Jerzy Jarmasz. *Towards the Integration of Perceptual Organization and Visual Attention: The Inferential Attentional Allocation Model.* Carleton University Cognitive Science Technical Report 2001-08. URL http://www.carleton.ca/iis/TechReports. December, 2001.

Towards the Integration of Perceptual Organization and Visual Attention: The Inferential Attentional Allocation Model.

Prospectus presented in partial fulfillment of the requirements for 07.699: Ph.D. Thesis in Cognitive Science

Jerzy Jarmasz, #253636

Ph.D. Program in Cognitive Science, Carleton University

Summer 2001

ABSTRACT

Object-based models of visual attention purport to explain why it is easier to process information within one object or perceptual group than across two or more groups. Perceptual groups are generally defined in terms of Gestalt grouping principles. These models of attention have been used to explain the phenomenon of cognitive tunneling within Heads-Up Displays (HUDs), on the assumption that the symbology of a Heads-Up Display (HUD) in a cockpit forms a single perceptual group and the outside scene forms another.

Despite extensive empirical support, object-based models have various shortcomings. In particular, the use of Gestalt grouping principles to define the notion of objects does not allow for an operational measure of what an object is to the visual system. Also, the Gestalt principles do not allow for a systematic distinction between spatial and object-based mechanisms of attention. Finally, it is generally assumed that Gestalt grouping occurs preattentively, whereas there is evidence that perceptual grouping requires attentional resources.

The proposed line of research aims to develop an account of object-based attention that does not rely on these premises. Rather, it is assumed that the cost of dividing attention between objects reflects the cost of perceptual organization itself. A qualitative model based on this assumption, called the “Inferential Attentional Allocation Model,” is given. A number of experiments are proposed to test key aspects of the model, in particular the effects of motion and top-down knowledge on perceptual organization and attention. It is expected that the results will facilitate the development of a quantitative model of object-based attention, based on a computational characterization of perceptual organization as inference to the best explanation. Finally, the implications of this research for HUDs with dynamic elements are discussed.

TABLE OF CONTENTS

ABSTRACT 2

1. INTRODUCTION 4
- **1.1 Some Basic Concepts on Attention** 6
- **1.2 The Spotlight Model of Attention** 6
  - 1.2.1Characteristics of the spotlight model 7
- **1.3 Evidence for object-based attention** 8
  - 1.3.1 Behavioural Evidence from Basic Research 8
  - 1.3.2 Behavioural Evidence from Aviation Psychology 11

2. WHAT IS WRONG WITH RESEARCH ON OBJECT-BASED ATTENTION 14
- **2.1 Gestalt principles of perceptual organization** 14
- **2.2 How many objects? One, two or more?** 15
- **2.3 The confounding of spatial and object-based factors** 18
- **2.4 Does perception really precede attention?** 20

3. TOWARDS AN INTERACTIVE, INCREMENTAL AND INFERENTIAL MODEL OF PERCEPTION AND ATTENTION 23
- **3.1 An interactive model of attention and perceptual grouping** 23
- **3.2 Visual inference** 27
  - 3.2.1 The nature of visual inference 28
  - 3.2.2 The premises of visual inference 30
  - 3.2.3 The rules of visual inference 32
  - 3.2.4 Validity: What is an object? 34

4. PROPOSED RESEARCH: GROUNDWORK FOR IAAM 38
- **4.1 Preliminary research** 40
- **4.2 Logic of proposed experimental research** 43

5. IMPLICATIONS FOR HUDS 45

REFERENCES 46

1. INTRODUCTION

Current theories of visual attention generally fall into two groups, spatial vs. object-based. Spatial, or "spotlight" (Posner, 1980) theories of attention are based on the assumption that attention is allocated to regions of the visual field before any perceptual organization or categorization occurs. Object-based theories take it as given that attention is allocated to perceptual groups or objects rather than undifferentiated regions of the visual field (Lavie & Driver, 1996). While these accounts are by no means incompatible, they are often assumed to be rival accounts of visual attention (Driver & Baylis, 1998). This debate is complicated by the fact that the notions of object and perceptual group have not been adequately operationalized, thus making it difficult to clearly distinguish between the object-based and spotlight hypotheses (Lavie & Driver, 1996).

The goal of the proposed research is to develop and test a new model of object-based attention. This model rests on the central assumption that visual attention and perceptual organization are concurrent and interactive processes. From this it follows that the processing cost of perceptual organization (i.e., the process of object formation) should be reflected in how attention is allocated in the visual field, resulting in slower or less accurate processing. This is in opposition to the view that currently prevails in most object-based attention research, viz., that attention is allocated to perceptual groups formed preattentively (Driver & Baylis, 1998; Feldman, 1997). On the standard view, attention is object based because attention selects perceptual groups. On the view I develop below, attention is object based because it participates in the process of perceptual organization.

A central limitation in research on object-based attention is the lack of a reliable, operational definition of objects and perceptual groups. This problem was not salient in early experiments on object-based attention (Driver & Baylis, 1989; Duncan, 1984; Treisman, Kahneman & Burkell, 1983) because the "objects" used in those experiments were quite simple, and contrived to form relatively unambiguous objects (Figure 1). Thus, the objects and groups

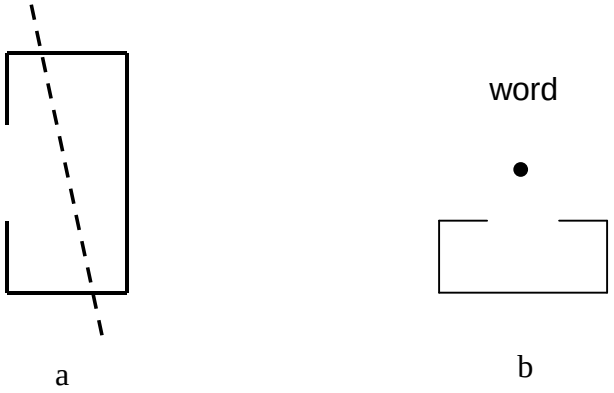

Figure 1: Examples of displays used in early research on object-based attention.
(a): Duncan (1984): (b): Treisman et al. (1983)

perceived in these displays were readily accounted for by the principles of perceptual grouping that were developed by Gestalt psychologists (see Koffka, 1935), which have become the standard account of pre-attentional grouping for most object-based attention theorists (Driver & Baylis, 1998). However, the need for better definitions of objects and perceptual groups is

obvious when examining the role of object-based attention in more rich and dynamic contexts. For example, object-based attention has been used to explain the phenomenon of cognitive tunelling, where a pilot focuses on computer-generated instrumentation (called a Head-up display, or HUD) to the point of missing crucial events in the outside scene (Fischer, Haines & Price, 1980). Wickens and Long (1994, 1995) have suggested that the HUD and the outside scenery form two separate perceptual groups. By assuming that attention is object-based, they claim that pilots can attend to the HUD or the scenery, but not to both at once, thus explaining cognitive tunneling. However, one cannot simply assume that a HUD is perceived as a single object, or even as a coherent group of objects, based on the Gestalt principles. Figure 2 shows a typical HUD. Clearly, the HUD consists of numerous elements that might not form an object. Thus, the visual displays used in applied research make it clear that a proper definition of "objecthood" is imperative.

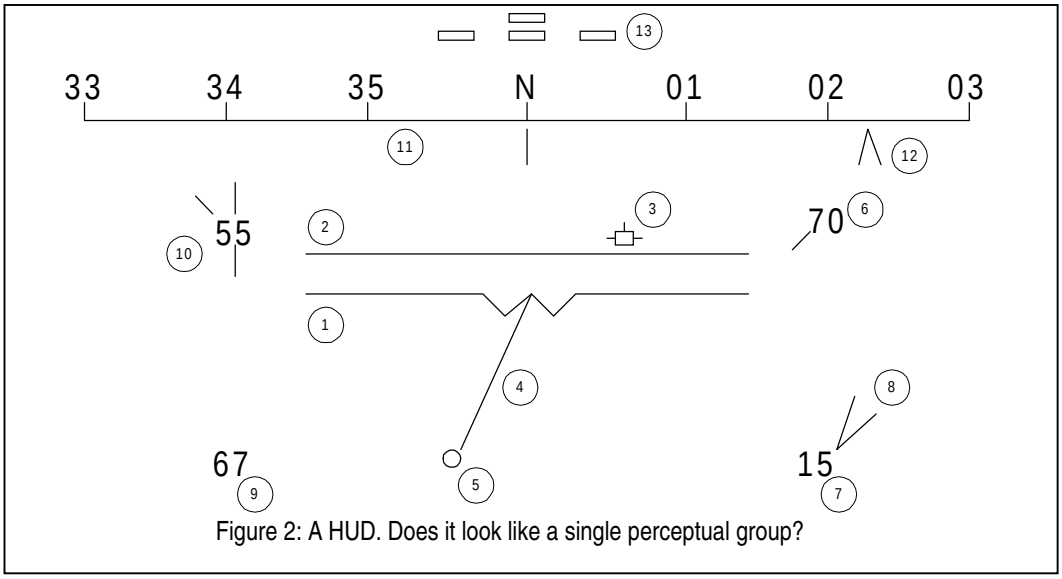

Figure 2: A HUD. Does it look like a single perceptual group?

Motion also poses a special challenge to theories of object-based attention. Driver and Baylis (1989) have shown that moving stimuli can hinder performance in an object-recognition task more than static stimuli that are near to the target stimuli. In another line of research, Spelke, Gutheil and Van de Walle (1995) have shown that motion plays a crucial role in the development of object perception in infants. From this it has been assumed that elements are grouped into objects and coherent groups by common motion (Valdes-Sosa, Cobo & Pinilla, 1998; Yantis, 1998), a principle known as "common fate" in Gestalt psychology. On this view, elements moving together are grouped together, and static elements, with null common motion, are also grouped together. However, a study by McLeod, Driver, Dienes and Crisp (1991) shows that in a display consisting of both elements with common motion and static elements, the moving elements are grouped together, whereas the static elements are not grouped. Thus, it is possible that scenes are not segregated into moving groups versus static groups based on common motion. It might not be the case that common fate separates elements into distinct objects as is commonly thought.

Object-based models seem to account for many important aspects of visual attention. However, they have not yet been fully developed. The role of the processes of object perception,

perceptual organization, and specifically motion, on attention need to be further clarified. In the current research, a model, called the Inferential Attentional Allocation Model (IAAM), is proposed that better takes into account the known processes of perceptual organization, especially motion and top-down factors. The main goal of the proposed research is to further refine IAAM and to test its validity in a series of experiments. The current research also proposes to further explore the benefit of applying IAAM to the study of HUDs.

1.1 Some Basic Concepts on Attention

Attention is a fundamental topic in cognitive psychology and a topic that has been raised extensively in human factors research. However, despite the numerous models of attention that have been proposed over the years, there is still much confusion as the nature and role of attention. Clearly, a primary role for attention is to selectively enhance the processing of "privileged" information. But this is by no means the only role. Attention serves to enhance neural information processing (Stelmach, Campsall & Herdman, 1997), modulate motor responses to stimuli (Tipper & Weaver, 1998), maintain working memory (Fernandez-Duque & Johnson, 1999), and control the sequencing of cognitive operations (Fernandez-Duque & Johnson, 1999). Attention also is necessary for the binding of perceptual features into a single phenomenal object (Treisman, 1998; Treisman et al., 1983; Treisman & Gelade, 1980). In addition, attention allows an organism to select a relevant mental representation of the environment to guide further action (Tipper & Weaver, 1998).

The contrast between the Treisman (1998; Treisman et al., 1983; Treisman & Gelade, 1980) account and the Tipper and Weaver (1998) view is informative. While both assume that attention operates on mental representations, Treisman contends that attention is necessary to form coherent representations, while Tipper and Weaver assume that attention selects a given representation, and take some form of completed representation for granted. This illustrates one of the fundamental tensions in the attention literature: does attention play a role in integrating perceptual information into coherent representations, or does it simply select "pre-formed" representations for further processing? This has been expressed in the literature in many ways, the most prominent being the debates on early vs. late selection and on spatial vs. object based attention. In both of these debates, one camp assumes that attention is required for perceptual processing; this is the case for the early selection and the spatial models of attention. The other camp assumes that some degree of perceptual processing occurs pre-attentively, and that attention simply serves the purpose of selecting one of these representations; late selection and object-based attention models tend to fall into this category.

Not surprisingly, experimental research lends some degree of support to both camps of models of attention. In what follows, the focus will be on data supporting the spatial, or "spotlight," and object-based theories of visual attention. While both accounts are useful in understanding the processing of visual information, object-based attention theories play a central role in the study of dynamic visual environments, both in the laboratory (Pylyshyn & Storm, 1988) and in applied settings (Wickens & Long, 1995). Thus, the emphasis in this research shall be on object-based attention.

1.2 The Spotlight Model of Attention

The earliest mechanistic model of attention is the filter model (Broadbent, 1958, cited in Fernandez-Duque & Johnson, 1999). This model grew mainly out of the application of information theory, as developed by Shannon (1938), to research on auditory attention. On this

model, attention is a cognitive structure that protects the system form information overload by filtering out information. Some information is let through the filter for further processing, whereas everything else is discarded. What is filtered out is determined on the basis of the physical characteristics of the stimulus, before any kind of conceptual processing takes place (i.e., before a stimulus is categorized or identified).

While the spotlight models have been useful in accounting for certain aspects of auditory information processing, these models have met with less success when applied to vision. For instance, it has been shown that using a visual cue to draw a subject's attention to a particular region of a display enhances processing of information at that location (Posner, 1980). In such cases, the stimuli to be processed are often the only ones in the display, and there is therefore nothing else to filter out. Nevertheless, responses are more rapid to cued than to uncued stimuli. These data cannot be accounted for with a model that simply assumes a passive filtering out of irrelevant data. Rather, they suggest, that the visual system has the ability to selectively enhance processing in particular regions of the visual field, much in the way a spotlight illuminates a particular region of a darkened stage. This conception of visual attention came to be known as the spotlight model (Fernandez-Duque & Johnson, 1999).

1.2.1 Characteristics of the spotlight model

There are two versions of the spotlight model, both of which are based on the notion that attention "highlights" a region of the visual field. One version of this model is similar to the filter model in that what falls outside of the attentional spotlight is assumed not to be processed (Posner, 1980; see also Fernandez-Duque & Johnson, 1999). In the second version, the spotlight serves to concentrate attentional resources to a particular region in space, thereby enhancing processing at that location, but without completely eliminating processing of the unattended regions (Downing & Pinker, 1985; Jonides, 1981).

The spotlight models, however, differ from the filter model in several important respects. First, the attentional filter is viewed as a structure through which information must flow. In contrast, the attentional spotlight is not a structure but rather a functional enhancement of information flow. On this view, the spotlight does not passively block the processing of extraneous information, rather it actively directs processing resources to the area within the spotlight (Fernandez-Duque & Johnson, 1999). Second, the spotlight has a spatial dimension that the filter lacks: the spotlight selects a region of space, whereas the filter is insensitive to the spatial layout of information. On this model, the latency to respond to two stimuli increases as the spatial separation between the stimuli increases, because the spotlight must travel from one to the other (Posner, 1980). These so-called "spatial effects" are often considered the marker effect of the spotlight model. Third, whereas the parameters of the attentional filter are assumed to be fixed (Fernandez-Duque & Johnson, 1999), the spotlight can be controlled. In particular, the spotlight can be moved to different parts of the visual field and, according to some specific models, the size and possibly the shape of the spotlight can be changed. This flexibility raises the issue of what controls the spotlight, which didn't emerge with the filter model.

In sum, the spotlight models improve on the filter model by introducing the notion of enhancement of information processing (crucial to cueing data), via mechanisms that control attentional parameters. Most of the early spotlight models included the assumption that the spotlight <u>cannot</u> be split between separate regions of space. Accordingly, attending to several regions of space required serial shifts of the spotlight. Later versions of the spotlight model include the assumption that the attentional spotlight can be split among two or three regions of

space (Fernandez-Duque & Johnson, 1999).

1.3 Evidence for object-based attention

Object-based attention is based on the assumption that attention is referenced to discrete objects in the visual field. On this view, it is more difficult to divide attention between two or more objects at once than to attend a single object. This is known as the "object effect" or the "object advantage" in the research literature (see Lavie & Driver, 1996). The contrast with the spotlight model is clear. Whereas on the spotlight model, one would predict that two nearby or overlapping objects would attended to as easily as a single object, on the object-based model this would not be the case. As discussed below, mounting evidence supports the idea that dividing attention between objects results in less efficient processing than attending to a single objects. It should be noted that spotlight and object-based attention theories are not contradictory but rather complementary (Lavie & Driver, 1996; Logan, 1996), despite the fact that they have often been described as rivals (Driver & Baylis, 1998). Nevertheless, in many cases the object-based theory explains many phenomena better than the spotlight model does.

1.3.1 Behavioural Evidence from Basic Research

A growing body of experimental evidence suggests that attention is object-based. Experiments by Duncan (1984), Treisman et al. (1983), and Baylis and Driver (1992, 1993; Driver & Baylis, 1989) stand out in this regard.

