Vision Research 51 (2011) 1538–1551

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Review

Visual cognition

Patrick Cavanagh

Centre Attention & Vision, LPP CNRS UMR 8158, Université Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France
Vision Sciences Lab., Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 11 August 2010
Received in revised form 22 January 2011
Available online 15 February 2011

Keywords:
Vision
Attention
Cognition
Motion
Object recognition

ABSTRACT

Visual cognition, high-level vision, mid-level vision and top-down processing all refer to decision-based scene analyses that combine prior knowledge with retinal input to generate representations. The label "visual cognition" is little used at present, but research and experiments on mid- and high-level, inference-based vision have flourished, becoming in the 21st century a significant, if often understated part, of current vision research. How does visual cognition work? What are its moving parts? This paper reviews the origins and architecture of visual cognition and briefly describes some work in the areas of routines, attention, surfaces, objects, and events (motion, causality, and agency). Most vision scientists avoid being too explicit when presenting concepts about visual cognition, having learned that explicit models invite easy criticism. What we see in the literature is ample evidence for visual cognition, but few or only cautious attempts to detail how it might work. This is the great unfinished business of vision research: at some point we will be done with characterizing how the visual system measures the world and we will have to return to the question of how vision constructs models of objects, surfaces, scenes, and events.

1. Introduction

A critical component of vision is the creation of visual entities, representations of surfaces and objects that do not change the base data of the visual scene but change which parts we see as belonging together and how they are arrayed in depth. Whether seeing a set of dots as a familiar letter, an arrangement of stars as a connected shape or the space within a contour as a filled volume that may or may not connect with the outside space, the entity that is constructed is unified in our mind even if not in the image. The construction of these entities is the task of visual cognition and, in almost all cases, each construct is a choice among an infinity of possibilities, chosen based on likelihood, bias, or a whim, but chosen by rejecting other valid competitors. The entities are not limited to static surfaces or structures but also include dynamic structures that only emerge over time – from dots that appear to be walking like a human or a moon orbiting a planet, to the causality and intention seen in the interaction of dots, and the syntax and semantics of entire events. There is clearly some large-scale information processing system that accumulates and oversees these visual computations. We will look at various mid-level visual domains (for example, depth and light) and dynamic domains (motion, intentionality and causality) and briefly survey general models of visual cognition. I will cover both mid- and high-level processing as equally interesting components of visual cognition: as rough categories, mid-level vision calls on local inferential processes dealing with surfaces whereas high-level vision operates on objects and scenes. Papers on high-level vision have been rare in this journal but papers on mid-level and dynamic aspects of vision are not and there are three other reviews touching on these area in this special issue (Kingdom, 2011; Morgan, 2011; Thompson & Burr, 2011). We start here by placing the mid- and high-level vision system within the overall processing architecture of the brain.

The descriptions of surfaces, objects, and events computed by mid- and high-level processes are not solely for consumption in the visual system but live at a level that is appropriate for passing onto other brain centers. Clearly, the description of visual scene cannot be sent in its entirety, like a picture or movie, to other centers as that would require that each of them have their own visual system to decode the description. Some very compressed, annotated, or labeled version must be constructed that can be passed on in a format and that other centers – memory, language, planning – can understand. This idea of a common space and a common format for exchange between brain centers (see Fig. 1) has been proposed by Baars (1988) and Dehaene and Naccache (2001) and others as a central bulletin board or chat room where the different centers post current descriptions and receive requests from each other like perhaps "Vision: Are there any red things just above the upcoming road intersection?" The nature of this high-level, visual description that can be exported to and understood by other centers is as yet, completely unknown. We can imagine that it might embody the content that we label as conscious vision if only

E-mail address: patrick.cavanagh@parisdescartes.fr

0042-6989/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2011.01.015

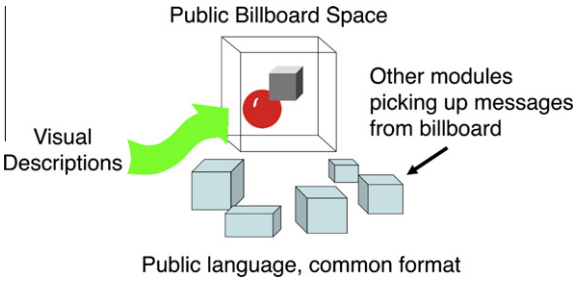

Fig. 1. Central billboard. Different modules post information on the billboard (or blackboard) and these become accessible to all. Vision would post high-level descriptions of visual events in a format that other brain modules understand (Baars, 1988, 2002; Dehaene & Naccache, 2001; van der Velde & de Kamps, 2006).

because consciousness undoubtedly requires activity in many areas of the brain so visual representations that become conscious are probably those shared outside strictly visual centers. The components of high-level visual representation may therefore be those that we can report as conscious visual percepts. That is not saying much, but at least, if this is the case, high-level vision would not be trafficking in some obscure hidden code and eventually we may be able to extract the grammar, the syntax and semantics of conscious vision, and so of high-level visual representation.

Saying that the components of high-level vision are the contents of our visual awareness does not mean that these mental states are computed consciously. It only means that the end point, the product of a whole lot of pre-conscious visual computation is an awareness of the object or intention or connectedness. In fact, what interests us here is specifically the unconscious computation underlying these products, and not the further goal-related activities that are based on them. A long, visually guided process like baking a cake or driving a car has many intermediate steps that make a sequence of conscious states heading toward some final goal but that higher-level production system (c.f., Anderson, Fincham, Qin, & Stocco, 2008; Anderson et al., 2004; Newell, 1990) is not visual in nature. We are interested in the rapid, unconscious visual processes that choose among many possible representations to come up the one that we experience as a conscious percept. Attention and awareness may limit how much unconscious inference we can manage and what it will be focused on but it is the unconscious decision processes that are the wheelhouse of visual cognition.

We can divide vision into two parts: measurement and inference (Marr, 1982). In the measurement part, neurons with receptive fields with enormous variation in specialization report spatially localized signal strengths for their particular parameter of interest. These receptive fields span signal classes from brightness all the way to face identity (Tsao & Livingstone, 2008; Turk & Pentland, 1991; see Ungerleider, 2011). They are reflexive, hardwired, acquired with experience, modulated by context and attention, but they give, at best, only hints at what might be out there. Research on the receptive fields forms the solid foundation of vision research. To date, the most influential discoveries in vision and the major part of current work can be described as characterizing this measurement component of vision. It is accessible with single cell recordings, animal research, and human behavior. It is understandable that this accessibility has led to impressive discoveries and successful research programs.

However, this is only the first step in seeing as the visual system must infer (see Figs. 2, 3) from these measurements a final percept that we experience. We do not get a sense of the world that is raw and sketchy measurement but a solid visual experience with little or no evidence of the inferences that lie behind it. Note that an inference is not a guess. It is a rule-based extension from partial data to the most appropriate solution. It is constraint satisfaction like real-world Sudoku or playing 20 questions with nature (Kosslyn, 2006; Newell, 1973). Guessing, even optimal guessing as specified by Bayes, is not a mechanism but only sets limits for any mechanistic approach. It is covered in a separate paper of this issue (Geisler, 2011). Deconstructing the mechanisms of inference is difficult and not yet very rewarding. There are too many plausible alternatives and too many flinty-eyed reviewers who can see the obvious shortcomings. So one goal of this review is to underline the difficulty of research in high-level vision as well as its importance. It did have a run of intense activity in the 1970s and 1980s during the days of classical, big picture, biological and computer vision. This synergy between physiology, biological and computation research peaked with the publication of David Marr's book in 1982 and Irv Biederman's Recognition-by-Components paper in 1987. Since then, there have been a few hardy and adventuresome contributors, whose work I will feature where possible. However, it became clear that many models were premature and underconstrained by data. Rather than risk the gauntlet of justifiable skepticism, most vision research turned to the more solid ground of how vision measures the world, putting off to the future the harder question of how it draws inferences.

In looking at the history of the inference mechanisms behind visual cognition, this paper will touch as well on conscious executive functions, like attention, that swap in or out different classes of measurements, and the memory structures that provide the world knowledge and heuristics that make the inferences effective. However, other sections of this issue cover these components in more detail (Carrasco, 2011). We will start by looking at the inferences themselves, beginning with the history of visual cognition and unconscious inference, an evaluation of the computational power of visual inference, a survey of the important contributions of the past 25 years and the basic components they lay out. We will end with a consideration of large-scale models of visual cognition and how it fits in with the overall architecture of the brain.

We give our respects to Helmholtz as a dazzling polymath of the late 1800s, a pioneer who along with Faraday, Cantor and others made staggering contributions. It was Helmholtz who gave us the concept of unconscious inference. Well, just a minute, actually it was not. In truth, he lifted it, as well as the anecdotes used to justify it, from ibn al-Haytham (1024, translation, Sabra, 1989; review Howard, 1996). Known as Alhazen in the west, ibn al-Haytham was the Helmholtz of his time, a well-known mathematician and pioneer contributor to optics (discovered the lens, the pinhole camera, and the scientific method) and mathematics. His books from the 11th century were translated into Latin and, until Kepler, they were the fundamental texts in Europe for optics. At least his first book of optics was. The second and third books where he outlined his theory of unconscious inference, the visual sentient, were much less well known. However, they were undoubtedly read by Helmholtz (who did cite Alhazen but only for the work of his first book) as he repeats Alhazen's concepts almost word for word. So to give credit where it is due, Alhazen is truly the father of visual cognition which will therefore in 2024 celebrate its 1000th anniversary. Since this review covers only the last 25 years of visual cognition and the 11th century falls a bit earlier, I will not say much about Alhazen other than to note that he had already outlined many of the ideas that fuel current research. As Jerry Fodor (2001) once said, "that's what so nice about cognitive science, you can drop out for a couple of centuries and not miss a thing. (p. 49)" Well, the basic ideas may not have changed much but the specifics are a lot clearer and the methods more sophisticated. That is what will be covered here.

Before reviewing the research itself, one question stands out that we should consider: cognition, does not the brain already do that? Elsewhere? How can there be a separate visual cognition?

