SEEING THE STRUCTURE OF OBJECTS

BY

EDWIN JAMES GREEN JR.

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Professor Brian P. McLaughlin

And approved by

____

____

____

____

____

New Brunswick, New Jersey

May, 2016

ABSTRACT OF THE DISSERTATION

Seeing the Structure of Objects

By EDWIN JAMES GREEN JR.

Dissertation Director:

Professor Brian P. McLaughlin

This dissertation is about our visual perception of objects and their geometrical properties. I offer an account of visual shape perception, and then apply this account in developing a theory of how vision secures reference to objects.

Chapter 1 provides an overview of the issues to be addressed. Chapters 2 and 3 concern our perception of shape. Specifically, chapter 2 argues that shape perception is *layered*: We perceive objects as having multiple shape properties, and these properties have varying degrees of abstraction. This picture contrasts sharply with certain views of shape representation in the philosophical and psychological literature, which I label *metric views*. Metric views claim, roughly, that vision only explicitly represents certain metric properties of objects, such as location, length, distance, and angle.

Chapter 3 argues that visual shape perception is *mereologically structured*: Roughly, we perceive an object's decomposition into parts, the intrinsic shapes of its parts, and the locations of the joints between parts. I argue that this forms the basis of a type of perceptual constancy—*structure constancy*. Moreover, I argue that this approach embodies a radical departure from views on which the visual experience of spatial properties is wholly viewer-centered.

Chapters 4 and 5 concern object perception. Chapter 4 considers the problem of how a visual representation secures reference to an external object. I argue that the two leading approaches to this problem (which I call the *pure causal view* and the *location-based view*) face serious difficulties. I then argue that part-based visual shape representation plays a crucial role in the mechanism of visual reference-fixing.

Chapter 5 addresses the question of what *counts* as an object for visual perception. More specifically, what types of things does vision pick out and track over time? On one recently popular view, visual processes of selection and tracking are specifically tuned to a class of entities called *Spelke-objects*. I argue that this view is problematic, primarily because it places excessively strong constraints on the geometry and topology of visual objects. I then defend a different account on which visual objects are (roughly) those things that satisfy traditional perceptual organization criteria.

Acknowledgements

This dissertation was written with the support of numerous friends, family members, teachers, and colleagues. I am especially grateful to the members of my dissertation committee: Brian McLaughlin, Susanna Schellenberg, Frankie Egan, Chris Hill, and John Morrison. Each of these individuals has provided invaluable comments, critique, and encouragement over the past several years. Brian deserves special mention here. In addition to directing my dissertation, Brian supervised my undergraduate honors thesis, and has thus offered more than seven years' worth of guidance and mentorship.

My gratitude also goes out to Jacob Feldman, who supervised my cognitive science project. Jacob's feedback has been consistently penetrating and constructive. Moreover, the empirical background he helped me to attain has proved indispensible in writing this dissertation. Chapter 5 grew largely from discussions with Jacob, Manish Singh, and other members of the Visual Cognition Lab.

I have also benefited greatly from discussions with (and, in some cases, written comments from) the following people: David Bennett, Ned Block, Robert Briscoe, Ben Bronner, Elisabeth Camp, David Chalmers, Eddy Chen, Andy Egan, Chaz Firestone, Randy Gallistel, Simon Goldstein, Gabe Greenberg, Zoe Jenkin, Ernie Lepore, Eric Mandelbaum, Bob Matthews, Lisa Miracchi, Casey O'Callaghan, Ron Planer, Zenon Pylyshyn, Jake Quilty-Dunn, David Rose, David Rosenthal, Mary Salvaggio, Manish Singh, Holly Smith, Daglar Tanrikulu, and Dean Zimmerman. Discussions with Ron Planer, in particular, were very helpful in refining my ideas. I'd also like to single out David Bennett, whose detailed comments led to substantive improvements of chapter 2.

Parts of chapter 3 were presented at the 2015 Minds Online Conference, where I received helpful commentaries from John Hummel and Jake Quilty-Dunn.

I’ve been lucky enough to attend graduate school in close proximity to my parents and siblings, and I am thankful for all of their support over the years. Furthermore, discussions with my father—a *distinct* Rutgers-affiliated Edwin Green—have been consistently fruitful (even if I still draw too heavily on null-hypothesis significance testing). I am also deeply grateful to the Sherbow family for providing my homes-away-from-home in Milford, CT and Oceanside, CA.

Finally, I cannot fully express the appreciation that I feel toward my wife, Mandy, for her constant emotional support and occasional editorial work. You’ve put up with me during numerous periods of stress and preoccupation, and have managed to remain loving and compassionate throughout. I couldn’t have done any of this without you.

Some material in this dissertation has been published elsewhere. Chapter 2 is adapted from Green (2015), in *The British Journal for the Philosophy of Science*, while chapter 3 is adapted from Green (forthcoming), to appear in *Mind & Language*.

Table of Contents

List of Figures

Chapter 1

Introduction

How are shape properties represented in visual perception? Moreover, what is the role of shape perception in establishing visual representations of individual objects? In the following chapters, I will address these and related questions.

1. Overview

This dissertation breaks down naturally into two parts. The first part, consisting of chapters 2 and 3, focuses on shape perception. The second part, consisting of chapters 4 and 5, focuses on object perception.

In chapters 2 and 3, I offer a twofold view of how visual experience represents geometrical properties. First, I propose that visual shape experiences represent shape properties in a *layered* manner. According to the this view, visual experience represents an object as having multiple shape properties, and these properties have varying degrees of abstraction. Second, I propose that visual shape experiences represent *compositional structure*. On this view, visual shape experience represents (*inter alia*) an object's *part decomposition*. I also argue that the representation of compositional structure is the basis of an important but rarely studied type of perceptual constancy, which I call *structure constancy*.

Chapters 4 and 5 together develop a view of visual object perception. First, I offer a view of how visual perception manages to *pick out* or *refer to* an object within a given perceptual context. This account draws heavily on the view of visual shape representation developed in Part I. More specifically, I argue that such shape representations play a *reference-fixing* role with respect to visual object representations. Second, I offer a view of the *types* of things that vision picks out. I argue that "visual objects" are best characterized by appeal to Gestalt principles of perceptual organization, rather than more restrictive principles such as three-dimensionality, cohesion, and boundedness (Carey 2009; Spelke 1990).

2. The Perception of Shape

2.1. Layering in Shape Experience

Detailed theories of shape experience are relatively scant in the philosophical literature. Of course, certain *puzzles* pertaining to shape perception have received a good deal of philosophical attention. Many have addressed Molyneux's question (the issue of whether a person blind from birth could, upon having her sight restored, immediately visually identify the shapes of objects). Similarly, many have addressed the nature of shape *appearances* (e.g., whether a slanted coin "looks elliptical" from one's perspective).[1] However, irrespective of how these issues are resolved, important questions about the nature of shape experience remain.

Note that while we often speak of "the shape" of an object, objects in fact have myriad shape properties. Some of these properties are highly determinable (e.g., the property of being a closed figure), while others are highly determinate (e.g., the property of being an equilateral triangle). In geometry, this idea is made rigorous by appeal to the sets of transformations under which shape properties remain invariant. Roughly, if shape

[1] On Molyneux's question, see, for example, Evans (1985), Campbell (1996), and Schwenkler (2013). On shape appearances, see, for example, Noë (2004), Briscoe (2008), Schellenberg (2008), and Hill (2014).

property *A* remains invariant under a wider class of transformations than shape property *B*, then *A* is more abstract than *B*. Thus, the property of being a parallelogram is more abstract than the property of being a square, because any transformation that preserves the latter property also preserves the former, while the converse is not the case.

Given that objects have myriad shape properties, a natural question concerns *which* of these properties we become aware of in visual experience. Do we *only* perceive objects as having highly determinate shape properties? Some version of this idea can be found in Berkeley, who famously suggested that perception does not furnish the "abstract general idea" of triangularity, but only presents us with ideas of particular triangles. On the other hand, do we *only* perceive objects as having more abstract shape properties? While this view is not popular among philosophers, variants of the position can be found in certain psychologists who suggest that visual perception recovers, at most, affine shape (e.g., Todd 2004). In chapter 2, I suggest that neither of these views is right. Rather, we perceive objects as having *multiple* shape properties, varying widely in their degree of abstraction.

For example, when you visually experience a square surface, I argue that you experience it simultaneously as: (i) a surface composed of points located in such-and-such a direction, at such-and-such a distance, and at such-and-such an orientation relative to your line of sight, (ii) a square, (iii) a parallelogram, and (iv) a closed figure. Likewise, when you visually experience an equilateral triangular surface, you experience it as: (i) a surface composed of points located in such-and-such a direction, at such-and-such a distance, and at such-and-such an orientation relative to your line of sight, (ii) an equilateral triangle, (iii) a triangle, and (iv) a closed figure.

