‘elementary features’ into a spatial master map, the implication being that the spatial location of stimuli is of central importance in visual attention. The idea of the importance of spatial location and the spatial properties of attention can also be seen in several theories of visual attention that have applicability broader than visual search tasks, and attempt to explain how humans represent and analyze visual scenes in general. These theoretical treatments will be looked at next, as they are also relevant to the empirical work of this thesis.

3.5 Attention as a Spotlight or a Zoom Lens?

William James suggested attention had a fringe or margin - in other words, that it was spatially limited (James, 1890/1950). A more modern, but related, conception is that attention acts like a spotlight over an internal, spatial representation of the visual scene. This notion was introduced by Posner et al. (1980) who stated,

“Attention can be likened to a spotlight that enhances the efficiency of detection\(^6\) of events within its beam.” (Posner et al., 1980) p.172.

This approach is reminiscent of the ‘internal screen’ discussed in the previous chapter, but has more detail to it. The key characteristics of the attentional spotlight is that it has a variable diameter, determined by the field being attended, and that it cannot be split to attend to stimuli in separate portions of the visual

\(^6\)The word ‘detection’ in Posner’s quote comes from the visual perception literature - it refers to a basic perceptual process, not a higher level cognitive process involving abstract coding (cf. ‘identification’).
field (Posner et al., 1980). If two stimuli require attention, the spotlight must be widened, rather than divided. This dichotomy is not trivial, as widening an attentional field may affect the resolving power within that field, whereas dividing it probably wouldn’t.

The spotlight account was originally based on experiments where observers had their reaction time recorded in conditions where they were required to identify the location of a light in their visual field - this location was sometimes cued and sometimes not cued (Posner et al., 1980). A similar, but more elaborate experiment was conducted by Eriksen and Murphy (1987), who presented a target adjacent to a distractor either 1, 2 or 3 degrees of visual angle away from fixation. In some instances, as in the experiments of Posner et al. (1980), there was a ‘pre-cue’ - a cross flashed at the location where the stimulus was going to appear. The distance between the target and the distractor was also varied trial-to-trial. It was found that the target-distractor distance was only important for trials in which there was a pre-cue. Specifically, the distractor had to be very close to the target on the pre-cue trials for it to have an effect (the ‘effect’ of the distractor being to slow reaction time). If it was not, the distractor did not have an effect. By contrast, in trials where there was no pre-cue, the distractor always had an effect. The upshot of this was that the pre-cue seemed to focus attention to a more limited spatial domain than did the absence of a pre-cue. Also, it was found that when there was a pre-cue, the distractor only had an effect if it was visually quite similar to the target. However, when there was no pre-cue, the distractor did not have to be as visually similar to the target.

These findings were integrated in a theoretical account that compared visual attention to a zoom lens - the resolving power and spatial extent share a complementary relationship such that an increase in one produces a decrease in the other (like a zoom lens) (Eriksen & Murphy, 1987). For attention, this means that a pre-cue focuses the ‘lens’ in that region, decreasing the spatial extent while increasing the resolving power. If there is no pre-cue, however, the lens is unfocused by default and so covers a wide spatial extent but has low resolving power. The differences in resolving power between the pre-cue and no pre-cue conditions explain the effectiveness of the distractors. In the pre-cue condition they had to be visually very similar to the target to have an effect because a high resolving power would require that they look similar for there to be confusion in the identification of the target, as a high resolving power means that more visual information is coming through the system and increasing the information available
to any system decreases uncertainty in that system\textsuperscript{7}. In the no pre-cue condition, however, they did not have to be similar, because the low resolving power means a whole range of visual stimuli would tend to look similar.

The theoretical position of attention as a ‘zoom lens’ has been backed up by Shulman and Wilson (1987), who found psychophysical responses to low-frequency gratings are enhanced when making global spatial discriminations, while responses to high-frequency gratings are enhanced when making finer, more local spatial discriminations (typical figures used to test local vs. global discriminations are large letters which are constructed from smaller versions of the same or different letters). The responses to gratings measured in these studies were detection responses - they were of a perceptual nature. Such detection processes are measured by requiring observers to adjust a physical, perceptual property until it matches a standard (Day, 1969). In this study, observers were required to modify the contrast of the gratings until they could only just see them. Because of the effect of attention on detection responses, these results would tend to indicate that the attentional ‘zoom lens’ is acting, at least to some extent, at a perceptual stage early in the visual hierarchy. This tends to support early-selection accounts of attention more than late selection accounts, but the findings may change if the observer's perceptual capacity was overloaded. Regardless of this, the zoom lens metaphor can explain such findings much better than can the spotlight metaphor.

However, the key ideas of the spotlight account are still present in that of the zoom lens - namely, that attention has a limited spatial extent but this spatial extent can be varied. Also, both accounts do not allow divisions of attention - the zoom lens account would state that if attention needs to be divided between two spatially separated stimuli, the resolving power of those stimuli will be much less, the extent of this decrease depending on the amount of spatial separation.

### 3.6 Object Based Visual Attention

Early research in visual attention, summarised above, has supported (sometimes implicitly), the idea that attentional processes in vision select information from a spatial map that is like an internal screen in its representation of visual space. However, more recent evidence has emphasises the flexibility of the way in which attention acts - suggesting that it can not only act on a spatial map, but can

\textsuperscript{7}This is a principle that comes from the union of thermodynamics and information theory, but is applicable to any dynamic system that uses information (Cover & Thomas, 1991). Its application to vision is simple - a more in-focus image will be more informative that one that is less in focus.
also act to select between a subset of visual objects, with object information being relevant to the process in addition to spatial information (Pylyshyn, 1998). In other words, an idea has emerged that visual attention can select between visual objects as well as spatial locations. When the selective processes of visual attention act on objects, the process is referred to as ‘object-centred’ or ‘object-based’ visual attention.

The crucial difference between object-based attention and spatially-based attention is that object-based attention selects from object representations, once they have been built by the visual system, whereas spatially-based attention acts before the stage where object representations have been formed and may assist in their very construction (Wright & Ward, 1998). For instance, feature-integration theory is primarily a spatially-based theory of attention that suggests different features are combined and this combination enables an object representation to be formed for a particular location on a spatial master map. Alternative approaches which adopt a more ‘object-oriented’ approach include that of Pylyshyn (1989), who suggests attention is a mechanism often used to track object representations across space. He suggests that in some cases object representations are created pre-attentively (e.g., perhaps when the objects are visually quite simple or familiar to the observer) and focused attention acts to track these objects across space. In such cases, Pylyshyn (1989) supposes that object representations can become ‘tagged’ in a spatial map, but they do not exist in that map per se. The ‘tag’ referred to by Pylyshyn (1989) is what he calls a Finger of Instantiation (FINST). The term ‘instantiation’ here has subsequently been used in the literature in reference to the process of object representations being kept separate from one another by attachment to a spatial map (e.g., see Bavelier, 1999). The connections from object representations to spatial locations is now referred to more often as a ‘token’ rather than a ‘finger’, though the basic idea remains (Bavelier, 1999).

The object-oriented approach has been developed to accommodate the idea that attention can select either spatial locations or objects on which to act - it emphasises the flexibility of attention, at the cost of making the picture more complicated. Although there is much evidence to support attention that is totally object-based, it comes mainly from experiments using quite simple stimuli and tracking tasks. Therefore, it may be the case that both spatial- and object-based attention can be used by the visual/cognitive system but that each is used under different task conditions.
3.7 Coherence Theory

A theory of visual attention that is not explicitly object- or spatially-based but still helps to elaborate the links between object recognition, visual attention and scene perception is coherence theory, developed by Rensink (2000a). Figure 3.6 depicts this theory in schematic form. Coherence theory explains visual attention and scene perception data using an attentional system and a non-attentional system. The attentional system serves the purpose of generating coherent objects from an assortment of ‘proto-objects’ across the visual field. These proto-objects are generated in the non-attentional system through perceptual processes such as grouping, image segmentation and surface completion\(^8\). Figure 3.7 shows the ‘early vision’ non-attentional part of the theory in more detail. Note that the proto-object stage is the first available to attentional processes. In older terminology, this would indicate the attentional bottleneck is located just above this stage - proto-objects are not the subjects of focused attention, but they are its direct input.

Proto-objects are volatile structures, meaning they are generated quickly and can disappear just as quickly. It is known that grouping and surface generation in early vision are very quick processes (Moore & Egeth, 1997). Therefore, these processes are candidates to take part in the formation of proto-objects. Also, there is little attentional influence on the generation of proto-objects - their generation is quite automated and proceeds quickly across the visual field. The attentional system, however, is under the volitional control of the observer and will select a small subset of proto-objects and make them coherent (i.e., they are transformed from volatile to coherent structures). The ‘coherence field’ is the portion of visual space-time (coherence indicates spatio-temporal coherence, not just spatial coherence) for which this process acts - the extent of this field is also under the control of the observer (much like the attentional ‘zoom lens’ described previously).

Perhaps the most important contribution of coherence theory is its formalised depiction of the inter-relationship between focused attention and scene perception. The analysis of an entire scene, as was mentioned previously, does not seem to occur via a spatiotopic, detailed representation that is built up across fixations, but instead occurs through the construction of a more simplistic and

---

\(^8\)These early visual processes are beyond the scope of the current thesis. However, they are perceptual processes in the sense defined previously - processes that deal with the physical visual qualities of objects and are only mildly, if at all, influenced by attention or any other top-down signals. See Moore and Egeth (1997) for more detail.
abstract representation of the scene ‘gist’. The gist of a scene is the basic information about its spatial layout (Irwin, 1993). In coherence theory, construction of this representation of scene gist is identified as a non-attentive process - focused attention acts only on the currently fixated portion of visual space, which can be processed in detail by early vision. This detailed input serves as the foundation on which to build a coherence field that can join together complex aspects of the visual input into a unified whole. This ‘unified whole’ is effectively an object representation which allows an observer to recognise and interact with an object. Although the construction of the representation of scene gist is non-attentive, the representation itself can influence the attentive system (indicated by the arrows in the diagram). This influence would take the form of an observer’s abstract knowledge of scene gist and layout influencing their perception of a particular object.

In an unfamiliar visual scene, the influence on the attentional system comes from implicit knowledge of scene gist and layout derived from fixations across that scene. In a familiar visual scene, these influences are present in addition to a scene schema, which is a form of top-down influence connected to abstract
memory processes - in more lay terms, a scene schema represents what an observer expects to see in a particular scene. However, it is extremely unusual for an observer to enter an environment for which they have no scene schema - the term ‘unfamiliar’ in this context refers to lack of familiarity in a visual sense, not a semantic or more abstract sense. For instance, just because an observer walks into someone’s backyard to which they have never been, does not mean they do not know what kind of objects and visual information they expect to encounter in a back yard. It is important to separate higher level semantic knowledge and expectations based on episodic memory from the representations involved in scene gist and layout - these are implicit, non-semantic representations that serve to guide attentional systems on a moment-to-moment basis - most notably, they can modulate the recognition of objects - a process Rensink (2000a) conceptualises as the integration of proto-objects into coherent entities. It is to these ‘coherent entities’ which we will now turn.