Fig. 2. Inference. In both these examples, the visual system assumes parameters for body shape and axes and fits these to the image measurements. Some of these assumptions are overly constraining and so occasionally wrong. The resulting errors demonstrate the inference underlying our perception. On the left, the gray goose is flying upside down, a maneuver known as whiffling. The front/back body orientation for the goose that we reflexively infer from the head orientation conflicts with actual body axis. On the right, the man is wearing shoes on his hands. We infer that the limbs wearing shoes are legs. We infer incorrectly. Errors such as these are evidence of inference and a window into the inference process.

Fig. 3. Ambiguous figure (from Rock, 1984). These amorphous shapes in white on black have very little information and yet they connect to object knowledge about human form. This recovers the possible shape of a woman sitting on a bench. No bottom up analysis can recover either of these elements. No image analysis based on parts or surfaces can work as shadow regions have broken the real object parts into accidental islands of black or white.

As Pylyshyn (1999) has detailed, yes, vision can have a independent existence with extraordinarily sophisticated inferences that are totally separate from standard, everyday, reportable cognition. Knowing, for example, that the two lines in the Muller-Lyer illusion are identical in length does not make them look so. Pylyshyn calls this cognitive impenetrability but we might see it as cognitive independence: having an independent, intelligent agent – vision – with its own inference mechanisms. Given that the brain devotes 30–40% of its prime cortical real estate to vision we can certainly imagine that the "visual brain" is a smart one, even if (or perhaps because) it does not give in to coercion from the rest of the brain. What is appealing about this separate visual intelligence is that its mechanisms of inference may be easier to study, unencumbered as they are with the eager-to-please variability of ordinary cognition as measured in laboratory settings. So when we look at what has been uncovered about visual cognition, we of course believe that these processes may be duplicated in the far murkier reaches of the prefrontal cortex for decision and conflict resolution at a broader conscious level of cognition. Visual cognition is a sort of in vivo lab preparation for studying the ineffable processes of all of cognition.

Some of the key works that defined visual cognition and advanced it over the years are by Irv Rock, who presented the core of visual cognition as the logic of perception (Rock, 1985). Shimon Ullman explored visual routines as a framework for computation by the visual system (1984, 1996). Donald Hoffman surveyed the perception of shape, light, and motion, demonstrating the raw intelligence required for each (Hoffman, 1998). Steve Kosslyn outlined an architecture for high-level vision (Kosslyn, Flynn, Amsterdam, & Wang, 1990). Kahneman, Treisman, and Gibbs (1992) proposed the influential concept of object files. Pylyshyn (1989, 2001, 2006, 2007) introduced the idea of indexing. Ken Nakayama, Phil Kellman, Shin Shimojo, Richard Gregory and others present the mid-level rule structures for making good inferences. Some topics related to visual inferences are part of the executive and data structures at the end of this review and are also covered in other reviews in this issue (Kingdom, 2011; Morgan, 2011).

Across all the different approaches to top-down, mid-, and high-level vision and visual cognition, the common theme is that there are multiple possible solutions. The retinal information is not enough to specify the percept and a variety of other information, generally called object knowledge, is called on to solve the problem. The literature to date consists of efforts to label the steps and the classes of process that make the call to extraretinal knowledge, but as yet there is little understanding or specification of the mechanisms involved. Basically, object knowledge happens, problem solved. What we would like to know is how the visual system selects the candidate objects that provide the object knowledge. We need to know the format of the input data that contacts object memory and the method by which the object data influences the construction of the appropriate model, not to mention what the format is for the model of the scene. In Section 1, we will survey papers on routines, executive functions, and architecture: how to set up image analysis as a sequence of operations on image data and on "object files" within an overall architecture for visual cognition. In Section 2 we will survey papers on the different levels of scene representation: object structure and material properties, spatial layout, and lighting. In Section 3, we cover dynamic scene attributes like motion, causality, agency, and events. Finally, we will look at the interaction of vision with the rest of the brain: information exchange and resource sharing.

This survey covers many topics chosen idiosyncratically, some straying outside vision as visual cognition is intimately interconnected with other high-level functions across the brain. Many important contributions have been undoubtedly omitted, some inadvertently, and others have fallen through the cracks between the many reviews in this issue. My apologies for these omissions. Several specialized and general texts have covered much of what is mentioned here and the reader is referred to Enns (2004), Gregory (2009), Hoffman (1998), Palmer (1999), Pashler (1998), and Ullman (1996), for example.

2. Visual executive functions: routines

Several point to the work of Marr (1982), Ullman (1984) and others as the introduction of the "computer metaphor" into vision research. But of course, it is not really a metaphor as brains in general and the visual system in particular do compute outcomes from input. We are therefore addressing physical processes realized in neural hardware that we hope eventually to catalog, locate and understand. Routines that might compute things like connectedness, belonging, support, closure, articulation, and trajectories have been the focus of small number of books and articles (Kellman & Shipley, 1991; Pinker, 1984; Rock, 1985; Roelfsema, 2005; Ullman, 1984, 1996; among others). These authors have proposed data structures that represent visual entities (Kahneman et al., 1992; Pylyshyn, 1989, 2001, 2006, 2007), processing strategies to construct them (Ullman, 1984), and verification steps to maintain consistency between the internal constructs and the incoming retinal data (Mumford, 1992).

2.1. Architecture

On a structural level, several dichotomies have been proposed for visual processing. Most notably, the processing in the ventral stream and dorsal stream were distinguished as processing of what vs where (Ungerleider & Mishkin, 1982), or action vs perception (Milner & Goodale, 2008). These anatomical separations for different classes of processing have led to numerous articles supporting and challenging this distinction. Similarly, Kosslyn et al. (1990) proposed a distinction between processing of categorical vs continuous properties in the left and right hemispheres respectively. Ultimately, these dichotomies should constrain how visual cognition is organized but little has come of this yet, other than to restate the dichotomy in various new data sets. Marr (1982) famously suggested tackling vision on three levels: computation, algorithm and implementation. It is on his computational level where the architecture is specified and, in his case, he argued for an initial primal sketch with contours and regions (see Morgan, 2011), followed by a 2½D sketch where textures and surfaces would be represented, followed by a full 3D model of the scene. Marr's suggestions inspired a great deal of research but his proposed architecture has been mostly superseded. Rensink (2000), for example, has proposed an overall architecture for vision that separates low-level visual system that processes features from two high-level systems, one attention based that focuses on the current objects of interest and one that is non-attentional that processes the gist and layout of the scene (Fig. 4). Rensink does not make any anatomical attributions for the different subsystems of this architecture.

2.2. Visual routines

Routines do the work of visual cognition, and their appearance in the psychological literature marked the opening of modern visual cognition, following close on earlier work in computer vision (c.f., Barrow & Tenenbaum, 1981; Rosenfeld, 1969; Winston, 1975). Shimon Ullman outlined a suggested set of visual routines (1984, 1996) as did Shimamura (2000) and Roelfsema (2005). These proposals dealt with the components of executive attention and working memory that are supported by routines of selection, maintenance, updating, and rerouting of information. Ullman pointed out examples of simple visual tasks that could be solved with an explicit, serially executed algorithm. Often the steps of Ullman's visual routines were not obvious, nor were they always available to introspection. In the tasks that Ullman examined (e.g., Fig. 5a), the subject responded rapidly, often within a second or less (Jolicoeur, Ullman, & Mackay, 1986, 1991; Ullman, 1984). The answer appeared with little conscious thought, or with few deliberations that could be reported. Is the red dot in Fig. 5a inside or outside the contour? We certainly have to set ourselves to the task but the steps along the way to the answer seem to leave few traces that we can describe explicitly. This computation of connectedness that follows the path within the contours of Fig. 5a was followed by several related tasks where explicit contours were tracked in order to report if two points were on the same line (Fig. 5b). Physiological experiments have evidence of this path tracing operation in the visual cortex of monkeys (Roelfsema & et al., 1998).

2.3. Indexing targets

The conclusion of this early work is that there is some active operation that follows a path and the operator is directly detectable in the cortex as it moves along the path (Roelfsema et al., 1998). Many agree that this operator is plausibly a movable focus of attention and that these results are directly linked to the other major paradigm of path tracking, multiple object tracking (Pylyshyn & Storm, 1988). The key point of these results is that attention is providing a continuously changing output during the task but without any meaningful access to how the tracking, path following, or region filling is accomplished.

Summing up, Ullman (1984, 1996) and others have suggested that a structure of routines lies behind the sophisticated and rapid processing of visual scenes. The overall architecture here remains only dimly defined. We could suggest some names for the routines like select, track, open object file, save to memory, retrieve, save to object file, as well as hierarchies of types of routines (Cavanagh, 2004). Clearly, for the moment at least, this exercise is a bit fanciful. In the absence of behavioral and physiological evidence for specific routines as actual processing components, wishing them to exist does not get us very far. Nevertheless, of the potential routines and components, the selection function of attention has received most research and continuing redefinition.