Duncan (1984) presented subjects with two overlapping figures, a box and a line drawn through it diagonally. The two objects could vary with regard to two properties: the box could be small or large, and it could have a gap in its right or its left edge; the line was either dotted or dashed (texture) or tilted to the left or to the right (orientation). The subjects' task was to identify two attributes at once; in one condition they were to report on attributes from a single object (either the line or the box); in the other, they were to report on one attribute from each object. On the assumption that attention is analogous to a spotlight, it would be expected that subjects' performance would not be affected by the distribution of the attributes to be judged across the objects, as they share the same location in space. However, the results showed that subjects' identification was more accurate when the two attributes were located on a single object compared to when one attribute was located on one object and another attribute was located on the other object.

Treisman et al. (1983) obtained a similar cost in performance when the targets to be identified belonged to different objects. Subjects were presented with a rectangular frame and a word which were configured in one of two ways. In one configuration, the frame and the word were presented apart (above and below a fixation point) representing two distinct objects. In the other, the word was presented within the frame, assumed to form a single object. In both cases the distance between the outline of the frame and the word was 1° of visual angle. The subjects' task was to read the word and to judge the location of a gap in the frame. The gap was always located the same distance from the word. The results clearly showed that performance was significantly facilitated when the word was presented within the frame (presumably forming a single perceptual object) compared to when the word and the frame were separate.

Driver and Baylis (1989) reported a ground-breaking experiment on object-based attention which was based on evidence previously interpreted as favouring a spotlight model of attention. This experiment examined response competition where interference of distractors on target response decreases in relation to increased spatial distance (Eriksen & Eriksen, 1974).

Driver and Baylis used the Eriksen and Eriksen paradigm (subjects responded to a central letter located in an array of five letters) but grouped the target and distractors together using common motion. When the outer letters moved with the target they interfered more with target identification than the nearby letters that had remained stationary. This demonstrated how a seemingly spatial effect breaks down when the target and the distractors are grouped together. Using the perceptual grouping principle of common fate (Koffka, 1935) the distant distractors were grouped with the target and as such produced more interference than distractors located closer to the target. The spatial models cannot account for these results, as the basic claim of the space-based hypothesis is that visual attention is allocated to contiguous regions in space, and that everything within such a region gets processed.

Clearly, the hypothesis that attention simply selects regions of the retinal image and favours the processing of everything in the selected regions is unable to account for these experimental results. This has lead researchers to formulate so-called object-based models of visual attention. The essence of these models is that attention operates on visual information that has already undergone some degree of processing which has organized the retinal image into objects or perceptual groups. It has been commonly assumed that the mechanisms which perform this perceptual organization follow the principles of perceptual organization as proposed and studied by the school of Gestalt psychology (Kanizsa, 1979; Koffka, 1935), or at the very least something close to this. The validity of this assumption shall be subsequently reviewed in more detail; for now, it will suffice to mention the grouping principles most commonly implicated in object-based attention by writers: grouping by similarity (elements of a display sharing a common attribute, such as shape or colour, are generally perceived as being grouped together); grouping by proximity (elements which are close to one another are generally perceived as forming a group); grouping by continuity, completion and closure (elements which only partly suggest a complete shape, such as collinear dashes, the corners of a box, or part of a disk, are generally perceived as forming the completed shape); and grouping by common fate (elements which move together are perceived as forming a group). Treisman et al.'s (1983) and Duncan's (1984) displays seem to rely mainly on the principles of completion and closure, whereas the stimuli used by Driver and Baylis (1989) rely on the principle of common fate.

Subsequent research has provided further evidence to support the claim that attention is object-based, in that it is often allocated on the basis of visual information that has undergone some degree of perceptual organization. For instance, Goldsmith (1998) showed that visual search is easier when features are linked to the same object than when they belong to different objects. Similarly, Duncan and Nimmo-Smith (1996) found that it is harder for subjects to discriminate between features that belong to different objects compared to a situation requiring discrimination between features that belong to a single object.

Kramer and Jacobson (1991) used a variation of the response competition paradigm (Eriksen & Eriksen, 1974) to test the object effect in a focused attention task. Subjects judged whether the target (a line) was dashed or dotted while ignoring distractors. Distracting lines (compatible or incompatible with target) were located to the left or right of the target. The distractors could be grouped according to the Gestalt principles with the target, or they could form a part of a different object. The distance between the target and the distractors was kept constant. According to the space-based models there should be no difference between the different conditions, because the spatial separation was constant. On the other hand, according to the object-based model there should be less interference when the distractors belong to a separate object. As predicted by the object-based model, the interference from distractors was drastically

reduced or eliminated when the distractors and the target belonged to different objects. Further, when the incompatible distractors formed part of the same object as the target, reaction time and accuracy was significantly reduced compared to when the compatible distractors belonged to the same object as the target.

In sum, the response competition that is produced by distractors cannot be explained by referring to the spatial hypothesis. Grouping the distractors and the target together by colour or good continuation causes significantly more interference with target response than distractors that are easily separated from the target. Spatial distance seems to have no influence here. This suggests that visual attention is directed to perceptual objects in the visual field that are segmented according to the Gestalt grouping principles.

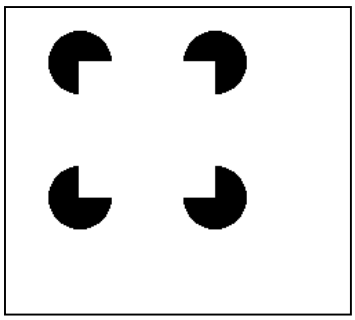

Figure 3: A "Kanizsa square" as used by Jordan & Tipper (1998)

Experimental evidence supporting object-based attention has also been adduced from inhibition-of-return (IOR) tasks. IOR was originally interpreted as supporting spatial models of attention, as certain locations in space are prevented from being constantly re-examined. However, if attention is object-based then the inhibitory mechanism should be directed towards structure in the visual field rather than location. To this end, recent evidence suggests that IOR is related to perceptual objects in the visual field rather than spatial location (Tipper, Driver & Weaver, 1991). Jordan and Tipper (1998) used a static display to examine the difference between cueing location vs. cueing visible objects in the display. The display consisted of black “pacmen” (discs with a quadrant missing) and lines. In one condition objects formed by illusory contours (so-called Kanizsa squares; see Figure 3) were visible whereas in another condition no such objects were visible. The IOR effect was much larger when an object was cued compared to when only a location was cued.

Perhaps the most compelling evidence for object-based attention comes from research done with moving displays. As we have seen, Driver and Baylis (1989) had already shown that common fate is a powerful grouping principle affecting attention. Along similar lines, Tipper and Weaver (1998) carried out IOR experiments using elements endowed with common motion. Research on IOR with static displays easily lends itself to a spatial explanation: attending to a particular location can inhibit attending to the same location a little while later. However, Tipper and Weaver reasoned that if an organism is to successfully survive in a dynamic and complex environment, then the IOR phenomenon must also apply to moving objects. In this case, it wouldn’t do to inhibit return to a specific location, because, by definition, moving objects are changing their location. Thus, it would make sense to expect that an object-based IOR mechanism exists to allow efficient scanning of moving scenes. Tipper and Weaver’s data does in fact suggest that there is an object-based component to IOR, in that subjects were slower to return attention to a pre-cued moving object. While their data does not exclude the possibility that IOR is at least in part location-based as well, it certainly argues for the notion that for attention to be allocated to moving stimuli, it must in some sense be object-based.

Pylyshyn and Storm (1988) tested the assumption, arising from a spatial model of attention, that people are able to track many elements by moving their attentional spotlight from element to element in rapid succession. They first tested subjects’ ability to track independently moving and similar elements within a display of many moving elements. This showed that people are able to successfully track about four or five elements for at least 10 seconds. A computer simulation of this task was developed using the actual trajectories that were shown to

subjects combined with a model of a spatial spotlight. This simulation showed that an attentional spotlight moving from element to element would not be able to keep track of the elements. Thus, Pylyshyn and Storm concluded that a spatial attention hypothesis cannot account for people's ability to perform multiple object tracking.

There are two object-based accounts of how people might perform multiple tracking of objects. One account is that the multiple elements are formed into a nonrigid polygon, with each element being one of the vertices of the polygon. On this view, Yantis (1992) found that tracking performance is affected by factors that facilitate the initial formation and maintenance of a perceptual group of elements to be tracked. Thus, Yantis argues that specific elements within a display of similar moving elements are tracked by grouping the elements into a single "superobject" (essentially a nonrigid polygon). The less polygon-like the nonrigid polygon is, the harder the task.

Pylyshyn and his colleagues have developed a different account, which assumes that each element being tracked is associated to a visual index, or a FInger of INSTantiation (Pylyshyn, 1989). On this view, the early visual system attaches indexes to the individual elements to be tracked. These indexes provide a way for the visual system to pick out specific elements of the visual field by referring to the elements themselves, and not to any properties of the objects (Pylyshyn, 1998). These indexes then allow the rest of the visual system to attend to those specific elements, to track them, to identify them, and so on (Scholl & Pylyshyn, 1999; Sears & Pylyshyn, 2000). Thus, visual indexes are a kind of representation or data structure within the visual system which function in a manner analogous to linguistic indexes and demonstratives (e.g., words such as "that" or "there")[1].

In sum, the ability to track moving objects appears to require that attention be object-based at least to some degree. Spatial models of attention are inadequate for explaining the experimental evidence obtained from multiple tracking tasks, as these models would require the spatial spotlight to visit each moving element in rapid succession. Successive tracking of this sort would mean that the visual system is able to predict the positions of the moving elements and to move the attentional spotlight at a speed that is beyond the capacities of the human visual system.

1.3.2 Behavioural Evidence from Aviation Psychology

Applied research, in particular from the area of aviation psychology, has shown that object-based attention is a construct that has external and ecological validity. Furthermore, object-based attention has been used to suggest ways in which HUDs can be improved. The aviation psychology research literature thus provides evidence that object-based attention is a very useful and powerful theory.

Fischer et al. (1980) described a simulator-based experiment where pilots were required to perform runway approaches flying an aircraft that was equipped with a HUD versus a traditional Heads Down Display (HDD). It was found that some of the pilots using the HUD failed to notice unexpected intrusions on the runway when they were also required to attend to events in the near domain (see also McCann & Foyle, 1996). This failure to notice runway intrusions was not experienced by pilots using the HDD. Although the results suggest that the

[1] On Yantis's (1992) account, attentional enhancement should be observed within the confines of the non-rigid polygon. Sears and Pylyshyn (2000) failed to find any such effects, thus arguing that attention is indexed specifically to the objects being tracked, and not to regions of the display defined by the objects, *contra* Yantis.

presence of the HUD is responsible for the degraded performance, the Fischer et al. study was flawed in that the location of the instrumentation (HUD vs. HDD) was confounded with the type of instrumentation.

Wickens and Long (1994) repeated the Fischer et al. (1980) experiment but with matched instrumentation across the HUD and HDD. In contrast to the Fisher et al. findings, pilots using the HUD were successful in noticing runway intrusions. However, these pilots were considerably (2.5 seconds) slower to respond to intrusions than were pilots using the HDD. In sum, both the Fisher et al. (detection accuracy) and the Wickens and Long (detection time) studies show a disadvantage for HUDs versus HDDs. This disadvantage seems to arise in situations where the pilot has to simultaneously attend to information located on the HUD and in the external scene. It should be noted that in these studies, the HUD is collimated to infinity, meaning that the HUD and the outside scene are at the same focal distance.

The disadvantages of HUDs are further illustrated in a study by Foyle, Stanford and McCann (1991), which required pilots to control their flight-path while maintaining a fixed altitude. Superimposing a HUD digital readout of altitude onto the flight path resulted in excellent control of altitude. However, when focusing on altitude, pilots tended to collide with the flight-path markers, such as buildings or landmarks. This tradeoff between using the HUD symbology (digital altitude) and processing of the external scene cannot be attributed to visual interference or masking: The same HUD symbology was presented across the various conditions. Instead, this evidence suggests that when the HUD symbology is required for performance, the symbology attracts the pilot's attention at the cost of not attending to object and events in the environment. That is, when focusing attention on one domain, information located on the other domain tend to go unnoticed. This phenomenon has been labelled cognitive tunnelling (Martin-Emerson & Wickens, 1997; Wickens & Long, 1995).

The evidence for object-based attention has led researchers to claim that the attentional problems experienced with HUDs, such as cognitive tunelling, are due to the near and far domains forming separate perceptual groups or objects. This claim is predicated on the notion that near and far domains differ along one or more of the Gestalt grouping principles. In particular, the HUD symbology is stationary relative to the pilot- or aircraft-centric view, whereas the external scene is in constant motion. Also, HUD symbology is usually displayed in a uniform colour, which may differ from the various colours of the external scene.

The claim that the near (HUD) and far (external scene) domains form separate perceptual groups provides a possible explanation of cognitive tunnelling when combined with the object-based hypothesis. On this view, when pilots attend to the near domain, all of the HUD symbols get processed quickly in parallel while processing of information in the far domain is delayed.

In addition to providing possible explanations for attentional difficulties in using HUDs, object-based models of visual attention have also been used to suggest possible solutions for these problems. These solutions have generally aimed at improving pilots' ability to integrate information from the near and far domains by fusing both domains (or at least some aspects of both) into a single domain or perceptual group. One way of doing this is to use conformal symbology. Broadly speaking, the definition of conformality used by HUD developers refers to the degree to which a symbol forms an object within the scenery. The idea is that a conformal symbol should serve as a virtual analog for far domain elements. In other words, symbology that is an accurate graphic representation of an actual object represented in the far domain, or that forms a one-to-one correspondence with the world is deemed to be conformal (Martin-Emerson & Wickens, 1997). On this view, conformal symbology can be a virtual runway overlaying the

actual runway or a scene-linked symbology where, for example, altitude is represented at the height of and possibly in actual objects in the far domain. Non-conformal symbology would be symbols such as a digital readout of the altitude or airspeed, path guidance information like glide slope, or localizer symbology (Wickens & Long, 1995). It should be noted that according to this definition, even traditional HUDs have some conformal symbology (e.g., a horizon line). In constrast, symbols representing VSI, airspeed, distance, and altitude etc. are usually non-conformal.

Experiments examining conformal symbology in HUDs have yielded promising results. For example, research by Foyle, Stanford and McCann (1991; see also McCann & Foyle, 1994) has shown that when using conformal symbology pilots are able to maintain altitude and follow a flight path without significant trade-offs in performance. In these studies, altitude symbology was rendered conformal by placing the symbology on virtual buildings along the flight path. In contrast, when the altitude indicator was superimposed onto the path (non-conformal) the task of maintaining altitude reduced flight path performance.

Varying the form of the conformal symbology does not seem to diminish the enhanced performance. McCann and Foyle (1996) and Shelden, Foyle and McCann (1997) have shown evidence for the same benefit of conformal symbology over non-conformal symbology regardless of whether the form of the symbology was analog ("clockface") or digital. These experiments are promising and suggest that the conformal character of the HUD symbology presumably enables parallel processing of information from the two domains. In accord with an object-based hypothesis, conformal symbology might allow for the creation of a single far domain (or object layer) of information. On this view, performance is enhanced because the pilot is able to allocate attention to the far domain without the need to switch to the near domain.

In sum, an object-based approach can be useful in assessing the attentional problems associated with the use of HUDs (i.e., cognitive tunelling). It can also suggest ways of dealing with these problems (i.e., conformal symbology).

2. WHAT IS WRONG WITH RESEARCH ON OBJECT-BASED ATTENTION

Although experimental evidence suggests that visual attention interacts with perceptual grouping mechanisms, there are problems in interpreting these results. The problems fall into two categories. One category concerns the nature of the perceptual grouping phenomena that are assumed to provide the units of object-based attention. As we have seen, the most common assumption is that these perceptual units can be accounted for with the principles of perceptual organization from Gestalt psychology. We shall see, however, that this approach is inadequate.

The other major difficulty is one that could be called "architectural." Theories of object-based attention generally rely on the assumption that attention and perception are two independent processes that operate serially and atomically (i.e., perceptual grouping must be complete before attention can be allocated to any part of the retinal image). While intuitively appealing, this view been challenged by a number of experimental results. After reviewing the problems with Gestalt-type perceptual grouping, the issue of the relationship between perception and attention is examined.

2.1 Gestalt principles of perceptual organization

Researchers who endorse object-based attention generally define an object as a perceptual group defined by the principles of perceptual organization established by Gestalt psychology (see Koffka 1935; Kanizsa, 1979). This definition has been explicitly supported by Baylis and Driver (1989; 1992; 1993; see also Lavie & Driver, 1996), and has been endorsed by other researchers (see Wickens & Long, 1994, 1995 among others for applications in aviation psychology).