I argue that this view is supported both by visual phenomenology and by the empirical evidence. Thus, note that certain shape changes are much more experientially salient than others. For example, a change from an equilateral triangle to a trapezoid is, other things being equal, more salient than a change from an equilateral triangle to a scalene triangle, even if the two changes involve the same amount of alteration to the figure's "local" features (e.g., the coordinates of individual edges or vertices). To a first approximation, I suggest that such differences in salience support the view that the visual experience of shape is layered.

There is also, I contend, compelling evidence that the visual system *extracts* and *uses* abstract shape properties, such as topology, in a number of processing tasks. Consider, for example, the problem of determining paths of apparent motion. Suppose that you see a pair of computer frames in succession. In frame 1, a solid disk occupies the center of the screen. In frame 2, the solid disk has disappeared, replaced by a solid triangle to the left and a hollow ring to the right. There is evidence that under these conditions, perceivers are most likely to see apparent motion from the disk to the triangle, rather than from the disk to the ring. Moreover, further variations suggest that this tendency is due to the fact that the disk and the triangle share the same *topology* (both are solid figures, while the ring is a one-holed figure).

I argue that evidence of this sort motivates the view that the visual system extracts and represents abstract shape properties. Accordingly, the symbolic format used by vision to represent shape must be adequate to explicitly encode such properties. This raises a critical difficulty for certain psychological views on which shape is coded via holistic templates (e.g., Ullman 1996; Edelman 1999). Roughly, the problem is that such views fail to make the explicit the respect in which a solid disk resembles a solid triangle (i.e., sameness of topology). But without a representation that makes this information explicit, we cannot explain why the visual system prefers motion paths that preserve topology to paths that preserve the geometry of a shape's bounding contour (which is shared by the disk and ring).

2.2. Compositional Structure in Shape Experience

A second problem about shape perception is raised by the fact that many objects do not *retain* a single determinate shape over time. Rather, they move *non-rigidly*. Thus, J. L. Austin writes:

> What is the real shape of a cat? Does its shape change whenever it moves? If not, in what posture *is* its real shape on display? (…) It is pretty obvious that there is *no* answer to these questions. (1962: 67)

This issue is entirely general. As a person moves, her arms swing about their joints, and her head often turns back and forth. Likewise, many artificial objects—such as scissors, truck-mounted cranes, staplers, and reclining chairs—move non-rigidly. Nevertheless, as with Austin's cat, there is *also* a sense in which such objects seem to retain their *overall structure* over time. Contrast these cases with the visual experience of someone squeezing a block of clay. In the latter, there seems to be very little of the clay's determinate geometric structure that is preserved. Of course, highly abstract properties like topology are left intact, but besides this, the change seems genuinely *arbitrary*. In chapter 3, I propose an explanation of why certain non-rigid changes strike us as "natural," while others do not.

I suggest that the ability to perceive an object as retaining overall structure across non-rigid changes ought to be considered a kind of perceptual constancy. I call this ability *structure constancy*. Structure constancy thus differs from the more familiar *shape constancy*. Shape constancy involves the ability to perceive an object's *shape* as remaining constant across *rigid* changes, such as rotations relative to the perceiver's line of sight.

I argue that when we exercise structure constancy, we perceptually experience a particular complex property, which I call *compositional structure*. The compositional structure of an object includes, roughly, (i) its decomposition into parts, (ii) the intrinsic shape of each of its parts, and (iii) the locations of the joints between parts. For example, the compositional structure of a human body would include (i) a decomposition of the object into (roughly) head, torso, arms, and legs, (ii) the shape of each of these parts, and (iii) the positions of the joints between these parts. Note that even as the human body moves non-rigidly, the properties specified in (i)-(iii) remain relatively constant. As a person walks, her global shape is constantly changing (due to movement of the arms and legs about their joints), but the intrinsic shapes of her parts and the positions of the joints between them are preserved. Accordingly, when we exercise structure constancy with respect to a moving human body, we perceive these properties as being retained.

As in the case of layering, the view that visual experience represents compositional structure is supported both by introspection and by empirical evidence. First, changes that alter an object's compositional structure are (other things being equal) more experientially salient than changes that preserve it. For example, a change that alters the intrinsic shape of a part is more salient that one that merely alters the precise angles between parts. Second, there is evidence that each of the properties that enter into an object's compositional structure are in fact extracted by the visual system. For instance, there are well-known Gestalt rules for parsing objects into components (Hoffman & Richards 1984). Furthermore, parsing appears to play a critical role in many visual processes, such as the perception of transparency (Singh & Hoffman 1998) and the distribution of visual attention (Vecera et al. 2000; Barenholtz & Feldman 2003). Moreover, observed patterns of shape discriminability also fit well with the view that vision represents compositional structure (Barenholtz & Tarr 2008).

If visual experience indeed represents compositional structure, then this has important consequences for the "code" by which visual experience represents shape. Specifically, I argue that we can draw interesting conclusions about both the *format* and *reference frame* of visual shape experience. Here I will merely state these conclusions (the arguments will come in chapter 3). First, the format of visual shape experience must be adequate to *prioritize* a particular part structure. This, I contend, rules out any strongly imagistic or template-based approach to experiential shape representation. Second, visual shape experience must be at least partly *allocentric*. This rules out a class of approaches to spatial experience that cast our visual awareness of spatial properties as wholly viewer-centered (Jackendoff 1987; Tye 1991, 1995; Prinz 2012).

My conclusion of chapter 3 is that the code of visual shape experience is best construed as a kind of *hierarchical description*. A hierarchical description explicitly represents the mereological relations between an object and its parts, and also explicitly represents the spatial relations that parts of an object bear to one another. I explain how hierarchical description permits the representation of compositional structure, and as such may subserve structure constancy.

3. The Perception of Objects

The second part of the dissertation (chapters 4 and 5) turns to the perception of objects. However, before outlining what I plan to accomplish in these chapters, I should offer some important background on the recent cognitive science of object perception.

3.1. Background: Object Representation in Vision

Object perception has been a popular area of research over the past couple of decades. Although many disputes persist, there is now wide agreement about several key features of visual object representation. Here I will focus on three.

First, there is compelling evidence that objects can be *targeted by visual attention*. In other words, we are able to selectively attend to an object, rather than to the location that object occupies. This hypothesis is supported by a number of different paradigms. For example, suppose that you view a computer display containing two rectangular objects. You are told that two letters will appear on the screen, and your task is simply to judge whether the letters are the same or different. Researchers have discovered that under these conditions, you will be faster if both the letters appear on the *same* rectangle than if they appear on *different* rectangles, even if the spatial distance between them is the same in both cases (e.g., Behrmann et al. 1998; Moore et al. 1998; Marino & Scholl 2005). A plausible explanation for this result is that attention *spreads throughout* a whole object, enabling speeded responses to any target that appears on the object.[2]

Second, there is compelling evidence that objects can be *selected* and *tracked over time*. In other words, the visual system is capable of both *representing* an individual object and *maintaining* that representation over time, even as the object moves about or

[2] In a similar vein, studies using the *flanker interference* paradigm have shown that visual attention cannot *help* but spread throughout an object, even when this interferes with a task in which the perceiver is engaged (see Kramer & Jacobson 1991).

changes features. For example, in the multiple-object tracking paradigm, it is found that perceivers can keep track of multiple objects as they move randomly about a computer screen, even in the presence of a field of distractor objects (Pylyshyn & Storm 1988).[3] Moreover, we are capable of tracking objects as they briefly disappear behind occluders, as long as the disappearance events are physically consistent with gradual occlusion (Scholl & Pylyshyn 1999). There is, moreover, evidence that objects can be successfully tracked across a wide range of feature changes (Bahrami 2003; vanMarle & Scholl 2003; Zhou et al. 2010).

Third, there is compelling evidence that vision often sets up *short-term memory files* for at least some of the objects that it represents. Consider, for instance, the *object-specific preview benefit*. In this paradigm, a target—say, a letter—first briefly appears on an object and then vanishes. Next, after some period of time, either the same letter or a different letter appears and the subject is asked to report its identity. It is found that subjects are fastest at performing this task when the letter that was seen previously reappears on the same object on which it appeared earlier, even if the object has shifted location in the interim (Kahneman, Treisman, & Gibbs 1992). This has been taken to indicate that attending to an object leads the visual system to open up a file in which features of the object can be stored. When a target initially appears on an object, it is automatically stored in the file for that object, and as a result, when the target reappears on the same object, subjects' threshold for reporting its presence is reduced.