2.4. Visual executive functions: attention

We might consider this movable point of information uptake – the focus of attention according to Ullman – as so far the key element in high-level vision. It selects and passes information onto higher level processes which we can assume include identification and what Kahneman et al. (1992) have called object files, temporary data structures opened for each item of interest. Pylyshyn (1989, 2001, 2006, 2007) has written about the closely related operation of indexing an object's location – a Finger of Instantiation – from which data can be selected. Ballard, Hayhoe, Pook, and Rao (1997) also describe similar properties of deictic codes that index locations while performing visually guided tasks. Pylyshyn (1989) proposed that his FINSTs were independent of attention and Kahneman, Treisman and Gibbs were originally agnostic on the relation between object files and attention. Nevertheless, the functions attributed to spatial attention overlap so extensively with the functions of indexing and those of the temporary data structures, that there seems to be no compelling reason yet to keep them separate (although see Mitroff, Scholl, & Wynn, 2005). The primary behavior consequences of indexing, selecting or attending to a location are the "attentional benefits" of improved performance and target identification. This localized attentional benefit was described by Posner (1980) as a "spotlight" and a vast field of research has followed the properties and dynamics of this aspect of attention (see Carrasco, 2011). Here we will look not at the benefits conveyed by attention but at the properties and limits of the system that controls it. We first look at how attended locations may be coded and the evidence that attentional benefits are con-

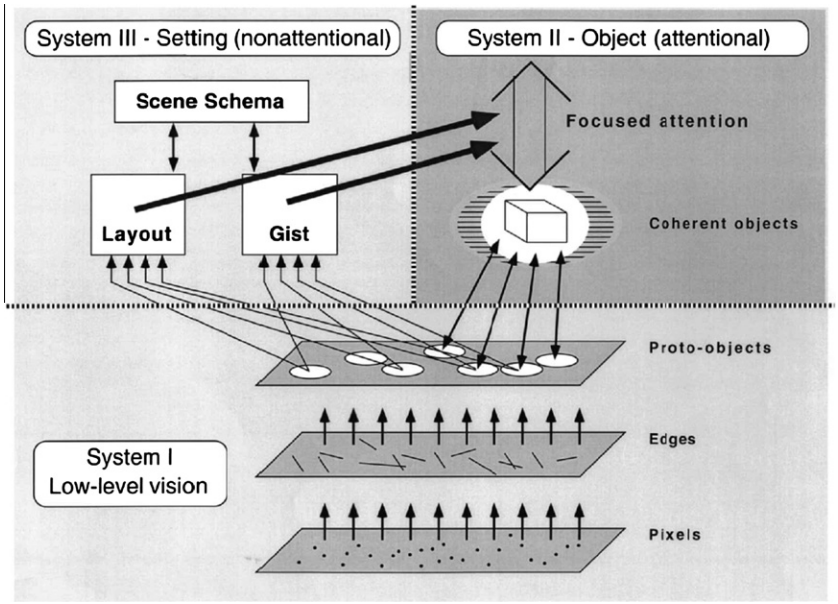

Fig. 4. Rensink's (2000) triadic architecture for the visual system. Early processes segment proto-objects from the background rapidly and in parallel across the visual field. Focused attention then can access these structures forming an individuated object with both temporal and spatial coherence. Information about the context or gist acquired outside of attention guides attention to various locations and sets scene priorities or salience.

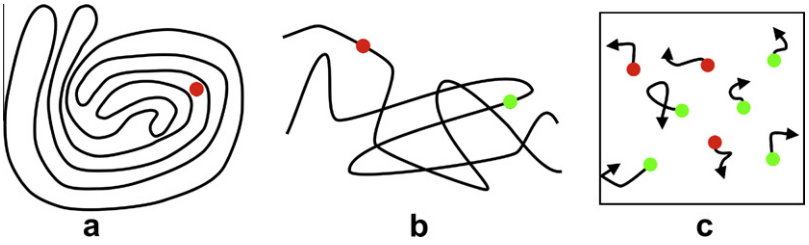

Fig. 5. Tracing and tracking. A movable indexing operator can trace through the paths of the figure in (a) to determine whether the red dot lies inside or outside a closed contour (Ullman, 1984). In (b) a similar operator can trace along the line from the red dot to see if the green dot falls on the same line (Jolicoeur et al., 1991). In (c) the three red tokens are tracked as they move (Pylyshyn & Storm, 1988). They revert to green after a short interval and the subject keeps tracking.

ferred on features at the corresponding location. We also consider why this architecture would impose a capacity limit to the number of locations that can be attended, as well as a resolution limit to the size of regions that can be selected. This location management system is only one part of attention's functions however, and we will end this section with a brief discussion of the non-location aspects of attention's architecture. Specifically, the attended locations need to be linked to the identity that labels the features at that location (Fig. 6) to form, as Kahneman et al. (1992) propose, object files. We will also need to allow for data structures, short term memory buffers, to keep track of the current task being performed, typically on an attended target, with links to, for example, current status, current target, subsequent steps, and criteria for completion.

2.5. Architecture of attention: location

2.5.1. Target location map, attention pointers

We begin with how attended locations are encoded. Numerous physiological, fMRI, and behavioral studies have shown that the spatial allocation of attention is controlled by a map (e.g., salience map, Itti & Koch, 2001; Treue, 2003) that is also the oculomotor map for eye movement planning (Rizzolatti & et al., 1987; see review in Awh, Armstrong, & Moore, 2006). Although the cortical and subcortical areas that are involved have been studied initially as saccade control areas, the activations on these maps do more than just indicate or point at a target's location for purposes of programming a saccade. Each activation also indexes the location of that target's feature information on other similarly organized retinotopic maps throughout the brain (Fig. 6). Overall, the link between these attention/saccade maps and spatial attention is compelling, indicating that activations on these maps provide the core function of spatial attention. In particular, attentional benefits follow causally from the effects these activations have on other levels of the visual system. The definitive evidence is given by a series of outstanding microstimulation studies. When delivering electric current to cells in saccade control areas with a movement field, for example, in the lower right quadrant, a high stimulating current triggers a saccade to that location. However, a slightly weaker stimulation that does not trigger a saccade generates either enhanced neural response for cells with receptive fields at that location (stimulating the Frontal Eye Fields and recording from cells in area V4, Moore, Armstrong, & Fallah, 2003) or lowered visual thresholds for visual tests at that location (shown for stimulation of superior colliculus, Muller, Philiastides, & Newsome, 2005). These findings indicate that the attentional indexing system is realized in the activity patterns of these saccade/attention maps and the effects of their downward projections. This anatomical framework does not provide any hints as to where or how the location, the features, and the identity information are combined (the triumvirate that constitutes an object file) nor where or how steps in visual and attention routines are controlled (see the end of this section for a discussion).

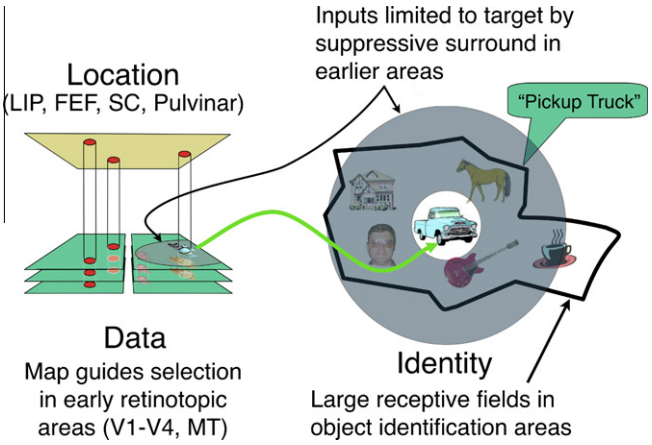

Fig. 6. Architecture of spatial attention (Cavanagh et al., 2010). A network of areas form a target map that subserves spatial attention as well as eye movements. Peaks of activity (in red) index the locations of targets and specify the retinotopic coordinates at which the target's feature data are to be found in earlier visual cortices which are shown, highly simplified, as a stack of aligned areas divided into right and left hemifields with the fovea in the center. In object recognition areas, cells have very large receptive fields shown here as a heavy black outline for the receptive field of one cell that specializes in identifying pickup trucks. These cells must rely on attention to bias input in favor of the target and suppress surrounding distractors so that only a single item falls in the receptive field at any one time. The surround suppression has to be imposed in early retinotopic areas as the large fields in object recognition cannot have local modulation of sensitivity.

2.5.2. Capacity of spatial attention

One of the classic definitions of attention (combined with flexible, localized performance benefits) has been its limited capacity. If two tasks execute simultaneously with lower performance than in isolation, they must call on a shared resource, attention (see Pashler (1998), for a review). This is the basis of the dual task paradigm used to evaluate attentional demands of different tasks. The limit is variously described as a bottleneck or limited attentional load, or cognitive resources. We can only attend to a few things at a time, we can only track a few things, we need attention to filter down the incoming information because there is just too much of it.

One of the principle paradigms to explore the capacity of spatial attention has been the multiple object tracking task (Pylyshyn & Storm, 1988; see Cavanagh and Alvarez (2005) for a review). In the initial experiments, accurate performance in this task was limited to tracking 4 or 5 items – a limit that was intriguingly close to other cognitive limits in apprehension and short term memory. Several studies have tested the nature of the information that is actually tracked. For example, targets are suddenly occluded and subjects must report location, direction, or identity of the targets. Location and direction are recalled best (Bahrami, 2003; Pylyshyn, 2004; Saiki, 2003) although some identity is retained if the task requires it (Oksama & Hyöna, 2004). However, further experiments showed that the tracking limit was not so fixed in value as it could range from 1 to a maximum of 8 as the speed of the items to be tracked slowed and the performance showed no special behavior near the magic number 4 (Alvarez & Franconeri, 2007). Not only was there no fixed limit (although a maximum of around 8), but the limit appears to be set independently in the left and right hemifields (Alvarez & Cavanagh, 2005): a tracking task in one hemifield did not affect performance in the other; however if the two tracking tasks were brought into the same hemifield (keeping the same separation between them and the same eccentricity), performance plunged. This hemifield independence seems most evident when the task involves location (Delvenne, 2005).

As a cautionary note, this dual tracking task shows that attention is not a single monolithic resource: performance in the two hemifields shows attentional independence. The concept that attention is a single, limited resource underlies the whole industry of dual task measures of attention demands but this dual tracking task (Alvarez & Cavanagh, 2005) puts this whole industry in question. For example, in an influential series of experiments by Koch and colleagues (Lee, Koch, & Braun, 1997; VanRullen, Reddy, & Koch, 2004), independence between two tasks was taken as fundamental evidence that one of the two tasks required little or no attentional resources. According to the authors, this second task did not affect the first because it did not draw on any attentional resources. However, the dual tracking tasks also show independence but now the two tasks are identical, so their lack of interference cannot be attributed to an asymmetry in their attentional demands. That would be equivalent to claiming that to accomplish the same task, one hemifield is taking all the resources and the other none. Logically impossible. Clearly, attention is not a unitary resource.