Object recognition is a process critically dependent on attention and one which sub-serves many important actions (both covert and overt) of human and animal

---

9It is important to clarify that ‘knowledge’ used in this context is referring to implicit knowledge that, despite being abstract, is not of a high-level semantic or episodic nature.
observers. Although the process of object recognition is critically dependent on attention, to fully understand it and the way it can sub-serve other cognitive processes it must first be studied, to some extent, in isolation. This is attempted in the following chapter, after which visual attention will be reconsidered in light of the new information.
Chapter 4

Object Recognition

4.1 Objects and Visual Attention

It was mentioned before that more recent theories of attention describe it as a resource-limited process in which these resources are shared between attended stimuli. A variation on this idea that relates specifically to visual attention suggests that the attended units need not be distinct stimuli separated in the visual field - instead, they can be different features of the same object (such as its orientation and colour) (Kahneman, Treisman, & Gibbs, 1992). Based on this idea, the ‘binding problem’\(^1\) of visual perception is often tackled with the argument that attention is used to bind the separate features of an object together into a unified percept. In FIT, the instantiation of the binding problem was that attention can create a combination of outputs from feature maps on a master ‘spatial map’. In coherence theory, by contrast, the instantiation was that the coherence field acts as a binder of proto-objects into a complete object. The difference between these two treatments of the binding problem is that FIT suggests primitive visual features such as colour and orientation need to be bound by attention, while coherence theory is suggesting that these features are bound already into proto-objects at an early (non-attentional) visual stage. Coherence theory also suggests that the attentional-based binding is taking place at a later stage, binding proto-objects, which are high-level visual features, into coherent visual objects.

The discussion of binding has been brought up because it provides a bridge between the discussion of visual attention and the discussion of object recognition.

\(^1\)The binding problem is common in the visual perception literature - it asks how different features of the one object can be bound together to create a unified perception of an object, given that the neurophysiology of early visual areas is known to act as a segregant of visual input, with different cortical areas being sensitive to different visual features such as colour, orientation and direction of movement. See Roskies (1999).
The bridge is that many theories of visual attention now tend to suggest the main purpose of attention in vision is to create coherent object representations - and these are what sub-serve object recognition processes. Whether the mechanism for creating such representations is more like a ‘spatial master map’, as it is in FIT, or more like a ‘coherence field’, as it is in coherence theory, is still an open question. However, the idea of a coherence field provides a starting point in defining what is meant by an ‘object’ - it is something in the real and visual worlds that is coherent over space and time.

4.2 Why Object Recognition is a Problem in Visual Perception

Recognising objects necessarily involves the extraction of a relatively constant perception of a singular spatio-temporal entity (the ‘object’) out of a dynamic retinal image or optic array. The retinal image of an object should be regarded as dynamic because it changes dramatically with viewing distance, viewing angle and extraneous factors within the scene itself (e.g., level of illumination, occlusion from other objects). The problem of determining how a constant perception is attained from a highly dynamic sensory input like this is referred to in the perception literature as a problem of ‘constancy’ (Gregory, 1997). Object recognition is a high-level constancy problem that requires, to some extent, a solution to multiple lower-level constancy problems. For instance, because our recognition of an object is robust to changes in the shape, size and orientation of the retinal image of that object, we can say the recognition process exhibits shape, size and orientation constancy (Ullman, 1996). Each of these problems is a difficult one to solve, as solving them typically requires elucidation of computational sub-problems, which each need to be solved in an explicit and mathematical manner\(^2\). Often then, most older theories of object recognition focus on explaining shape constancy exclusively, but there are several prominent theories now emerging that seek to explain most, if not all, of the constancies involved in recognising an object (Ullman, 1996). Regardless of differences in approach, however, it can be said that all research into object recognition seeks to explain how humans rapidly and accurately identify objects in the visual scene (Wallis & Bülthoff, 2002).

\(^2\)In a computational context, it may be more appropriate to think of ‘constancy’ as ‘invariance to a particular transformation or series of transformations’. For instance, colour constancy is invariance to illumination conditions (or invariance to a surface colour being transformed by an illuminant).

\(^3\)The contrast between ‘perception’ and ‘identification’ mentioned in relation to attention applies here also.
4.3 Approaches to the Study of Object Recognition

In the context of visual perception, an object can be defined as a physical entity in space that has some meaning to the person viewing it. The use of the term ‘meaning’ here not only indicates that the object may be automatically associated with some semantic label, but also that the object may be useful, physically, in some way to the viewer (these two different forms of meaning could be termed explicit and implicit, respectively). Thus, objects ultimately sub-serve purposes of semantic identification and/or physical interaction by the viewer (Logothetis, 1996). Despite this accepted definition of the objective properties of real-world objects, several quite different theoretical approaches have developed to explain the process of object recognition.

4.3.1 Structural Descriptions

First of all, the ‘structural description’ approach has maintained that objects are represented in the mind/brain\(^4\) as models composed of separate parts in three-dimensional space and, furthermore, that the spatial relationships between parts are described symbolically in the mind/brain by some form of semantic tagging (e.g., ‘left-of’, ‘above’, ‘inside-of’) (Ullman, 2000). This theory suggests an abstractly-coded structure or model is used as a representation and incorporated into psychological functioning. Being structural rather than pictorial, it was proposed that mental operations could move these internal representations in real-time (Shepard & Metzler, 1971). This process would facilitate the matching of different views to the same object perception - what is termed view invariance (Marr & Nishihara, 1978).

A particular instantiation of the structural description approach was championed by Biederman (1987) and colleagues. It held that any conceivable object was represented in the brain by a combination of fundamental object model-parts (‘geons’), of which there were only 36 but which, together, could represent a huge range of objects as a series of connected component parts forming a three-dimensional model for a whole object (see Figure 4.1 for example geons). It was also suggested by these researchers that symbolic, semantic descriptions were used internally to identify the relationships between different model-parts (or geons).

\(^4\)The term ‘mind/brain’ is used here because structural description approaches originated in an era where psychological phenomena could only be loosely tied to brain states. Therefore, the extent to which the author intended them to be structural (pertaining to the brain) and/or functional (pertaining to the mind) is ambiguous.
The key property of this approach is that object representations are built up of components, each of which consists of visual properties that are invariant over a multitude of visual transformations. These parts are connected by more abstract representational components. For instance, a telephone can be represented as several geons, placed one on top of the other. The relative placement of these geons is determined by abstract coding (i.e., semantic coding), while the geons themselves are more explicitly coded (i.e., coded in Euclidean space).

The abstract parts of such a representation are more variable - allowing for larger changes, while still creating the same object perception/recognition. Using our example, the flexible coding of connections between object parts would allow a consistent recognition response of ‘telephone’ to a whole range of pictures or presentations of different telephones, each varying in height/width, and in the proximity of different components to one another. However, this invariance of response would only be maintained if the component geons were identical between the different instances of the ‘telephone’ object.

Figure 4.1: Geons (left) and other stimuli used in the study of object recognition. From Wallis and Bülthoff (1999).

4.3.2 View-Based Approaches

Also prominent in the field of object recognition is the ‘view-based’ approach, which is founded on the idea that objects are represented as collections of two-dimensional views, each one corresponding to a different projection of the same object onto the human retina (Wallis & Bülthoff, 1999). Most researchers who support the view-based approach do not maintain that all possible views of an object must be represented in the brain, only that a sufficient number must be represented to allow interpolation of views. Interpolation of two pictorial views, it is thought, will facilitate recognition of novel views.

In evidence of the view-based approach, Bülthoff and Edelman (1992) used a novel object class of paper clip-like stimuli (see Figure 4.1) to train observers,
before testing them in a recognition task using the same class of objects. Two different views (or images) of a particular stimulus were presented to an observer and it was found that recognition was better for views lying between those two learned views than for views lying outside of them. It is unlikely that this data could be explained by a structural description process, as performance would be approximately equal for all views if an object model is created to represent the unfamiliar ‘paperclip’ object and then this model is manipulated internally. By contrast, a view-based approach would suggest object recognition performance is non-uniform across object views - more familiar views will be recognised more quickly and easily. This is exactly what Bülthoff and Edelman (1992) showed. Using a similar experimental setup, Logothetis and Pauls (1995) trained monkeys to recognize novel stimuli and subsequently recorded from inferotemporal (IT) cortex, revealing neurons preferentially active for specific views of these newly trained objects. They also found that different neurons were active for different views.

Both of these studies appear to favour a view-based approach to object recognition and have been supported by numerous other studies which have used similar methodologies (see Tarr & Gauthier, 1998). Like the structural description approach, the view-based approach specifies what manipulations occur to representations located in the mind/brain.

4.3.3 Invariant Properties and Feature-Based Approaches

The third approach could be called the ‘invariant properties’ approach - it has been developed around the idea that properties of objects that do not change across viewpoints are automatically detected by the brain (e.g., by self-organising neural networks tied to visual input) and it is combinations of these invariant properties that form the representations of individual objects (Ullman, 2000). In such a theory, the nature of the object representations are not of central importance - it is more about the capacity of the neural system to carry out the relevant computations. The invariant properties approach has been supported in recent years by neurophysiological studies, which have shown that responses of particular neurons are invariant to transformations of a particular object in visual space, but are also selective to that object, or to a class of visually similar objects (see Wallis and Bülthoff (1999) for a review). Research has also shown the strong role perceptual learning plays in building up such transform-invariant neural responses (Logothetis & Sheinberg, 2000).

The ‘feature based’ approach incorporates ideas mainly from the ‘invariant
properties’ approach, but also draws from other approaches. Its main proposition is that objects are decomposed into invariant features and that these features may be connected in three-dimensional internal space if the recognition task necessitates it (Tarr & Gauthier, 1998). Furthermore, this approach maintains that the learning of invariant features proceeds initially through the comparisons of different views - that is, the invariant features are learnt by the detection of correspondences across different two-dimensional views of an object. Depending on the amount of exposure an observer has to an object, and consequently the amount of familiarity that observer has with the object, it could be represented in a number of different ways, and the ‘features’ that are being extracted could be more or less abstract. It is thought that the computations at high levels of the visual system may best be considered adaptive, meaning they act in the most computationally efficient way given the demands of the current task/situation. In visual tasks, this would equate to object representations being constructed in less-abstract, more image-based terms when objects are relatively unfamiliar. Likewise, it would also correspond to more detailed and abstract representations being built up for familiar objects. If an object becomes familiar, it could be because the observer has had to interact with it. In these cases, a representation like a ‘structural description’ would allow the observer better physical interaction with that object than would an ‘image-based’ representation.