Change detection studies have provided further evidence for such short-term memory files. Thus, suppose a subject is initially shown a set of objects, and then, after a

[3] While the majority of studies have uncovered a parallel-tracking limit of about four objects, recent work has shown that at least in certain conditions, perceivers can track up to eight objects (see Scimeca & Franconeri 2015).

brief delay, shown a second set. The second display is either identical with the first, or differs from it by a single feature of one of the objects. The subject's task is simply to indicate whether the items in the two displays are the same or different with respect to a certain feature (e.g., color, orientation, or shape). If a subject is asked to monitor for a change in, say, color alone, then her accuracy declines for set sizes of greater than 4. Interestingly, however, it is found that even when subjects are asked to monitor for changes in *multiple* features (say, color and orientation) accuracy *also* starts to fall off with set sizes of greater than 4 (Vogel, Woodman, & Luck 2001). This indicates that once *one* feature of an object is encoded, there is little cost to encoding *further* features of the same object. This result can be explained on the hypothesis that visual short-term memory is primarily limited by the number of object-specific *files* at its disposal (about 4). However, there is little cost to storing multiple features of an object in the same file.

To explain these and related findings, many have proposed that vision contains a subsystem for representing and tracking particular objects. This subsystem is often called the *object file* system.[4] An object file—as I will use the phrase—contains two components: a *referential* component and a *storage* component. The referential component of an object file is akin to a name or natural language demonstrative. It picks out the object, and can continue to pick out the object over time. The storage component of an object file is akin to a receptacle within which information about the object can be stored. Thus, we can think of an object file as analogous to a labeled folder: The label (referential component) picks out a particular thing, and the folder (storage component) stores information about that thing.

With this as background, I'll now outline the positive aims of Part II.

[4] See, for instance, Carey (2009), Kahneman et al. (1992), Pylyshyn (2007), and Recanati (2012).

3.2. How does Vision Pick Out an Object?

Object files pick out individual objects. But how does this occur? In chapter 4, I develop a theory of how the referential component of an object file—which I call an *attentive visual object representation* (AVOR)—has its reference secured.

One possible view is that an AVOR secures reference to whichever object appropriately causes it to be deployed. Thus, if a coffee mug causes my visual system to deploy an AVOR, then, simply by virtue of this causal connection, the coffee mug counts as the referent of that AVOR. Versions of this view have recently been endorsed by Jerry Fodor and Zenon Pylyshyn (Fodor 2008; Pylyshyn 2007; Fodor & Pylyshyn 2015). Other authors share with Fodor and Pylyshyn the view that visual reference is secured without the aid of descriptive information (e.g., Dretske 1995; Recanati 2012).

I argue that any view of this sort faces critical difficulties. The most obvious problem is *referential indeterminacy*. Thus, what determines that my AVOR refers to the *coffee mug*, rather than to any of the other links in the causal chain leading from the mug to the AVOR? In response to this difficulty, Fodor (2008) and Dretske (1995) offer maneuvers for at least ruling out *proximal links*, such as patterns of retinal stimulation. I argue, however, that such moves will not succeed in resolving a much more difficult (but equally pervasive) type of referential indeterminacy—namely, an indeterminacy between an object and its *parts*. Roughly, a purely causal account of visual reference-fixing has no way of distinguishing visual reference to the *coffee mug* from visual reference to the *handle* of the coffee mug.

If we reject a *purely* causal account of visual reference-fixing, the natural alternative is to develop a *jointly* causal and descriptive model. On such approaches, an

AVOR secures reference to object *both* because the object causes its deployment *and* because it is associated with certain descriptive information about the object. However, I argue that the existing approaches in this vein are inadequate. Almost all such models appeal to some kind of *localization* constraint on visual reference. In other words, to succeed in visually referring to an object, your AVOR must be associated with accurate information about the object's *location*.[5] However, I argue that this proposal is both theoretically implausible and empirically inadequate. In particular, there is strong evidence that perceivers can retain the ability to select and track objects in the absence of accurate location information.

One might think that this result is fatal for the causal-descriptive approach to visual reference-fixing. However, I believe that this would be a mistake. I propose that rather than appealing to descriptive information about *location*, we should instead appeal to descriptive information about *shape*.

Note, first, that it is intuitive that shape perception and object perception should bear a close relation to one another (cf. Schwenkler 2012). Whenever you single out a coffee mug in vision, it also seems that you have *some* experience of the coffee mug's shape. Furthermore, it seems that shape perception plausibly places *constraints* on differentiating an object. For instance, it is difficult to see how a perceiver could succeed in differentiating a coffee mug while representing it as having the size and shape of a Mack truck. In line with this, I believe that at least *approximate* accuracy in shape representation is critical to segregating an object from its surroundings. Moreover, an accurate representation of an object's shape also enables information received from that object can be packaged separately from information received from distinct objects.

[5] See Strawson (1963), Evans (1982), and Clark (2000).

Finally, there is evidence that when perceivers *lose* the ability to process shape (as in severe visual form agnosia), the ability to select objects in visual attention is lost as well (De-Wit, Kentridge, & Milner 2009).

But what *type* of shape representation is adequate to fix the reference of an AVOR? To answer this question, I call on the hierarchical shape descriptions introduced in chapter 3. I argue on the basis of empirical data that such shape representations are plausibly generated preattentively. Moreover, hierarchical shape descriptions also enable us to resolve the troublesome indeterminacy between objects and their parts. Thus, I propose that by integrating our theory of object representation with a viable theory of shape representation, we can make significant progress on the problem of how visual perception manages to pick out an individual object.

3.3. What Counts as an Object for Vision?

In chapter 4, I discuss how the visual system fixes reference to objects. But what *is* a "visual object" in the first place? What *kind* of thing does the visual system select and track over time? I turn to this issue in chapter 5.

Many who have theorized that vision contains an object file system have also suggested that the object file system internalizes certain *principles of objecthood*. These principles are supposed to specify both what it takes for something to count as an object, and also what it takes for two things to count as different objects. The object principles are held to be operative both during object selection (i.e., picking out an object at a time) and during object tracking. In chapter 5, I consider and critique a recently popular view of the object principles. On this view, the object file system is keyed to *Spelke-objects* (e.g., Burge 2010; Carey 2009; Rosenberg & Carey 2009; Spelke 1990). More specifically, the view alleges that the object file system selects and tracks in accordance with (*inter alia*) the principles of *three-dimensionality*, *cohesion*, and *boundedness*. I call this *restrictive view* of visual objects.

The restrictive view can be contrasted with a different approach that has been historically popular in perception research. On this approach, visual objects are best characterized by appeal to traditional principles of perceptual organization. These include both the principles of perceptual *grouping* and perceptual *parsing*. Perceptual grouping principles specify rules for *composing* smaller elements into larger ones, while perceptual parsing principles (also discussed in chapters 3 and 4) specify rules for *decomposing* larger elements into smaller ones. As I explain, such principles are far more permissive than the Spelke-object principles. They permit certain things to count as objects that would not be permitted under the restrictive view. These include both *groups* of things (e.g., flocks of birds or swarms of geese) and *parts* of things (e.g., a person's arm or the handle of a coffee mug).

I argue that the permissive view is superior to the restrictive view. The former fits better both with phenomenology and with the existing empirical evidence. In particular, the permissive view can account for all of the evidence standardly cited in support of the restrictive view, but can also account for additional evidence that the latter cannot.

Chapter 2

A Layered View of Shape Perception[*]

1. Introduction

The ability to conceptualize objects in the scene before our eyes depends in large part on seeing their shapes. It is by seeing the shapes of cars, buses, and motorcycles that you are able to cognize them as cars, buses, and motorcycles, respectively. As such, the question of how visual perception presents shape properties to thought deserves close philosophical scrutiny. In this chapter I'll propose a view of how shape properties are represented both in visual experience and in subpersonal visual processing. My thesis is that, in both cases, shape is represented in a *layered* manner: An object is represented as having multiple shape properties, and these properties have varying degrees of abstraction. Call this the *layered view* of shape perception.

The plan for this chapter is as follows. In section 2, I introduce the distinction between a *metric* property and an *abstract shape* property. Roughly, metric properties depend essentially on certain distance and/or angular measurements, while abstract shape properties do not—they are more qualitative. In section 3, I discuss some views of shape perception in the psychological and philosophical literature. I suggest that on several psychological views, the visual system's *subpersonal representation* of shape is wholly metric (i.e., the visual system only explicitly encodes the metric properties of objects), and that on some philosophical views, the representation of shape in visual *experience* is wholly metric (i.e., only metric properties figure in visual shape phenomenology). In

[*] This chapter is adapted from Green (2015).

section 4, I argue that the visual experience of shape is layered, rather than metric. To preview, my argument is that such layering is necessary in order to explain patterns of salience in the differences among various shape experiences. In section 5, I discuss a host of evidence indicating that the visual system extracts and uses information about abstract shape in a variety of processing tasks. In section 6, I argue that such evidence vitiates the proposal that the subpersonal representation of shape is wholly metric, and weighs in favor of the view that the visual system encodes shape in a layered manner. In section 7, I discuss some evidence concerning the neural underpinnings of abstract shape perception. In section 8, I suggest that the layered view has important implications for the process of concept acquisition.