If the tracking limit reflects the capacity of spatial attention to index multiple locations, then this flexible value (Alvarez & Franconeri, 2007) and independence between hemifields (Alvarez & Cavanagh, 2005) rule out the classic notion that there is a fixed number of slots for attention (or awareness), at least for attention to locations. In any case, there does appear to be some resource that limits the number of locations that we can attend to or be aware of. We might ask, what is this resource? How can we get more of it? Could we lighten our attentional load? One possibility is a physical rather than metaphorical resource: real estate, specifically cortical real estate. On the attention/saccade maps (Fig. 6) each activity peak – each attentional focus – selects a spatial region for processing benefits and engages surround suppression (Bahcall & Kowler, 1999; Cutzu & Tsotsos, 2003; Mounts, 2000) to prevent selection of nearby distractors. There is a finite amount of space on the attention map and if there is more than one attended target, these suppressive surrounds can produce mutual target–target interference if one activity peak falls in the suppressive surround of another. This target–target interference may be a key factor limiting the number of locations that can be simultaneously attended (Carlson, Alvarez, & Cavanagh, 2007; Shim, Alvarez, & Jiang, 2008). The limited resource is therefore the space on the attention map over which attended targets can be spread out without overlapping their suppressive surrounds. Once the suppressive surrounds overlap target locations, performance is degraded and the capacity limit has been reached.

2.5.3. Resolution of spatial attention

An additional limit to selection arises if two objects are too close to be isolated in a single selection region. When items are too close to be individuated – when they cannot be resolved by attention – they cannot be identified, counted or tracked (He, Cavanagh, & Intriligator, 1996; Intriligator & Cavanagh, 2001). Attentional resolution is finest at the fovea and coarser in the periphery, like visual resolution, but 10 times or so worse so that there are many textures where we can see the items, they are above visual resolution, but we cannot individuate or count them. Our attentional resolution is so poor that if our visual resolution were that bad, we would be legally blind. There is an equivalent, coarse limit for the temporal resolution for attention as well. Events changing at rates of higher than 7 Hz cannot be individuated (Verstraten, Cavanagh, & Labianca, 2000) even though the presence of changes can be detected up to 50 Hz or more (see Holcombe, 2009).

2.6. Architecture of attention: non-location aspects

2.6.1. Features

Spatial attention is intensively studied at behavioral and physiological levels because of its accessible anatomical grounding in

the saccade control centers. Feature-based attention is less studied but equally important (see Carrasco (2011) and Nakayama and Martini (2011) for more details). Feature attention provides access to locations based on features but does so across the entire visual field (Maunsell & Treue, 2006; Melcher, Papathomas, & Vidnyánszky, 2005; Saenz, Buracas, & Boynton, 2002). Aside from this intriguing property of spatial non-specificity and a great deal of research on which features can drive this response, little is known yet about the centers that control it, the specificity of the projections from those centers to earlier visual cortices, nor how those projection then promote the locations of the targeted features to activity on the saccade/attention salience map (producing "pop-out", see Wolfe & Horowitz (2004)).

2.6.2. Binding and object files

An attention map (Fig. 6) may specify where targets are, and so provide access to that target's features, but is that all there is to the "binding problem"? Treisman (1988) proposed that this binding – the bundling together of the various features of an object – was accomplished by attention on a master map of locations that famously glued together the features found at those locations on independent feature maps. This suggestion was followed by many articles that supported and challenged it (see Nakayama & Martini, 2011). Indeed, some authors proposed that co-localization was all that was happening (Clark, 2004; Zeki, 2001; Zeki & Bartels, 1999). Specifically, features that were concurrently active in different, specialized areas of the visual cortex were "bound" together by default, by virtue of being co-localized – having the same position on the various retinotopic cortical maps for different features (Melcher & Vidnyánszky, 2006). Our description of attentional pointers (Cavanagh, Hunt, Afraz, & Rolfs, 2010) provides a location to be co-localized to (Fig. 6) and the set of features within the attended location specified by the pointer are "bound" in the sense that they are read out or accessed together. This version of binding by co-localization with an attentional pointer differs from Treisman's original proposal only in the absence of some "glue", some sense in which the features are linked together by more than just retinotopic coincidence. Indeed, there may be more going on than just co-localization and the extra piece to this puzzle is provided by another suggestion of Kahneman et al. (1992), that of the object file. This is a temporary data structure that tallies up the various features of an object, specifically an attended object. Once the location of an object is specified, its characteristics of location, identity and features can be collected. This is a hypothetical construct but critically important for bridging the gap between a target's location and its identity. This data structure, wherever and whatever it is (see Cavanagh et al., 2010) may provide the difference between simple localization, perhaps equivalent to the "proto-objects" of Rensink (2000), and truly bound features. Multiple item tracking tasks appear to depend more on the localization functions of attention, perhaps the "proto-object" level, than on the bound position and features of the tracked targets. Tracking capacity is reduced dramatically if subjects must keep track of identity as well as location (Oksama & Hyöna, 2004). Some behavioral evidence of the properties of previously attended (Wolfe, Klempen, & Dahlen, 2000) or briefly attended (Rauschenberger & Yantis, 2001) items also suggests that something observable actually happens to these co-localized features once the "object file" is finalized.

2.6.3. Buffers for task execution

There are many functions lumped together in the current literature as "attention". This is certainly a great sin of simplification that will appear amusingly naïve at some future date, but it is what we do now. We include the processes that maintain contact with the target object – tracking, tracing, and solving the correspondence problem – as part of attention. Many authors also include the data structures and short term memory buffers that keep track of the current task being performed as components of the attentional overhead. Overwhelm any of these with too much "attentional load" (c.f., Lavie, 2005) and processing suffers. At some point these different components and functions will have their own labels.

For the moment, I only point out that these are necessary elements of visual cognition. Object files are candidates for one type of buffer that holds information on current targets. Processing also needs a temporary buffer for other task details required to run the visual routines that do the work. These buffers may reside in the prefrontal cortex or span frontal and parietal areas (Deco & Rolls, 2005; Lepsein & Nobre, 2006; Rossi, Pessoa, Desimone, & Ungerleider, 2009). We can assume that these details – current operation, current sequence of operations, criteria for terminating – take space in a short term memory that may be visual or general. None of the papers of this special issue deal with visual short term memory nor its interaction with attention, an extremely active field but several recent reviews cover these topics (Awh, Vogel, & Oh, 2006; Deco & Rolls, 2005; Funahashi, 2006; McAfoose & Baune, 2009; Smith & Ratcliff, 2009).

To make a little more sense of the very vague notion of routines, I previously proposed that we can divide them (Cavanagh, 2004) into three levels: vision routines, attention routines, and cognitive routines. Let's put vision routines on the bottom rank, as automated processes that are inaccessible to awareness. Some of these might be hardwired from birth (e.g. computation of opponent color responses), others might emerge with early visual experience (e.g. effectiveness of pictorial cues), and still others may be dependent on extensive practice (e.g. text recognition). Attention routines, in contrast, would be consciously initiated by setting a goal or a filter or a selection target and they have a reportable outcome but no reportable intermediate steps. Their intermediate steps are a sequence of vision routines. Examples of attention routines might be selecting a target (find the red item), tracking, binding, identifying, and exchanging descriptions and requests with other modules (Logan & Zbrodoff, 1999). Finally, at the top level of the hierarchy, cognitive routines would have multiple steps involving action, memory, vision and other senses where there are several reportable intermediate states. They are overall much broader than vision itself. Each individual step is a call to one attention routine. Examples might be baking a cake, driving home, or brain surgery. Attention routines divide the flow of mental activity at its boundaries where the content of awareness changes: new goals are set, new outcomes are computed and these enter and exit awareness as one of the key working buffers of these mental tasks. If attention routines are a real component of visual cognition, this accessibility will help catalog and study them.

Summing up, Ullman and colleagues' work on path tracing and region filling and then Pylyshyn and colleagues' work on tracking moving targets have brought new approaches to the study of attention. Various experiments have measured capacity and information properties of this particular type of attention and laid out physiological networks that would underlie their operation (Cavanagh et al., 2010). Kahneman, Treisman and Gibbs's (1992) proposal of object files has filled another niche as a necessary, much desired function with little, as yet, supporting evidence either behavioral or physiological. These many new branches of attention research have shown significant growth over the past 25 years, and are currently the most active area of high-level vision.

3. Surfaces, depth, light and shadow

From the highest level of visual system architecture, we move to the lowest level that may still rely on inference and so can still

be labeled visual cognition: object and scene properties like surfaces, materials, layout, light and shadow. The use of inference here is open to debate however. Some of the analysis at this level could call on bottom up processes like the sequence of filters (receptive fields) underlies holistic face recognition (Tsao & Livingstone, 2008; Turk & Pentland, 1991) and the cooperative networks that converge on the best descriptions of surfaces and contours (Grossberg & Mingolla, 1985; Marr, 1982). These would process retinal input directly, without branching to alternative, context-dependent descriptions based on non-retinal information. There are nevertheless many examples where object knowledge does play a role and these suggest that, at least in some cases, inference is required to, for example, link up surfaces (Fig 7c), or differentiate shadow from dark pigment (Fig. 3).

The first step in piecing together the parts of an object is to put together its contours and surfaces, a process called completion by many if there is only partial information in the image. Many of the properties of grouping and good continuation, studied for a century, contribute to these early steps. Gregory (1972) and others pioneered the use of sparse images that led to filling in with the "best" explanation; cognitive and subjective contours (Fig. 7). Avoiding the label of mid-level vision, Gregory referred to these influences as rules that were neither top-down nor bottom up, but "from the side" (Gregory, 2009). Solid conceptual work in this area was introduced by Kellman and Shipley (1991) in their papers on unit formation: the lawful relation between contours that lead to joining various bits together across gaps and occluders. Nakayama and colleagues (Nakayama, He, & Shimojo, 1995; Nakayama, Shimojo, & Silverman, 1989) underlined the importance of attributing ownership to a contour: it belongs to the nearer surface and pieces of contour of the far surface can link up underneath that near surface (Sadja & Finkel, 1995). Qiu and von der Heydt (2005) added spectacular physiological evidence to this aspect of border ownership showing that some cells in area V2 responded to a line only if it was owned by the surface to its, say, left; whereas other cells would respond to the same line only if it belonged to the surface on the right (see Fig. 8). This is one of the most impressive pieces of physiological evidence for visual functions that depend on the overall visual scene layout remote from the receptive field of the cell.