Consistent with an approach emphasising that different forms of object representation could exist for different objects and different circumstances, Tarr and Gauthier (1998) have suggested that the most viable model of object recognition may be ‘one that incorporates the most appealing aspects of both image-based and structural-description theories’ (p.1). However, in selecting the most appealing aspects of the structural description approach, they are careful to point out that a recognition-by-components (RBC) model such as that proposed by Biederman (1987) relates non-local features in a qualitative fashion. Tarr and Gauthier (1998) propose that the ideal structural description, by contrast, would operate between more local image features and would be one that ‘captures the positional certainty between image measurements’ (p. 14). They call this type of structural description ‘implicit’ as it does not provide a global description of object shape (as does the RBC theory of Biederman, 1987), but instead codes relations between local features. Furthermore, they suggest that the structural description relating image features together would be probabilistic in nature, meaning that recognition of the object will be maintained even if the configuration of features in the image does not conform exactly to the stored structural description (the
The probabilistic nature of this structural description type means it is probably more computationally efficient in the human neural substrate, and therefore more likely to be implemented. However, image-based computation is probably always more efficient than structural descriptions for basic recognition tasks and is probably implemented neurally as long as exposure to the object is limited (as there are less 'stored views') (Ullman, 2000). After greater durations of exposure to the object-learning stimulus, some form of structural description may be necessitated, as it is the most appropriate for allowing object interaction and more sophisticated analysis of it higher in the cortical/psychological hierarchy (Tarr & Gauthier, 1998). In summary, view-based and structural descriptions may exist for different objects because of different task demands when that object is encountered and/or because of different lengths of exposure (learning time) to different objects. Because the most appropriate approach to object recognition appears to be one taking into account the constraints of the neural system in which the recognition process must operate, the neurophysiology of the area will be discussed in some more detail now.

### 4.4 Neurophysiology of Object Recognition

It is well known that V1 (primary visual cortex) is divided up into many columnar sub-units, each one selectively responsive to a particular stimulus dimension (e.g., motion, colour, orientation), or a subset thereof. This property is also present in V2 and V3, but to a lesser extent (Livingstone & Hubel, 1988). Figure 4.2 shows these visual brain areas as well as those more directly involved in object recognition. It has been suggested that vision is segregated in the cortex through two streams that differ in the way they respond to visual input and the parts of it to which they are selective. The 'ventral stream' (V4-PIT-CIT-AIT) is involved with identifying the detailed structure of the visual input and recognising visual objects, while the 'dorsal stream' (V1-V2-V3-intraparietal areas) is involved with locating visual stimuli in space (Ungerleider, 1982). As we are focusing on object recognition, the neurophysiology of the ventral stream will be discussed more than that of the dorsal stream.

Figure 4.3 shows parts of the temporal cortex to which the following discussion refers. IT cortex receives input from cortical areas V4, V4t, DP, VOT as well as areas of the parahippocampal gyrus (TF, TH) and areas FST and STP (Felleman & Essen, 1991). It has been shown that part of the inferior temporal cortex (TE3, TEO) has a columnar architecture resembling that of V1, but the cells within each
column are sensitive to different complex visual configurations, rather than simple bars, edges or coloured blobs (Tanaka, Saito, Fukada, & Moriya, 1991). Examples of these complex configurations are blobs on sticks and striped patterns. The neural responses to these stimuli are maintained even if the configuration of these features is changed somewhat. However, these responses disappear if the entire stimulus is transferred to another part of the visual field. Cells in these areas have also had their responses to real objects recorded, as presented in photographic images (Desimone & Duncan, 1995). In this study, object-selective responses were maintained if an object was presented from different viewpoints, but not when they were moved to a different location in the visual field. These results are more similar to those of Tanaka et al. (1991) than one would initially think. For instance, changing the location of a bar attached to a blob from leaning 45 degrees left to leaning 45 degrees right would result in the same change, at a basic image level, that rotating some object like a chuppa-chup (a ball on a stick) in depth would - and a rotation in depth produces the same change in the retinal image as does a change in viewpoint. Because these cells are robust to such changes, but not to changes in location across the visual field, it is said they exhibit viewpoint invariance, but not translation invariance (translation meaning a shift to another place in the visual field).

Although areas TEO and TE3 do not exhibit translation invariance other areas, that are more anterior in IT, do. One of the most cited properties of anterior IT neurons is their ability to respond to an object at various locations in the visual field. Evidence shows that the receptive field size of neurons in temporal cortex is much larger than those in occipital cortex - this may sub-serve translation invariance across the visual field. More specifically, receptive field size increases as a function of a neuron’s distance along the temporal lobe - being 4° just after V4, 16° in TEO (also called PIT) and 150° in more anterior sections.
of IT (TE1) (Wallis & Rolls, 1997). As well as translation invariance, some of these neurons exhibit invariance to transformations of size, spatial frequency and orientation (Rolls, 1992). Therefore, cells in TE1, which is higher in the cortical hierarchy than TEO and TE3, exhibit both viewpoint-invariance and translation-invariance (Rolls, Tovee, Purcell, Stewart, & Azzopardi, 1994). This collection of results from recording in temporal cortex appears to fit well with an information processing hierarchy view of early brain areas, where different transformations are carried out in subsequent neural stages.

More recent studies of neurophysiology probe the neural response by converting the magnetic properties of blood flow in the brain into a colour-coded image. This technology is termed functional magnetic resonance imaging (fMRI), as it is thought to reveal aspects of brain function by the differential magnetic resonance created by oxygenated vs. de-oxygenated haemoglobin molecules in the blood (Ungerleider, 1995; Farah, 1990). Therefore, this technology does not directly measure the electrical neural response, like the single-cell recordings discussed previously, but it instead measures an indirect correlate of the neural response - namely, changes in blood oxygenation that are thought to occur because firing
neurons need relatively more nutrition from the blood than do non-firing ones. Because it is an indirect measure of neural activity, the activity picked up on the fMRI is typically of longer duration than the casual neural activity and is picked up a short time after the neural response actually occurs (Ungerleider, 1995). Despite this short coming, the technology has been used to compare the neural response of humans to different classes of object (e.g., faces, birds, cars, laboratory-created objects) - see Figure 4.4 for an example.

![Figure 4.4: Brain areas involved in face recognition compared with those involved in expert greeble recognition. Also shown are the responses to non-expert stimuli (birds and cars, in this case). FFA is activated in face and greeble tasks, but not in bird or car tasks. Therefore, FFA is activated in the expert tasks, but not the non-expert tasks. From Tarr and Cheng (2003).](image)

The question guiding most fMRI research into object recognition is: are different brain areas responsible for representing and responding to different object types or are they all dealt with in an object ‘centre’ or ‘module’? Initially, studies like this tended to indicate there was a ‘face’ module for recognising faces (specifically, the fusiform face area - FFA, corresponding roughly to TEa and TEm) that was separate to a module responsible for the recognition of all other objects. More recently, evidence has indicated that area FFA simply deals with expert recognition tasks where very fine discriminations are required (Tarr & Cheng, 2003). Very fine discriminations are required for faces because humans need to, and so learn to, differentiate between (usually) several thousand to function adequately in a social context. For most people, there is no other class of visual object where this kind of discrimination is needed. However, for ornithologists

\[\text{As an illustration of this principle - autistics are known to perform worse in facial recognition tasks and this is often attributed to them not being able to build up neural representations of appropriate complexity as they are not emotionally compelled to look at other people’s faces (Malach, Levy, & Hasson, 2002).}\]
and car enthusiasts, fine visual discrimination between different birds or different cars is required - fMRI conducted on such people has found their FFA to be markedly activated when viewing an object belonging to their class of expertise, compared to objects that don’t belong to it and compared to other people (i.e., non-experts) viewing that same object belonging to the expert’s class of expertise (Tarr & Cheng, 2003). This study suggests FFA may not be an area specifically geared for facial recognition, but may instead be an area capable of performing the very fine discriminations required for expert recognition.

The study conducted by Tarr and Cheng (2003) can be considered a quasi-experiment as participants were pre-selected based on their visual expertise in a particular area. A more controlled experiment to examine the effect of expertise on recognition-related brain activation would be one in which observers were not pre-selected but were trained to create a particular type of visual expertise. To meet this need, Tarr and colleagues created an artificial class of objects called ‘greebles’ (see Figure 4.5), which are similar in gross form but have numerous small variations (like faces, birds or cars). Using these stimuli, recognition learning proceeds by an observer (be it macaque or human) classifying different instances of this object class. This classification need not be explicit (most expert object recognition, such as that involving faces, is not) - instead, categorisation can occur by the observer grouping similar objects together and being rewarded for correct trials. Studies using greebles have confirmed that activation of FFA is not specific to any object class but will occur when observers are required to distinguish between members of a class with which they have been trained to an expert level. In addition to establishing which brain areas are responsible for which types of object recognition, it is also important to establish the timecourse of activation in these areas in order to complete the neurophysiological data relating to object recognition.

4.5 The Timecourse of Object Recognition Processes

Psychophysical experiments have shown that reaction times required to recognise an object increase if that object appears out of context in a visual scene (Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce & Pollatsek, 1992). Delays were found both when the object appeared in a scene where it normally wouldn’t be

6This paradigm is most important for monkeys, who can’t speak, but is probably also important for humans so as not to confound the variables of ‘recognition expertise’ (non-expert vs. expert) and ‘type of learning’ (explicit vs. implicit) in the design of the experiment.
located (e.g., a couch next to a swimming pool) and when an object appeared in an unusual spatial location within a scene in which it would normally be located (e.g., a couch inverted in a living room). More recently, the timecourse of object recognition processes have been examined using EEG technology. Thorpe, Fize, and Marlot (1996) presented participants with a series of natural scenes, for 20 msec each, some of which contained an animal and some of which did not. Observers were required to respond as soon as possible if an animal was present, but were not to respond if one was not present\(^7\). Despite the short presentation time, observers were very accurate in identifying whether or not there was an animal in the scene (the mean percentage of correct responses for an observer was 94\%). Using the electroencephalogram (EEG), which measures gross electrical activity from the scalp\(^8\), Thorpe et al. (1996) compared the average brain potentials generated on correct ‘go’ trials (i.e., trials where the participant correctly indicated an animal was present) with those generated on correct ‘no-go’ trials (i.e., trials where participants correctly did not make an indication that an animal was present). This difference was most marked in frontal lobe recording sites, 150 msec after stimulus onset. Training observers in this task over a one

\(^7\)Such a task, where observers make a response if a criterion is met and none if it is not, is called a ‘go/no-go’ task (Pashler, 1998).

\(^8\)EEG analysis of brain function involves receiving electrical activity through electrodes distributed across the scalp. The electrical activity from each of these recording sites can be analysed separately or in concert. Initially, EEG activity was thought to represent the summation of cortical action potentials, but it is now thought to represent the summation of slower, post-synaptic potentials (Caspers, 1993).
week period failed to decrease response times, and so Fabre-Thorpe, Delorme, Marlot, and Thorpe (2001) concluded that it represented the operation of an ‘incompressible recognition mechanism’ that can help categorise unfamiliar scenes as fast as familiar ones. In contrast, Johnson and Olshausen (2003) suggest that there are two EEG components elicited in this type of task - one is a presentation-locked\(^9\) signal lasting 135 msec and a second signal lasting between 150-300 msec that co-varies with reaction time (and so is not presentation-locked). They suggest further that the presentation-locked signal relates to the processing of basic image features, while the variable signal relates to complete object recognition. Therefore, as long as images are distinct at a basic level, rapid categorisation can occur very quickly but when complete object recognition is needed for categorisation, it takes longer (150-300 msec). The key difference between the experiment of Johnson and Olshausen (2003) and that of Fabre-Thorpe et al. (2001) was that Johnson and Olshausen (2003) controlled the lower-level properties of the presented images (e.g., their Fourier spectra) while Fabre-Thorpe et al. (2001) did not, meaning they could not be sure recognition performance was not helped by differences in low-level image properties.