2. Metric Properties and Abstract Shape Properties

It is common to arrange shape properties according to their relative *stability*, where the stability of a shape property is given by its invariance under geometrical transformation (change).[1] On this construal, shape property *A* is less stable than shape property *B* iff the transformations under which *A* is invariant form a proper subset of the transformations under which *B* is invariant. Thus, for example, the property of being a square is less stable than the property of being a rectangle, which in turn is less stable than the property of being a quadrilateral.

If an object *o* has shape properties *A* and *B*, and *A* is less stable than *B*, then an asymmetric entailment holds between the two: *o*'s having *A* entails that it has *B*, but not vice versa. Thus, *o*'s being rectangular entails that *o* is quadrilateral, but the converse is

[1] This general approach traces back to the mathematician Felix Klein's innovative work in the 1870s (known as the Erlangen program) on the stratification of geometries according to the relative stability of the properties they examine.

not the case. Furthermore, if *o* has shape properties *A* and *B*, and *A* is less stable than *B*, then a transformation cannot cause *o* to lose *B* without also causing it to lose *A*. Thus, if *o* starts out as a rectangle and so as a quadrilateral, then any transformation that causes *o* to cease to be quadrilateral must also cause *o* to cease to be rectangular.

Formally, I'll define the notion of a *metric property* as follows: A property *F* is a metric property iff *F* is invariant only under some subset of the similarity transformations. The similarity transformations include translation (simple change of position), rotation, reflection (change in "handedness"), and uniform scaling (simple change in size). Less formally, we can think of metric properties as properties that fail to survive changes in distances, lengths, and/or angles. They include, for example, being a square (which depends on having four angles of exactly 90°), being a square with a 20-inch perimeter, and being a circle with a 10-foot radius. Metric properties also include features that are much less stable, such as an object's precise location within a frame of reference, which fails to survive even translation or rotation. Thus, one type of metric property that will be particularly important in what follows is the location of a visible surface patch within a frame of reference centered on the viewer (i.e., viewer-centered distance and direction). This property is invariant under *none* of the similarity transformations, so trivially it is invariant under a subset of them.

Correspondingly, I'll define the notion of an *abstract shape property* as follows: A property *F* is an abstract shape property iff *F* is invariant under some proper superset of the similarity transformations. As such, we can think of abstract shape properties as ones that *can* survive at least some changes in distances, lengths, and angles. Properties like being a parallelogram or being a triangle are abstract, because they survive stretching and shearing, both of which alter a figure’s constituent edge lengths and angles.[2] For instance, suppose a parallelogram is stretched along its horizontal axis so that its top and bottom edges double in length. After this transformation, its edge lengths and angles are different, but it remains a parallelogram. Thus, abstract shape properties are more stable than metric shape properties, and asymmetric entailments obtain between the two—e.g., something’s being square entails that it is a parallelogram, but not vice versa.

I’ll concentrate on two types of abstract shape property here: *topological properties* and *affine shape properties*. A topological property is any property that is preserved under all topological (i.e., one-to-one, continuous) transformations. Topological transformations are often called “rubber sheet” transformations, because they include all the deformations one can apply to a rubber sheet—e.g., twisting, stretching, bending, etc. However, they do not include tearing an object in two, poking holes in an object, “filling in” the holes of an object, or “gluing” pieces of the object together. Topological properties include connectedness, an object’s number of holes, and an object’s property of being inside or outside another object. Because of the generality of topological transformations, any two solid figures—e.g., a ball and a block—are topologically equivalent.

An affine shape property is any property that survives affine transformations. Roughly, affine transformations include the similarity transformations along with stretching and shearing along an arbitrary direction.[3] Affine shape properties include:

[2] In a stretch transformation, all points are moved in a direction perpendicular to a fixed axis, and move by an amount proportional to their initial distance from the axis. In a shear transformation, all points of an object are moved in a direction parallel to a fixed axis, and move by an amount proportional to their initial distance from the axis. A shear transforms, e.g., a rectangle into a (non-rectangular) parallelogram.

[3] Formally, an affine transformation is a function $f(x) = \mathbf{A}x + \mathbf{b}$, such that $\mathbf{A}$ is an invertible matrix, x is a coordinatized point, and $\mathbf{b}$ is a position vector. Affine transformations thus include linear transformations with the addition of translation.

collinearity, being straight vs. curved, parallelism, ellipticality, triangularity, being a parallelogram, coplanarity of lines, the number of sides in a polygon, and signs of curvature (concave vs. convex) along the surface of an object (Todd 2004).[4] Since distances and angle magnitudes are not preserved under affine transformation, affine shape properties are more stable than metric properties. Thus, if one surface is a stretching of another surface, then the two are affine equivalent, even though they are metrically distinct. Moreover, since the affine transformations form a subset of the topological transformations, it follows that any topological property also counts as an affine shape property. However, when I refer here to affine shape properties, I'll have in mind properties that are affine invariant but *not* topologically invariant, such as those listed above. Figure 2.1 shows examples of topological transformation and affine transformation.

[4] A helpful way to visualize the types of changes possible under affine transformation is via the close relation between affine transformation and parallel projection: Any affine transformation is equivalent to a composition of at most two parallel projections (Brannan et al. 2012: p. 84). Moreover, many such transformations (though not uniform scaling) can be expressed as a single parallel projection. Thus, affine transformations can be visualized by imagining a parallel projection mapping one plane to another. If a figure *A* is specified on the plane that is the preimage of the mapping, then the figure on the projection plane will differ from *A* by at most an affine transformation.

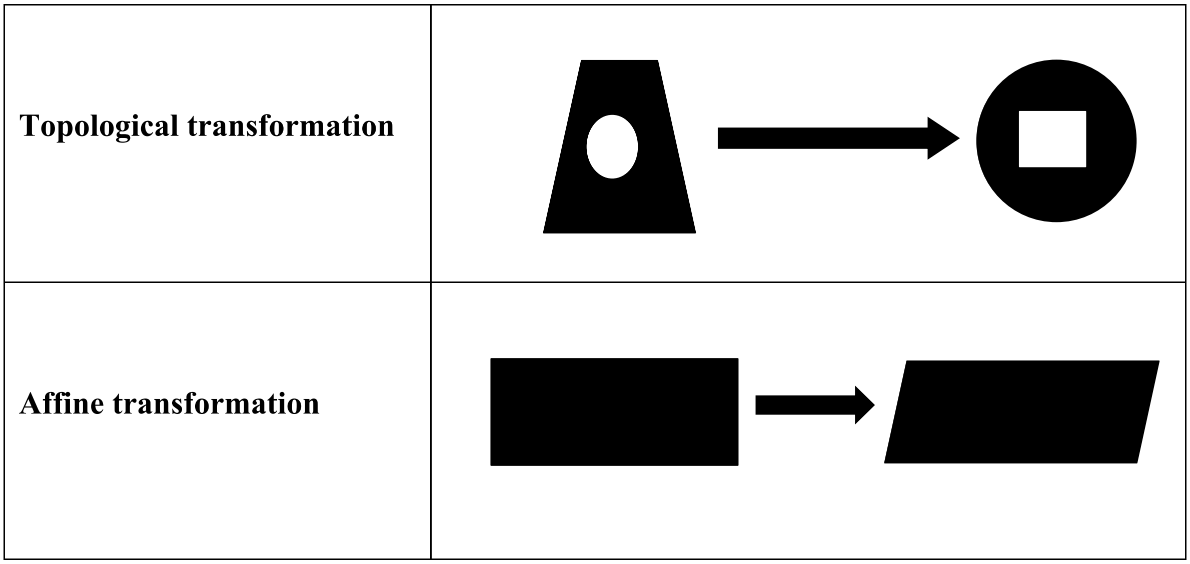


Figure 2.1. Examples of topological transformation and affine transformation

3. Metric Views

Many have been committed to what I'll call *metric views* of either visual *representation* (at the subpersonal processing level) or visual *experience*. This section introduces these positions, in preparation for arguing against them.