The claim of Gestalt psychology is that the human visual system tends to organize stimuli into perceptual groups according to certain principles of spatial organization. A few representative examples of these principles are: similarity; proximity; closure; good continuation; common fate (see Figures 4 and 5). These principles are motivated by an underlying overarching principle of "figural goodness," or *Prägnanz* as it is called by Gestalt psychologists. This principle states that the visual system organizes stimuli so as to form the simplest, most elegant,

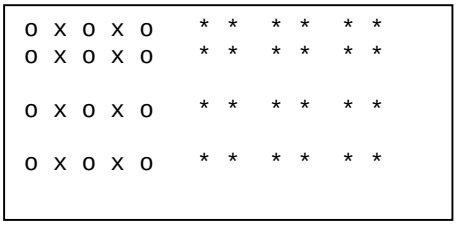

Figure 4: Grouping into columns by similarity and proximity

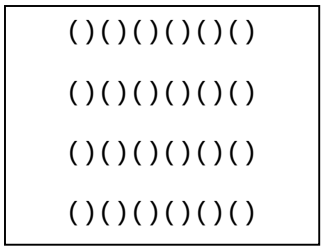

Figure 5: Grouping by closure

and intuitively most satisfying configuration. All the other principles are supposed to be particular instances of this one.

While these principles are useful in describing perceptual grouping phenomena, particularly in carefully contrived situations, they have a major flaw: the principle of *Prägnanz*, while intuitively appealing, is too vague and subjective to support a principled characterization of the mechanisms of perceptual organization. Thus, as Palmer (1999) states, the Gestalt

grouping principles are little more than *ceteris paribus* rules: particular rules apply only in particular contexts, when they are clearly the only rule that can apply. When two or more grouping principles are applicable at once, there is no way of predicting *a priori* which principle will in fact determine the configuration the observer sees (though some work has been done studying the interaction between the various grouping principles; see Baylis & Driver, 1992). Thus, the ultimate determinant of "figural goodness" is the observer's subjective experience, and whichever set of grouping principles correspond to a particular observer's subjective experience is claimed to explain a particular perceptual configuration.

Gestalt grouping principles are problematic in other ways, beyond their inability to provide objective determinants of perceptual grouping. The first is that they do not provide a distinction between perceptual groups and objects, which could themselves be comprised of many perceptual groups. Second, the Gestalt principles do not allow for a clean separation of so-called "spatial" and "object-based" effects in visual attention. It is assumed in spotlight models that multiple stimuli are inspected serially, and thus the time it takes to attend to multiple stimuli is a function of their spatial separations. Conversely, one would predict from object-based models that if multiple stimuli are located within a single object or perceptual group, they are processed in parallel, and thus the time in takes to process stimuli is independent of their spatial separations. However, the Gestalt principle of grouping by proximity predicts that the closer elements are, the more likely they are to be perceived as belonging to a single group or object. Thus, the rapid processing of elements that are close together could be explained either in terms of the spotlight model or the object-based model (Lavie & Driver, 1996).

It is clear that the Gestalt principles per se cannot be the basis for a reliable, operational characterization of the concept of object. However, as we have seen, much research on object-based attention takes for granted that these principles do account for the grouping mechanisms that underlie attention. The conclusions that have been drawn from much object-based attention research must therefore be viewed with some caution. In particular, the results of many experiments are ambiguous and can often be interpreted as supporting either spotlight or object-based theories. Also, research on HUDs that claims to show that pilots' performance was improved by somehow fusing the near (HUD) and far (external scene) domains might be re-interpreted as showing that the performance benefits might be due to some other aspect of object-based attention. Both of these issues shall be reviewed shortly.

It is worth noting that only one of the Gestalt principles speaks to the issue of objects in motion (i.e., the principle of common fate). Similarly, most research on visual attention, object-based or otherwise, has used static (or mostly static) displays. This is a significant limitation for experimental research. Most typical environments, including HUDs and helmet-mounted displays (HMDs), are highly dynamic. Thus, the application of basic research on visual attention to HUDs will have to involve more research on visual attention in dynamic displays. We have already seen a few examples of such work. The line of research proposed here will further explore this issue as well.

2.2 How many objects? One, two or more?

As a consequence of not having an operational definition of objectness, there is a lack principled, objective criteria for individuating objects in a display. In many experiments, this is not a serious issue because the displays are so simple that there is no ambiguity as to which elements should be taken to be objects. For instance, the displays used by Duncan (1984) consisted only of a box with a dashed line running through it. This display is fairly unambiguous.

There are, however, examples of research where the assumptions as to the number of objects being displayed or perceived by the observer are questionable. For instance, Treisman et al. (1983) constructed displays that consisted of a word and a box-shaped frame shown simultaneously. The frame had a gap in one of its sides. In one condition, the word and the frame did not overlap spatially; this configuration is most likely perceived as two objects. However, in the other condition, the word was displayed inside the frame. In this case, Treisman and her colleagues assumed that the display was perceived as a single (albeit complex) object. In both conditions, the distance between the gap in the frame and the word was the same. Since reaction times were lowest in the "one object" condition, even though there was no change in the distance between the targets the subjects had to judge, it was concluded that the overlapping display was indeed perceived as a single object, and that attention is more easily allocated to a single object than to two objects. This explanation depends on the assumption that a spatial or spotlight model would necessarily predict that reaction time would be a function of the distance between the targets. However, it could be the case that the distance between the centroids of the box and the word would determine reaction times under a spatial model (for more on this see Driver & Baylis 1993). Thus, the assumption that the overlapping display forms a single object also requires a particular reading of spatial models of visual attention. None of these assumptions are necessarily warranted.

There are cases where a configuration that appears to be made up of separate objects might be perceived as a single object with internal structure. Lavie and Driver (1996) used a display consisting of two dashed lines intersecting at their mid-points and forming an "X". Each of these lines was displayed in a different colour (red and green). The working assumption was that each line would be perceived as an object in its own right, and that dividing attention between both lines would produce a performance cost that would override any costs due to spatial separation. This assumption was again supported by the reaction times of the subjects performing the experimental task: comparing two elements located on the same line took less time than comparing two elements located on different lines. However, an experiment by Herdman, Jarmasz and Johannsdottir (2000) where both lines were displayed in the same colour suggests that under some circumstances the two lines are indeed perceived as two separate objects, while in other circumstances the two lines appear to be perceived as a single "X".

Research on applying object-based models to HUD or HMD environments illustrates more vividly how assumptions about the number of objects or perceptual groups in a display can have a significant practical impact. The assumption that has driven much of this research is that there are two relevant perceptual groupings in a HUD: the near domain (the instrumentation) and the far domain (the scenery). Many of the performance problems associated with HUDs have been explained as difficulties in dividing attention between both of these perceptual groups. For instance, Martin-Emerson and Wickens (1997) examined the difference in the use of HUDs versus HDDs across different visibility levels in terms of dividing attention between the near and far domains. As pilots came in for an approach to land under different visibility conditions they had to hold a stable altitude and control lateral and vertical tracking. Flight path guidance was superimposed onto the path for the HUD condition and located below the windshield for the HDD condition. The results showed that for the HUD condition pilots were faster to respond to events within the HUD display and to control altitude when under zero visibility as compared to full visibility conditions. In full visibility pilots attend to the far domain, thereby making it more difficult to control the altitude and respond to events occurring in the HUD. However, lateral and vertical tracking errors also decreased in the full visibility condition.

Martin-Emerson and Wickens interpreted these results as showing that pilots had difficulty dividing attention between two perceptual "objects," that is, the near and the far domains. This is consistent with an object-based hypothesis. According to this line of reasoning, in the full visibility conditions, pilots are attending to both domains, as shown by the fact that their performance on tasks requiring information from the far domain improves when this domain is visible. Because the far domain is now drawing some attention away from the near domain, pilots' performance in tasks which require information from the near domain (i.e., the HUD) is degraded in full visibility conditions relative to the zero visibility conditions.

However, there is an alternative explanation which also makes use of an object-based attention model. Because pilots had to maintain their altitude and respond to events within the near domain, it is unlikely that they switched attention completely toward the far domain. Indeed, it is possible that the pilots successfully integrated the HUD flight-path symbology with the external environment but experienced difficulty attending to other information located in the HUD. That is, given that the near domain is itself made up of various symbols, the pilots might simply have experienced difficulty in integrating information within a single domain, rather than difficulty in dividing attention between domains.

As discussed above, it has been suggested that conformal symbology might be used to reduce cognitive tunneling by fusing the near and far domains into a single perceptual group. Foyle, Stanford and McCann (1991), McCann and Foyle (1994; 1996) and Shelden, Foyle and McCann (1997) have studied the effects of using conformal symbology in a HUD environment. Although the results are promising, there were a number of methodological problems with these experiments. In particular, adding elements onto the flight path, whether those elements are virtual buildings or numbers, increases the number of cues that the pilot can use to control the aircraft's flight path. Accordingly, the enhancement in timesharing the flight path task and the symbology-based tasks may not be due to a reduced requirement to switch attention across domains. Instead, this enhanced dual-task performance may be attributed to the reduced load associated with controlling the aircraft's flight path when more path cues are present.

In another experiment, Martin-Emerson and Wickens (1997) tested the difference between conformal and non-conformal HUDs using symbology that differed only in terms of path guidance information. Both conditions included non-conformal symbology such as VSI, heading, speed, and distance. In the conformal condition, a virtual runway overlaying the actual runway provided path guidance. In the non-conformal condition, path guidance was represented by a localizer and a glide slope, a fixed aircraft symbol and a reference line. The subjects' task was to approach to land under different visibility conditions. The results showed that for the non-conformal condition there was large variance in lateral tracking errors depending on visibility. In comparison, when pilots used the conformal symbology the lateral tracking errors were undifferentiated across the different levels of visibility, thus suggesting that conformal symbology helps pilots perform the task equally well in all visibility conditions. Interestingly, the condition of symbology type had no effect on vertical tracking errors; this is not accounted for in Martin-Emerson and Wickens (1997).

Martin-Emerson and Wickens (1997) concluded that the improvements in pilot performance were due to the conformality of the symbology. However, the object-based attention hypothesis can be applied to the Martin-Emerson and Wickens results although not necessarily under the assumption that conformal symbology fuses the two domains into a single perceptual group. For example, it could be argued that conformal symbology leads to better performance because the symbology forms a coherent object in and of itself. In accord with the

object-based attention hypothesis, a sense of objectness presumably makes it easier for pilots to attend and process information within the conformal symbology. That is, the advantage of conformal symbology may not be due to the notion that the symbology is integrated into the external scene, as is commonly assumed, but instead attributed to the facilitating effects of object-based attention.

In sum, the lack of operational criteria for the enumeration and individuation of objects is a major problem in research on object-based attention. The results of much basic and applied research relying on the object-based framework can be interpreted in a number of ways, depending on which elements of the display are taken to be objects, or simply parts of objects. Thus, a goal of the proposed research is to further develop operational criteria for determining what counts as an object. Finally, the lack of an adequate definition of objecthood leads to difficulties in distinguishing object-based and spatial factors in visual attention, a topic we now turn to.

2.3 The confounding of spatial and object-based factors

McCann, Foyle and Johnston's (1993) finding that responses to targets were significantly delayed when a cue was presented in the nontarget domain could be interpreted as showing that near (HUD) and far (external scene) domains form separate visual objects. In accord with the object-based attention hypothesis, target responses are slower when the cue occurs in the nontarget domain because it takes time to switch attention from one object (domain) to the other. However, a careful review of the McCann, Foyle and Johnston procedure leads to an alternative explanation. In this experiment, pilots were required to perform an approach to land. As they approached landing, a three-letter cue ("IFR" or "VFR") appeared either on the HUD or on the runway (external scene). The cue indicated where to look for a target among several geometric symbols that appeared on both the HUD symbology set and the runway. "IFR" (for instrument flight rules) indicated that the set of symbols on HUD was relevant. "VFR" (for visual flight rules) indicated that the set of symbols located on the runway was relevant. The pilots were required to identify whether one of the symbols (the target) was a diamond or a stop sign: A landing was allowed only if the target was a diamond. Four boxes were located on the HUD to flank either side of the runway and another four boxes were superimposed onto the far domain in a similar position. The distance between the boxes was equal for both domains. The three geometric symbols appeared in three of the boxes on each domain and the cue would appear in the fourth box on either the near or the far domain. If the cue appeared on the HUD it filled the box in either the bottom left or the bottom right corner. If the cue appeared on the runway it filled either the top left or the top right box.

The results showed that subjects were significantly slower in responding to the relevant target when the target and the cue where located on different domains. For example, when the "IFR" (indicating that target on the HUD is the relevant one) appeared in the display, pilots were faster to respond to the target when it was located on the HUD than on the runway.

As noted above, the object-based attention interpretation of the McCann, Foyle and Johnston (1993) result is that the near and far domains form separate visual objects: It takes time to switch attention from one object to the other. However, another plausible interpretation is that the slower responses in the cross-domain condition are due to there being two different types of cues. The sudden onset of the three letters is a form of exogenous cueing that immediately draws attention to that location. In contrast, the interpretation of the three letters is a form of symbolic, or endogenous, cueing: the participants had to interpret the meaning of the three letters to

determine the relevant target location. Attentional allocation is much slower with endogenous cues than with exogenous cues. Whereas attentional allocation can occur within 100 ms with exogenous cues (Wright & Ward, 1998), the allocation of attention with endogenous cues can require 300 ms and longer (Stelmach el al., 1997; Wright & Ward, 1998). On this view, when "IFR" was shown on the HUD, the pilots would be able to determine almost immediately whether the target on the HUD was a diamond or stop sign. There would be no need to interpret the cue itself. Accordingly, when the symbolic cue concurred with the direct cue, pilots were fast to respond. When the symbolic cue did not concur with the direct cue (a different location of the relevant target was indicated versus the direct cue), then responses were slow.

The conclusion reached by McCann, Foyle and Johnson (1993) would be unequivocally valid only if it were the case that the near and far domains were being processed as two separate objects. However, their interpretation is not the only plausible one, which casts doubt on the assumption that the visual system is indeed dealing with only two perceptual objects in this case. Thus, it is not so straightforward to explain attentional effects in HUDs solely in terms of object-based attention, particularly when the criteria for individuating objects in a display are ambiguous.

A second, and related, problem in differentiating between spatial and object-based factors concerns the size of the display. In Duncan's (1984) experiment the display was less than 1°. This allows for only a small and potentially insignificant variation in spatial distances. It could therefore be argued that object-based factors are only important within an attentional spotlight. Similarly with the paradigm used by Treisman et al. (1983), the spatial area relevant when the word was presented within the frame was much smaller than the area relevant when the word and the frame were presented separately. Therefore, the Treisman et al. results might have reflected the benefit of spatial proximity. Also, the onset of the frame may have exogenously grabbed attention making it difficult to read the word in the condition where the frame and word were presented separately.

Spatial factors other than location or separation between targets are also frequently confounded with object-based factors. Targets are often so different that the same-object effect can be the result of some artifact (e.g., spatial frequency differences). In Duncan's (1984) experiments, the two attributes of the line are available at a high spatial frequency whereas the two attributes of the box are available at low spatial frequency. The results may therefore reflect difficulties in processing or attending to different spatial frequencies. Furthermore, the targets used and the instructions given often indicate objects prior to the actual task. To wit, Duncan's procedure in which subjects are required to identify the height of a box and the texture of a line may bias subjects toward processing the stimuli as objects.

To summarize, results from both basic and applied research on visual attention show that object-based and spatial factors are often confounded. This is largely due to the fact that models of attention are not mutually exclusive, as it is sometimes implied in the experimental literature. Rather, they reflect two parallel and complementary systems. Evidence from neuroscience shows that there are two main pathways, one for spatial processing and one for object-based processing, in the visual system (Mishkin et al., 1983). Further, there are models of attention that integrate both spatial and object-based operators of attention (Logan, 1996). Clearly separating these two attentional systems in an experimental setting requires valid and reliable operational definitions of objects and perceptual groups, which are still lacking. It will be argued below that such operational definitions will be provided only when the notion of perceptual organization as inference to the best explanation is applied to object-based attention.

2.4 Does perception really precede attention?

It has traditionally been assumed that perceptual grouping is pre-attentive (Driver & Baylis, 1998; Feldman, 1999). On this view, grouping does not require, and happens prior to the allocation of, visual attention. This has been one of the key assumptions of the object-based model. For attention to be allocated preferentially to objects, these objects must presumably be made available to the attentional system. As noted above, the objects implicated in visual attention are usually defined in terms of mechanisms of perceptual grouping. Thus, perceptual grouping is thought to be a prerequisite, logically and temporally, to attentional processing.

The view that perceptual grouping is pre-attentive is being challenged. Mack et al. (1992) report a series of studies in which subjects were required to focus on a cross in the middle of a patterned display and report on the length of the arms of the cross. As a secondary task, subjects were required to report on the perceptual grouping in the patterned display. The background patterns consisted of small squares organized either into rows or columns. This was achieved using grouping by proximity or grouping by colour. Mack et al. found that when participants were required to focus on the cross in the centre of the display, their ability to identify the grouping pattern, and even to detect any grouping at all, decreased significantly. The cross and the background pattern occupied the same location in the display, in that the cross clearly overlaid the background pattern. The displays had an average diameter of approximately 7.6°, while the visual angle of the squares in the pattern was typically 0.37° by 0.37°, and the spatial separation between the squares tended to be in the 0.3° to 0.9° range. Thus, it cannot be argued that the patterns were so small that participants were unable to see them, nor that they were so large that participants were unable to visually integrate the pattern due to drop-off of visual acuity. What these results suggest is that it is not enough for a grouping pattern to be foveated, but rather that if a pattern is not specifically attended, it will not be identified or even detected.