The variability in the recognition mechanism hypothesised by Johnson and Olshausen (2003) would appear to be more reconcilable with examining object recognition and visual attention in concert, as the need for attention to bind together primitive visual features (or proto objects) into a coherent object would likely be a process that takes a different amount of time for different objects and for different settings in which those objects need to be recognised. This is because the time course of recognition differs depending on the familiarity of an object and the expectation one has of it being in a particular scene (the scene ‘gist’ and ‘layout’ of Rensink, 2000a). Therefore, the two mechanisms suggested by Johnson and Olshausen (2003) may correspond to the proto-objects vs. coherence field (or non-attentional vs. attentional) distinction of Rensink (2000a).

### 4.6 Integrating Object Recognition and Visual Attention

From the above discussion, it is clear the area of object recognition is tackled with a number of different theoretical approaches. Basic neurophysiology suggests, however, that all theoretical approaches must take into account a cortical

\(^9\)Presentation-locked’ means a signal/response having a time course that is constant across trials and so occurs a fixed or ‘locked’ time after the presentation of the stimulus.
hierarchy with increasing receptive field sizes and signal latencies at subsequent stages (see Figure 4.6). Also, the different approaches can be unified somewhat when consideration is given to visual attention - a process on which object recognition is often critically dependent. Coherence theory is a framework that deals with both object recognition and visual attention. To re-iterate, the central tenet of coherence theory is that pre-attentive processes, operating in parallel across the visual field, group primitive visual features into proto-objects, which are volatile structures present in the activity of the neural substrate only as long as the observer is fixating on the relevant portion of the visual field. Once visual attention is directed towards a particular group of proto-objects, however, a coherence field is formed which integrates proto-objects into a singular perception of an object - an entity coherent over space and time. As long as visual attention remains engaged, this coherent object can be rapidly recognised. Furthermore, the observer may move their fixation point somewhat while still maintaining this same coherence field (i.e., attention can be directed to a particular object slightly off-centre of an observer’s gaze, and the perception of that object will still remain coherent - cf. covert attentional shifts discussed by Wright & Ward, 1998).

Figure 4.6: Receptive field sizes and EEG/electro-physiological signal latencies at different stages of the object recognition hierarchy in the ventral visual stream. From Fabre-Thorpe et al. (2001).
Luck and Beach (1998) and other researchers have suggested a slightly different integration of attention and object recognition. Their idea is that attention can facilitate object recognition via the temporal tagging of object representations. The temporal tagging hypothesis suggests that each object is given a temporal tag which differentiates it from other objects. Similar to the FINST idea of Pylyshyn (1989) already mentioned (which is essentially a spatial-tagging hypothesis), this idea suggests different object representations are kept separate from one another by integrating their identifying information (the visual representation) with information that relates to their place in time (the temporal tag). Evidence for this approach comes mainly from the repetition blindness (RB) experiments discussed in Chapter 3.

The idea of temporal tags relates to coherence theory because it suggests the temporal tag gives an object representation its individuality and so keeps it separate from other objects. However, coherence theory suggests the ‘coherence field’ involves spatial, as well as temporal, coherence. Spatial coherence is not directly addressed by the temporal tagging hypothesis, but its authors suggest that spatial coherence is addressed by theories such as FIT and FINST (see Chapter 3). Hypotheses dealing with spatial and temporal individuation of object representations have important implications for how object representations are learned and/or transferred to memory. It is this transfer of visual object information into memory which is of primary interest in the next chapter.
Chapter 5

Integration of Visual Information in the Frontal Lobe

As was mentioned in the previous chapter, the visual system can be divided functionally and neurologically into two distinct systems: the ventral and dorsal processing streams (Goodale & Humphrey, 1998). Because the processes involved in maintaining visual information can involve both of these streams (Fuster, 1997a), they will be looked at in detail in the first part of this chapter.

5.1 Cortical Connectivity Underlying the ‘What’ and ‘Where’ Streams

A segregation of visual input occurs as early as the retina, where photo-receptive cells (cones and rods) connect to anatomically and neurophysiologically distinct ganglion cells (type A and type B cells, respectively) which each integrate input from a number of photo-receptors (Kolb, 2003). Type A cells have transient neurophysiological response patterns, in contrast to the sustained pattern exhibited by type B cells (Gouras & Kruger, 1979). A sustained pattern is one that takes longer to establish, is maintained for longer and takes longer to reduce to baseline levels than does a transient pattern. These patterns of activity are functionally significant - they reflect the fact that cones collect relatively static visual input related to the more ‘picture-like’ aspects of retinal input (e.g., colour, fine spatial detail). Because this input is relatively static, it only needs a relatively static neural response to code it. Likewise, rods respond to visual input of a more active and dynamic nature (e.g., fast movement at low spatial scales) and the transient response pattern of the connected neurons is best suited to this type of visual input.

The spatial distribution of rod and cones across the retina also demonstrates
segregation - most cones are concentrated in the fovea, whereas rods are spread out far into the periphery. The non-homogeneity across the retina of rods and cones, and hence A and B ganglion cells, is easily explained in functional terms. As was noted in Chapter 3, visual transients serve to attract visual attention and produce eye movements that result in foveation of the part of the visual scene in which the transient was produced. Therefore, the transient responses of A cells, which can code this transient visual information, can be seen as sub-serving visual processes higher in the optic/cortical hierarchy that orient the fovea (and hence B cells) to the location of the transient for further, more detailed processing through sustained neural responses.

The transient/sustained division is maintained in the lateral geniculate nucleus (LGN), the first body inside the cranial cavity where visual neurons synapse. Around 90% of the cell bodies in the LGN belong to one of two classes: magnocellular or parvocellular. Magnocellular neurons, mostly connected to A cells in the retina, exhibit transient responses reminiscent of A cells (Hubel & Wiesel, 1968). Likewise, parvocellular neurons exhibit sustained responses like the B cells to which they are connected. The two cell types are segregated into different layers in the LGN and also project to different layers within V1 (Hubel & Wiesel, 1968). From their initial termination points in V1, the M and P streams project to different cortical layers of that same area, with M cells projecting to a layer of neurons (4B) exhibiting responses to direction of motion and orientation of visual stimuli, while P cells project to layers (2,3) exhibiting sensitivity to colour and orientation (Livingstone & Hubel, 1988).

From V1, the connections from cells originally in the M stream seem to terminate in areas sensitive to binocular disparity (V2), global aspects of visual form (V3) and global aspects of visual motion (V5/MT) (Livingstone & Hubel, 1988). The cells originally in the P stream, however, become more anatomically separated from these and terminate in areas sensitive to static aspects of visual input such as surface properties (V4) and complex configurations of visual form (IT). Figure 5.1 shows these cortical areas and their locations relative to one another.

From analysis at a cellular level, it is clear the two visual streams serve very different purposes in the processing of visual information. The ventral visual stream is thought to resolve information about object identity (‘what’ something is) and the dorsal stream is thought to resolve the spatial location of visual stimuli (‘where’ something is) (Ungerleider, 1982). Grossberg (2000) suggests that this arrangement into complementary, parallel streams is common in brain organisation and proposes that its purpose is to resolve uncertainty and ambiguity.
in the visual input efficiently. However, he also suggests the complete resolution of such ambiguity can only come about by interactions between the two streams. One such interaction has already been hinted at - the locus of attention can be changed by ‘where’ information in order to resolve ‘what’ information at a particular location. In order to further examine the interactions between these two streams, it is necessary to first look at the broader functions served by different cortical areas.

The cortical areas posterior to the central sulcus (post-central areas) are primarily sensory in nature while those anterior to it (pre-central areas) sub-serve mainly motor functions (Fuster, 1997a). This anterior-posterior division corresponds, and relates embryo-logically, to the motor (ventral) and sensory (dorsal) division in the spinal cord. However, as the human brain sub-serves cognitive
functions as well as those involved in basic sensory or motor processes, Fuster (1997b) suggests that the frontal cortex must be conceptualised as a motor cortex in a broader sense. For instance, although planning a speech is a very abstract process not tied to immediate movement, it eventually does result in movement output, controlled by primary motor areas. Logical thought and reasoning, also correlated with pre-frontal activity, will likely result in motor activity eventually. Even if they do not result in action, they can be considered as ‘potential’ actions or planned actions. These more abstract and integrative functions are correlated with anterior frontal activity and more basic functions are correlated with activity in posterior areas (Fuster (1997b), Chapter 8). This same ‘gradient of function’ is present in posterior cortex, but the direction of the gradient is reflected around the central sulcus (basic unimodal sensory areas are more posterior than the polymodal association areas).

Because behaviour can ultimately be described in terms of a perception-action (sensory-motor) cycle, there must be links between posterior and frontal cortical areas to produce this cycle. Jones and Powell (1970) summarised anatomical data that looked at connections between sensory areas and pre-frontal cortex (PFC). For cortex involved in the visual modality, reciprocal connections between area 20 in IT and areas 21 (IT) and 8A (pre-frontal) were found. Such connections were also found between area 21 and area 46 (see Figures 5.2 and 5.3). Also, projections have been found from the sulcus principalis to the superior temporal sulcus - these are part of the uncinate fasciculus (Pandya, Dye, & Butters, 1971). From this evidence, it is clear that different temporo-frontal connections begin at different stages of the hierarchy in the ventral visual stream. This ‘staggering’ of connectivity throughout the hierarchy also occurs in somatosensory and auditory cortex (Fuster, 1997b). However, connections from different parts of a sensory hierarchy generally go to different parts of the pre-frontal cortex (see Fuster (1997b), Chapter 3). It may be the case, therefore, that as more information is resolved higher in a sensory hierarchy, the quality of the information sent to the frontal lobe is greater and can sub-serve more abstract functions. However, lower quality information can be resolved quicker and can therefore sub-serve an appropriate motor function if a more immediate behavioural response is required.

To summarise, the advantage of having information of a lower quality sent earlier in the hierarchy is that this information can be used by frontal centres to guide basic motor responses while higher quality information can be used, in parallel, as input to more complex cognitive functions, but at the expense of greater processing time. Furthermore, certain visual information will still reach
the frontal lobe in the event of damage to an isolated portion of IT. Some of the 'higher fidelity' temporo-frontal connections (e.g., 21 to 46) are thought to sub-serve the maintenance of visual information (Fuster, 1997b; Miyashita, 1993). To examine how such information is maintained in neuro-computational terms, it is first necessary to revisit the concepts of memory and representation in more general terms.

Figure 5.2: A map of the cerebral cortex showing Brodmann areas - areas defined by different cytoarchitectures.