I'll construe a *metric representation* of shape as one that only explicitly encodes metric properties, such as locations, distances, lengths, and angles. I won't attempt to offer a reductive analysis of the notion of explicit representation here, but the notion can be intuitively cashed out as follows. When a representation makes certain information explicit, that information is made immediately available for use by the system that uses the representation. By contrast, when a representation leaves certain information implicit, further computations are necessary in order to extract that information (Kirsh 2003). An illustration of this difference is due to David Marr (1982: 20): The Arabic numeral system makes explicit a number's decomposition into powers of ten (e.g., "63" in the Arabic system is equal to $6*10^1 + 3*10^0$), while leaving its composition into powers of two implicit. The binary numeral system, on the other hand, makes explicit a number's decomposition into powers of two (e.g., "1011" in the binary system is equal to $1*2^3 + 0*2^2 + 1*2^1 + 1*2^0$), while leaving its decomposition into powers of ten implicit.

A representative kind of metric representation is Marr's 2½-D sketch, which can be construed as a type of depth map. It is an array specifying the viewer-centered distance, direction, and local orientation at each point (up to a certain resolution) for all visible surfaces in the scene (see Marr 1982: 275-9).[5] The 2½-D sketch is a metric representation because it only explicitly encodes viewer-centered locations and angles (specifically, the locations of small surface patches and the angles of surface normals relative to the line of sight[6]), and these are metric properties.[7] Moreover, substantial computation is needed in order to extract (most) non-metric properties on the basis of a 2½-D sketch. (This is also partly because of the *local* character of the 2½-D sketch—see note 7.) For instance, to extract the abstract shape property 'parallelogram' on the basis of a 2½-D representation of a surface, the system must perform computations to verify, *inter alia*, that the surface has *four straight edges*, that those edges are *connected*, and that two pairs of those edges are *parallel*. None of this information is made explicit by the 2½-D sketch—indeed, the 2½-D sketch doesn't even have the resources for representing parallelism or number of sides. Thus, if the 2½-D sketch encodes abstract shape properties at all, it does so only implicitly.

[5] Because the 2½-D sketch is limited to describing the geometry of *visible* surfaces, it does not include any description of the way surfaces complete behind occluders.

[6] A surface normal at point *p* on a surface is a line segment perpendicular to the plane that is locally tangent to (i.e., "just grazes") the surface at *p*. Marr proposed that local surface orientation at a point *p* is specified in the 2½-D sketch by encoding the angle formed by the surface normal at *p* and the viewer's line of sight.

[7] The 2½-D sketch representation is also *local*—geometrical properties (e.g., location, orientation) are ascribed only to very small elements of the scene, such as small surface patches and edge segments. Some subsequent theorists have rejected the local assumption (see Jackendoff 1987: 331-8), suggesting extensions of the 2½-D sketch that explicitly segment the scene into objects, surfaces, backgrounds, etc.

Call a view on which the visual system represents shape only via metric representation a *metric view* of visual shape representation. Marr himself did not hold a metric view. In Marr's theory, the 2½-D sketch was followed by a 3-D structural description, which represents certain abstract shape properties of objects and their parts (see Marr 1982: ch. 5; Marr & Nishihara 1978; see also Biederman 1987, 2013).[8] Thus, the 3-D model might simply represent an object's part—say, a person's leg—as "roughly cylindrical." Nonetheless, while Marr clearly thought that visual shape analysis was not exhausted by the 2½-D sketch, several subsequent accounts of visual shape representation—primarily in the object recognition literature—have closely resembled the 2½-D sketch in important ways.[9] The most common position in this vein is the so-called *view-* or *image-based* approach (Tarr & Pinker 1989; Ullman & Basri 1991; Ullman 1996, 1998; Edelman 1997, 1999; Williams & Tarr 1999; Riesenhuber & Poggio 2002; Graf 2006). According to most such proposals, the visual system represents an object's shape simply by specifying the numerical coordinates of certain local features of the object (or the object's projected image).[10] For instance, on Ullman's (1996; 1998)

[8] It is interesting to observe, however, that Marr called the 2½-D sketch "the end, perhaps, of *pure* perception" (1982: 268, emphasis added). Steven Pinker (1997: 260) appears to endorse a similar view. It is unclear what Marr had in mind by "pure perception."

[9] Vision scientists *outside* the object recognition literature have often rejected the metric view of shape representation. Notable opponents to the metric view include of course the researchers whose work is discussed below (e.g., Biederman, Chen, Koenderink, Todd, and Wagemans). Furthermore, many psychologists working on perceptual organization have placed emphasis on the perceptual recovery of affine properties such as collinearity and parallelism, and topological properties such as closure and connectedness, since these are important cues to perceptual grouping, figure-ground segregation, and/or amodal completion (see, for example, Feldman 2007; Hoffman 1998; Kellman 2003; Nakayama & Shimojo 1992; Palmer 2003; and Tse 1999; see Wagemans et al. 2012 for review). Vision scientists who study the processes of extracting shape from line drawings, shading, or texture also generally reject the metric view, sometimes in favor of a view on which vision represents affine shape (see Belhumeur et al. 1999; Cole et al. 2009; Koenderink et al. 2001; Phillips et al. 2003; Todd 2004).

[10] Though all such views agree that the coordinate system in which features are specified is viewer-centered, they differ on whether it is 2-D (Ullman 1998; Edelman 1999) or 3-D (Williams & Tarr 1999). Ullman (1996: 110-2) suggests that depth values are used when they are available, but the model he adopts does not require them. This distinction between 2-D and 3-D view-based schemes will not matter for current purposes, since either type of representation is metric in nature.

approach, the representation of shape that serves as input to object recognition is a vector specifying the viewer-centered 2-D locations of simple image elements like edges, vertices, and contour inflection points. An input vector *v* of this sort is recognized as deriving from a particular object *o* if the visual system can obtain *v* by linear combination of a small number of vectors stored in memory that are known to correspond to distinct images of *o*.[11] This proposal, and others like it, shares a critical feature with the 2½-D sketch—namely, it entails that vision only explicitly represents certain metric features of objects, such as the viewer-centered locations of their edges, vertices, etc. (after normalizing for position, rotation, and scale). Transformations of an object outside of the similarity group (e.g., stretching, shearing, or bending) will alter these locations.

The metric view of visual shape representation is a theory about subpersonal visual processing. However, it suggests a counterpart in the domain of phenomenology, which I'll call the metric view of visual shape *experience*. On this proposal, the only geometrical properties represented in visual experience are metric properties, such as the locations and orientations of small surface patches.

Although detailed theories of shape phenomenology are relatively scant in the philosophical literature, the metric view can be found in some authors. For instance, certain passages suggest that Evans (1985) endorsed a version of this view. Evans proposes that visual experience represents shape solely by egocentrically locating the points of visible surfaces in "behavioral space"—a specification of space common to each modality. Thus: "To have the visual experience of four points of light arranged in a

[11] Ullman and Basri (1991) proved that, under certain conditions (e.g., when an object is rigid, all its points are visible in each view, and points are correctly "matched" across images), the vectors of X- and Y-coordinates of points in a specific image of an object (under parallel projection) can be expressed as linear combinations of such vectors in three distinct images of the same object.

square amounts to no more than being in a complex informational state which embodies information about the egocentric location of those lights" (Evans 1985: 339).[12] This indicates that, for Evans, the experience of shape amounts, roughly, to representing the viewer-centered locations of visible surface points.[13]

Peacocke's (1992) notion of *scenario content* bears some similarity to Evans's proposal. According to Peacocke, at the most fundamental level, visual experience represents a *positioned scenario*. This is described as a way of filling out space relative to an origin and axes fixed on the center of gravity of the perceiver's body (1992: 63). More specifically:

> In picking out one of these ways of filling out the space, we need to do at least the following. For each point…identified by its distance and direction from the origin, we need to specify whether there is a surface there and, if so, what texture, hue, saturation, and brightness it has at that point, together with its degree of solidity. The orientation of the surface must be included. So must much more in the visual case: the direction, intensity, and character of light sources; the rate of change of perceptible properties, including location; indeed, it should include second differentials with respect to time where these prove to be perceptible. (1992: 63)

Again, this essentially amounts to a point-by-point representation of surface depth and orientation (though other local features are included as well). And as such, scenario content specifies, in the first instance, metric properties—point-wise distance, direction, and orientation relative to the viewer. However, Peacocke recognizes the need to enrich the scenario content approach in order to account for certain well-known perceptual phenomena.[14]

[12] Page references for Evans (1985) correspond to the reprint found in Noë and Thompson (2002).

[13] A caveat: Evans does not explicitly claim that the approach to shape experience that he endorses for configurations of points of light also holds for experiences of solid figures or surfaces. Thus, it is possible that he would have rejected the metric view as a complete account of shape perception.