One might argue that grouping patterns are identified pre-attentively and then attentionally filtered out by focusing on a particular aspect of a display. Ben-Av et al. (1992) have looked into whether perceptual grouping requires attentional resources, or whether these processes are inhibited by attentional mechanisms. They employed a dual-task paradigm, in which participants were required to carry out a grouping pattern identification and/or detection task concurrently with a singularity detection task. The grouping patterns were based on grouping by proximity and grouping by similarity of shape. The singularity detection task involved detecting the presence of a single predetermined shape which did not match the shapes used in the grouping patterns. One task was always designated as primary, the other as secondary. The relative performance on each task was measured as a function of masking at different stimulus-onset asynchronies (SOAs) and of which task was designated as primary. Ben-Av et al. found that at long SOAs, participants reached optimal performance on both concurrent tasks, regardless of which task was primary. However, when singularity detection was the primary task, participants' performance on the grouping tasks was significantly reduced at short SOAs. Conversely, when one of the grouping tasks was primary, performance on neither task suffered significantly at short SOAs. Thus, the results of the experiments where the singularity detection was the primary task replicate the findings of Mack et al. (1992), whereas the experiments where the grouping tasks were primary suggest that focusing attention on one aspect of a display is not enough to interfere with the performance of another. Focusing on the grouping pattern did not interfere with the task of detecting a singularity, but focusing on the singularity interfered with the detection of a grouping pattern. These findings undermine the suggestion that

perceptual grouping happens pre-attentively and is simply filtered out by focusing attention on the similarity, for, by the same reasoning, the singularity should also be filtered out when the grouping pattern is focused. The more likely explanation is that perceptual grouping does in fact require visual attention, whereas singularity detection doesn't, and might even happen preattentively. Thus, deliberately focusing attention on the singularity starves perceptual grouping of the attentional resources it needs, whereas focusing on the grouping pattern does not interfere with singularity detection, as this task doesn't require attention.

What, if anything, is perceived without attention? Rock et al. (1992) examined this question. They used a method that was similar to that used in Mack et al. (1992). The main task was again to determine which arm of a cross was longer. A number of secondary tasks were used, including the detection of a single stimulus in the background, counting the number of elements presented in the background, and identifying some attribute of a single element, such as its shape or colour. Rock et al. found that when subjects were focusing on the cross (i.e., the primary task) and were exposed to the background stimulus for the first time, they were able report the number of elements in the stimulus, as well as report their colour. However, they were unable to identify the shape of the elements that were presented. This suggests that a few basic features of visual stimuli, such as their existence, location, and colour, are processed preattentively, but that the perception of shape requires attention.

As noted above, much of the research on object-based attention relies on the assumption that perceptual grouping is a pre-requisite for visual attention. There is, however, much evidence that undermines this assumption. It should be noted that there is another influential view on object-based attention that holds that visual attention is required for perceptual wholes to be formed, not the other way around. Perhaps the best exponent of this view in recent times is Treisman (1980, 1988, 1998). In brief, Treisman suggests that the early visual system analyses stimuli into basic features, (e.g., shape, colour, etc.), and that attention is required to bind these features into a single percept at a specific location in space (but see Goldsmith, 1998, for an account of object-based feature integration). Here, attention is viewed as a kind of spatial map that integrates information from different feature maps by binding features that share the same location in visual space. Supporting evidence for this theory comes from visual search experiments, where a single element in a display is found much more quickly, and in a sense "pops out," when it is distinguished from all the other elements by having a unique feature (e.g., it is the only element with a given colour or shape) than when it is distinguished by a unique combination of features which are found elsewhere in the display (e.g., many of the elements are crosses and many of them are red, but the element to be detected is the only red cross in the display). Treisman's interpretation of these findings is that when an element is distinguished by a unique feature, it is easily detected because it is the only element which is associated with the "map" for that feature. In contrast, when an element is distinguished by a combination of features, it is not uniquely associated with any given feature map, and thus it can only be distinguished by the visual system by fusing the information from various feature maps into a perceptual whole at a particular location through visual attention.

On Treisman's (1980, 1988, 1998) interpretation, visual attention is mainly a spatial integrator, which does little to explain findings showing that it is easier to attend to a single object than to divide attention between many. However, Treisman's theory, along with the other research discussed in this section, remind us that the role of attention in perception cannot be ignored or underestimated. Likely, attention and perception are mutually dependent, at least to some degree. That is, visual perception and attention are likely interactive and concurrent

processes. On this view, perceptual grouping is not pre-attentive. Rather, the cost of dividing attention between objects might reflect the cost of organizing a scene into objects and groups instead of, or in addition to, any costs due to switching attention between objects. The following section examines evidence that attention and perceptual grouping occur interactively, and discusses possible implications for an integrated theory of perceptual grouping and object-based attention.

3. TOWARDS AN INTERACTIVE, INCREMENTAL AND INFERENTIAL MODEL OF PERCEPTION AND ATTENTION

Research on object-based attention is generally based on the following assumptions: (1) the relation of perceptual organization to attention is serial, in that visual stimuli are first organized into groups, then the grouped stimuli are selected by the attentional system for further processing; and (2) perceptual organization can be characterized in terms of Gestalt-type principles. As noted earlier, these assumptions are inadequate. Gestalt grouping principles are unable to explain and predict actual grouping performance in all but the simplest cases, and attention appears to be as involved in perceptual grouping and grouping itself is involved in attention. On the Inferential Attentional Allocation Model (IAAM) view, it is argued that the following claims are more appropriate base assumptions for the study of object-based attention:

1.) Attention and perceptual organization are interactive and incremental processes. As each process progresses, it provides more information to the other process, which in turn further aids the first process in mutually supporting feedback loop.
2.) Perceptual organization is an inferential process, whereby the visual system attempts to (re)construct which three-dimensional coherent distal structures that most likely gave rise to the proximal stimulus. Under this view, perceptual grouping, and vision in general, is a type of inference to the best explanation (Hoffman, 1998; Palmer, 1999; Pomerantz & Kubovy, 1986).
3.) These "explanations," or representation of physical objects, both guide attention and are themselves partly influenced by attentional mechanisms.
4.) ***It follows, then, that that attention is object-based precisely because it has a relationship of mutual dependence with the processes of perceptual organization*.**

These claims reflect current opinion in visual object perception, from the domains of computational vision (Lowe, 1985), the psychology and psychophysics of visual perception (Hoffman, 1998; Palmer, 1999), and neuroscience (Grossberg, 1994; Grossberg, Mingolla, & Ross, 1994). Yet they have been largely ignored in object-based attention. Nevertheless, if these claims were integrated into theories of object-based attention, they would allow these theories to meet many of the criticisms that have been raised against them. The goal of the proposed research is to show how these claims about visual attention and perception can be used to further develop and strengthen object-based accounts of visual attention.

To flesh out these claims, two areas of research need to be discussed. First, research in the neurology of vision has yielded models of integrated perceptual organization and attentional processes. However, these models only briefly touch on the processes of perceptual grouping themselves. Therefore, a review of current research on perceptual grouping, which shows that vision is largely an inferential process, is also required.

3.1 An interactive model of attention and perceptual grouping

Grossberg and his colleagues (1994; Grossberg & al., 1994) have synthesized their research on the neurology of visual perception, visual attention, and search into a model called the "Spatial and Object Search Algorithm" (or SOS algorithm), depicted in Figure 6. The model consists of five modules. There are two attentional modules, an "Object Recognition System," which is meant to "explicate [...] concepts of object attention," and a "Multiplexed Spatial Map," which "explicates concepts of spatial attention" (Grossberg et. al, 1994, p. 476). The other three

modules are perceptual: a Static Boundary Contour System which extracts edges between

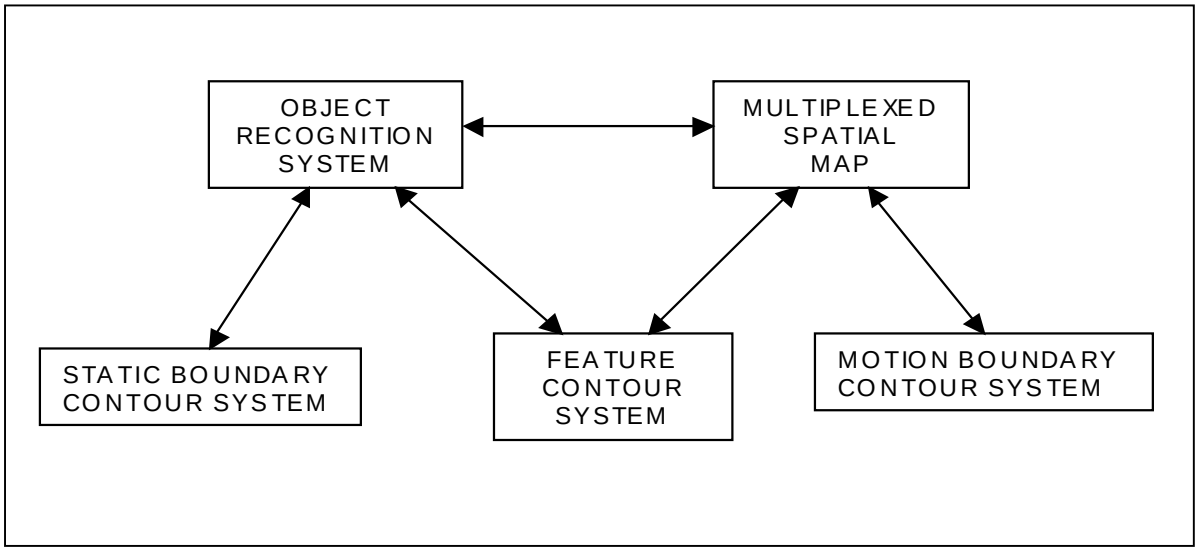

Figure 6: Components of the SOS system (adapted from Grossberg & al., 1994)

regions based on the static properties of the retinal image, a Feature Contour System which detects surfaces with a uniform attribute or conjunction of attributes (e.g., colour, texture, etc.), and a Motion Boundary Contour System which detects region boundaries (i.e., edges) from dynamic properties of the retinal image such as optic flow (Gibson, 1979). The system is highly interactive. The static Boundary Contour System and the Feature Contour System interact with the Object Recognition system to extract static information about objects. The Feature Contour System and the motion Boundary Contour System interact with the Multiplexed Spatial Map to extract information about object motion. The Object Recognition System and the Multiplexed Spatial Map also interact, thus defining "spatially invariant ORS [Object Recognition System] recognition categories" (Grossberg et al., p. 477 figure caption) and controlling "visual search" (Grossberg et al., p. 477 figure caption). Through these interactions, the model accomplishes four logical steps: (1) a retinotopic map of stimulus features is established and provides input to the perceptual organization/search system; (2) the retinotopic map is segmented into groups by extracting boundaries (with the Boundary Contour Systems) and surfaces (with the Feature Contour System); (3) one of the groups defined in step (2) is selected for further processing using the two attentional mechanisms; the factors influencing which group gets selected include both stimulus saliency (bottom-up information) and certain target features that are being searched for (top-down information); (4) the group selected in step (3) is compared to stored representations of objects in order to determine whether the group is a recognized object and/or an object containing the target features. Failures at steps (2), (3) or (4) make the system revert to the previous step to continue the process. This model defines visual search as a process of recursively forming and selecting groups until the target feature is found, or until all possibilities are exhausted. Note, however, that this algorithm defines a logical sequence only, and does not imply a temporal sequence for information processing in the SOS model.

The SOS model (Grossberg, 1994; Grossberg & al., 1994) has a number of interesting and desirable characteristics. Importantly, it integrates object-based and spatial attentional processes. Much evidence (see Logan, 1996 and Lavie & Driver, 1996, among others) suggests

that object-based and spatial attentional factors, while distinct, are intimately related. The SOS model also shows how attention and perceptual grouping might interact in a parallel and continuous manner. The authors in fact characterize the model as a "heterarchy of neural networks with continuous and asynchronous dynamics," implying that no "module" takes precedence over any other (Grossberg & al., p.478). This reconciles the findings of Rock et al. (1992), Mack et al. (1992) and Ben-Av et al. (1992), that perceptual grouping cannot happen without attention, with object-based attention studies showing that visual attention is sensitive to perceptual organization (e.g., see Duncan, 1984; Lavie & Driver, 1996; Triesman et al., 1983). Another important feature of the SOS model is that it shows how many different types of information are simultaneously required in the processes of visual attention and perception. As we wil see later, visual perception makes use of many types of information, ranging from bottom-up to top-down factors. The SOS model identifies four broad types of information: edge and boundary information, surface property information, information from motion, and "top-down," or conceptual, information. We shall see that even computational systems that only perform perceptual grouping require information at least from these various levels. Furthermore, the SOS model is organized as a number of neural networks working asynchronously and in parallel. The neural networks use an algorithm based on adaptive resonance theory (Carpenter & Grossberg, 1991, cited in Grossberg & al., 1994), which is closer to the real dynamics of the cortex than the backpropagation algorithms still being used in many connectionist systems.

Perhaps most interestingly, the SOS model is consonant with a model of perceptual organization developed independently by Palmer and Rock (1994). Their model is based on psychological and psychophysical evidence, as well as considerations on the "logic" of perception, without the benefit of data from neuroscience. Nevertheless, the Palmer and Rock (1994) model shares certain features with Grossberg's SOS model. The Palmer and Rock model (see Figure 7) has three stages. First, boundaries between elements are determined from the retinal image. Second, regions with uniform features are formed from boundary information and

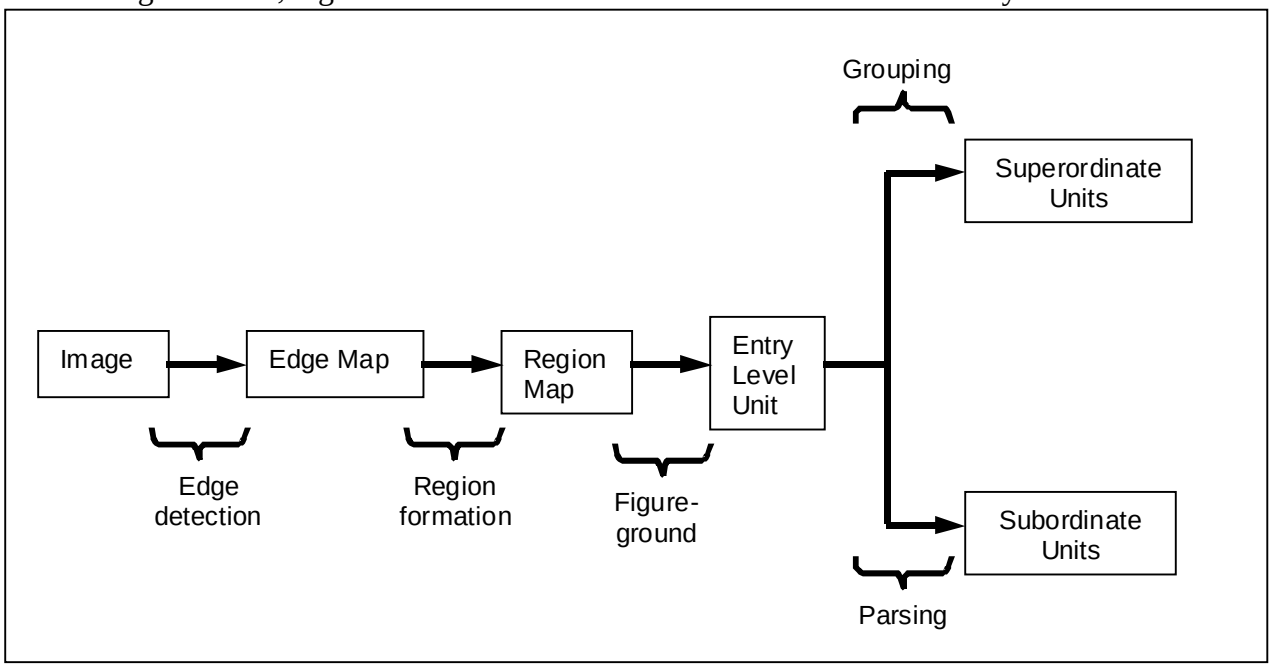

Figure 7: Flowchart for perceptual organization, from Palmer and Rock (1994, p. 42)

surface properties from the retinal image. Finally, the regions with common properties are separated into figure and ground, and the figure regions serve as the "entry level" units for perceptual organization proper, which consists in grouping and parsing[2] the regions of the image based on various types of information (i.e., region properties and top-down information). Note that the Palmer and Rock model leaves out attentive processes. However, the work of Grossberg (1994; Grossberg & al., 1994), Ben-Av et al. (1992), Rock et al. (1992), and Mack et al. (1992) suggests that both attention and perceptual grouping occur at the level of processing. Furthermore, Yantis, writes that "between early and high-level vision is intermediate or middle-level vision. Among
the operations at this level of the visual system are the processes of perceptual organization and attention" (1998, p. 187). By assuming that attention and perceptual organization occur at roughly the same level of visual processing, the Palmer and Rock model can be mapped onto the SOS model with a fairly good fit despite the omission of attentional processes. The first "box" in the Palmer and Rock flowchart, corresponds to Step 1 in the SOS algorithm, the second, third, and fourth boxes in the Palmer and Rock model (edge map, region map, entry level units) correspond to Step 2 of the SOS algorithm. The perceptual organization boxes, with the inferred attentional processes, would correspond to Step 3. Whatever follows these last boxes (i.e., object recognition and identification) would be Step 4 in the SOS algorithm. While it is unclear whether the Palmer and Rock flowchart is meant to be a temporal model or a logical one, it is nevertheless noteworthy that by starting purely from "logical and empirical constraints" (Palmer & Rock, 1994, p.42), Palmer and Rock found essentially the same "modules" of visual processing as did Grossberg et al., (1994), starting from neuroscientific evidence.