5.2 Working Memory and the Maintenance of Visual and Spatial Information

Working memory (WM) is a theorised limited capacity system that temporarily maintains and stores information to support ongoing thought processes and pro-

\[1\] Fidelity in this context refers to the quality of information transmission - in other words the signal to noise ratio of this transmission. A higher signal/noise ratio corresponds to a higher fidelity connection.
vide an interface between perception, long-term memory and action (Baddeley, 2003). It is said that there are two subsystems of working memory: the visuo-spatial sketchpad and the phonological loop; and that these are overseen by a ‘central executive’ that controls the flow of information to and from these subsystems (Baddeley, 1986) (see Figure 5.4). The central executive has also been described as a ‘supervisory attentional system’. This description emphasises the importance of attention in the role and operation of the executive and once again demonstrates the inseparability of attention and memory processes.

Figure 5.4: Working memory and the central executive. From Baddeley (2003).

The visuo-spatial aspect of working memory is of most interest to this thesis. However, this area has been researched far less than other aspects of working memory. Furthermore, as stated by Baddeley (2003), visuo-spatial WM is poorly
integrated with areas of research such as visual attention and object recognition, to which it is obviously related. Also, the use of the term ‘visuo-spatial’ may indicate that memory for visual and spatial information is inseparably integrated in working memory. However, Sala, Gray, Baddeley, Allamano, and Wilson (1999), in a clinical study, have demonstrated a double dissociation between visual and spatial working memory, meaning that some neurological patients display a deficit for a spatial WM task (manually rearranging a series of blocks to resemble a pattern presented 30 seconds earlier) but not a visual WM task (recognising a pattern 30 seconds later amongst similar distractors) and vice versa. Such a dissociation is also supported in single unit studies in monkeys performing WM tasks with spatial and non-spatial (but still visual) stimuli (Fuster, Bauer, & Jervey, 1982). This dissociation of visuo-spatial WM function has been explained in two ways. Firstly, a ‘domain-specificity’ model suggests that the ventral and dorsal aspects of PFC process different types of information - the dorsal aspect processing spatial information and the ventral aspect processing visual form information (cf. the ventral-dorsal division in posterior cortex, discussed previously) (Baddeley, 2003). Secondly, a ‘process-specificity’ model suggests instead that these two areas both receive the same information, but simply process it in different ways (Baddeley, 2003). Unfortunately, the experimental work in this area has not yet been able to resolve the dispute between these two models (Curtis & Esposito, 2003).

Because of the evidence showing a dissociation between visual and spatial WM, it may be more useful to conceptualise the visuo-spatial WM subsystem as one that can maintain either visual or spatial information independently or both together, depending on task demands. This dynamic perspective of WM can be better incorporated in a neurophysiological framework. For instance, Fuster (1997b) suggests functional sub-systems in WM may be better conceptualised as neural networks of memory rather than localised modules that interact with one another. In this context, WM can be regarded as the temporary activation of sensory and motor networks associated with a particular task. To better understand the view of memory as temporary network activation, it is necessary to understand the fundamental principles of memory at the level of individual neurons and small networks of neurons.

5.3 Memory in Neural Networks

Neurophysiologically, memory is considered as the selective facilitation and inhibition of synaptic strengths (greater synaptic strength corresponding to easier
transfer of activity from pre- to post-synaptic neurons). Hebb (1949) proposed the principle of *synchronous convergence* to depict this:

“When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased.” (p. 62)

This has subsequently become known as *Hebb’s postulate*. The basic conception is on a cellular level, and so is inadequate for understanding functional subsystems in the brain. The notion of changes in synaptic strength can be generalised to groups of cells, however. For this purpose, Hebb (1949) also introduced the notion of a ‘cell assembly’ - a collection of cells that have become associated together because of the temporal coincidence of their activity. Such an assembly can act as a closed system after stimulation has ceased. The notion of cell assemblies, therefore, extends Hebb’s postulate from interactions between two cells to the formation of functional sub-systems in the brain. Because of the abundance of long cortico-cortical connections in the brain (see Fuster (1997b), Chapter 2) and the nature of Hebb’s postulate, the association of two cells or systems need not be localised to a restricted portion of cortex - hebbian cell assemblies can form across *distributed* portions of cortex (Fuster, 1997b). Hebbian cell assemblies can explain the formation of functional subsystems and even long-term memories. However, it is still necessary to differentiate this type of information maintenance from the more short-term maintenance with which this thesis is concerned.

Zipser, Kehoe, Littleworth, and Fuster (1993) and Fuster (1997b) differentiate between two neurophysiologically and functionally distinct types of memory - *active* and *passive*. Active memory involves the maintenance of neural activity during the period where information is retained. Passive memory does not. Instead, passive memory is embodied in the collection of synaptic strengths at a particular time\(^2\). The distinction between active and passive memory can help explain the observation that mild head injury and electrical treatments (e.g., ECT) can disrupt short-term memory but not long term memory - it is thought these events disrupt the neural activity required to maintain information in active memory (short-term) networks (Russel, 1971). More traumatic brain injury can, of course, disrupt long-term memory and it is thought that this is because of the

\(^2\)The term ‘synaptic strength’ refers to how well electrical activity in one neuron elicits electrical activity with another, through contact at the synapse. The synapse is critical to this process because factors such as its size, the locations of contact and the type and rate of neuro-transmission across it, will all affect the flow of information between neurons (Wang, 2001).
physical destruction of networks in which the synaptic strengths have changed (Miyashita, 1993).

In terms of stimulus familiarity, the distinction between active and passive memory is that active memory is more concerned with information novel to the organism and passive memory is concerned with information that is not novel. In other words, if information needs to be maintained in active memory, it means that it is not available in passive memory because the observer has not encountered that information enough in the past for the neural activity associated with active memory to persist long enough to have caused physical, structural synaptic change allowing the maintenance of that information passively throughout time (without the need for continuing neural activity). Therefore, the distinction between active and passive memory can be conceptualised in terms of the way information is maintained in each. Active memory requires information maintenance by neural activity while passive memory does not, as the information is embodied in the neural structure. This thesis is concerned with active memory and, more specifically, with active memory that requires the maintenance of information for up to 30 seconds (what is thought to be the upper limit of active memory systems - see Curtis & Esposito, 2003). Physiological studies using tasks requiring the maintenance of sensory information typically find some form of persistent activity that occurs during the period after the stimulus is presented, and persists despite the absence of any new sensory stimulation.

In order to explain persistent activity in cortical areas that occurs without incoming sensory stimulation, it is necessary to extend Hebb’s postulate somewhat. This extension occurs by introducing the notion of recurrent networks. Recurrent networks are formed by neurons connecting to themselves and to other neurons within a localised network (Fuster, 1997b). Presumably, the formation of these networks occurs through learning governed by Hebbian learning rules (like Hebb’s postulate), or rules that are similar. However, once these networks are formed, they enable neural activity to persist in a localised portion of cortex for a time well beyond that predicted by simple sensory stimulation. Figure 5.5 depicts the structure of a recurrent network model and the delay activity that results. This recurrent network is part of a model created by Zipser et al. (1993) to explain delay activity in IT cortex that occurs during what is called a Delayed Match to Sample (DMS) task.
5.4 The Delayed Match to Sample (DMS) Task

Physiological studies of maintained (delay) activity in cortex use tasks where an observer is required to view a visual stimulus and, after a delay period, indicate whether a test stimulus is the same (match) or different (distractor) to the stimulus presented initially (sample). These types of tasks are generally referred to as Delayed Match-to-Sample (DMS) tasks, as the observer is required to match a stimulus presented after a delay to one presented before the delay. The components that make up a basic DMS task are as follows (Fuster, 1997b):

1. Display a discrete item of sensory information
2. Enforce a delay of a few seconds or minutes
3. Simultaneously present different sources of information, one of which was the discrete item initially presented
4. Allow observer to make a choice.
The type of response required by the observer can be varied, and consequently there are two sub-types of the DMS task. A delayed response task is one where the observer must make a spatially-directed response (e.g., right or left) to the matching stimulus. A delayed matching task is one where the observer simply indicates (in a yes/no) fashion if/when the matching stimulus is present. The key feature of this dichotomy is that a delayed-response task requires a spatial response and a delayed-matching task does not (Fuster, 1997b). Usually then, delayed response stimuli are of a spatial nature and delayed matching stimuli are visually abstract with no distinguishing spatial information, or they belong to discrete visual categories determined by basic visual features (e.g., colour). It is also possible to use tasks involving spatial and non-spatial information. For instance, sample stimuli could be presented in different spatial locations and match stimuli could all be presented in the same location - the observer simply has to indicate whether they were the same visually, regardless of spatial location. The reverse could also be done (matches are spatially distributed, samples are presented in the same area). Presenting stimuli in different spatial locations equates to translating them across the visual field - different spatial transformations could also be applied to the match or sample stimuli to examine the robustness of the memory performance/delay activity to these different transformations (c.f. robustness to transformations in repetition blindness, examined in Chapter 2). In general though, all such tasks are considered DMS tasks. Typically, a DMS experiment also has an Immediate-Match-to-Sample (IMS) control condition, where sample and match are presented without any delay (Fuster, 1997b). If the task requires a spatial response, there may also be a control condition which uses only a single target, but still uses a delay period (Curtis & Esposito, 2003). As was mentioned before, these tasks are designed primarily to examine the neural activity that invariably occurs in their delay period. An example of a DMS task and the delay activity produced is shown in Figure 5.6.

5.5 Delay Activity in DMS Tasks

Neural activity that persists in sensory cortices and frontal cortex during DMS tasks has been consistently reported for 30 years. Since the discovery of this ‘delay activity’, it has been thought that it is generated by recurrent, reverberating (i.e., self-excitatory) circuits within the neural substrate (Wang, 2001). Regardless of the specific causes and maintaining factors behind delay activity, it is clearly related to information retention in a DMS or WM task, as a decrease in
Figure 5.6: The delayed match to sample (DMS) paradigm (A) and the Delayed Activity Distributions (DADs) that result. This figure shows the activity of a unit in PFC (B) that is still somewhat sensitive to the sensory stimulus (i.e., it fires when the monkey needs to remember the red stimulus, but not when it needs to remember the blue one - hence the difference between activity resulting from presentation of a red stimulus compared to that sensitive to the blue stimulus (C)). From Fuster (1997b).
this activity is correlated with a decrease in performance on these tasks and the
time course of delay activity corresponds to the retention interval in such tasks
(Wang, 2001; Curtis & Esposito, 2003; Fuster, 1997a). Because the cue and the
response in DMS tasks can vary from trial to trial, it is possible that neural activ-
ity measured during such tasks is related to the sensory stimulus itself. However,
searching for neural activity common to trials with different cues can reveal gen-
eral purpose mechanisms for the maintenance of visual information. In fact, it
is found that some units that display activity are sensitive to properties of the
stimulus, and others are largely independent of the stimulus. For instance, Rao,
Rainer, and Miller (1997) recorded from 195 neurons in the PFC of the macaque
and found some exhibited delay activity sensitive only to what a stimulus was,
others displayed activity sensitive only to where it was, while the delay activity of
the remainder was sensitive to both. They hypothesised the neurons with delay
activity sensitive to both served as a link between those sensitive to only one, by
integrating the two types of visual information. Earlier studies usually focused on
a smaller subset of PFC neurons than did Rao et al. (1997) and so usually found
sensitivity only to what a stimulus was (see Fuster & Jervey, 1982a) or only to
where a stimulus was (see Funahashi, Bruce, & Goldman-Rakic, 1989).