[14] Consider, for instance, Mach's tilted-square/regular-diamond figure, which can be seen either as a diamond or as a tilted square. Both of these percepts are compatible with precisely the same scenario content (e.g., viewer-centered distance, direction, and orientation). Because of this, Peacocke introduces a

The layered view is consistent with (but does not entail) the view that visual experiences have scenario content. But if the layered view is right, visual experiences must also have much *more* than scenario content. In particular, visual experiences must represent abstract shape properties in addition to metric properties.

Finally, I should note that the metric view of visual shape experience is quite pretheoretically attractive. It is natural to think of visual experience as simply delivering a pixilated map of the environment specifying the distances and directions of individual surface points. On this picture, it is the job of cognition to "carve up" this map in certain ways and extract abstract shape categories (e.g., triangle, quadrilateral, solid figure, etc.) on that basis.[15]

In what follows I will argue against both types of metric views. I'll first argue that the layered view of visual shape experience does a better job than the metric view of explaining patterns of salience in the differences among shape experiences. Then I'll argue that the metric view of visual shape representation cannot explain the visual system's ability to put information about abstract shape properties to use in a number of processing tasks.

further layer, which he calls "protopropositional content." Protopropositional content is truth-evaluable, and it consists of individuals, properties, and relations. Peacocke suggests that protopropositional content includes the properties of being square, diamond, collinear, curved, parallel, and symmetric. The layered view is consistent with this proposal, though it is consistent with other views as well.

[15] Though this is not the place for historical exegesis, it is worth asking whether Berkeley (1710/1982) held a metric view of visual shape experience, given his famous rejection of the "abstract general idea" of triangularity (Introduction, §13). Though Berkeley allows that a determinate idea of a particular triangle may on occasion function to "stand for and represent" the property of triangularity (Introduction, §15), he seems to have believed that this requires a cognitive act on the part of the subject—one must *use* the idea in a certain way. Thus, though the matter is by no means clear-cut, it seems fair to assume that Berkeley would have agreed that visual experience itself presents us only with determinate metric properties.

4. Against Metric Views of Visual Shape Experience

This section addresses the question of which geometrical properties are represented in the *conscious visual experience* of shape. For present purposes, I will simply assume that visual experiences attribute properties to objects in the environment, and that the representation of such properties makes a difference to visual phenomenology.

As noted above, the metric view of shape experience holds that visual experience presents us *only* with metric properties of objects, such as point-wise distance and orientation. Another view—the one I'll defend here—is that states of visual experience represent geometrical properties at multiple levels of abstraction simultaneously. Call this the *layered view* of visual shape experience. Thus, for example, when viewing a triangular surface, you might simultaneously experience it as: (i) a surface composed of points located in such-and-such a direction, at such-and-such a distance, and at such-and-such an orientation relative to your line of sight, (ii) a triangle, and (iii) a solid figure.

How can we determine which (if either) of these views is correct? Perhaps the most obvious way would be to simply introspect one's experience and see whether it reveals the representation of abstract shape properties in addition to metric properties. Unfortunately, however, the method of introspection faces a number of well-known problems (see Schwitzgebel 2011). Moreover, if I introspect my experience and claim to encounter abstract shape properties while you introspect yours and claim to encounter only metric properties, how can we determine who is right?

A more promising option, it seems, would be to employ the method of *phenomenal contrast*, recently championed by Susanna Siegel (2010). This method works as follows. First, formulate the hypothesis that a property *F* is represented in visual experience. Next, examine two overall experiences, *A* and *B*, such that (i) *A* is a candidate for representing *F*, (ii) *B* is not such a candidate, and (iii) *A* and *B* are as similar as possible in other respects. Then check whether *A* and *B* contrast phenomenally. If they do, then determine whether the proposal that *A* includes a visual experience that represents *F* provides the best explanation of this phenomenal contrast. Critically, this last stage can invoke empirical considerations (see Block 2014), though Siegel does not generally do so.

Siegel uses the phenomenal contrast strategy to defend the view that certain "high-level" properties, such as causation, natural kinds, etc., are represented in visual experience, alongside the usual suspects (color, shape, motion, etc.). Thus, for evaluating this hypothesis, the method of phenomenal contrast recommends that we examine two experiences that are essentially identical in respect of the colors, shapes, etc., that they represent, but perhaps differ in the representation of such high-level properties. If the two experiences differ phenomenally, then (perhaps!) the best explanation is that one represents high-level properties while the other does not.

In the current case, we wish to compare the hypothesis that visual experiences *only* represent metric properties with the hypothesis that visual experiences *also* represent abstract shape properties. Thus, the most straightforward application of the method of phenomenal contrast would be to examine two experiences that are essentially identical in the metric properties that they represent, but perhaps differ in their representation of abstract shape properties. Any difference in shape phenomenology between these two experiences could be taken to support the layered view.

But now we face a problem. Given that any difference in abstract shape entails *some* difference in metric properties (as discussed in section 2), it is *prima facie* plausible that any change in the visual experiential *representation* of abstract shape will entail *some* change in the visual experiential *representation* of metric properties.[16] Accordingly, any pair of phenomenally contrasting experiences that even potentially differ with respect to their representation of a particular abstract shape property (e.g., triangularity) will also plausibly differ with respect to their representation of numerous metric properties (e.g., locations of surface points, edge lengths, angles, etc.). So it seems unlikely that we will be able to find two experiences that uncontroversially agree in their representation of metric properties, but perhaps differ in their representation of abstract shape properties. So how can we identify a pair of experiences that allow us to appropriately compare the two hypotheses of interest?[17]

This is a tricky situation, but I suggest that there is a maneuver available. Rather than looking merely at two individual shape experiences, we can examine *pairs of changes* in visual shape phenomenology, one of which clearly involves a change *only* in the representation of metric properties, and the other of which is a candidate for *also* involving a change in the representation of a given abstract shape property. The hypothesis recommended by the layered view is that, other things being equal*,* changes of

[16] I do not actually endorse the latter entailment, and in fact I suspect that it does not hold (though I admit that it has pretheoretic plausibility). But I aim to show here that even if we grant the entailment, we still have strong reasons to suppose that abstract shape properties are represented in visual experience.
[17] This problem in applying the method of phenomenal contrast is liable to arise in any situation where one wishes to compare two hypotheses, *P* and *Q*, where *P* claims that only *determinates* within a particular category (e.g., scarlet) figure in conscious experience, while *Q* claims that *determinables* within that category (e.g., red) *also* figure in conscious experience. Moreover, I suspect that analogs of the method discussed next—viz., examining patterns of salience among changes in experience, rather than simply comparing two individual experiences—for overcoming this difficulty may be applied in other cases.

the latter type should be *more salient* (i.e., more noticeable or striking) than changes of the former type.

However, on any view of shape experience—including the metric view—certain shape changes should be expected to be more salient than others. For instance, a transformation that stretches a rectangle by a factor of 2 should be more salient than a transformation that stretches it by a factor of 1.5, simply because, e.g., point locations are altered more in the former case. Thus, the claim is *not* that the metric view cannot predict that certain changes will be more salient than others—trivially, it can. Rather, the claim is that the layered view offers a *better* explanation of the *specific* patterns of salience associated with shape changes. This is because the metric view on its own does not predict that the salience of a given shape change should be sensitive to whether or not that change crosses the boundary of an abstract shape category. The layered view, on the other hand, does predict this.

One method, then, would be the following: First formulate a hypothesis about the experience of abstract shape—e.g., "Some visual experiences represent abstract shape property *F*." Then consider the visual experience of a *base stimulus* that has *F*. Next, consider experiences of two stimuli—which we can call *test stimuli*—that meet the following conditions: Both test stimuli differ from the base stimulus in their metric properties, but test stimulus 1 shares *F* with the base stimulus, while test stimulus 2 does not. According to the layered view of visual shape experience, the experience of the base stimulus differs from the experience of test stimulus 2 as regards the representation of (at least) *two types* of properties—both metric properties and abstract shape property *F*—while it differs from the experience of test stimulus 1 only in the representation of metric properties. As such, the layered view would predict that—other things being equal—the former difference will be more salient than the latter.

Holding other things equal, however, is no easy task. Roughly, we want to ensure that the two changes (base stimulus to test stimulus 1 vs. base stimulus to test stimulus 2) are approximately comparable, aside of course from the critical geometrical difference (viz., one crosses the boundary of a relevant abstract shape category, while the other does not). Most importantly, we want to ensure that any difference in the salience of the two changes has to do with perception of abstract shape properties, rather than with detecting differences in very local features, such as point or pixel locations. Perhaps more intuitively, we want to ensure that the change from the base stimulus to test stimulus 2 isn't more salient simply because the two figures have less "overlap" in their constituent points than the base stimulus and test stimulus 1.