Finally, research closer to the domain of attention also argues for a incremental and interactive conception of perception and attention. Di Lollo, Enns, and Rensink (2000) argue that re-entrant neural connections (i.e., feedback between the various stages of visual processing) are necessary to account for the timecourse of various masking phenomena in which a figure that shares a location with another figure without occluding it can nevertheless obliterate the perception of the second figure, as long as the display of the first figure temporally overlaps and extends beyond the display of the second figure. This indicates that a later image can "erase" an earlier image from iconic memory before it can be fully processed as a figure in its own right. This can only happen if perceptual and attentive processes can mutually affect each other as well as affecting "earlier" visual processing more closely related to the retinal image.

In sum, evidence from visual perception shows that attention and perceptual grouping occur in parallel and interactive ways. This goes against the prevailing assumptions in object-based research that perceptual grouping is pre-attentive, and that it is a prerequisite for object-based attention. These findings do not invalidate the notion of object-based attention. Rather, they support the claim made above that attention is object-based because of its relationship of mutual dependence with perceptual organization. This claim is the main assumption underlying the proposed research. Thus, an understanding of the mechanisms of perceptual organization will lead to a better understanding of the phenomenon of object-based attention.

---

[2]The focus in research on object-based attention has mainly been on grouping. However, it stands to reason that parsing could also play a role in "object formation," and thus in the deployment of attention.

3.2 Visual inference

As noted previously, Gestalt psychologists attempted to characterize perceptual organization purely in terms of certain physical characteristics of visual stimuli, on the assumption that the human brain was particularly attuned to these characteristics. However, these characteristics were defined by the vague and ultimately subjective notion of figural goodness or *Prägnanz* (Koffka, 1935). The attempt to explain the objective causes of the phenomenal experience of perceptual grouping lead back in a vicious circle to the very phenomenology it hoped to account for.

The fact that the Gestalt psychologists encountered this problem points to the central difficulty in the study of vision, and perception in general. Namely, that vision is neither a purely objective nor a purely subjective phenomenon. Attempts to characterize vision in purely subjective terms (e.g., the structuralist approach of Wilhelm Wundt, which attempted to introspect the "sensory atoms" that were alleged to make up perceptions; see Palmer, 1999) have failed, as have attempts to characterize vision in purely objective, "organism-free" terms (e.g., the ecological optics of J.J. Gibson).

The most successful approaches to vision are constructivist approaches, which recognize that successful visual perception depends both on properties of the environment and properties of the visual system, including information gained through experience and learning. All of these approaches are in some way indebted to Helmholtz's insight that vision is largely a process of "unconscious inference" (Palmer, 1999). On this view, the task of vision is to infer objects in the environment and their characteristics from the image projected onto the retina. Furthermore, we generally have no conscious access of the inferential process and the information it uses (Palmer, 1999). An example of this in everyday life is the perception of whole objects based only on a partial view of them. When you see only two or three sides of a cardboard box, you automatically perceive it as a whole box, and not as just the three flaps of cardboard that are visible to you.

However, the 2-D image available to the visual system is ambiguous, in that a particular 2-D image corresponds to any number of possible 3-D environments. This is known as the "inverse problem" in vision: how does the visual system "reverse" the projection of a 3-D environment onto the 2-D plane of the retinal image so as to "reconstruct" the original (distal) 3-D stimulus, or something like it? For instance, why is it that we will perceive a whole box every time, and not just three flaps of cardboard? The constructivist approach suggests that, due to some combination of innate abilities and experience, we quickly learn to "infer" a complete box from the three flaps. That is, our visual system "explains" the retinal image consisting of three flaps as being due to the existence of a complete box.

This approach to vision has become the de facto standard in vision research, both when it comes to studying natural vision systems (see Hoffman, 1998; Palmer, 1999; Pomerantz and Kubovy, 1986) and computer vision (see Kellman & Shipley, 1991; Lowe, 1985; Marr, 1982). Nevertheless, most research on object-based attention has ignored this approach to visual perception. Instead, it refers to Gestalt grouping principles when describing and explaining the perceptual units that attention is supposed to operate on. Object-based attention research needs to integrate the constructivist approach to perceptual organization. The first step is to get a grasp on the constructivist approach. On the view that perception is inference, the relevant literature can be reviewed in terms of the following questions:

- What is inference, and what type of inference does the visual system use?
- What are the premises (i.e., the "raw materials") of visual inference?

- What rules does visual inference use to get from its premises to a reconstruction of the world?
- What are the characteristics of a valid visual inference?
- What are the conclusions the visual system arrives at — fully integrated objects, a whole scene, etc.?

3.2.1 The nature of visual inference

On the IAAM view, attention and perceptual organization are mutually dependent. Attentional allocation should in some way reflect the process of perceptual organization. To see how this might be the case, the inferential nature of perceptual organization must first be explained.

Inference is generally subdivided into three general areas: deduction, induction and abduction (Peirce, 1931-1958). In deduction, the soundness of a conclusion always follows strictly from the truth of the premises and the form of the whole argument. The classic example is "All men are mortal; Socrates is a man. Therefore, Socrates is mortal." As long as the premises are true, and they follow certain forms that are known to be valid, the conclusion of a syllogism is guaranteed to be true. We can see that this model is unlikely to be appropriate for visual inference. The "inverse problem" in vision discussed earlier shows that visual stimuli are ambiguous. Thus, the premises that the visual system receives from the environment do not provide enough information for an interpretation of a visual scene to be logically deduced.

The term "induction" is often used to refer to all non-deductive reasoning. However, strictly speaking, induction is a type of reasoning where properties of a subset are generalized to the whole set in a probabilistic manner. An example is a poll: from a small sample of the population, it is estimated with a certain degree of certainty (typically 95%, the "nineteen times out of twenty" we hear of so often) that a given fraction of the population as a whole backs a certain political party, and that another fraction supports another party, and so on. In effect, this type of reasoning says "Based on the subset of the population we've observed, we're 95% certain that 30% of the population supports party X; but there's a 5% chance we're wrong." This seems a more realistic account of visual processing, given the ambiguous and incomplete nature of visual stimuli. A particular type of statistical inference that has been used with some success in the study of perceptual organization is Bayesian inference, in which predictions about the environment are made on the basis of conditional probabilities — i.e., by determining the likelihood that such-and-such is the case when a given condition is fulfilled. An example of this would be a system that determines with a likelihood of 87% that a given configuration of 2-D elements indicates the presence of a tabletop given that it is known the visual system is in fact looking at some kind of furniture. Bayesian inference has proven to be a powerful tool in probabilistic decision making (many texts present a Bayesian approach to decision-making; an example is Lindley, 1985) and seems like a promising approach for certain visual processes (see Lee, 1995, among others).

However, induction has two important limitations. One is that is requires quantitative information in the form of sample statistics and numerical probabilities. It is not clear that vision, or cognition in general, is a quantitative process in this way. Furthermore, generalizations from statistical induction do not go much beyond initial data. For instance, induction can be used to infer the likelihood that the population mean falls within certain limits, or the likelihood that the sample mean is different from another sample mean, but not much else. This is not an appropriate model to explain how three-dimensional volumes are reconstructed from partial

information, as when we perceive a complete box from only three cardboard flaps. A more extreme case is shown by our ability to construct illusory (and sometimes impossible) three-dimensional shapes from black-and-white line figures, as seen in Figure 8.

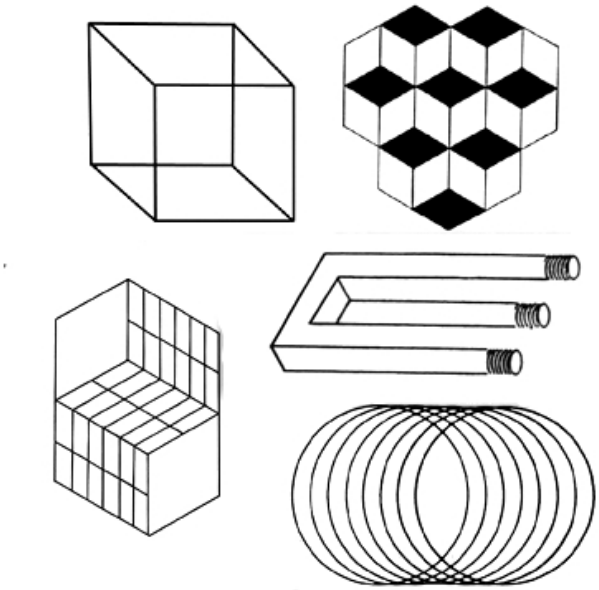

Figure 8: Line-drawing figures that elicit the impression of 3-D structures (IllusionWorks, 1997)

The third type of inference is generally known as abduction, a term originally coined by C. S. Peirce (1931-1958). This form of inference is a catch-all category, grouping any method of inference that doesn't fit the patterns for deduction and statistical induction. This includes reasoning using less formal methods such as analogies and making creative leaps from data to theory (Holyoak & Thagard, 1995), as well as spontaneously generating hypotheses in order to explained observed facts through what might be called "educated guesses". Indeed, abduction is generally considered to be synonymous with inference to the best explanation, in which inferences are motivated by a desire to provide the best possible explanation for a phenomenon (Lipton, 1991). This terminology is in line with the way vision is characterized by many researchers (Hoffman, 1998; Leyton, 1994; Palmer, 1999). Intuitively, it seems more plausible that the visual system infers three-dimensional volumes from the images in Figure 8 because such volumes would best explain the shapes, rather than because those volumes are the structures most plausibly inferred using statistical induction. Inference to the best explanation also seems to be the best model for the way humans actually reason, because it doesn't suffer many of the shortcomings associated with deduction and induction (Lipton, 1991).

One of the central claims of the present research is that perceptual organization is a form of inference to the best explanation. But, to justify this claim, inference to the best explanation must first be characterized. Abductive reasoning happens to be the most poorly characterized type of reasoning. Lipton (1991) points out that while the term "Inference to the best explanation" has intuitive appeal, it has not so far been adequately defined. It is not enough to say that perception infers states of the world that would best explain the proximal stimulus. An account of how these explanations are generated, and of what counts as the "best explanation" is required. Vision scientists have uncovered a number of rules that govern visual inference (Hoffman, 1998; Palmer, 1999). These rules might provide a basis for a systematic account of how possible explanations are generated. These rules are examined in section 3.2.3. What counts

as a valid explanation is discussed in section 3.2.4. Nevertheless, if inference to the best explanation is to be used as a model for visual inference, then an account of inference to the best explanation is necessary. Furthermore, the implications of characterizing vision as inference to the best explanation also need to be examined. Attempts to build computerized abductive reasoning systems have lead to the conclusion that abduction is generally intractable computationally (Selman & Levesque, 1996). However, if the visual system is performing inference to the best explanation, this type of inference must somehow be tractable. Further considerations about the visual system might elucidate how perception as a type of inference to the best explanation might be computationally tractable.

In summary, out of the three forms of inference discussed above (deduction, induction and abduction), abduction best characterizes the inferential nature of perceptual organization. However, even this characterization leaves many questions about perceptual organization unanswered. The following sections attempt to answer some of these questions. Nevertheless, we can already note the advantages that this approach provides over the typical assumptions that models of object-based attention are based on, namely that grouping is pre-attentive and is governed by the Gestalt grouping principles. If perception organization is based on inference to the best explanation, and the best explanation of most perceptual groupings can be given in terms of the physical stimuli that the groupings are meant to reconstruct, then we have an objective basis for characterizing what counts as an object and for developing metrics of perceptual organization. Furthermore, we know intuitively that certain explanations are more difficult to provide than others. A characterization of perceptual organization as abduction should therefore provide a natural account of the processing cost involved in organizing a visual scene into objects and groups. I will argue below that the processing cost of perceptual organization is an important index of attentional allocation in a scene.

3.2.2 The premises of visual inference

It is well known that reasoning, both scientific and mundane, starts from at least two kinds of premises: observational facts and conceptual-theoretical postulates. Can this aspect of inference be extended to perceptual organization? As mentioned above, vision seems to require both bottom-up, stimulus driven information (corresponding to facts) and top-down, conceptual inferences (corresponding to concepts and theories). In the following the nature of the bottom-up "premises" of, as well as the role of general-purpose conceptual information in, perceptual organization are discussed. The following section examines a particular type of top-down influence on vision, viz. the rules of visual inference.

It is often assumed that the starting point, the "raw data," as it were, of visual inference and perceptual organization are lines and edges detected by specialized cells in the very early stages of the visual system, mainly the retina and the subcortical visual system (Kosslyn, 1994). Many computer vision systems are based on this principle, in that they first attempt to construct a line-and-edge version of a scene before attempting to organize it and parse it into individual objects (cf. Biederman, 1995; Marr, 1982).

However, Palmer and Rock (1994) argue that regions of uniform connectedness might be more natural starting points for visual inference (see also Palmer 1999). This principle of perceptual organization (but not, as the authors stress, of perceptual grouping) states that regions of an image which share a common property, such as texture or colour, will tend to be perceived as constituting a unified element in a display. Any resemblance to the Gestalt principle of grouping by similarity is only superficial; whereas the principle of similarity claims to account

for why certain elements are perceived as belonging together, the principle of uniform connectedness makes a claim about how these elements are formed in the visual system in the first place. In addition to experimental evidence, support for this notion is adduced from the fact that computational models for extracting lines and edges from a scene often produce spurious shape boundaries, whereas algorithms that segment images based on regions with common attributes generally produce more accurate segmentation and thus provide a better starting point for perceptual organization (see Palmer, 1999, pp. 269-273, for a comparison of these approaches.) Accordingly, as noted above, Palmer and Rock (1994) propose a model of perceptual organization wherein organization (i.e., grouping and parsing) do not occur until basic segmentation of the image, based on the principle of uniform connectedness, has occurred. Once this has happened, they argue, the visual system can use this information as a basis for grouping and parsing the elements in the scene, as well as inferring the existence of certain objects.

Given that visual processing is to a large degree dedicated to inferring which elements belong together and constitute objects, or at least coherent wholes, it should also be expected that motion would play an important role in visual inference. Accordingly, research by Spelke, Gutheil and Van de Walle (1995) suggests that children first start individuating elements in a display on the basis of motion information. Children seem to very quickly develop the expectation that two elements, even if they appear to belong to different objects, constitute two parts of the same objects if they move in unison and are partly occluded in a way that suggests that a physical connection between the parts is hidden behind the occluder. Also, children expect an object which disappears behind an occluder to reappear on the other side with motion that is consistent with the original motion of the element. Spelke et al.'s experiments suggest that if the element re-appears with a trajectory that is not consistent the original trajectory (e.g., an element disappearing behind the occluder from the left reappears on the top edge of the occluder and moves upwards), the child reacts to it as if it were a new object. Similarly, Yantis (1998) has shown that apparent motion can induce observers into grouping elements into specific patterns, which change as a function of the speed of the apparent motion. These data suggest a natural and logical extension of the principles of uniform connectedness and similarity to motion. Indeed, the Gestalt principle of common fate, which states that elements sharing temporal dynamics will be grouped together, is essentially a temporal version of grouping by similarity. However, while in principle motion cues and static information should both be equally important for perceptual organization, in many cases, as in the Spelke et al. (1995) study, motion information seems to have the ability to override static information. Thus, it is possible that certain types of information are more important for visual inference.

Finally, it would be a mistake to ignore the role of conceptual, top-down knowledge in visual inference. As noted above, the "inverse problem" of vision tells us that the proximal stimulus alone is insufficient for a unique reconstruction of the distal stimulus. Thus, some information, other than raw sensory information, is required to constrain possible visual inferences from retinal stimulation. We can see this in extreme cases such as figure, a classic image by R. C. James (Figure 9). In it, most people initially see a number of black splotches on a white background, perhaps representing leaves lying on the ground. Once people are told that the image contains a dog, they tend to notice the dog rather quickly and will generally be able to pick it out every time they see the image again. In other words, the knowledge that there is a dog in the picture allows an observer to construct a unique interpretation of the image that

Figure 9: Can you find the dog?

would have been impossible otherwise.