The choices for how the surfaces are combined are not always logical – a cat may come out impossibly long, for example – but these choices appear to be driven by the priority given to connecting collinear segments that both end in T-junctions (e.g. Kellman & Shipley, 1991). Given this very lawful behavior, we might ask

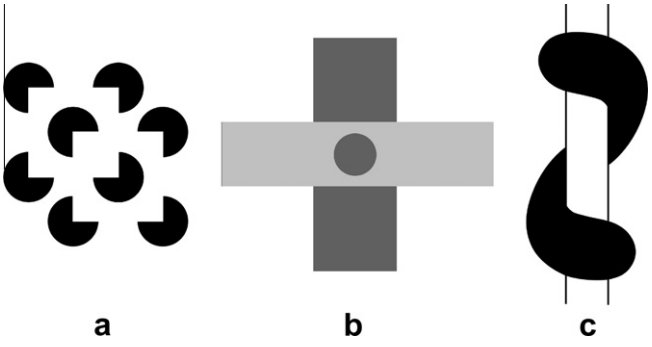

Fig. 7. Cognitive contours, unit formation and relatedness. (a) Gregory (1972) pointed out that we may perceive a shape covering the disks to most easily explain the missing bits of disks. This figure suggests that the collinear edges may be more important than the "cognitive" shape as here the shapes and their depth order are unstable but the subjective contours remain. (b) Kellman and Shipley (1991) proposed a set of principles underlying relatedness that drives the linking of contours and surfaces. (c) Tse (1999a, 1999b) showed that the relatedness was not necessarily at the level of image contours as in this example a volume appears to link behind the cylinder in the absence of any collinear contours.

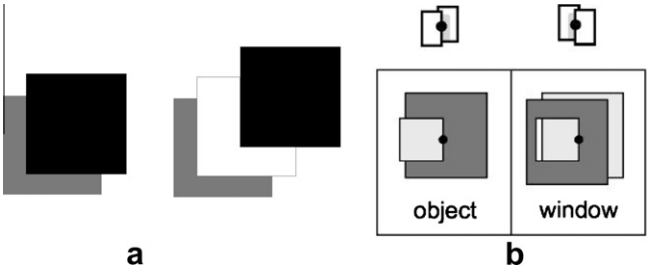

Fig. 8. Border ownership. (a) The front surface owns the border, allowing the back surface to extend under it as amodal completion. The T-junctions here establish the black square as in front, owning the border between the black and gray areas. The gray area completes forming an amodal square so that searching for the image feature – the L shape – is actually quite difficult (He & Nakayama, 1992) (b) Qiu and von der Heydt (2005) report that some cells tuned to orientation are also selective to which surface, left or right, owns the border. One cell may preferentially fire to the border with the object in front on its left whereas another cell may prefer the front surface, as defined only by contour cues, even without disparity information, to be on the right.

whether there is anything really inferential here. Indeed, Grossberg and Mingolla (1985) and Grossberg (1993, 1997) have modeled the majority of these examples within a cooperative neural network that requires no appeal to "object knowledge". However, these straightforward examples give a restricted picture of the range of completion phenomena. Tse (1999a, 1999b) has shown that there is quite good completion seen for objects that have no collinear line segments and that appear to depend on a concept of an object volume even though it is an arbitrary volume. Clearly, there is more going on here than can be explained by image-based rules (Fig. 7c). Some consideration of potential volumes has to enter into the choice. According to Tse (1999a, 1999b) object knowledge here can be as minimal as the property of *being* an object – having a bounded volume – and not necessarily on characteristic properties of a recognized, familiar object.

One critical principle contributing to the inferences of 3D surface structure is the distinction between generic vs accidental views (Freeman, 1994; Nakayama & Shimojo, 1992). One surface that overlaps another will be seen to make T-junctions at the points of occlusion from the great majority of viewing angles. A cube has a generic view with three surfaces visible, the side (2 surfaces) or end views (1 surface) are accidental directions and of much lower frequency from arbitrary viewpoints. This generic view principle helps reduce the number of possible (likely) interpretations for a given image structure.

Similar examples of the importance of object or scene knowledge are seen in the processing of shadows. In extreme examples like Mooney faces or other two-tone images (look at Fig. 3 again), these are simply dark regions with nothing that particularly specifies whether they are dark pigment or a less well illuminated part of the scene. In this case, a first guess of what object might be present is required to break the ambiguity of dark pigment vs dark shadow as no other image analysis based on parts or surfaces can work as shadow boundaries have broken actual object parts into accidental islands of black or white (Cavanagh, 1991). Two-tone representations do not occur in nature scenes but they are nevertheless readily recognized by infants (Farzin, Rivera, & Whitney, 2010) and by newborns (Leo & Simion, 2009). This suggests that the objects are not recovered by specialized processes that have been acquired to deal specifically with two-tone images, which newborns are unlikely to have encountered, but by general purpose visual processes capable of disentangling dark shadow and dark pigment based on object knowledge. These processes would evolve for ordinary scenes where there are often redundant cues to help dissociate dark shadow from dark pigment. In the case of two-tone images, however, only object-based recovery is capable

of extracting shadowed objects. Two-tone images are useful tools that can give us access to these mid-level inferential processes in isolation.

Once a shadow has been identified as such, it provides information both about spatial layout and illumination. The separation between the object and its shadow influences the object's perceived 3D location in the scene as shown by Pascal Mamassian and colleagues (Kersten, Knill, Mamassian, & Bülthoff, 1996; Mamassian, Knill, & Kersten, 1998). The processes linking the shadow and the object are, however, quite tolerant of discrepancies (Fig. 9) that are physically impossible (Cavanagh, 2005; Ostrovsky, Cavanagh, & Sinha, 2005). The information that a dark region is a shadow also contributes to processes that recover the surface reflectance (Gilchrist et al., 1999; see Kingdom, 2011). Correcting for the illumination only recovers relative reflectance – which area inside the shadow may have similar reflectance to areas outside the shadow. An additional process is required to assign absolute reflectance – which area actually looks white as opposed to gray. Gilchrist has shown that certain image properties lead to an assignment of white in general to the most reflective surface and this acts as an anchor so that other surfaces are scaled accordingly (Gilchrist et al., 1999).

Summing up, Gregory and others established sparse figures, subjective contours and completion phenomena as a fruitful workshop for studying principles of surface and object construction. Kellman and Shipley (1991) demonstrated how contour relatedness could support the specification of which surfaces belonged together, a process they called unit formation. Nakayama and Shimojo (1992) emphasized the concept of border ownership and generic views as a key step in understanding surfaces and how they are arranged and joined. van der Heydt (Qiu, Sugihara, & von der Heydt, 2007; Qiu & von der Heydt, 2005) demonstrated that there was evidence in the visual cortex for these processes of extracting subjective contours and assigning border ownership. Grossberg (1993, 1997; Grossberg & Mingolla, 1985) showed that neural networks could solve many of these same surface completion puzzles based on simple boundary and surface systems that interact. Tse (1999a, 1999b) demonstrated that completion extended to more complex situations, relying on object properties that went beyond image-based continuity. Gilchrist extended the resolution of image ambiguity into the domain of lightness (Gilchrist et al., 1999).

4. Objects

What is an object? An object is the fundamental component of visual processing; it is the lynchpin on which so much else hangs. But, embarrassingly, no one has a good definition (see Feldman, 2003; Palmeri & Gauthier, 2004; Spelke, 1990; Scholl, 2001). The definition may be lacking but the research is not (see excellent review in Walther & Koch, 2007). In addition to objects, we may also need a category for "proto-objects" (see Rensink, 2000), the status of segmented potential objects available prior to selection by attention. The necessity of this level of representation is clear when we consider that object-based attention can only exist if objects exist so that attention can access them (Moore, Yantis, & Vaughan, 1998). A second piece of evidence for proto-objects is the ability of humans and other species to make rapid judgments of approximate number of elements in a scene (Dehaene, 1992, 1997; Halberda, Sires, & Feigenson, 2006). The judgments of number are not affected by large variations in the sizes, brightness or shapes of each item suggesting that each item must be segmented from the background and treated as an individual element (Allik & Tuulmets, 1991) prior to access by attention and independently of whether the inter-element spacing allows individuation of the elements by attention. It is not clear yet what the differences are between this pre-attentive object representation and the post-attentive representation.

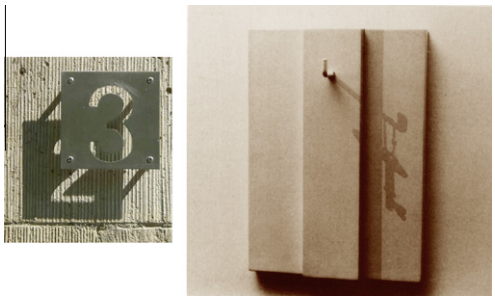

Fig. 9. Shadows. A shadow region is taken as a change of illumination not a change in pigment. These inferences of light and reflectance are made in these two examples even though the two shadows are obviously impossible (Ostrovsky et al., 2005).

4.1. Object structure

Several authors in computer vision proposed that the various junctions on solid and curved objects form a set of constraints that determine the final volume bounded by these contours and junctions (c.f., Barrow & Tenenbaum, 1981; Malik, 1987). This approach was very much bottom up, making no call on knowledge of potential objects, only on the regularities of the junctions and constraints they impose on 3D structure. The work in this area was detailed and analytical but despite the clarity of the proposals, or perhaps because of it, the difficulty in extracting the initial contour description from the image ended the efforts in this area (although see Elder, 1999). Others have worked on the fundamental nature of objects whereby the concave and convex extrema around an object boundary are a diagnostic code of the object shape. Richards and Hoffman (1985) called this the codon theory and the importance of these two boundary features has been followed up more recently by Barenholtz, Cohen, Feldman, and Singh (2003).