Psychologists and computer scientists who have faced this problem have developed measures of the degree to which two stimuli overlap in their local features (see Veltkamp & Latecki 2006). Thus, suppose that we represent figures within a coordinatized frame of reference. A given figure can then be represented by a binary vector indicating, for each point *p* within the reference frame, whether *p* "belongs" to the figure ("1" if the point belongs, "0" if it does not). Given this scheme, one simple way to measure the difference between the overall "point distribution" of two figures (and thus the change between them), called the Hamming distance (Ullman 1996: 5), would be to first normalize two figures to a standard position and orientation, then add up the number of places in which the vectors for the two shapes differ. Another method would be to find, for each point *p* belonging to one figure, the distance from *p* to the closest point belonging to the other shape, and take the maximum of these distances (known as the Hausdorff metric). Yet another method would be to simply sum the distances between each point of one figure and its nearest neighbor in the other figure, which would give a measure of the overall "point displacement" from one figure to the other (again, following normalization).[18]

While it is fortunate that such measures exist, their disparateness may make it seem impossible to "hold other things equal" across two shape changes. Nonetheless, there is a way forward. We can ensure that, no matter which of these measures is used, test stimulus 2 is at least as—if not more—different from the base stimulus in its local features. And luckily, stimuli that obey this restriction have been used in a recent shape discrimination study by Todd, Weismantel, & Kallie (2014). Todd et al. set out to compare the detectability of shape changes at varying levels of abstractness. Subjects were first shown a stimulus—the base stimulus—for 300 ms. Then, after a brief delay, they were shown two other stimuli in succession, each for 300 ms. One of these stimuli was metrically equivalent to the base stimulus, while the other was metrically distinct. The subjects' task was simply to indicate which of these two objects was equivalent to the base stimulus. The metrically distinct stimulus could differ from the base in one of four ways: It could involve a stretching (change in contour length), a skewing causing very slight convergence of contours that were parallel in the base stimulus (loss of parallelism), the addition of a bump to the base stimulus's contour (loss of collinearity), or the introduction of a hole (change in topology). The first type of change disrupts only metric shape, leaving affine shape and topology intact. The second and third types of

[18] Alternatively, we might sum the *squared* distances between corresponding points of the two figures, and take the square root of this sum (known as the Procrustean distance).

changes disrupt affine shape, but leave topology intact.[19] The fourth type of change disrupts topology. Examples of these changes are shown in figure 2.2.

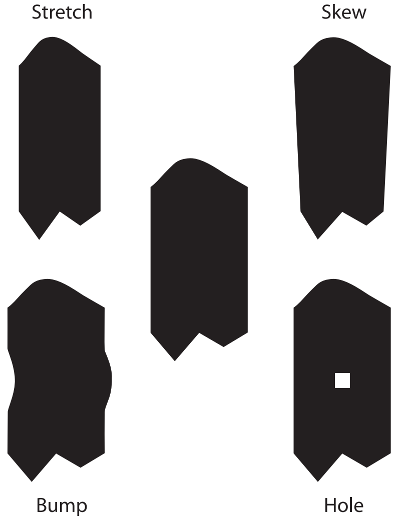


Figure 2.2. Examples of the shape changes used in Todd et al. (2014). The base stimulus is shown in the center. Reproduced with permission of the Association for Research in Vision and Ophthalmology.

Before covering the results of this study, I recommend that you consider your experiences of the stimuli in figure 2.2, and try to decide which changes are most phenomenologically salient. For me at least, the result is fairly clear. The changes that disrupt abstract shape (skewing, adding a bump, or adding a hole) are more salient than the change that disrupts only metric shape (stretching). Indeed, they strike me as 'qualitative' in a way that the latter change does not, even though the overall point displacement (for instance) is actually greater in the stretching transformation. This should make initially plausible the view that abstract shape properties (e.g., parallelism,

[19] Loss of collinearity also alters an object's *projective* properties. A projective property is a property that is preserved under projective transformations. Since the affine transformations form a subset of the projective transformations (Brannan et al. 2012), any property that is preserved under all projective transformations is also preserved under all affine transformations. Thus, every projective property is an affine property, but not vice versa. For present purposes, I focus on the larger set of affine properties, but it is possible that the two types of properties have different degrees of salience in visual phenomenology. Indeed, I find loss of collinearity to be more phenomenologically salient than loss of parallelism. This is borne out in the results of Todd et al. (2014).

collinearity, and number of holes) figure in shape phenomenology alongside metric shape properties (lengths, angles, and curvature).

The results of the experiment comported with this intuition.[20] Todd et al. analyzed subjects' performance in cases where—by almost all common measures of local feature differences between figures, such as those discussed above—the topologically distinct stimulus was *less different* from the base than the affine distinct stimuli, and the affine distinct stimuli were in turn less different from the base than the merely metrically distinct stimulus. It was found that, even in these conditions, subjects were better at performing the task in the affine change conditions (when one of the stimuli involved skewing or adding a bump to the base stimulus) than in the mere metric change condition (stretching), and were better still in the topological change condition (addition of a hole). This indicates that, given two shape changes *A* and *B* such that (i) *A* disrupts an abstract shape category while *B* does not, and (ii) by all or most available measures, the magnitude of local feature difference is either roughly comparable or somewhat greater in the case of *B*, *A* tends to be more salient than *B*.[21]

Now, assuming that subjects perform discrimination tasks like this on the basis of their *visual* shape phenomenology (whether this is the case will be considered shortly), these results raise a challenge for metric views of visual shape experience. For if visual

[20] Moreover, this is by no means the only study to document increased salience for changes that cross the boundary of an abstract shape category. See also the study by Kayaert & Wagemans (2010) described below, along with Amir et al. (2014), Biederman & Bar (1999), and Todd et al. (1998). Comparisons across these studies are admittedly difficult, however, because slightly different measurements of local feature differences across figures were used.

[21] Condition (ii) is crucial, of course. If the metric change (stretching) were made very extreme (e.g., compressing the object to only a few pixels) then it would likely be more salient than the changes in affine shape or topology shown above. But this is not a problem for the layered view. On the layered view, the salience of a particular shape change is predicted to be a complex function of differences in geometrical properties at varying degrees of abstraction—*including* metric properties. As such, if the change in metric properties (lengths, angles, point locations, etc.) is extreme enough, then it should be expected to be more salient than a given transformation that disrupts abstract shape, if the change in metric properties in the latter case is much smaller.

experience does *not* represent abstract shape properties, and instead only represents, e.g., the viewer-centered locations of surface points, then we have no obvious explanation of why changes between objects that alter abstract shape should be especially salient in visual shape phenomenology. But the layered view offers a natural explanation for this.

I'll now consider two potential responses on behalf of the metric view.

First, one might suggest that a version of the metric view could predict the results of the Todd *et al.* experiment without invoking the representation of abstract shape properties if the view simply posited an appropriate subjective similarity function *R* over experiences of metric properties (assuming that discriminability tracks subjective similarity). That is, perhaps experiences represent *only* metric properties such as length, distance, location, and angle, but, by *R*, experiences of metrically distinct but affine equivalent objects turn out to be (other things being equal) more subjectively similar than experiences of affine distinct objects. (Of course, however, *R* could not be based on any of the measures of local feature difference given above.)

I suspect that an appropriate subjective similarity function could indeed predict the results of the Todd et al. study (though to have general applicability the measure would likely need to be forbiddingly complex and context-sensitive[22]). Note, however, that a similarity function *R* over shape experiences would be compatible with *either* the metric view *or* the layered view of the contents of shape experiences. But upon reflection, I think we still have reason to favor the layered view, because the layered view offers a better account of why *R* is the "right" indicator of similarity between shape experiences. On the layered view, the reason why—other things being equal—experiences of objects

[22] There are, it should be noted, numerous factors that seem to affect how similar two shapes are seen to be. One important contributor, which I will not discuss here, is whether two shapes can be decomposed into parts with roughly the same metric structure (e.g., Barenholtz & Tarr 2008).

within the same abstract shape category are subjectively more similar than experiences of objects from different abstract shape categories is because the latter objects are represented in experience to *differ* in that shape category, while the former are not represented to so differ. The metric view, on the other hand, does not have any ready explanation of what *grounds* these facts about subjective similarity. Rather, on the metric view the relevant similarity function would be left brute and unexplained.