Top-down information is essential for vision in general, not just for the perception of ambiguous displays. If, as it has been suggested, vision is a form of inference to the best explanation, then top-down knowledge is absolutely essential. As note above, inference to the best explanation is considered to be computationally intractable. This is largely due to the fact that the task of selecting an appropriate set of initial assumptions, which is termed a "support set," to guide the generation of possible explanations is itself computationally intractable. Conceptual information might provide such a support set, thus potentially making inference to the best explanation a computationally tractable model for vision.

Much work clearly needs to be done on what kinds of top-down information are used by the visual system, and how that information is used. However, even at this early stage a distinction can be made between abstract, context-free information on the one hand, and concrete, context dependent information on the other. The image of the dog in Figure 9 is an example of the latter. An example of the former type of information might be a very general category, such as surfaces, edges, or even objects themselves. It might be argued that edges and surfaces are already provided by, or at the very least extracted directly from, the proximal stimulus, as suggested above. However, Feldman (1999) argues that certain primitive assumptions, what he calls "existential axioms," are required for an inferential approach to perceptual organization to work. Without these, the problem of assigning a unique (or almost unique) grouping to a set of initial elements cannot be adequately constrained, which is simply another form of the "inverse problem." Thus, it might be the case that the visual system starts with certain innate, very general categories that are used to constrain visual inference, and that these categories are supplemented with others that are learnt through experience as the individual develops. The problem of innate knowledge versus experience notwithstanding, some type of top-down knowledge is needed to constrain the "inverse problem."

In sum, the visual system bases its inferences about what is in the environment on many different types of information: lines, edges, regions of uniform connectedness, motion, general abstract categories and past experience. Top-down information is required to constrain possible interpretations of bottom-up information, in line with the constructivist view described above. Interestingly, these findings are in agreement with those of Grossberg et al. (1994). Visual perceptual organization, and vision in general, relies on the interaction of many different types of information. The next step is to determine how these types of information are in applied by the visual systems to "draw conclusions" about what we see and to allocate attention.

3.2.3 The rules of visual inference

Inference involves the derivation of conclusions from premises according to specified rules. Rules are crucial they determine and provide methods for the derivation of valid conclusions. The relationships between the rules are also important. In situations where more than one rule might apply, there must be some systematic way of deciding which of the rules will apply. There must be rules on how the rules apply.

A necessary step in studying vision as inference is the study of the rules of visual inference. Indeed, much research is being carried out in this direction, as can be seen in the excellent reviews by Palmer (1999) and Hoffman (1998). While researchers have described a relatively large number of rules (Hoffman, 1998, lists 35 of them throughout his book and suggests that this is only a fraction of the total), there appears to be an overall pattern to these rules. This pattern is suggested by the term "inference to the best explanation." Most of the rules

of vision seem to be rules for determining the 3-D structure that best explains the proximal (retinal) stimulus. What counts as the "best" explanation has yet to be clearly determined (Lipton, 1991). However, in the case of vision, is seems that the "best" explanation is determined by a few general principles, namely the genericity principle (Albert & Hoffman, 1995; Feldman, 1999) and the principle that certain basic regularities govern the physical world around us (Hoffman, 1998).

The genericity principle was proposed by Albert & Hoffman (1995). This principle states that the 3-D structure that best explains the proximal stimulus is one that is as generic as possible. For instance, a straight line in the retinal image could be the result of the projection of a straight line in 3-D space onto the retina, or it could be the result of the projection of some flat surface – say a disk – being viewed edgewise. In the first case, no change in viewing angle will change the perception that the line is in fact a line. In the second case, a very minor tilt in the angle of the disk would very quickly reveal that what looked like a line was in fact a disk, not a line. The disk produced the perception of a line only under very specific circumstances, whereas the line produces the perception of a line under almost all conditions. That is, it produces the perception of a line generically. Hoffman (1998) therefore gives a number of rules of visual inference that are applications of the genericity principle, such as

- Always interpret a straight line in an image as a straight line in 3D
- If the tips of two lines coincide in an image, then always interpret them as coinciding in 3D

A number of rules also seem to encode some knowledge about the basic regularities of our physical environment (Hoffman, 1998). In many cases, these rules can be subsumed under the likelihood principle (Pomerantz & Kubovy, 1986), which is simply the idea that perceptual organization should reflect likely states of affairs in nature. This is similar to the genericity principle, although the two principles can conflict, as Hoffman has shown (1998)[3]. Nevertheless, likelihood is often a very good explanation for why certain things might be perceived as being organized in a certain way. For instance, physical objects tend to be relatively coherent and solid masses, and thus when a number of elements move in unison, it is natural that they be perceived as part of the same rigid object. This is the intuition behind the Gestalt principle of common fate. In fact, it has been shown that most, if not all, of the principles of perceptual grouping as described by Gestalt psychologists can often be re-interpreted in terms of the likelihood principle rather than the principle of *Prägnanz*, often with more credible results (Pomerantz & Kubovy, 1986). Also, basic knowledge about gravity and the position of natural light sources seems to play a role in visual inference. Elements in a visual scene are rarely perceived as floating in mid-air (unless the elements are balloons or other objects which are naturally found in the skies), even if this requires the visual system to go against the rules of genericity (see Hoffman, 1998, p.30, for an example). Similarly, the fact that natural light sources are almost invariably above the observer's head leads the visual system to interpret shading in specific ways, to the point of re-organizing an image to make the shading consistent with an elevated source of light. The visual system also seems to make use of information about 3-D structure from depth cues such as texture gradients, perspective, and stereo cues, among others (Hoffman, 1998; Palmer, 1999). Thus, it would seem that visual inference is guided to a large degree by vast stores of (mostly

[3] Both the likelihood and genericity principles have, on the face of it, clear adaptive value, and are likely the result of evolutionary pressures on the primate visual system. Such considerations, however, are beyond the scope of this document.

tacit) knowledge about the physical properties of the world.

In contrast to the principle of *Prägnanz*, the likelihood principle potentially provides a way of obtaining objective criteria for which configurations of two-dimensional elements will be perceived as objects. Objects are perceived when there exists a most likely or most generic three-dimensional interpretation of the proximal stimulus. Furthermore, such an interpretation can be matched against a real physical structure for validation, unlike Gestalt interpretations which can only be matched against the interpretations of other observers. A complete validation of the likelihood and genericity principles would require a full "ecological survey" to match real-world 3-D structures to the 2-D retinal images to which they most reliably give rise, as Pomerantz and Kubovy (1986) suggest. Nevertheless, the "likelihood approach" might provide a solid starting point for determining how simpler, man- or computer-generated images are organized and perceived in terms of objects. We shall return to this matter later on.

As we have seen, it appears that not all "visual rules" are equal. Rules about how gravity affects objects can override rules about the genericity of an image. Similarly, motion appears to be able to override static, figural grouping factors (Driver & Baylis, 1989). It is therefore necessary to study how the rules themselves are organized and applied. Specifically, it must be determined whether there is a hierarchy for the rules of vision, whether this hierarchy is flexible or rigid, and how this hierarchy is enforced. Some insights might be gained form computational models of rule-governed systems. For instance, the typical production system uses a number of methods for ensuring that two conflicting rules do not fire at once (e.g., rules are applied in a particular sequence) or that the conflict is resolved according to some other rule (e.g., rules about motion might always supplant other rules in a computer vision system). However, much work is still required in this area. The rules we have for vision are still tentative. We have a large number of rules which likely interact in ways we still don't understand, so a production system is likely not a good model for vision, as is probably true of any system based on a finite set of explicit rules, for that matter. Furthermore, there still isn't an explicit, operational definition of what it is that visual inference produces as its conclusions. In other words, we still don't know how to define what an object is.

3.2.4 Validity: What is an object?

The issue of the validity of the inferences of the visual system has already been discussed in part in the sections on inference — if valid rules are applied in a valid manner, the conclusion should be valid. However, this only covers the internal consistency of the inferences. Vision must have external validity. That is, the visual system must produce a mostly veridical perception of the environment, which allows an organism to survive in and negotiate its environment.

The matter of veridical perception is not simple. It is not the case that vision is veridical simply when there is a one-to-one mapping between an "object" inferred by the visual system and an object in the environment. First of all, it is not clear that the visual system produces representations of objects as output. Second, it is not clear what gets to count as an object, either in the visual system or in the environment.

What does the visual system output? This is in fact an ill-formed question, because vision has many functions. At lower levels, vision appears to perform such tasks as extracting edges and surfaces from a scene, texture segregation, colour identification, etc. At higher levels, it can be said that vision carries out object identification, object recognition, comparison of many objects, motion perception, scene organization and parsing, directing action towards objects in

the environment, and so on. Accordingly, it is probably more correct to say that the visual system produces many "outputs," some of which serve as inputs to further stages of visual processing, and some of which are used in other cognitive functions (e.g., planning movements). It might therefore make sense to restrict the issue of external validity to only one of the functions or "modules" of the visual system. Given that the main goal of this research is to study how perception and attention interact, and that the attentional model under consideration is the object-based attention model, it is reasonable to focus on the mechanisms of perceptual grouping.

Most object-based attention research assumes that attention interacts with perception at, or perhaps right after, the level of the grouping processes. However, there are two potential problems with this approach. First, there is no reason in principle to suppose that attention is referenced to perceptual groupings rather than phenomenal objects (i.e., parts of the retinal image that an observer actually recognized as an object). Second, the common account of what constitutes a perceptual grouping, based on the Gestalt principles, is itself inadequate and thus cannot help in determining what type of entity attention actually selects and operates on.

This is further reflected in the experimental work of Rock and his colleagues (Rock & al., 1992; Mack & al., 1992) and of Ben-Av et al. (1992) which was examined earlier. These experiments showed that perceptual grouping requires attention. However, perceptual organization clearly has an effect on visual attention, as is shown by the extensive research in object-based attention. The results reported by Rock and his colleagues and by Ben-Av et al. do not invalidate the basic conclusion of research on object-based attention, viz. that attention can be affected and directed by perceptual structure. Nevertheless, they underscore the difficulty of adequately defining the "units" of attention, particularly when these are defined using the Gestalt principles.

Part of the problem with defining perceptual groups or units is that it is not clear that perceptual organization happens only at one "level" of visual processing. Much research on object-based attention uses stimuli that are essentially geometric shapes, when their "objectness" depends mainly on their geometric properties, rather than conceptual properties. However, object-based attention might be referenced to objects defined by more conceptual information. For instance, in Figure 9, if attention is directed to Gestalt-type perceptual groups, or to groups defined only by bottom-up properties of the proximal stimulus, then perceiving the dog would not affect how attention is allocated in the image. However, it might that the set of spots that are perceived as a dog are attended to as a single group, allowing an observer to process the information from the set of spots more quickly. There appears to be no experimental evidence to help decide between these claims. Rather, it seems that many researchers have tacitly assumed that such high-level, conceptually-driven perceptual organization simply cannot play a role in the allocation of attention. Presumably, this type of perceptual organization would itself require attention. Again, there is no reason in principle to make this assumption.

Defining an object is a complex task, even when the definition is constrained to the domains of perception and attention. This might be because perceptual objects might not be "constructed" at a single stage of visual processing, but rather the perceptual representation of an object might involve many stages of processing. This is also complicated by the following facts: (1) many physical objects are complex wholes made up of many other elements, which, given the right context, could be taken as objects in their own right (e.g., a tire from a car); and (2) many of the "objects" of everyday experience are not physical objects *per se*, but rather images of objects on television and computer screens, movie screens, and the printed media. These have no three-dimensional properties, and their only physical attributes are luminance, chrominance,

contrast, and whatever geometric properties they have. They seem to have object status derivatively, in that they present our visual systems with stimuli that are very close to the stimuli produced by the physical objects they represent.

The inevitable conclusion is that "objecthood" cannot be defined in an absolute, context-independent manner (Smith, 1996). But this does not mean that there cannot be an "objective" definition of an object. We can take a cue from the etymology of the word "object": an object is an object for something, as in the phrase "she is the object of my desires." Thus, objecthood can be defined as a relational, but nevertheless real and objective, phenomenon. ***An object is something that has a relation to a subject. Objects are objects because they are manipulated, acted upon, and thought about, in certain ways, by a subject***.

An implication of the relational view of objecthood is that what is taken to be an object depends very much on what the subjects goals are. For instance, if you want to pick up a pop bottle, you take the whole bottle as an object – including its label, cap and contents. If you want to take the cap off, then you also take the cap to be an object. If you want to read the label, then you take the label to be an object, and so on. Any attempt to link perceptual representations to physical objects must take into account the context-dependent nature of objects. Just because a coherent blob of matter could be an object doesn't mean it is (see Smith, 1996). Rather, blobs of matter can be objects, but they don't really become objects until some agent interacts with them.

We might now be in a position to start outlining the concepts involved in "objecthood" a bit more clearly. Consider the following definitions:

Physical object. A coherent whole ("blob of matter"), which can be more or less rigid, that is being or could be acted upon or experienced by a subject (in operational terms, objects are task-dependent).

Virtual object. A two-dimensional image (computer screen, print, etc.) of a physical object; it has the same visual stimulus properties as physical objects (i.e., physical and virtual objects are indistinguishable at the retinal image). Virtual objects are also differentiated and individuated in terms of the subject's actions, goals and desires.

Note that the definitions of these types of objects, external to the observer, are nevertheless expressed in relation to the observer. This might seem to make these objects *less* objective. However, this is only true when the word "objective" is simply taken to mean "independent of the observer," full stop, as it is often mistakenly understood (this definition is often seen in dictionaries, for instance). If we return to the original meaning of "objective" — i.e., something that exists as the object of a subject's actions and experiences — then these definitions make objects *more* objective. They restore (or rather, bring into sharper focus) to objects the property of being something that can be acted upon, manipulated, represented, experienced by a subject, precisely because they are not the subject (or rather, they are not being treated as the subject by the subject). Admittedly, this topic pushes into some more arcane aspects of metaphysics. Nevertheless, many authors are converging on a similar view of objects and "objectivity" (a prime example is Smith, 1996; see also Varela, Thompson, & Rosch, 1991). Nevertheless, the practical consequence of these considerations is that any mechanism which identifies and individuates objects must do so not only on the basis of the proximal stimulus produced by the object, but also on the basis of the use the mechanism puts the object to. Words and letters on a page become objects to you only if you read them; otherwise, they are only squiggles, local variations of pigmentation that are properties of a sheet of paper.

Perceptual object. A mental or neural representation of a physical or virtual object. Its purpose is to bind elements of the retinal image together into a phenomenal whole. Perceptual

objects are essentially a mental model of the physical object or structure which gave rise to the retinal image. They are inferences as to what physical objects best explain the proximal stimulus, based on the information available to the visual system at a given time. This information includes various principles relating the retinal image to physical objects, such as genericity and likelihood, as well as knowledge about what object might likely be found in a given environment. Note that in the case of virtual images, the perceptual object will nonetheless tend to be that of a three-dimensional object. In other words, the perceptual object is, in this case, a representation of the physical object that might have given rise to the virtual object as well as the retinal image, since physical and virtual objects give rise to retinal images that are the same in nature. In each case, because the perceptual object is an inference, it can be either a valid or invalid inference; perception can be veridical or non-veridical ("external" objects do not have this normative aspect to them.) Perceptual objects are the units of attention. It remains to be seen whether the visual system assigns objecthood to parts of the retinal image without also allocating attention to that object.

The definition given above does not specify the "level" at which the object exists, in that it doesn't specify whether a perceptual object contains merely structural information and whether that information corresponds to parts of wholes or to wholes, or whether it contains conceptual or semantic information. This omission is deliberate. There is increasing converging evidence, from neuroscience (Churchland, Ramachandran & Sejnowski, 1994), from psychophysics (Di Lollo et al., 2000) and from cognitive psychology (Pylyshyn, 2000), that the perception of objects is not a sequential process, but rather an iterative, interactive one. As we have seen, Di Lollo et al. have shown with their "object substitution" paradigm that the perception of one figure can disrupt the perception of another figure without actually occluding the other shape, due to the fact that the figures have overlapping time courses. This suggests that perceptual objects are progressively built up over time. However, the application of conceptual knowledge might not be mandatory for perceptual organization and attentional allocation to occur. Even partial representations of objects provide information to other parts of the visual system, as Churchland, Ramachandran and Sejnowski (1994), and Grossberg (1994; Grossberg & al., 1994) argue. Thus, a perceptual object that is selected by attention might be anything from a "surface slab" (i.e., a region of common surface properties in Grossberg's SOS model) to a segment of an object, to a whole object, to a group of objects. The "product of the perceptual process," as we might call it, that guides attention will be different depending not only on the constraints which guide the formation of the object (figural information, motion information and prior experience) but also on the time course of visual processing itself. For example, it might be the case that a person's attention is captured by a shiny doorknob before they are able to construct a percept of, and attend to, the whole door.