It is now known that many parts of the dorso-lateral pre-frontal cortex (DLPFC)
exhibit activity at the start of a trial in a DMS task, as well as during the de-
lay. The earliest component of this activity is non-specific (i.e., does not change
with cue) and is present 150 msec after the presentation of the cue in areas of
the DLPFC (Fuster & Jervey, 1982b). Some of these cells, however, also exhibit
a later component in their response that is cue-related and this component can
vary with the type of response required (Fuster & Jervey, 1982b; Funahashi et al.,
1989). Overall, however, these pre-frontal units are less specific to the nature of
sensory stimuli than areas in posterior cortex (Fuster, 1997b). Fuster and Jervey
(1982b) suggests that the component of activity that is invariant across trials is
related to the ‘anchoring’ or focusing of attention, while the variable component
is related to encoding of the sensory stimulus. This thinking is demonstrative
of the widely acknowledged link between attention and mnemonic encoding (see
Baddeley, 2003).

As well as being observed in DLPFC, delay activity has also been recorded
from IT. For instance, Miyashita (1988) presented 97 fractal images\(^3\) to macaque
monkeys in a fixed order (see Figures 5.8 and 5.7). The cycle of 97 images was

---

\(^3\)These fractals were generated using an algorithm that resulted in them being uncorrelated
with one another in terms of spatial/image/geometric structure (Miyashita, Higuchi, Sakai, &
Masui, 1991)
repeated several times before recording from area IT occurred. It was found that neurons sensitive to a particular pattern were also sensitive to patterns presented nearby in the presentation sequence (i.e., patterns that had a close serial position number, or SPN). This effect was interpreted as being due to the build up of associations between different fractal patterns, as there was no such correlation of responses between neighbouring stimuli (i.e., ‘clustering’ of responses along the SPN axis) when a set of unlearned fractal patterns was used. Miyashita (1993) suggests this associative memory is more likely to be a stable type of memory (like the ‘passive’ type mentioned previously) than a labile (‘active’) type of memory. This is supported by the fact that selectivity to these patterns did not change significantly over several hours, after the initial associations had been established.

![Figure 5.7: A schematic of Miyashita’s 1991 DMS task. a) The presentation of two stimuli in sequence. b) The overall presentation sequence, consisting of 97 images and repeating. From Wallis (1998).](image)

A task differing slightly from the DMS format, called a paired associate task, was also used to examine the long-term selectivity of IT neurons (Sakai & Miyashita, 1991). This task used 24 fractal stimuli grouped into 12 pairs, where each pair had no apparent geometric similarity (to ensure no predictability, in visual terms, of what fractals were paired with one another). Monkeys were presented with one fractal pattern for 1 sec, followed by a delay of 4 sec, followed by presentation of two fractals, one that was the pair to the initial one and one from a different pair. Monkeys were rewarded for selecting the paired associate from the two fractals presented after the delay. Gradually then, the monkeys learnt which fractal went with which in each pair. This is similar to a DMS task, except that the observer does not choose the same (matching) stimulus, but instead chooses one paired with the initial stimulus. Neurons recorded from in

---

4These unlearned patterns were matched with the learned ones in terms of their Fourier spectra. Responses from these unlearned stimuli were obtained simultaneously with the ones from the learned stimuli.
monkeys performing this task showed that, once monkeys were performing the task correctly, paired pictures were the most effective stimuli for many neurons in IT, even if those neurons responded to more than two of the fractals (i.e., there was always a particular pair for which they were most selective).

Miyashita (1993) suggests that the associational mechanisms of IT neurons revealed in these experiments may help in setting up view-invariant representations of objects by associating different views of the one object together, given that in natural circumstances, different views of the one object will be encountered in close temporal sequence, as an observer changes viewing angle (see Wallis (1998, 1999); Wallis and Bülthoff (1999, 2001) for more discussion on this issue). Because of this consideration, these data support the conception of IT as not only a centre that responds to complex visual features, but also as a storehouse for representations of visual objects (see Mishkin, 1982; Weiskrantz, 1990). Several researchers suggest these two functions may be complementary - complex visual features are learnt (and, hence, neurons become responsive to them) because they are invariant across several views of an object and so the combination of these features can come to represent this object (see Logothetis, 1998; Wallis, 1998).

The existence of delay activity in IT cortex during visual association and DMS tasks is not in dispute. Tomita, Ohbayashi, Nakahara, Hasegawa, and Miyashita (1999) built on this evidence and looked at how this activity relates to activity found in PFC during similar tasks. They found evidence for a functional link between IT and PFC that is involved in the learning of visual object associations.
Their experiment involved severing the posterior portion of the corpus callosum, meaning only top-down signals from PFC could transfer between hemispheres, and bottom-up (sensory) signals could not. After this surgery, monkeys were presented with stimuli in the visual field ipsilateral or contra-lateral to the hand with which they were required to make a response. They had previously been through a training regimen which taught them to associate particular stimuli with one another and with a particular method of response. It was found that this associational knowledge only had an effect on their response if they were presented in the field ipsilateral to the hand from which a response was required. This was interpreted as meaning that the associational knowledge was mediated by a top-down signal from pre-frontal areas to IT. A control experiment, involving the lesioning of sub-cortical structures, showed that this link was cortical and did not pass through any sub-cortical areas. This experiment suggests that both PFC and IT are important in explaining performance on visual association tasks.

The importance of both PFC and IT in visual association tasks was also promoted by Wallis (1998) who theorised that the learning of visual stimulus patterns and objects takes place via a feed-forward neural link between IT and pre-frontal cortex, and that the maintenance of information about objects takes place in PFC rather than IT. This was proposed to take place in a recurrent network in PFC, which sent signals back to IT periodically. Amit (1999) explained the same set of results by suggesting that the recurrent network was located in IT. Amit’s main reason for disagreeing with Wallis was, he said, because the maintenance of the delay activity is more likely to occur through recurrent networks local to that activity rather than recurrent networks located further afield in pre-frontal cortex. There is no clear empirical evidence to support either one of these hypotheses. However, Fuster (1997a) supports the notion of delay activity in DMS tasks being maintained by reciprocal connections between IT and area 8 in pre-frontal cortex. This support comes from a study in which pre-frontal cortex was cooled (i.e., reversibly lesioned) and micro-electrode recordings were taken from IT, while a macaque monkey was performing a standard DMS task (see Figure 5.9) (Fuster & Jervey, 1982a). This study showed that the pre-frontal cortex was instrumental in generating the delay activity in IT that is present during a DMS task, as this activity was significantly reduced when the pre-frontal cortex was cooled compared to when it was not. The reduction could have only occurred in this experiment because the area of PFC that was cooled was somehow involved in maintaining the activity in IT. It is clear, therefore, that any recurrent network that may be present in IT is not sufficient, in and of itself, to maintain the delay
activity observed there in DMS experiments. This does not resolve the dispute between Wallis (1998) and Amit (1999), but makes it clear that temporo-frontal interactions are important in maintaining delay activity, and the neural networks local to IT are insufficient for this purpose.

Figure 5.9: Neural activity in IT measured during a DMS task, in conditions where pre-frontal cortex had and had not been cooled. From (Fuster, 1997b).

In summary, it is clear that DMS tasks create a characteristic pattern of delay activity in anterior temporal and frontal cortices. One way to better examine the neural mechanics underlying the generation of this activity is to use technology looking at gross activity in these regions, during performance of the same task. This has been done, to some extent, using EEG technology.

5.6 EEG Analyses of the Maintenance of Visual Information

As was mentioned in Chapter 4, EEG records electrical activity from the scalp and this activity is thought to be due to summation of post-synaptic potentials in cortical and sometimes sub-cortical areas. It is thought, furthermore, that EEG activity could not occur without some degree of synchronisation of cortical potentials within the recording area and that EEGs during complex cognitive activities will represent superimposed oscillations from a number of different brain areas (Basar, 1998). As with the analysis of individual cells' neural activity, it is
thought the *frequencies* of EEG signals is their most important property (i.e., as opposed to amplitude). The frequency domain is therefore of primary importance when analysing EEG activity, especially that related to more complex cognitive tasks. EEG activity is typically divided into five frequency bands: delta (1-3Hz), theta (3-8Hz), alpha (8-13Hz) and gamma (above 15Hz) (Basar, Basar-Eroglu, Roschke, & Schurmann, 1993). EEG rhythms in the theta range are the most studied in relation to cognitive function - Basar et al. (1993) states that this is not because these are the only ones involved in cognitive function, but because the *distributed theta system* is the one that is most correlated with cognitive and associative functions\(^5\). Functions of distributed systems are best analysed in the context of the EEG using a statistical parameter known as *coherence*. Coherence is the normalised estimate of the cross-correlation function of EEG signals from two different areas with respect to frequency (Basar, 1998). Coherence gives an idea of how electrically coupled two brain areas are - that is, how much they are within the same frequency range and how in-phase their oscillations are (Petsche, 1993).

In EEG analyses of experiments using simple visual stimuli and tasks, it is found there are oscillations of EEG activity local to the occipital and temporal cortices. In tasks requiring some form of memory retention, however, there is typically a more distributed oscillations occurring across different brain areas, notably frontal cortex. The activity associated with such higher-level tasks is mostly in the theta range, but alpha activity is found to occur before engagement of distributed memory systems in such tasks (Basar et al., 1993). However, Ward (2003) suggests memory processes are mostly related to theta and gamma rhythms, while attention is related to alpha and gamma rhythms. However, very little memory consolidation or retrieval can occur without the use of attentional resources (Baddeley, 2003), so this distinction might not be quite so clear cut. These are important points to consider when analysing processes involved in the performance of DMS and WM tasks using EEG.

Sarnthein, Petsche, Rappelsberger, Shaw, and Stein (1998) conducted a study where the coherence between frontal and temporal scalp sites was measured during two different DMS tasks. One used textual characters as stimuli, the other used abstract line drawings. In both cases, there was a significant enhancement of coherence in the theta range between the pre-frontal and temporal electrode sites, \(^5\)Because different frequency bands of EEG activity usually appear at characteristic sites, it is hypothesised that parts of the cortex can be considered ‘generators’ for these particular EEG frequencies. A distributed system, therefore, is simply a collection of such generators distributed across the cortex. Parts of the activity may also arise from sub-cortical areas.