4.2. Object recognition

Others worked on the structure of an object and its parts as a code for known objects, allowing retrieval of more object knowledge to fill in details of the object missing in the image. Marr and Biederman among others have stressed the power of an object-description format that can be easily extracted from the image and compared to memory. They considered objects to be a compendium of parts: either simple cylindrical volumes (Marr, 1982) or a set of basic volumes (Biederman, 1987) or more flexible volumes (superquadrics, Pentland, 1987). The object description was given by the spatial relation among these parts: what was joined to what and where. These simplified objects captured some inner essence of objects and were often quite recognizable, in the same way that Johansson's animated point-light walkers were compellingly walking humans. There were again issues about how exactly to get to the object descriptions from the image data but the importance of this part-based level of object description was clear and these proposals have had enormous influence.

The basic approach of these volumetric object schemes is to have an object description that is view invariant. The parts are detected independent of view direction and their structure is coded in an object-centered reference frame. The code therefore solves the problem of how to identify objects from many different viewpoints. On the other hand, there is evidence that object recognition by humans shows viewpoint dependence (Rock & DiVita, 1987). Some proposals do suggest viewpoint dependent representations

and these proposals base object recognition on 2D views (Bülthoff, Edelman, & Tarr, 1995; Cavanagh, 1991; Fukushima, 1980; Logothetis, Pauls, Bülthoff, & Poggio, 1994; Poggio & Edelman, 1990; Sinha & Poggio, 1996). This of course requires that multiple views of each object can be stored and can be matched to image data independently of size or location.

4.3. Context

One consistent result is that objects (and scenes) appear to be processed from a global level to local. According to Bar (2004) some low spatial frequency information is sufficient to generate some gist or context (Oliva & Torralba, 2006) that acts as a framework to fill in the rest (Henderson & Hollingworth, 1999). Bar has demonstrated this progression with priming studies as has Sanocki (1993). This order of processing effect is perhaps different from the order of access effect – the reverse hierarchy proposal (Hochstein & Ahissar, 2002) – whereby high-level descriptions are more readily available for visual search and/or conscious inspection. For example, we see a face before we can inspect the shape of its constituent features (Suzuki & Cavanagh, 1995). The reverse hierarchy proposal does not require that high-level descriptions are computed first, although it does not rule that out either.

4.4. Object benefits

Finally, others have explored behavioral consequences of "objecthood". Scholl, Pylyshyn, and Feldman (2001) used a multiple object tracking task to examine what features are essential for good tracking – with the assumption that good tracking required good objects (Fig. 10). They found that targets that were connected to others or that flowed like a liquid (VanMarle & Scholl, 2003) were difficult to track. Franconeri, Bemis, and Alvarez (2009) followed a similar approach but asked what properties led to more accurate numerosity estimates. Judgments of numerosity are very relevant because they call on an early segmentation of the scene into objects or proto-objects so that the numerosity is independent of the perceptual properties of the items: their size or brightness or shape or organization. Numerosity was affected, however, by the same manipulations that influenced tracking – objects that appeared to connect to others appeared to be less numerous. Finally a series of studies examined what constituted an object so that it could cast a shadow or have a highlight (Rensink & Cavanagh, 2004). The studies exploited visual search tasks that showed a search cost for detecting an odd angled shape when it was seen as a shadow compared to when it was seen as pigment. The cost was eliminated if the object casting the shadow had no volume that could cast a shadow. These studies show that even though we do not know what an object is, we may be able to catalog the instances where "object-like" entities produce processing advantages (or disadvantages).

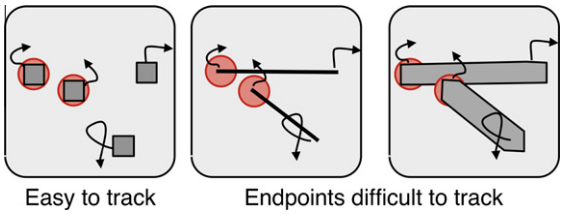

Fig. 10. What is an object that it can be tracked (Scholl et al., 2001). What happens if you try to track a part of an item? Can that part be considered an "object" so that you can track it without interference from the rest of the item? This study took a standard tracking display like that on the left where subjects tracked the items initially marked with red (actual trials presented four targets and four distractors). To test the nature of the objects that could be tracked, pairs of targets and distractors were joined as lines or bars (right hand two panels). The end points of the lines or bars moved on exactly the same trajectories as in the standard display and if an end point could be considered an object, tracking should not be affected. In fact, performance plummeted, suggesting that there is an intrinsic object that is the minimum unit on which attention can operate. In this case, the minimum object appeared to be the full line or bar so that a target endpoint had to be defined as a specific end of a particular tracked bar, perhaps doubling the information required for tracking.

To sum up, the concept of an object is notoriously difficult to define. Nevertheless, several very influential proposals have been made to specify how 3D structure of an object can be decoded from its 2D contours, through sets of junction types, or non-accidental features, or convex and concave extrema. Independently of the retrieval of 3D structure, other proposals have addressed the possibility of object identification based on volumetric modeling of the object's part structure or view-dependent prototype matching and this work has led to scores of articles and applications in biological and computer vision. This area has been among the most fruitful domains of vision research in the past 25 years. Others like Bar (2004) have extended the schemata (Bartlett, 1932; Neisser, 1967), frames and scripts (Minsky, 1975; Schank & Abelson, 1977) of context to show how low-spatial frequencies can provide the global, contextual information that facilitates object recognition. Finally, several studies have reverse-engineered object-superiority and object-inferiority effects to explore the space of objecthood: what is an object that it may be counted or tracked or cast a shadow.

5. Motion, action, causality, agency, events

There is more to vision that just recognition of objects in static scenes. The true power of vision is its ability to be predictive, to see things coming before they happen to you. And the most useful information for prediction is of course the motion of objects in the scene. In fact, it is so useful that two separate motion systems appear to have evolved quite independently, one a reflexive low-level system and the other an active, attention-based, high-level system (Anstis, 1980; Braddick, 1974, 1980; Cavanagh, 1992; Lu & Sperling, 1996). The low-level system does not call on inference or other advanced processing strategies but the high-level system does. Rock (1985), for example, showed how ambiguous apparent motion stimuli could be seen in more than one organization depending on cues in the stimulus or instructions. As he suggested, this demonstrated that there was a logic underlying the percept. Like subjective contours, there was a "subjective" motion path, a space–time contour that best explained the partial image data. Other examples of momentum and organization in apparent motion have been used to make similar points (Anstis & Ramachandran, 1987). If the object seen in motion has certain properties these can constrain the interpretation. For example, Chatterjee, Freyd, and Shiffrar (1996) have shown that the perception of ambiguous apparent motion involving human bodies usually avoids implausible paths where body parts would have to cross through each other.

Motion can tell us more than where an object is going, it can also tell us what the object is. The characteristic motions of familiar objects like a pencil bouncing on a table, a butterfly in flight, or a closing door, can support the recognition of these objects. In return, once the object and its stereotypical motion are recognized, knowledge of that motion can support the continuing percept. Like the first notes of a familiar tune, our knowledge can guide our hearing of the remainder of the melody, filling in missing notes. Selfridge (1959) had argued that shape recognition was supported by legions of "daemons" each of which searched for its matching pattern in the scene and signaled when it showed up. In a related paper (Cavanagh, Labianca, & Thornton, 2001), we proposed dy-

namic versions of these agents, "sprites" that would underlie the processing of characteristic, stereotyped motions. "Sprites" are routines responsible for detecting the presence of a specific characteristic motion in the input, for modeling or animating the object's changing configuration as it makes its stereotypical motion, and for filling in the predictable details of the motion over time and in the face of noisy or absent image details. Point-light walkers make this point most compellingly. A human form is easily recognized from the motions of a set of lights attached to a person filmed while walking in the dark (Johansson, 1973; Neri, Morrone, & Burr, 1998). Johansson (1973) proposed that the analysis relied on an automatic and spontaneous extraction of mathematically lawful spatiotemporal relations. However, in the paper on sprites, visual search tasks showed that point-light walkers could only be analyzed one at a time. Perception of this compelling, characteristic motion required attention.

The idea that there is a story behind a motion percept is a simple version of the even more intriguing effects of intentionality and causality. The original demonstrations by Michotte (1946) for causality and by Heider and Simmel (1944) for intentionality have captivated students of vision for decades. These effects demonstrate a level of "explanation" behind the motion paths that is, to say the least, quite rich. It suggests that the unconscious inferences of the visual system may include models of goals of others as well as some version of the rules of physics. If a "Theory of Mind" could be shown to be independently resident in the visual system, it would be a sign that our visual systems, on their own, rank with the most advanced species in cognitive evolution. Well, that has not yet been demonstrated and there have only been a few articles on causality in visual research over the past 25 years (Scholl & Tremoulet, 2000; Scholl & Nakayama, 2002; Falmier & Young, 2008). Many more studies have focused on the perception of intention, agency and the animate vs inanimate distinction, especially in children (Blakemore & Decety, 2004; Rutherford, Pennington, & Rogers, 2006; Schlottmann & Ray, 2010).

Beyond the logic, the story and the intentions implicit in perceived motion lies an entire level of visual representation that is perhaps the most important and least studied of all. Events make up the units of our visual experience like sentences and paragraphs do in written language. We see events with discrete beginnings, central actions and definite end points. This syntactic structure of the flow of events undoubtedly influences how we experience the components within an event as closely spaced in time just as the Gestalt laws describe how we see grouped items as closer together in space than they are. One lab has been responsible for the major portion of research on visual events (Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Zacks & Tversky, 2001) and has been able to show a number of fundamental properties arising from our processing of elements grouped together over time as events.

Summing up, the phenomenology of motion perception has been one of the richest sources of examples for high-level vision: bistable organizations that undergo dramatic reorganization under the influence of object knowledge, attention and instruction. There is evidence of high-level motion codes that participate in the recognition of objects and the animation of perceived motion. Finally, there is great promise for new research in causality and agency and event perception. In other words, not much has happened yet, but these areas are nevertheless at the center of high-level processes and will clearly get more attention in the coming years.