In sum, object-based attention and perceptual organization can be characterized in terms of an inferential process that integrates three broad types of information: bottom-up sensory information, conceptual information about objects and their properties, and information about the goals and expectations of the agent attending to objects. It is expected that this characterization will enable the operational definition of objecthood, as well as the development of an integrated model of perceptual organization and object-based attention. The following section outlines such a model and discusses the issues that still need to be addressed in order to develop the model more fully.

4. PROPOSED RESEARCH: GROUNDWORK FOR IAAM

The "received" view on object-based attention is that attention is allocated to perceptual groups that are formed pre-attentively. These groups are generally assumed to be formed according to the Gestalt grouping principles (Driver & Baylis, 1998). I have argued that these assumptions are incorrect and cannot be maintained in the face of evidence. First, there is evidence that perceptual grouping is at least in part a process that requires attention (Ben-Av & al., 1992; Mack & al., 1992; Rock & al., 1992). Moreover, there is neurological evidence that perceptual and attentional processes operate in parallel at a neural level (Grossberg, 1994; Grossberg & al., 1994). Finally, perceptual grouping and organization appears to be better explained as an inferential process rather than a process governed by the Gestalt principles (Hoffman, 1998; Palmer, 1999; Pomerantz & Kubovy, 1986).

I propose IAAM as an alternative account of object-based attention. On this view, perceptual organization and attention are concurrent, mutually interactive processes. Attentional allocation is a function of the "effort" required by the visual system to organize a visual stimulus into groups and objects. The object effect obtains because perceptual organization itself requires and engages attention. A consequence of this hypothesis is that attentional allocation should vary as a function of the strength and presence of various grouping factors. The following example illustrates the contrast between the standard object-based attention framework and the one I am proposing. Compare the two images in Figure 10: the one on the left is composed of two elements that are defined by the principle of uniform connectedness (Palmer & Rock, 1994), whereas the image on the right contains two elements that are defined by the principle of good continuation (Koffka, 1935).

Figure 10: Two groups of circles.
On the right: circles defined by uniform connectedness.
On the left: circles defined by good continuation.

On the typical object-based attention model, the size of the objet effect (the advantage for processing within objects relative to across objects) should be the same for the group on the left and the group on the right, since grouping and object formation is supposed to be pre-attentive. At most, there might be a processing cost associated with grouping the segments on the right into two circles, which would affect reaction times in an additive manner (Sternberg, 1969). On the IAAM view, however, there would be a difference in the size of the object effect between the two groups of circles. The group on the right in Figure 10 is presumably more difficult to group into two circles, rather than one large group, than the group on the left. If the visual system has a tendency to persist in forming individual objects, then on this account we can expect that the object effect will be stronger for the group on the right. But if the visual system allocates more attention to objects that are easily formed, then we can expect the object effect to be weaker for the group on the right. These differences are shown in Figure 11.

A further implication of IAAM is that top-down information plays an important role in modulating attentional allocation. As noted in section 3, what counts as a perceptual object is determined in part by a person's goals and beliefs. For instance, the elements of composite objects will only be seen as objects if an observer intends to manipulate those elements, either physically or mentally. Thus, in IAAM the degree to which the visual system persists in forming

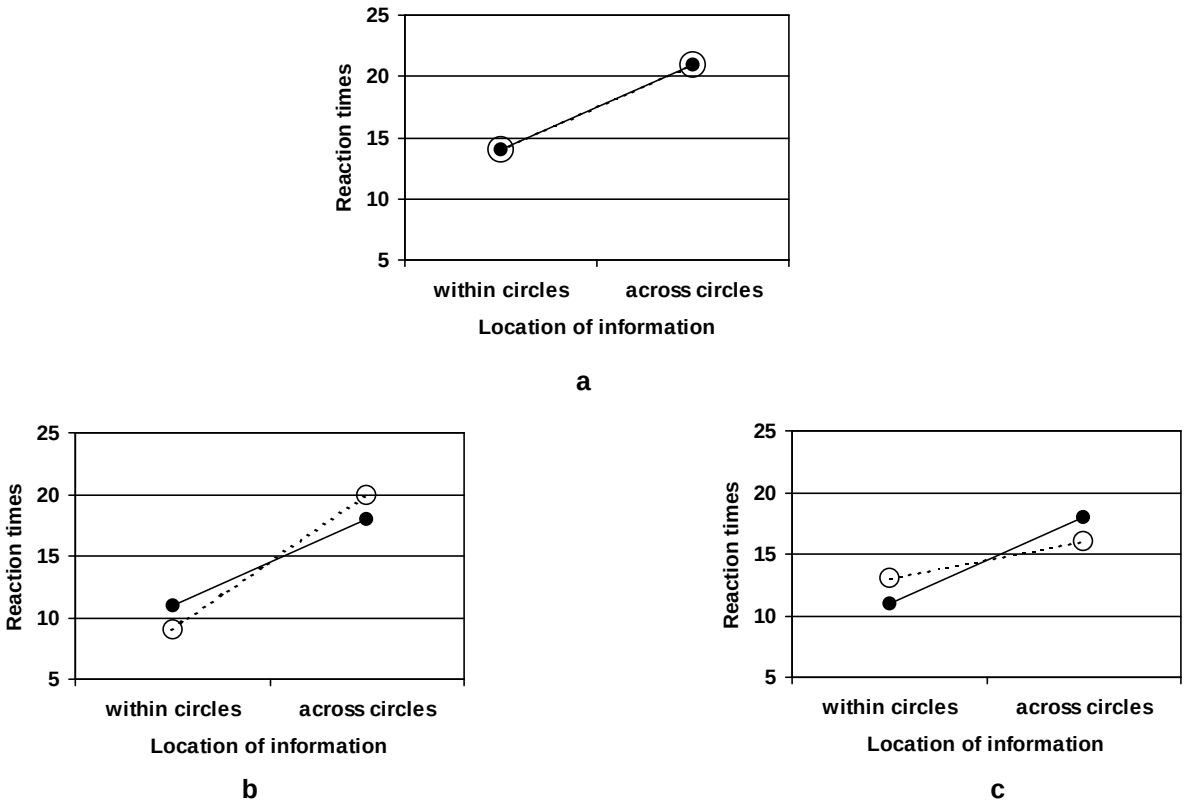

Figure 11: Variations in the size of the object effect between solid circle and segmented circle conditions. Solid lines: solid circles. Dashed lines: segmented circles

Panel a: Standard object-based attention theory Panel b: Revised theory, high motivation to form objects. Panel c: Revised theory, low motivation form objects

objects is a function of top-down constraints. On this view, if an observer is motivated to perceive each circle in Figure 10 as an object in its own right (e.g., by being instructed to do so, or by performing a task that requires interrogation of global attributes of the circles, such as size or colour), then the disparity between information processing within a single object and processing across objects will be larger for the segmented circles than the solid circles. Conversely, if an observer has no such motivation, or is instructed to attend to local features of the circles (e.g., line width or length or the dashes), then the object effect will be smaller for the segmented circles. This would yield reaction times similar to those illustrated in Figure 11.

The considerations mentioned above yield the following constraints on IAAM:

- Attentional and perceptual processes should operate in parallel, as in the SOS model (Grossman, 1994; Grossman & al., 1994)
- The attentional processes should integrate both spatial and object-based operators, as in the SOS model
- The perceptual mechanisms should embody a process of "inference to the best explanation" that integrates bottom-up and top-down information (Hoffman, 1998; Palmer, 1999)
- The bottom-up information available to perception should be extracted pre-attentively according to the principles of uniform connectedness and common region, as proposed in the model by Palmer and Rock (1994)
- Allocation of attention to objects and perceptual groups should be a function both

of the ease of grouping and of the motivation for grouping. That is, more attention will be allocated groups and objects that are difficult to organize, but only if the observer is in some way motivated to perceive those objects and groups

- The difficulty of organizing the proximal stimulus into objects should be primarily a function of bottom-up factors
- The motivation to perceive certain groups or objects rather than others should be largely a function of top-down factors
- The output of this system should be mental or neural representations of objects or perceptual groups that are readily available for processing by other systems, such as object-recognition, decision-making, or planning of action, as well as for further processing by the perceptual systems for more detailed perceptual organization.

The model outlined above yields two main empirical predictions. First, the size of the object effect should vary as a function of the difficulty of organizing the proximal stimulus into objects or groups. Second, the variation of the size of the object effect should itself be a function of top-down factors which motivate the visual system to organize the proximal stimulus into particular configurations. The aim of the proposed research is to test these predictions experimentally. Doing so requires further clarification of the inferential nature of perceptual organization. Specifically, the functioning of the principles of genericity and likelihood discussed in section 3.2.3 need to be better understood in the framework of inference to the best explanation. Also, the notion of "top-down factors" needs to be clarified. Further review of the literatures on these topics will doubtless be of use in this. However, the focus of the proposed research is experimental. This requires operational definitions of ease of perceptual organization, motivation to form certain perceptual organizations, bottom-up factors, and top-down factors. Reviewing the literature will help refine these definitions, but the process of operationalization occurs mainly in experiments. Thus, a main goal of this research is to develop and refine experimental paradigms to test the empirical implications of the model given above. Some preliminary research has been carried out in this regard. After reviewing this work, I shall propose a set of experiments which constitute the framework I will use for testing my model empirically.

4.1 Preliminary research

The literature on perceptual organization and object perception (see Hoffman, 1998; Palmer, 1999) suggests three main factors that influence perceptual organization: motion, static grouping principles, and conceptual factors. On the view proposed in this research, static factors and motion, as bottom-up factors, determine the processing cost involved perceptual organization. The conceptual factors – the top-down factors – modulate the degree to which the visual system persists in organizing bottom-up elements into specific objects and groups. Thus, static grouping factors and motion determine how much attention certain perceptual groupings require, whereas conceptual factors determine how much attention is actually allocated to those groupings. An experimental paradigm for testing the model object-based attention outlined above should therefore involve the concurrent manipulation of bottom-up grouping factors (motion and static grouping cues) on the one hand, and top-down factors (e.g., instructions to attend to groups of elements as groups) on the other in a factorial design.

A pilot study on the interaction between the bottom-up factor of common fate (or common motion; Koffka, 1935) and the top-down factor of attentional focus has been carried out

(Jarmasz, Herdman & Johannsdottir, in press). The typical paradigm for studying object-based attention contrasts performance on comparisons tasks for local object attributes (e.g., size, shape or colour of individual components) within objects and across objects (see Duncan, 1984; Lavie & Driver, 1996; Treisman et al., 1983). Jarmasz et al. (in press) extended this paradigm in two ways. First, elements were grouped by common fate rather by the static Gestalt principles. Second, attentional focus was used to modulate grouping effects.

The combination of common fate and the standard object-based attention comparison task in the Jarmasz et al. study is notable, because existing studies of the effect of motion on grouping have not used this type of information processing task. Pylyshyn and Storm (1988) studied observers' ability to track independently moving elements. Baylis and Driver (1989) showed that common fate modulates interference from distractors in a letter recognition task. McLeod, Driver, Dienes and Crisp (1991) examined the effects of motion on conjunction search tasks. The experimental design used by Valdes-Sosa, Cobo and Pinilla (1998) most closely resembles the standard comparison paradigm. In that study, observers saw two groups of dots defined by common motion and had to identify two global attributes (direction of movement of the group, and speed of the group) for either one or both groups. However, they were not required to compare the attributes. Pilots must regularly compare information from various parts of their instrumentation in order to fly safely. Thus, the task in the Jarmasz et al. (in press) study represents information processing in the cockpit more accurately than other studies that have studied the effects of common fate on attention.

Jarmasz et al. (in press) conducted two experiments using the comparison task paradigm outlined above. In the first experiment, 11 participants were instructed to compare the colour of two dots. The dots appeared in a group of moving dots, a group of static dots, or in two groups (one static, one moving) at once. Latencies to trials where both target dots were the same colour were significantly lower when they appeared in the same group (static or moving) than when they appeared across groups, by about 20 milliseconds. No significant differences were found for trials where the targets were different (see Figure 12). Significance was assessed with 95% within-confidence intervals calculated using the method given by Loftus and Masson (1994). This suggests that common fate alone is a weak grouping principle, and is consistent with the hypothesis that weak grouping factors lead to a small object effect when there is no top-down inducement to perceive groups as wholes.

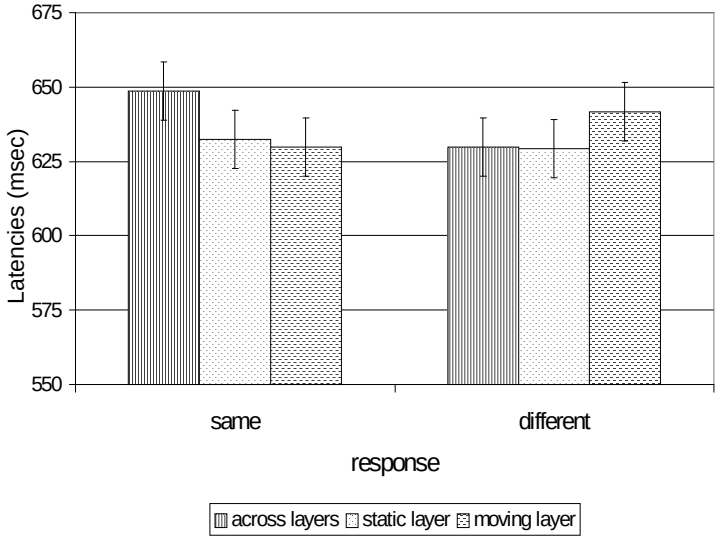

Figure 12: Latencies for Experiment 1, from Jarmasz et al. (in press).

Experiment 2 used the same task and stimuli as Experiment 1, with the following change: participants were instructed to focus their attention on a particular group of dots for a whole block of trials. They were to change the group they attended from block to block. It was shown that reaction times in the comparison task were fastest when both targets were in the group being attended (see Figures 13 and 14). This is consistent with the hypothesis that the size of the object effect increases when there is inducement from top-down factors to group elements.

The Jarmasz et al. (in press) study successfully established an experimental paradigm

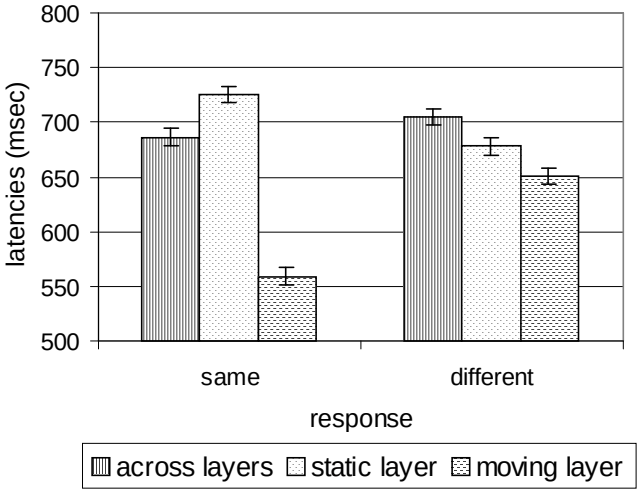

Figure 13: Latencies from Experiment 2 (Jarmasz et al., in press); focus on moving layer

combining the standard object-based comparison target task and grouping by common fate. It also provides some preliminary support for the model of object-based attention I am proposing here. The Jarmasz et al. paradigm is the framework I will use to test the theory that the object effect varies as a function of both the processing cost of perceptual grouping based on bottom-up factors and the "motivation" to group elements into particular objects based on top-down factors.

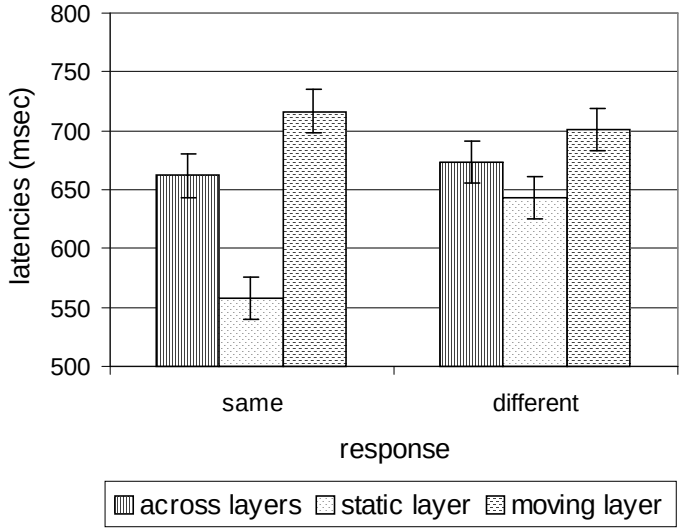

Figure 14: Latencies from Experiment 2 (Jarmasz et al., in press); focus on static layer

4.2 Logic of proposed experimental research

There are three main experimental hypotheses that flow from the model proposed in section 4. First, elements in a visual scene that are difficult to group into a coherent object as a function of bottom-up factors require more attention that elements that are easy to group. Second, more attention is available to group elements into an object when an observer is motivated to perceive the elements as an object by top-down factors. Third, the difficulty of perceptual grouping and the motivation to group elements into objects will interact, such that the visual system will allocate more attention to elements that are difficult to group only when the visual system is motivated to do so.