71
during the 4 sec delay interval. The authors suggested that these results indicate that WM involves the synchronisation in activity between temporal and frontal centres by low frequency brain activity (theta activity). Like Fuster (1997a) and Miyashita (1993), these authors are suggesting there is a cortical link between PFC and areas of temporal cortex that is revealed by results showing correlation in activity, and inter-dependency, between the two areas. The question is, then, how exactly does this link work? This is one of the questions addressed by the empirical work of this thesis, outlined in the next chapter.
Chapter 6

Proposed Method

6.1 General Method for Experiments 1-4

The first two experiments combine the paradigms of visual search and change blindness to examine the allocation of attention across individual objects in the visual field. The presentation parameters used in these experiments are very similar to those employed by Rensink (2000b) who presented two search fields in alternation, with a blank field in between (to mask the change). In Rensink’s experiment, the search fields contained outlined black and white rectangles arranged in an invisible 5 x 4 grid, and the total display subtended approximately 15° x 12° of visual angle. The number of rectangles presented on any one trial (i.e., the set size) was either 2, 6 or 10. Each search field was presented for 500 msec, while the blank screen was presented for 120 msec. On half of the trials (chosen at random), one of the items in the second search field was changed relative to the first.

In each trial, all search stimuli were defined by one of two values on one of two visual dimension (vertical or horizontal oriented bars, positive or negative contrast polarity). The changing item (the target) changed from one of these values to the other (e.g., a vertical bar became horizontal) across search fields, while the other items (the distractors) remained the same in both search fields. However, half of the distractors were defined by the first value on the dimension of change and half were defined by the second. Therefore the two possible values on the dimension of change were distributed equally across the distractors. For example, in trials where the orientation of one of the elements was changed between horizontal and vertical, half of the distractors were horizontally oriented and half were vertically oriented.

This methodology is essentially what is used in Experiments 1 and 2. However, in these experiments four stimulus dimensions are being used instead of
two and there are multiple possible changes within each dimension. The stimulus dimensions being used in these experiments are velocity of motion, colour saturation, orientation and size. The shapes used in the search fields are ellipses. Examples of these ellipses undergoing the various types of change are shown in Figure 6.2. Because size is one of the visual dimensions of change, these ellipses can be different sizes between trials and within the one trial. The median ellipse size is defined by a bounding rectangle that is $1.2^\circ \times 0.4^\circ$ of visual angle, the same size used by Rensink (2000b) (see Figure 6.1).

![Figure 6.1: The median ellipse size is bounded by a rectangle $1.2^\circ \times 0.4^\circ$ of visual angle.](image)

Instead of using a blank field to mask the change like the Rensink (2000b) study, Experiments 1 and 2 use the ‘mud splash’ technique (O’Regan, Rensink, & Clark, 1999). A mud splash is used in order to preserve the continuity of moving elements on the screen. If a blank field were used to mask the change, a visual transient (i.e., a luminance onset) would be caused by the moving element being displaced on the second search field relative to the first, assuming motion were continued throughout the blanking period. Even if motion were halted for the blanking period, the observer may experience problems tracking the moving objects, because it is not in the location expected and so could interrupt the continuity of tracking eye movements. The mud splash technique provides a good alternative to blanking here, as it can preserve the continuity of moving objects in the visual field, while still preventing attention from automatically detecting the transient associated with the change (Rensink et al., 1997). Bahrami (2003) outline a similar rationale for their use of the mud splash technique.

As in Rensink’s experiment, the set sizes used for Experiments 1 and 2 are 2, 6 and 10 and these are varied from trial to trial at random. In Experiment 1, the main dependent measure is accuracy, while in Experiment 2 it is Reaction Time (RT). In both experiments, participants are required to indicate, via a key press, if and when they notice a change. If the participant cannot detect any change, they are required to press a different key. The four types of change will each occur
on an equal number of randomly selected trials. Overall, however, change will only occur on half of the trials for a particular subject, as in the study conducted by Rensink (2000b) and many other change blindness studies.

![Figure 6.2: Examples of the ellipses used as search stimuli and the effects of various stimulus variables changing.](image)

All manipulations in these experiments occur within-subjects. This also occurred in the Bahrami (2003) study, which used seven observers. In the study by Rensink (2000b), 12 observers were used per condition, but each condition involved several within-observer comparisons. Going by these studies then, a number of 7 participants for Experiment 1 and 14 for Experiment 2 should be sufficient. The difference in numbers is because of the additional manipulations in Experiment 2 as compared to Experiment 1. Experiment 3 should have an even higher minimum (about 20) as it involves more manipulations than the others.

Before a participant begins one of the following experiments, they are tested with a computerised version of the Ishihara Color Blindness test\(^1\) and the Freiburg Acuity Test (Bach, 1996).

### 6.1.1 Experiment 1: Thresholds for Detecting Changing Objects

In each of the trials of Experiment 1, only one stimulus changes per trial and this stimulus changes on only a single stimulus dimension (colour saturation, velocity of motion, orientation or size). The magnitude of change variable is used so that a change threshold can be established for each observer viewing a particular type of change with a particular set size. This threshold is calculated using the psychophysical method of constant stimuli, which requires the experimenter to show the observer, in a random order, various values of the stimulus variable (in this case, the magnitude of change) and, for each presentation, the observer is required to answer yes/no to whether they detect the stimulus (see Green & Sweets, 1974). The number of trials in which the observer correctly detects the

\(^1\)This test is a composite of a number of tests available at [http://colorvisiontesting.com](http://colorvisiontesting.com)
stimulus (detects the change) is recorded. The value for which the observer detects
the change 75% of the time is then taken as the detection threshold. Using this
method allows a psychometric function to be plotted for each observer, at each
set size, for each type of change (see Figure 6.3 for an example of a psychometric
function). This will show the percentage of trials where change was detected, as
a function of the magnitude of change.

Figure 6.3: An example of a psychometric function with hypothetical measurement
units.

Literature applying psychophysical analysis to visual search has found that
thresholds changed significantly with set size but were not significantly different
between observers (Palmer, Verghese, & Pavel, 2000). Statistical tests for sig-
nificance will also be applied to data from the current experiments and, if the
results are similar, a single logistic function will be plotted for all observers relating
change threshold to set size. The plotting of these functions will also enable
comparison between the thresholds required for the detection of different types of
change - different logistic functions will be plotted for each type of change.

Table 6.1 shows the variables involved in Experiment 1 and their possible ma-
nipulations. From trial to trial, the set size, the type of change and the magnitude
of change are manipulated. All changes in Experiment 1 occur from the middle of
the range of the changing stimulus variable (e.g., colour hue changes up or down
from its initial value of 50%). This is because the same magnitude of change will
not necessarily be equivalent if it occurs from different points along a stimulus
dimension, and so changes must occur from a constant point along a stimulus di-
mension for thresholds to be meaningful. In support of this approach, literature
has found that the detectability of values along particular stimulus dimensions
varies non-linearly. For instance, Hohnsbein and Mateeff (1998) have shown that
detecting the same change in velocity occurs more readily if the starting velocity
is higher than if it is lower. This type of non-linear response is also evident in the perception of colour (Gegenfurtner & Kipper, 2003) and presumably exists for many other stimulus variables.

The variables in Table 6.1 are manipulated at random from trial to trial\(^2\). There are eight possible changes for each stimulus variable - four are decrements and four are increments (see Table 6.2). These decrements and increments are multiples of 10% of the entire range of the stimulus variable. Although these manipulations are made at random, the presentation is controlled so that by the end of the experimental session, there have been equal numbers of trials for each level of a particular variable. Because there is an equal number of trials for each level of each manipulated variable, there are 2 x 4 x 3 x 6 = 144 unique trials. In addition, each of these trials is presented 5 times to ensure the accuracy of the established threshold value. Therefore, there are 720 trials in total for a single participant. Each presentation goes through four change cycles, each cycle is structured as follows:

- 500 msec presentation of unchanged search field
- 120 msec mud splash
- 500 msec presentation of changed search field
- 120 msec blank.

Therefore, each cycle takes 1240 msec. Each trial includes four back-to-back cycles and so each trial takes 4960 msec. Because there are 720 trials, each experimental session lasts 59 minutes and 32 seconds. However, each session with a participant is started with a training phase in which a single stimulus changes on a particular stimulus variable and then this change is shown embedded in a search field. This training occurs for all four stimulus variables and training is complete when the observer reaches a criterion (at least 75% correct).

### 6.1.2 Experiment 2: Detecting Multiple Changes in Objects

The thresholds computed in Experiment 1 will be used as the minimum magnitude of change for a given set size in Experiment 2. The main dependent measure in Experiment 2 is the participant’s Reaction Time (RT) to detecting change, rather

\(^2\)Although these manipulations are made at random, the computer code driving the experiments ensures that the number of trials per condition (i.e., level of a particular manipulation) is equal for a particular participant. This is the same for all experiments.
### Table 6.1: Manipulations in Experiment 1

<table>
<thead>
<tr>
<th>Manipulation</th>
<th>No. Levels</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change</td>
<td>2</td>
<td>Absent, Present</td>
</tr>
<tr>
<td>Type of Change</td>
<td>4</td>
<td>Motion Velocity, Colour Saturation, Orientation, Size</td>
</tr>
<tr>
<td>Set Size</td>
<td>3</td>
<td>2, 6, 10</td>
</tr>
<tr>
<td>Magnitude of Change</td>
<td>6</td>
<td>-30, -20, -10, +10, +20, +30</td>
</tr>
</tbody>
</table>

### Table 6.2: Stimulus Variables: Experiments 1-4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>Point from which Change is Made</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colour Saturation</td>
<td>0-100%</td>
<td>50%</td>
</tr>
<tr>
<td>Speed of Motion</td>
<td>-1.5 - 1.5°/s</td>
<td>0.0°/s</td>
</tr>
<tr>
<td>Size</td>
<td>(0.6° x 0.2°) - (2.4° x 0.8°)</td>
<td>1.2° x 0.4°</td>
</tr>
<tr>
<td>Orientation</td>
<td>0-360°</td>
<td>90°</td>
</tr>
</tbody>
</table>

than their accuracy in detecting change. However, a constraint relating to the participants’ accuracy is placed on them in this experiment - they must make errors on no more than 10% of the trials. This is to ensure the validity of the RT data and is a constraint used in many visual search experiments (see Palmer, 1998). The way this constraint is implemented is that participants are given error feedback (are informed by a 1000 msec visual message at the end of each trial which shows a tick or cross symbol) and are trained in the task before results are recorded. If a participant still does not meet the criterion in the actual experiment, they are excluded from the analysis.

Because thresholds are applicable best to the conditions in which they were measured, the conditions of Experiment 2 are as similar as possible to those of Experiment 1 in all aspects other than the presence of error feedback. However, each trial in Experiment 2 is terminated when observers make a response, so the length of each trial is variable. Also, trials do not terminate until the observers make a response. Furthermore, Experiment 2 involves a manipulation of the number of changes undergone by a particular object, as well as the types of changes co-occurring in any particular trial. The manipulation of the number of changes enables us to measure the way in which multiple changes, from different stimulus domains, combine to influence the detection of change.

Experiment 2 will be analysed with RT as the dependent measure and set size, type of change, number of changes and magnitude of change as the independent variables. The type of change variable can be single or dual, and different types of dual change (e.g., colour + motion, orientation + colour) will have their effect
on RT assessed independently and then have these effects compared. As was mentioned in the Introduction, this methodology will enable a comparison that can examine the extent to which different types of stimulus change are equivalent in terms of their influence on change detectability. For example, it is possible that different types of change may have an additive effect when they are combined - they may additively facilitate change detection. In this case, the extent of the facilitation could be easily predicted \textit{a priori}. Alternatively, they may interact non-linearly but still have a lawful facilitation effect. In some circumstances, however, combining two types of change may suppress the detectability of change somewhat. These issues will be explored further after data from this experiment is analysed.