6. Conclusions

While there has been remarkable progress in high-level vision over the past 25 years, it is perhaps worthwhile pointing out that many of the major questions were identified much earlier. They certainly formed the core of Gestalt psychology (see Rock & Palmer, 1990). These phenomenological discoveries – subjective contours, ambiguous figures, depth reversals, visual constancies – have filled articles, textbooks, and classroom lectures on philosophy of mind and perception for the last 100 years and in some cases much more. What has changed over the past 25 years is the degree to which implementations and algorithms have been developed to explain these high-level effects. In particular, by the mid-1980s, the pioneering work in computer vision (Barrow & Tenenbaum, 1981) and the cognitive revolution (Neisser, 1967) had ignited a ground fire of exciting advances and proposals. These peaked with the publication of Marr's book in 1982 and Irv Biederman's Recognition-by-components paper in 1987. Work on object structure, executive function (memory and attention) and surface completion have kept mid- and high-level vision active since then but the pace has perhaps slowed between the mid 1990s and 2010. In its place, driven by brain imaging work, many labs have focused on localization of function and on the interactions of attention and awareness. Attention itself attracts an ever increasing amount of research, triggered by early work of Posner (1980) and Treisman (1988) and the active attention contributions of Pylyshyn (1989) and others and now the ever more detailed physiological work (Awh et al., 2006; Treue, 2003). At some point, we will have to become a bit more clear on what exactly is attention and then it is likely that mid- and high-level vision approaches will more fully participate in the vast enterprise of attention research.

So what is visual cognition? On the large scale, visual processes construct a workable simulation of the visual world around us, one that is updated in response to new visual data and which serves as an efficient problem space in which to answer questions. The representation may be of the full scene or just focused on the question at hand, computing information on an as-needed basis (O'Regan, 1992; Rensink, 2000). This representation is the basis for interaction with the rest of the brain, exchanging descriptions of events, responding to queries. How does it all work? Anderson's work on production systems (c.f. Anderson et al., 2004, 2008) is a good example of a possible architecture for general cognitive processing. This model has sets of "productions", each of them in an "if X, then Y" format, where each production is equivalent to the routines mentioned earlier. These respond to the conditions in input buffers (short term memory or awareness or both) and add or change values in those buffers or in output buffers that direct motor responses. This production system architecture is Turing-machine powerful and biologically plausible. Would visual processing have its own version of a production system that constructs the representation of the visual scene? Or is there a decentralized set of processes, each an advanced inference engine on its own that posts results to a specifically visual "blackboard" (van der Velde & de Kamps, 2006) constructing, as a group, our overall experience of the visual world? This community approach is currently the favored hypothesis for overall mental processes (Baars, 1988; Dehaene & Naccache, 2001) and we might just scale it down for visual processes, calling on multiple specialized routines (productions) to work on different aspects of the image and perhaps different locations. On the other hand, the very active research on visual attention hints that there may be one central organization for vision at least for some purposes.

Clearly, the basic architecture for vision remains a central prize for the next 25 years of vision research. More specifically, that is the challenge if there is a true inferential architecture for vision. The alternative is that high-level vision is executed as a vast table-look up based on and interpolated from stored 2D views (e.g. Bülthoff et al., 1995). Something like this is found for face recognition (see Ungerleider, 2011) where filters and the closest match in a face space, perhaps biased by expectations, seem adequate to ex-

plain the recognition of individuals (Freiwald, Tsao, & Livingstone, 2009; Quiroga, Kreiman, Koch, & Fried, 2008). In other words, as one intrepid reviewer of this paper pointed out, low-level vision approaches may eventually subsume all the functions of visual cognition, for lunch. The game is afoot.

Acknowledgment

The author was supported by a Chaire d'Excellence Grant and a NIH Grant (EY09258).

References

Al-Haytham, I., (1024/1989). *The optics of Ibn Al-Haytham books I–III*. Translated by A. I. Sabra. London: The Warburg Institute.

Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception & Psychophysics, 49*, 303–314.

Alvarez, G. A., & Cavanagh, P. (2005). Independent resources for attentional tracking in the left and right visual fields. *Psychological Science, 16*, 637–643.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7*(13). 14.1–14.10.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036–1060.

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences, 12*, 136–143.

Anstis, S. (1980). The perception of apparent movement. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 290*, 153–168.

Anstis, S., & Ramachandran, V. S. (1987). Visual inertia in apparent motion. *Vision Research, 27*, 755–764.

Awh, E., Armstrong, K. M., & Moore, T. (2006). Visual and oculomotor selection: Links, causes and implications for spatial attention. *Trends in Cognitive Sciences, 10*, 124–130.

Awh, E., Vogel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience, 139*, 201–208.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Oxford: Oxford University Press.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences, 6*, 47–52.

Bahcall, D. O., & Kowler, E. (1999). Attentional interference at small spatial separations. *Vision Research, 39*, 71–86.

Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition, 10*, 949–963.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*, 723–742 (discussion 743–767).

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*, 617–629.

Barenholtz, E., Cohen, E. H., Feldman, J., & Singh, M. (2003). Detection of change in shape: An advantage for concavities. *Cognition, 89*, 1–9.

Barrow, H. G., & Tenenbaum, J. M. (1981). Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence, 17*, 75–116.

Bartlett, S. F. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Blakemore, S.-J., & Decety, J. (2004). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*, 561–567.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research, 14*, 519–527.

Braddick, O. J. (1980). Low-level and high-level processes in apparent motion. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 290*, 137–151.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 5*, 247–260.

Carlson, T. A., Alvarez, G. A., & Cavanagh, P. (2007). Quadrantic deficit reveals anatomical constraints on selection. *Proceedings of the National Academy of Sciences, USA, 104*, 13496–13500.

Carrasco, M. (2011). Visual attention. *Vision Research, 51*, 1484–1525.

Cavanagh, P. (1991). What's up in top-down processing? In A. Gorea (Ed.), *Representations of vision: Trends and tacit assumptions in vision research* (pp. 295–304). Cambridge, UK: Cambridge University Press.

Cavanagh, P. (1992). Attention-based motion perception. *Science, 257*, 1563–1565.

Cavanagh, P. (2004). Attention routines and the architecture of selection. In Michael Posner (Ed.), *Cognitive neuroscience of attention* (pp. 13–28). New York: Guilford Press.

Cavanagh, P. (2005). The artist as neuroscientist. *Nature, 434*, 301–307.

Cavanagh, P., & Alvarez, G. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9*, 349–354.

Cavanagh, P., Hunt, A., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in Cognitive Sciences, 14*, 147–153.

Cavanagh, P., Labianca, A. T., & Thornton, I. M. (2001). Attention-based visual routines: Sprites. *Cognition, 80*, 47–60.

Chatterjee, S. H., Freyd, J. J., & Shiffrar, M. (1996). Configural processing in the perception of apparent biological motion. *Journal of Experimental Psychology: Human Perception & Performance, 22*, 916–929.

Clark, A. (2004). Feature-placing and proto-objects. *Philosophical Psychology, 17*, 443–469.

Cutzu, F., & Tsotsos, J. K. (2003). The selective tuning model of attention: Psychophysical evidence for a suppressive annulus around an attended item. *Vision Research, 43*(2), 205–219.

Deco, G., & Rolls, E. T. (2005). Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology, 76*, 236–256.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1–42.

Dehaene, S. (1997). *The number sense*. New York: Oxford University Press.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Delvenne, J. F. (2005). The capacity of visual short-term memory within and between hemifields. *Cognition, 96*, B79–B88.

Elder, J. H. (1999). Are edges incomplete? *International Journal of Computer Vision, 34*(2/3), 97–122.

Enns, J. T. (2004). *The thinking eye, the seeing brain*. NY: WW Norton.

Falmier, O., & Young, M. E. (2008). The impact of object animacy on the appraisal of causality. *American Journal of Psychology, 121*, 473–500.

Farzin, F., Rivera, S. M., & Whitney, D. (2010). Spatial resolution of conscious visual perception in infants. *Psychological Science* (Epub ahead of print).

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences, 7*, 252–256.

Fodor, J. (2001). *The mind doesn't work that way*. Cambridge: MIT Press.

Franconeri, S. L., Bemis, D. K., & Alvarez, G. A. (2009). Number estimation relies on a set of segmented objects. *Cognition, 113*, 1–13.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature, 368*, 542–545.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience, 12*, 1187–1196.

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

Funahashi, S. (2006). Prefrontal cortex and working memory processes. *Neuroscience, 139*, 251–261.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51*, 771–781.

Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., et al. (1999). An anchoring theory of lightness perception. *Psychological Review, 106*, 795–834.

Gregory, R. L. (1972). Cognitive contours. *Nature, 238*, 51–52.

Gregory, R. L. (2009). *Seeing through illusions*. Oxford, UK: Oxford University Press.

Grossberg, S. (1993). A solution of the figure-ground problem in biological vision. *Neural Networks, 6*, 463–483.

Grossberg, S. (1997). Cortical dynamics of three-dimensional figure-ground perception of two-dimensional pictures. *Psychological Review, 104*, 618–658.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review, 92*, 173–211.

Halberda, J., Sires, S., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science, 17*, 572–576.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of awareness. *Nature, 383*, 334–338.

He, Z. J., & Nakayama, K. (1992). Surfaces versus features in visual search. *Nature, 359*, 231–233.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–249.

Helmholtz, H. von (1867/1962*). Treatise on physiological optics* (Vol. 3), New York: Dover, 1962; English translation by J P C Southall for the Optical Society of America (1925) from the 3rd German edition of Handbuch der physiologischen Optik (first published in 1867, Leipzig: Voss).

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243–271.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*, 791–804.

Hoffman, D. D. (1998). *Visual intelligence: How we create what we see*. NY: Norton.

Holcombe, A. O. (2009). Seeing slow and seeing fast: two limits on perception. *Trends Cogn Sci., 13*, 216–221.