Attentional allocation will be operationalized with two-alternative forced-choice (2AFC) tasks of the type described in section 4.1. The dependent measures will be reaction times and error rates as a function of the location of the targets being compared. The main comparisons will be between the dependent measures for targets displayed within a single object and targets displayed across two or more objects.

Difficulty in grouping elements based on bottom-up factors will be manipulated by varying the type and strength of static and dynamic grouping factors used to create the displays (for an overview of these factors, see section 3.2.3). The experiments will focus primarily on common fate and degree of connectedness between elements as the bottom-up factors, although others may be used as well. Common fate was chosen because of its central importance in object

perception as demonstrated by Spelke et al. (1995). Furthemore, it is still unclear how motion affects performance in the 2AFC task (Jarmasz et al., in press). Degree of connectedness was chosen because it also plays a central role in object perception. On the view developed by Palmer and Rock (1994), uniform connectedness (e.g., a solid line between elements) is one of the mechanisms involved in the initial parsing of an image that occurs before attention and various grouping principles are applied to a visual stimulus. Thus, grouping by uniform connectedness should place no demands on attention, whereas grouping by partial connectedness should require attention.

Motivation to group elements into particular objects will be varied by requiring participants to process groups as wholes before or while they perform the 2AFC task. For instance, participants might be required to report or inspect a global attribute of a group (size, overall shape, number of elements) before comparing the two target elements. Participants will also be given varying degrees of control over the movement of groups of elements, via a joystick or a head-tracked HMD. It has been suggested that a characteristic feature of objects is that they can be interacted with (Smith, 1996). Thus changing the amount of control participants have over a group of elements should alter the degree to which the group is seen as an object in its own right.

Most of the proposed experiments will be at least two-factor designs, with degree of grouping difficulty and motivation to group as main factors. In some cases the factors will be varied between subjects or between experiments for practical reasons.

5. IMPLICATIONS FOR HUDS

As noted in the introduction, HUDs improve performance in many flying tasks, but they interfere with information processing under certain conditions. As Fischer et al. (1980) showed, the presence of a HUD can prevent the effective allocation of attention to the outside scenery. The development of HUDs with head-referenced, moving elements raises the danger of pilots' attention being captured by an additional layer of information (Herdman el al., 2001). A major goal in HUD design is to display relevant information to pilots in such a way that attending to the far domain is not impaired by the HUD, and that information processing within the HUD is not adversely affected by the perceptual characteristics of the HUD itself, such as attentional capture by particularly salient parts of the HUD.

Object-based attention has proven to be useful in the study of information processing with HUDs. It has been used in order to explain the cognitive tunneling observed with HUDs (Fischer et al., 1980; Wickens & Long, 1995). It has also been used to improve the design of HUDs, in the form of conformal symbology (Wickens & Long, 1995). It is therefore expected that IAAM will be of benefit in the evaluation and design of advanced dynamic HUDs, especially HUDs containing moving elements. The experiments proposed above establish an experimental paradigm for studying object-based attention effects in dynamic displays. It is also expected that these experiments will provide some insights into trade-off that HUD designers might face. The grouping factors that are manipulated in the experiments – motion, connectedness, source of trajectory control, and grouping by concept – likely affect information processing in different ways. For instance, it might be the case that grouping by motion afford pilots enhanced information processing within a group without significantly impairing information processing between groups; however, further grouping by concept or by connectedness might further enhance within-group processing while adversely affecting between-group processing. Furthermore, different types of missions might require that information in the display be integrated or grouped in different ways; studying how different grouping factors affect attention in different situations might allow HUD designs to be tailored to the needs of specific missions.

Ultimately, the design and evaluation of HUDs would most benefit from systematic, quantitative tools for predicting the effects of specific design decisions on attention and information processing. The commonly-held assumption that object-based attention is referenced to Gestalt grouping principles (Driver & Baylis, 1998) has not yielded such tools, and I argue above that the Gestalt principles are not suited to quantitative metrics of objectness anyway. It is much more likely that quantitative measures will result from an approach that takes perceptual organization to be an inferential process, and that assumes that the cost of computing perceptual groupings is reflected in the allocation of attention within and across groups. It is expected that the theoretical and experimental work proposed in section 4 will contribute to the development of a model of object-based attention that will sustain quantitative measures of objectness.

REFERENCES

Albert, M. K., & Hoffman, D. D. (1995). Generictiy in spatial vision. In R. D. Luce, M. D'Zmura, D. D. Hoffman, G. J. Iverson, & A. K. Romney (Eds.), *Geometric representations of perceptual phenomena; papers in honor of Tarow Indow on his 70th birthday.* Mahwah, NJ: Erlbaum.

Baylis, G. C., & Driver, J. (1992). Visual parsing and response competition: The effect of grouping factors. *Perception & Psychophysics*, *51*, 145-162.

Baylis, G. C., & Driver, J. (1993). Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 451-470.

Ben-Av, M. B., Sagi, D., & Braun, J. (1992). Visual attention and perceptual grouping. *Perception & Psychophysics*, *52*, 277-294.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn and D. N. Osherson (Eds.), *An Invitation to Cognitive Science, Volume 2: Visual Cognition.* Cambridge, MA: MIT Press.

Broadbent, D. E. (1958). *Perception and Communication*. New York: Pergamon Press.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern Recognition by Self-Organizing Neural Networks.* Cambridge, MA: MIT Press.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch and J. L. Davis (Eds.), *Large-Scale Neuronal Theories of the Brain.* Cambridge, MA: MIT Press.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General, 129,* 481-507.

Downing C. J., & Pinker, S. (1985). The spatial structure of visual attention. In M. I. Posner and O. S. M. Marin (Eds.), *Attention & performance: Vol. XI* (171-188). Hillsdale, NJ: Erlbaum.

Driver, J., & Baylis, G. C. (1989). Movement and visual attention: The spotlight metphor brakes down. *Journal of Experimental Psychology: Human Perception and performance*, *15*, 448-456.

Driver, J., & Baylis, G.C. (1998). Attention and visual object segmentation. In R. Parasuraman (Ed.), *The Attentive Brain.* Cambridge, MA: MIT Press.

Duncan, J., (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*, 501-517.

Duncan, J., & Nimmo-Smith, I. (1996). Objects and attributes in divided attention: Surface and boundary systems. *Perception & Psychophysics*, *58*, 1076-1084.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*, 143-149.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*, 583-597.

Feldman, J. (1997). Regularity-based perceptual grouping. In *Computational Intelligence, 13,* 582-623.

Feldman, J. (1999) The role of objects in perceptual grouping. *Acta Psychologica,102,* 137-163.

Fernandez-Duque, D., & Johnson, M. L. (1999). Attention Metaphors: How metaphors guide the congnitive psychology of attention. *Cognitive Science, 23,* 83-116.

Fischer, E., Haines, R. F., & Price, T. A. (1980). Cognitive issues in head-up displays. *NASA Technical Paper 1711*, NASA Ames Research Center, Moffett Field, CA.

Foyle, D. C., Stanford, B., & McCann, R. S. (1991). Attentional issues in superimposed flight symbology. In R. S, Jensen, (Eds.), *Proceedings of the Sixth International Symposium on Aviation Psychology* (577-582). Colombua OH: The Ohio State University.

Gibson, J. J. (1979). *The Ecological approach to visual perception.* Dallas, TX: Houghton Mifflin.

Goldsmith, M. (1998). What's in a location? Comparing object-based and space-based models of feature integration in visual search. *Journal of Experimental Psychology: General, 127*, 189-219.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. In *Perception & Psychophysics*, *55*, 48-120.

Grossberg, S., Mingolla, E., & Ross, W. D. (1994). A neural theory of attentive visual search: interactions of boundary, surface, spatial and object representations. In *Psychological Review, 101*, 470-489.

Herdman, C. M., Jarmasz, J., and Johannsdottir, K. (2000). *Research on Heads Up Displays and Helmet Mounted Symbology.* Summary Report. PWGSC File Number: WW7711-9-7577.

Herdman, C. M., Johannsdottir, K. R., Armstrong, J., Jarmasz, J., LeFevre, J., & Lichacz, F. (2001). Mixed-up but flyable: HMDs with aircraft- and head-referenced symbology. *Engineering Psychology and Cognitive Ergonomics: Ashgate.*

Hoffman, D. D. (1998). *Visual Intelligence.* New York, NY: W. W. Norton & Company.

Holyoak, K. J., & Thagard, P. (19965. *Mental Leaps: Analogy in Creative Thoug*ht. Cambridge, MA: MIT Press.

Humphreys, G. W., (1993). Interaction between objects and space systems revealed through neuropsychology. In D. E. Meyer & S. Kornblum (Eds.), *Attention & performance XIV* (143-162). Cambridge, MA: MIT Press.

IllusionWorks (1997). *Depth Ambiguity* [On-line]. Available HTTP: www.illusionworks.com/html/depth_ambiguity.html.

Jarmasz, J., Herdman, C. M., & Johannsdottir, K. R. (2001). Object layers in HUDs: The role of motion in grouping symbology. *Engineering Psychology and Cognitive Ergonomics: Ashgate.*

Jonides, J. P. (1981). Voluntary versus automatic control over the mind's eye. In J. Long and A. Baddeley (Eds.), *Attention & Performance IX*. Hillsdale, NJ: Erlbaum.

Jordan, H., & Tipper, S. P. (1998). Object-based inhibition of return in static displays. *Psychonomic Bulletin & Review*, *5*, 503-509.

Kanizsa, G. (1979). *Organization in Vision; Essays on Gestalt Perception.* New York: Praeger.

Kanwisher, N., & Driver, J. (1992). Objects, attributes, and visual attention: Which, what and where. *Current Directions in Psychological Science*, *1*, 26-31.

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. In *Cognitive Psychology, 23*, 141-221.

Kosslyn, S. M. (1994). *Image and Brain: The Resolution of the Imagery Debate.* Cambridge, MA: MIT Press.

Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt, Brace & World.

Kramer, A. F. & Jacobson, A. (1991). Perceptual organization and focused attention: The role of objects and proximity in visual processing. *Perception & Psychophysics*, *50*, 267-284.

Lavie, N. & Driver, J. (1996). On the spatial extent of attention in object-based visual selection.

Perception & Psychophysics, 58, 1238-1251.
Lee, T. S. (1995). A Bayesian framework for understanding texture segregation in the primary visual cortex. In *Vision Research, 35,* 2643-2657.
Leyton, M. (1992). *Symmetry, Causality, Mind.*Cambridge, MA: MIT Press.
Lindley, D. V. (1985). *Making Decisions, Second Edition.* London, UK: John Wiley & Sons Ltd.
Lipton, P. (1991). *Inference to the Best Explanation.* London: Routledge.
Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review, 103,* 603-649.
Lowe, D. G. (1985). *Perceptual Organization and Visual Recognition.* Boston, MA: Kluwer Academic Publishers.
Mack, A., Tang, B., Tuma, R., Kahn, S., & Rock, I. (1992), Perceptual organization and attention. *Cognitive Psychology, 24,* 475-501.
Marr, D. (1982). *Vision; a Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco, CA: Freeman.
Martin-Emerson, R., & Wickens, C. D. (1997). Superimposion, symbology, visual attention, and the head-up display. *Human Factors, 39,* 581-601.
McCann , R. S., & Foyle, D. C., & Johnston, J. C. (1993). Attentional limitation with head-up displays. In R. S. Jensen (Ed.), *Proceedings of the 37th Annual Meeting of the Human Factor Society* (1345-1349). Columbus, OH: The Ohio State University.
McCann, R. S., & Foyle, D. C. (1994). Superimposed symbology: Attentional problems and design solutions. *SAE Transactions: Journal of Aerospace, 103,* 2009-2016.
McCann, R. S., & Foyle, D. C. (1996). Scene-linked symbology to improve situation awareness. *AGARD Conference Proceedings CP 575* (16-1 - 16-11). Brussels, Belgium.
Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. In S. M. Kosslyn and R. A. Andersen (Eds.), *Frontiers in Cognitive Neuroscience*, p. 19-23. Cambridge, MA: MIT Press, 1992.
Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.
Palmer, S. E., & Rock, I. (1994): On the nature and order of perceptual organizational processing: A reply to Peterson. In *Psychonomic Bulletin & Review, 3,* 515-519.
Peirce, C. S. (1931-1958). *Collected papers of Charles Sanders Peirce.* C. Hartshorne & P. Weiss (Eds.). Cambridge, MA: Harvard University Press.
Pomerantz, J. R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization. In , K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), *Handbook of Perception and Human Performance, Volume II*. New York, NY: John Wiley & Sons, Inc.
Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3-25.
Posner, M. I., Walker, J. A., Friedrich, F. A., & Rafal, R. D. (1987). How do the parietal lobes direct covert attention? *Neuropsychologia, 25,* 135-145.
Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32,* 65-97.
Pylyshyn, Z. (1998). Visual indexes in spatial vision and imagery. In R. D. Wright (Eds.), *Visual Attention* (215-231). Oxford, NY: Oxford University Press.
Pylyshyn, Z. (2000). Visual indexes, preconceptual objects, and situated vision. Manuscript under review for a special issue of *Cognition* on "Objects and Attention," July 2000.
Pylyshyn, Z., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 1-19.

Reuter-Lorenz, P. A., Drain, M., & Hardy-Morais, C. (1996). Object-centered attentional biases in the intact brain. *Journal of Cognitive Neuroscience, 8*, 540-550.
Robertson, L. C., & Rafal, R. (2000). Disorder of Visual Attention. In M. S. Gazzaniga (Eds.), *The New Cognitive Neuroscience* (633-649). Cambridge, MA: MIT Press.
Rock, I., Linnett, C. M., Grant, P., & Mack, A. (1992). Perception without attention: Results of a new method. *Cognitive Psychology, 24*, 502-534.
Sears, C. R., & Pylyshyn, Z. W. (2000). Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology, 54*, 1-14.
Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, *38*, 259-290.
Shannon, C. E. (1938). "A symbolic analysis of relay and switching circuits." Master's thesis, Massachusetts Institute of Technology; published in *Transactions of the American Institute of Electrical Engineers*, *57*: 1-11.
Shelden, S. G., Foyle, D. C., & McCann, R. S. (1997). Effects of scene-linked symbology on flight performance.
Smith, B. C. (1996). *On the Origin of Objects.* Cambridge, MA: MIT Press.
Spelke, E. S., Gutheil, G., & Van de Walle, G. (1995). The development of object perception. In S. M. Kosslyn & D. N. Osherson (Eds.), *Visual Cognition: An Invitation to Cognitive Science Vol 2* (297-330). Cambridge, MA: MIT Press.
Stelmach, L. B., Campsall, J. M., & Herdman, C. M. (1997). Attentional and Occular Movements. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 823-844.
Sternberg, S. (1969). Memory-scanning : Mental processes revealed by reaction-time experiments. In *American Scientist, 57*, 421-457.
Tipper, S. P., Driver, J., & Weaver, B. (1991). Short Report: Object-centered inhibition of return of visual attention. *The Quarterly Journal of Experimental Psychology*, *43 A*, 289-298.
Tipper, S. P., & Weaver, B. (1998). The medium of attention: Location-based, object-based, or scene-based? In R. D. Wright (Ed.), *Visual Attention* (77-107). Oxford, NY: Oxford University Press.
Treisman, A. (1988). Features and objects: The fourteenth Barlett memorial lecture. *Quarterly Journal of Experimental Psychology*, *40A*, 201-237.
Treisman, A. (1998). Feature binding, attention and object perception. In *Philosophical Transactions of The Royal Society of London*, *353*, 1295-1306.
Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.
Treisman, A., Kahneman, D., & Burkell, J. (1983). Perceptual objects and the cost of filtering. *Perception & Psychophysics, 33*, 527-532.
Valdes-Sosa, M., Cobo, A., & Pinilla, T. (1998). Transparent motion and object-based attention. *Cognition*, *66*, B13-B23.
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge, MA: MIT Press.
Wickens, C. D., & Long, J. (1994). Conformal symbology, attention shifts, and the head-up display. *Proceedings of the 38th Annual Meeting of the Human Factor and Ergonomics Society*, (6-10) Nashville, TN: Human Factors and Ergonomics Society.
Wickens, C. D., & Long, J. (1995). Object versus space-based models of visual attention: Implications for the design of Head-Up Displays. *Journal of Experimental Psychology:*

Applied, *1*, 179-193.
Wright, R. D., & Ward, L. M. (1998). The control of visual attention. In R. D. Wright (Ed.), *Visual Attention* (132-186). Oxford, NY: Oxford University Press.
Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, *24*, 295-340.
Yantis, S. (1998). Objects, attention, and perceptual experience. In R. D. Wright (Eds.), *Visual Attention*, 187-214. Oxford, NY: Oxford University Press.