6.1.3 Experiment 3: Detecting Changes in Multiple Objects

Experiment 3 extends the methodology of Experiment 2 by having multiple items in the one search field change simultaneously. Furthermore, the items that may change are explicitly identified to the observer at the start of each trial (by an illuminated circle surrounding each of them). The following manipulations are made, at random, across trials in Experiment 3:

- Whether a change occurs
- Number in the target (changeable) group
- Whether the changing group is identified at the start of the trial
- Number of items in the target group that do change
- Total number of items in search field (2,6,10)
- Number of visual dimensions of change (1 - 4: size, colour, orientation, direction of motion).

The condition where a group of search stimuli that can potentially undergo change is identified will be referred to as \textquote{group identified}, the other condition will be referred to as \textquote{group not identified}.

When observers detect a change in Experiment 3, they are required to press a key. In the \textquote{group identified} condition they must then identify the group that was flagged at the beginning of the trial and then identify the members of that group that changed. This identification occurs by the observer selecting the relevant
search elements with the mouse pointer and clicking on them. This method of combining a multiple object tracking (MOT) paradigm with a change detection paradigm was also used by Bahrami (2003).

In the ‘group identified’ condition, observers can focus their attention on only the target group. Because of this, it is expected that performance will be better overall for this condition as compared to the ‘group not identified’ condition, both in terms of reaction times (RT) and the proportion of changes identified in any one trial. Furthermore, it is predicted that the degree of homogeneity amongst the non-changing distractors will have more of an effect in the ‘group not identified’ condition than in the ‘group identified’ condition. However, the effect of the distractors in ‘group identified’ will likely become apparent when observers are required to track several objects for changes. The extent to which the distractors affect a participant’s performance in the multiple object tracking conditions will be discussed in the context of the spatial indexing theories discussed in Chapter 3.

Analysis of results for Experiment 3 will be very similar to Experiment 2. However, in theoretical terms, the results of Experiment 3 may have implications for the Guided Search model and spatial indexing theory (see Chapter 3).

6.2 General Considerations for Experiments 4-7

Experiments 4-7 use the electroencephalogram (EEG) as a dependent measure rather than behavioural measures, which are used in experiments 1-3. Figures 6.4 shows electrode sites in a typical EEG setup. Experiments 4-7 look primarily at output from the sites T3 (temporal) and F3 (frontal). The onset of alpha and theta rhythms in these sites will be examined with reference to the presentation of stimuli, in order to verify existing data on the time course of such responses. The data from Experiments 6 and 7, however, will be analysed in more depth using the measure of coherence discussed in Chapter 5. Coherence will be used to provide a measure of how electrically coupled these frontal and temporal areas are during the maintenance of visual information in the DMS task. This data will be compared with the single-cell data already reviewed to provide a broader picture of these cortico-cortical interactions and give us more insight into the neural substrate of this active visual memory.

In similar experimental paradigms (i.e., working memory and DMS tasks) using EEG as a dependent measure, previous authors have included between 10 and 30 subjects in a particular experiment (see Sarnthein et al. (1998) and
Therefore, it is estimated that 25 subjects will be needed for each of the experiments 4-7.

6.2.1 Experiment 4: EEG of Change Detection

Experiment 4 uses the stimuli introduced in experiments 1 and 2 but utilises an EEG as the dependent measure, rather than a behavioural index. Because of the sensitivity of EEG to movement artifacts from eye and neck muscles, stimuli in Experiment 4 are presented in a more central portion of the screen (the field subtends 10° x 8° rather than 15° x 12°) than in experiments 1-3 and do not undergo any movement. In all other respects, this experiment is essentially the same as Experiment 2 and so is looking at single and multiple changes within a single search element. However, this experiment will use the data from Experiment 2 to create an index of difficulty for the different combinations of change types and change magnitudes. This index of difficulty is then used as an independent variable with which to track changes in the EEG. Like the later experiments, the EEG data focussed on in this experiment is that coming from electrodes over anterior temporal areas and doso-lateral frontal areas (see Data Analysis, below, for more information). It is expected that greater coherence, or phase-locking, will be exhibited between these two areas for more difficult change detection tasks. This, however, is a very tentative hypothesis. It is expected that, at the very least, there will be some systematic effect of change detection difficulty on the fronto-temporal EEG.
6.2.2 Experiment 5: EEG of Repeated Presentation

In Experiment 5, a series of pictures are presented to participants in a sequence. The aim of this experiment is to look at the time course of EEG responses in the occipital, temporal and frontal lobes as participants are presented with objects and required to maintain information about them across a delay. The experiment presents participants with a sequence of fractals (see Figure 5.8 in Chapter 5) and recognisable objects (see Figure 6.5). Each stimulus is presented for 100 msec, in keeping with the RSVP and repetition blindness (RB) experiments discussed in Chapter 2. However, the time delays between stimuli are manipulated systematically as is the repetition (and temporal proximity in the presentation sequence) of a particular stimulus. Also, stimuli can undergo elementary transformations (see Figure 6.6). The use of various visual transformations should help establish the level of invariance at this level of visual representation (i.e., how robust the EEG is to visual transformations of the same object/stimulus). This would extend the work on invariance of the temporal lobe (see Chapter 4) to invariance of temporo-frontal interactions.

Figure 6.5: Recognisable objects used in Experiments 5-7. From Tarr and Cheng (2003).

The delay between two neighbouring stimuli is dictated by the trial block. There are six blocks, corresponding to six different delay times: 0 msec, 100 msec, 200 msec, 500 msec, 1 sec, 5 secs and 10 secs. The first purpose of Experiment 5 is to examine the accuracy of the behavioural response as stimuli undergo different transformations across time and space. The second purpose is to examine how
6.2.3 Experiments 6 and 7: EEG of Delayed Match-to-Sample

Experiments 6 and 7 use the Delayed Match-to-Sample paradigm discussed in Chapter 5. They use the same stimuli as Experiment 5 and require an observer to view an initial stimulus, wait a period, then view a series of four stimuli and match one of these to the initial stimulus. The experiments differ in the nature of the stimuli used and the response required. Experiment 6 uses untransformed stimuli (see Figure 6.7) while Experiment 7 uses transformed stimuli (see Figure 6.8). Both experiments require the participant to make a ‘yes/no’ response at the end of each trial, indicating if there was a match present. Both experiments manipulate the delay time (levels: 5 sec, 10 sec, 15 sec, 20 sec, 30 sec). Experiment 7 includes manipulations regarding the type and number of visual transformations while Experiment 6 does not. The accuracy of behavioural responses in these experiments, like those of Experiment 5, should be quite high and, like that experiment, the main data under consideration here is the EEG. The EEG will be looked at to examine the fronto-temporal interactions involved in a delay task such as this. As this task is very similar to the ones in the experiments of Miyashita (1993), it is expected that his single-cell data can be reconciled with our EEG data to begin building a theoretical model of fronto-temporal interactions in the maintenance of visual information. As with Experiment 5, the use of visual information in these experiments is interesting.
transformations in Experiment 7 is designed to probe the level of invariance that exists for the representations under consideration.

![Figure 6.7: The fractal presentation paradigm with untransformed stimuli. The dots represent the delay period.](image)

Figure 6.7: The fractal presentation paradigm with untransformed stimuli. The dots represent the delay period.

![Figure 6.8: The fractal presentation paradigm where the match and some distractors have had a single transformation applied to them (orientation, size, contrast).](image)

Figure 6.8: The fractal presentation paradigm where the match and some distractors have had a single transformation applied to them (orientation, size, contrast).

### 6.3 Summary of Experiments

In summary, the first four experiments are integrating the paradigms of visual search and change blindness in order to establish more objective and useful measures of the ease/difficulty with which changes in various features are detected by the attentional system. Experiments 5-7 look at how the processes of object recognition and visual memory interact in a neural context. These experiments are primarily concerned with uncovering how object representations in the visual system are maintained through interactions with the frontal lobe. Together, the two areas of experimental work in this thesis represent a psychophysical and neurophysiological approach to understanding crucial links between attention, recognition and memory. More specifically, the initial experiments look at the way in which elementary visual features are combined into coherent representations (Experiments 1-3) while the later experiments examine the neural correlates of this combination (Experiments 4-5) and how integrated representations are used in higher-level cognitive tasks (Experiments 6-7).
Chapter 7

Results

7.1 Results for Pilot Study

Results of a pilot study (based on experiments 1 and 2) were planned for inclusion in this confirmation document. However, unforseen problems with the experimental software meant they were not ready at the time of printing. Therefore, this data will be presented as part of the seminar presentation given on March 8, 2005.
Appendix A

Budget

A.1 Summary of Funding Requested

Table A.1 shows the equipment necessary to conduct the experiments outlined in this document and the relevant costings. These costings are based on a quote received from ADInstruments Pty. Ltd. (see A.2).

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Quoted Price ($AUS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol Swabs</td>
<td>1000</td>
<td>161</td>
</tr>
<tr>
<td>Electrode Paste</td>
<td>6</td>
<td>93</td>
</tr>
<tr>
<td>EEG Cap with Electrodes</td>
<td>1</td>
<td>980</td>
</tr>
</tbody>
</table>

The quote also included a freight cost of $30. Therefore, the total amount requested for this budget is $161 + $93 + $980 + $30 = $1264. Below is the email received as a quote for the cost of these products.

A.2 Actual Quote E-Mail

From: Helen Lalevski <hlalevski@adi.com.au>
To: aburmester@hms.uq.edu.au
Subject: RE: Quotation request
Date: Fri, 18 Feb 2005 10:14:34 +1100 (09:14 EST)

Dear Alex,

Thank you for visiting our web site and for your interest in our product range. Please find following the price information you requested.

Requested products:
MLA1094 Alcohol Swabs (1000 Pack): $161.00 per pack (GST Inclusive)
MLA1095B Electrode Paste (6 pack): $93.00 per pack (GST Inclusive)
MLAEC1 EEG Electro-cap System 1 (medium cap): $980.00 each (GST Inclusive)
Freight Handling: $30.00

Please note, these items can also be ordered via our Australian Online Store at http://www.adinstruments.com/webstore/index.php?region=Australia
If you have any further questions please feel free to contact me on +612 9899 5455 or by reply email. I look forward to hearing from you again.
Kindest regards,
Helen

Helen Lalevski
Applications Scientist
ADInstruments Pty Ltd
Unit 6, 4 Gladstone Rd
Castle Hill NSW 2154
Ph: +612 9899 5455
Fax: +612 9899 5847
Email: hlalevski@adi.com.au
Web: www.adinstruments.com
References


Gouras, P., & Kruger, J. (1979). Responses of cells in the foveal visual cortex of
monkeys to pure colour contrast. *Journal of Neurophysiology, 42*, 850–860.


94