Howard, I. P. (1996). Alhazen's neglected discoveries of visual phenomena. *Perception, 25*, 1203–1217.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology, 43*, 171–216.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201–211.

Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition, 14*, 129–140.

Jolicoeur, P., Ullman, S., & Mackay, M. (1991). Visual curve tracing properties. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 997–1022.

Kahneman, D., Treisman, A., & Gibbs, D. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*, 175–219.

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

Kersten, D., Knill, D. C., Mamassian, P., & Bülthoff, I. (1996). Illusory motion from shadows. *Nature, 4*, 31.

Kingdom, F. A. (2011). Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research, 51*, 652–673.

Kosslyn, S. M. (2006). You can play 20 questions with nature and win: Categorical versus coordinate spatial relations as a case study. *Neuropsychologia, 44*, 1519–1523.

Kosslyn, S. M., Flynn, R. A., Amsterdam, J. B., & Wang, G. (1990). Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition, 34*, 203–277.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Science, 9*, 75–82.

Lee, D. K., Koch, C., & Braun, J. (1997). Spatial vision thresholds in the near absence of attention. *Vision Research, 37*, 2409–2418.

Leo, I., & Simion, F. (2009). Newborns' Mooney-face perception. *Infancy, 14*, 641–653.

Lepsien, J., & Nobre, A. C. (2006). Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. *Brain Research, 1105*, 20–31.

Logan, G. D., & Zbrodoff, N. J. (1999). Selection for cognition: Cognitive constraints on visual spatial attention. *Visual Cognition, 6*, 55–81.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4*, 401–414.

Lu, Z.-L., & Sperling, G. (1996). Three systems for visual motion perception. *Current Directions in Psychological Science, 5*, 44–53.

Malik, J. (1987). Interpreting line drawings of curved objects. *International Journal of Computer Vision, 1*, 73–107.

Mamassian, P., Knill, D. C., & Kersten, D. (1998). The perception of cast shadows. *Trends in Cognitive Sciences, 2*, 288–295.

Marr, D. (1982). *Vision*. San Francisco, CA: W H Freeman.

Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29*, 317–322.

McAfoose, J., & Baune, B. T. (2009). Exploring visual-spatial working memory: A critical review of concepts and models. *Neuropsychological Review, 19*, 130–412.

Melcher, D., Papathomas, T. V., & Vidnyánszky, Z. (2005). Implicit attentional selection of bound visual features. *Neuron, 46*, 723–729.

Melcher, D., & Vidnyánszky, Z. (2006). Subthreshold features of visual objects: Unseen but not unbound. *Vision Research, 46*, 1863–1867.

Michotte, A. (1946/1963). *La perception de la causalité*. (Louvain: Institut Supérior de Philosophie,1946) [English translation of updated edition by T. Miles, E. Miles *The Perception of Causality*, Basic Books, 1963].

Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia, 46*, 774–785.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–280). New York: McGraw-Hill.

Mitroff, S. R., Scholl, B. J., & Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition, 96*, 67–92.

Moore, T., Armstrong, K. M., & Fallah, M. (2003). Visuomotor origins of covert spatial attention. *Neuron, 40*, 671–683.

Moore, C. M., Yantis, S., & Vaughan, B. (1998). Object-based visual selection: Evidence from perceptual completion. *Psychological Science, 9*, 104–110.

Morgan, M. J. (2011). Features and the 'primal sketch'. *Vision Research, 51*, 738–753.

Mounts, J. R. (2000). Evidence for suppressive mechanisms in attentional selection: Feature singletons produce inhibitory surrounds. *Perception and Psychophysics, 62*, 969–983.

Muller, J. R., Philiastides, M. G., & Newsome, W. T. (2005). Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences, USA, 102*, 524–529.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In S. Kosslyn, D. N. Osherson (Eds.), *Frontiers in Cognitive Neuroscience* (2nd ed., pp.1–70 Cambridge, MA: MIT Press.

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research, 51*, 1526–1537.

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science, 257*, 1357–1363.

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception, 18*, 55–68.

Neisser, U. (1967). *Cognitive psychology*. New York: Prentice Hall.

Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature, 395*, 894–896.

Newell, A. (1973). You can't play 20 questions with nature and win. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Oksama, L., & Hyöna, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition, 11*, 631–671.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of lobal image features in recognition. *Progress in Brain Research, 155*, 23–36.

O'Regan, J. K. (1992). Solving the 'real' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology, 46*, 461–488.

Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies in scenes. *Perception, 34*, 1301–1314.

Palmer, S. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5*, 1–13.

Pashler, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Pentland, A. (1987). Recognition by parts. In *Proc. First Int. Conf. Comput. Vision*, London, UK (pp. 612–620).

Pinker, S. (1984). Visual cognition: An introduction. *Cognition, 18*, 1–63.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343*, 263–266.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*, 3–25.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32*, 65–97.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brian Sciences, 22*, 341–365.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*(1/2), 127–158.

Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition, 11*, 801–822.

Pylyshyn, Z. W. (2006). *Seeing and Visualizing: It's not what you think*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2007). *Things and places. How the mind connects with the world*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*, 179–197.

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience, 10*, 1492–1499.

Qiu, F. T., & von der Heydt, R. (2005). Figure and ground in the visual cortex: v2 combines stereoscopic cues with gestalt rules. *Neuron, 47*, 155–166.

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences, 12*, 87–91.

Rauschenberger, R., & Yantis, S. (2001). Masking unveils pre-amodal completion representation in visual search. *Nature, 410*, 369–372.

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

Rensink, R. A., & Cavanagh, P. (2004). The influence of cast shadows on visual search. *Perception, 33*, 1339–1358.

Richards, W., & Hoffman, D. D. (1985). Codon constraints on closed 2D shapes. *Computer Vision, Graphics, and Image Processing, 31*, 265–281.

Rizzolatti, G. et al. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25*, 31–40.

Rock, I. (1984). *Perception*. New York: W.H. Freeman.

Rock, I. (1985). *The logic of perception*. Cambridge, MA: MIT Press.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology, 19*, 280–293.

Rock, I., & Palmer, S. (1990). The legacy of Gestalt psychology. *Scientific American, 263*, 84–90.

Roelfsema, P. R. et al. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*, 376–381.

Roelfsema, P. R. (2005). Elemental operations in vision. *Trends in Cognitive Sciences, 9*, 226–233.

Rosenfeld, A. (1969). *Picture processing by computer*. New York, NY: Academic Press.

Rossi, A. F., Pessoa, L., Desimone, R., & Ungerleider, L. G. (2009). The prefrontal cortex and the executive control of attention. *Experimental Brain Research, 192*, 489–497.

Rutherford, M. D., Pennington, B. F., & Rogers, S. J. (2006). The perception of animacy in young children with autism. *Journal of Autism and Developmental Disorders, 36*, 983–992.

Sadja, P., & Finkel, L. H. (1995). Intermediate-level visual representations and the construction of surface perception. *Journal of Cognitive Neuroscience, 7*, 267–291.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience, 5*, 631–632.

Saiki, J. (2003). Feature binding in object-file representations of multiple moving items. *Journal of Vision, 3*, 6–21.

Sanocki, T. (1993). Time course of object identification: Evidence for a global-to-local contingency. *Journal of Experimental Psychology: Human Perception and Performance, 19*(4), 878–898.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding*. New Jersey: Lawrence Earlbaum Associates.

Schlottmann, A., & Ray, E. (2010). Goal attribution to schematic animals: Do 6-month-olds perceive biological motion as animate? *Developmental Science, 13*, 1–10.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*, 1–46.

Scholl, B. J., & Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science, 13*, 493–498.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multi-element tracking. *Cognition, 80*, 159–177.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*, 299–309.

Selfridge, O. G. (1959) Pandemonium: A paradigm for learning. In D. V. Blake, A. M. Uttley. *Proceedings of the symposium on the mechanisation of thought processes*. Her Majesty's Stationary Office.

Shim, W. M., Alvarez, G. A., & Jiang, Y. V. (2008). Spatial separation between targets constrains maintenance of attention on multiple objects. *Psychonomic Bulletin & Review, 15*, 390–397.
Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition, 9*, 313–323.
Sinha, P., & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature, 384*, 460–463.
Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review, 116*, 283–317.
Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*, 29–56.
Suzuki, S., & Cavanagh, P. (1995). Facial organization blocks access to low-level features: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 901–913.
Thompson, P. & Burr, D. (2011). Motion perception. *Vision Research*.
Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 40*, 201–237.
Treue, S. (2003). Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology, 13*(4), 428–432.
Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31*, 411–437.
Tse, P. U. (1999a). Volume completion. *Cognitive Psychology, 39*, 37–68.
Tse, P. U. (1999b). Complete mergeability and amodal completion. *Acta Psychologica (Amst), 102*, 165–201.
Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*, 71–86.
Ullman, S. (1984). Visual routines. *Cognition, 18*, 97–159.
Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT Press.
Ungerleider, L. G. (2011). Object and face perception and perceptual organization. *Vision Research*.
Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). MIT Press.
van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences, 29*, 37–108.
VanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science, 14*, 498–504.
VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience, 16*, 4–14.
Verstraten, F. A. J., Cavanagh, P., & Labianca, A. T. (2000). Limits of attentive tracking reveal temporal properties of attention. *Vision Research, 40*, 3651–3664.
Walther, D., & Koch, C. B. (2007). Attention in hierarchical models of object recognition. *Progress in Brain Research, 165*, 57–78.
Winston, P. H. (1975). *The psychology of computer vision*. New York, NY: McGraw-Hill.
Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*, 495–501.
Wolfe, J. M., Klempen, N., & Dahlen, K. (2000). Postattentive vision. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 693–716.
Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133*, 273–293.
Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*, 3–21.
Zeki, S. (2001). Localization and globalization in conscious vision. *Annual Review of Neuroscience, 24*, 57–86.
Zeki, S., & Bartels, A. (1999). Toward a theory of visual consciousness. *Consciousness and Cognition, 8*, 225–259.