A second response for the metric view is recommended by closer attention to the view-based models of object recognition discussed above. According to several view-based models, the representation of shape is *sparse*—it involves simply representing an *n*-tuple composed of the viewer-centered coordinates of "critical features" like vertices, edges, curvature extrema, and inflection points. Perhaps, then, visual experience is sparse in the same way—only the coordinates of such critical geometrical features are represented. Now, importantly, some of the changes in abstract shape used by Todd et al. (2014) (viz., adding a bump or a hole), involved adding *extra* vertices or curvature extrema. As such, on some view-based proposals, this would result in the addition of extra elements to the *n*-tuple specifying object shape. Stimulus stretching, on the other hand, did *not* involve adding an extra critical feature. Perhaps, then, it will be claimed that—other things being equal—changes that result in the addition or subtraction of critical features are more experientially salient than changes that do not. This would explain why some of the abstract shape changes were more salient than the stretching change.

There are a couple of things to note in response. First, observe that the skewing change did *not* increase the number of vertices or curvature extrema in the object, yet was still more salient than stretching. Second, other studies have shown independently that models on which object shape is encoded simply as an *n*-tuple of critical feature coordinates do a relatively poor job of explaining patterns of salience in shape discrimination. Generally, such views predict that the dissimilarity of two shapes should be a function of the distances between their critical feature coordinates. However, studies specifically testing this prediction have found that discriminability is instead more strongly influenced by abstract shape properties of objects (e.g., whether the objects' axes are straight vs. curved) and abstract (or "categorical") relations among parts of the overall shape, such as whether one part intersects another part above vs. below the latter part's midpoint (e.g., Hummel & Stankiewicz 1996; Biederman & Bar 1999). Still, the issues in this area are complicated, so I leave open whether a "sparse" metric view may be able to predict the specific patterns of discriminability found in Todd et al. (2014).[23]

I noted above that studies of shape discriminability seem to favor the layered view, but only on the assumption that subjects perform such tasks on the basis of *visual phenomenology*. However, this assumption may be questioned. Perhaps the difference in salience is rooted not in visual experience, but rather in the way shape properties are cognized. Doubtless we generally categorize objects in thought according to abstract shape properties (e.g., their number of sides), so perhaps shape changes are especially salient when they are accompanied by differences in postperceptual categorization.

[23] Nevertheless, the evidence discussed in the next section and the arguments in section 6 provide, I think, strong reasons to doubt that the view-based approach (including the sparse versions of this approach) can provide a complete account of shape representation at the subpersonal level, although it may give part of the story. If these arguments succeed, then the defender of the metric approach to shape *experience* would then be in the position of explaining why *only* the metric components of subpersonal shape representation subserve visual phenomenology, while other components do not.

Though it is quite difficult to conclusively rule out an alternative explanation of this sort, there are reasons to be skeptical of it.

First, the difference in salience between changes that preserve certain abstract shape properties (e.g., parallelism or solidity) and those that do not simply *feels* perceptual, rather than cognitive. Plausibly, the stimulus with a bump *visually appears* more different from the base stimulus than does the stretched stimulus. This does not seem to be a matter merely of how those stimuli are grasped in cognition.

Moreover, roughly the same patterns of salience have also been obtained with young infants. Kayaert and Wagemans (2010) used a dishabituation paradigm to study affine shape perception in infants and toddlers. The children were repeatedly shown either a triangle or trapezoid. After they habituated to this stimulus, they were presented with a display containing two test stimuli: (i) an object that differed from the original by only a metric change, and (ii) one that differed in affine structure (see figure 2.3). The former was a stretching of the habituation stimulus (but preserved its number of sides), while the latter transformed it either from a triangle into a trapezoid or vice versa. However, these two changes were constrained such that either they involved the same overall point displacement, or the change that preserved affine shape involved a larger difference than the change that failed to preserve affine shape. It was found that even the youngest infants (approximately 3 months) looked significantly longer toward the affine-distinct stimulus than the merely metrically-distinct stimulus, and the size of this effect did not differ significantly between younger and older children.

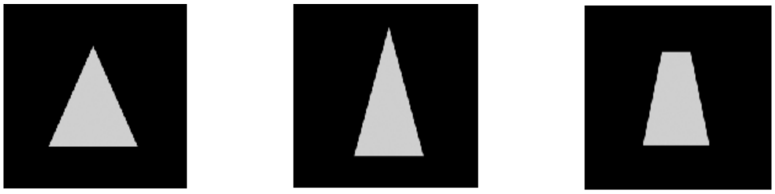


Figure 2.3. Stimuli used by Kayaert and Wagemans (2010). The triangle on the left differs from the triangle in the middle by a mere metric change (stretching) that preserves affine shape, while it differs from the trapezoid on the right in its affine shape. *Source*: Kayaert & Wagemans (2010).

The fact that abstract shape changes are already more salient (other things being equal) in infancy lends some support to the view that this contrast in salience is rooted in perception, because it suggests that the tendency to experience changes in abstract shape as more salient is present very early and is likely involuntary. These are hallmark features of a perceptual process (Fodor 1983; Pylyshyn 1999). Nevertheless, it should be admitted that this evidence is confirmatory, but not conclusive.

However, there is also a large amount of evidence that information about abstract shape is extracted and put to use in a number of *paradigmatically visual* processes, such as apparent motion perception, structure-from-motion, and object tracking. I contend that this, in conjunction with the above observations, provides good reason to believe that abstract shape properties are represented in *visual* experience, and not just in postperceptual phenomenology. I discuss this evidence next.

5. The Visual System Uses Abstract Shape Properties

There is now a great deal of evidence that both topological and affine properties play an important role in visual processing.[24] I begin with topological properties.

One way to test whether a property is extracted during early visual processing, rather than in, say, postperceptual cognition, is to use very short presentation times (e.g., Sekuler & Palmer 1992). The idea is that early removal of a stimulus "interrupts" the processing of that stimulus. Thus, to probe for the perception of topological properties, Lin Chen (1982; 1990) gave subjects a discrimination task in which they were shown pairs of figures for just 5 milliseconds and then asked to indicate whether the figures were the same or different in shape. In one experiment, the pair of figures was drawn from the following set: solid square, solid circle, ring, or solid triangle (see figure 2.4). Crucially, while the circle and the ring are very similar with respect to local metric properties such as contour curvature, they have different topologies (viz., one figure has a hole while the other does not). On the other hand, the solid circle is topologically equivalent with both the solid square and the solid triangle. The crucial measure was the rate of correctly reporting that two figures of different types were in fact different in shape. For if the visual system represents the topological properties of objects (such as their number of holes), then these properties should serve to distinguish the ring from the other figures, but should not serve to distinguish the other figures from one another. As such, one might expect to find *better* discrimination performance in the case of, say, the ring and solid circle than in the case of, say, the solid circle and triangle. And this was indeed found: subjects were significantly better at distinguishing the ring and solid circle

[24] For an overview of evidence in favor of the perception of topological properties, see Chen (2005). For overviews of evidence in favor of the perception of affine properties, see Todd (2004) and Bennett (2012).

(64.5% correct) than either the solid circle and square (43.5% correct) or the solid circle and triangle (38.5% correct). Moreover, this pattern of results continued to hold after differences in spatial frequency, luminous flux (i.e., the total amount of light energy provided by the figures), and area were held constant across topologically equivalent and topologically distinct pairs of stimuli (see Chen 1990).

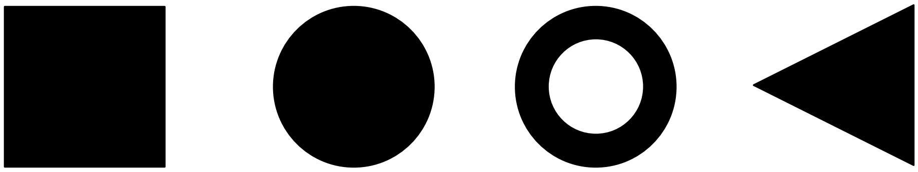


Figure 2.4. Stimuli similar to those used by Chen (1982)

Chen has also examined the role that topological properties play in the perception of apparent motion. As is well known, when one stimulus is flashed and then another is flashed in a different location, then with a suitable spatiotemporal gap between the flashes, the viewer will have a visual experience as of a single object moving continuously from one location to the other. One interesting variant on this paradigm involves presenting multiple stimuli, rather than one, in the second frame. For example, the first stimulus *A* may be followed by a pair of stimuli *B* and *C*. In this case, the visual system faces the problem of "choosing" whether to represent motion from *A* to *B*, from *A* to *C*, from *A* to both *B* and *C* (i.e., "splitting"), or no motion at all. A heavily examined issue concerns which properties the visual system exploits in solving this problem (see Green 1986). If a property is exploited in determining matches in apparent motion, this provides good evidence that the property is represented in vision, because it indicates that the visual system uses representations of the property during motion processing.[25] Thus,

[25] It should be noted, however, that visual motion perception is not a single process, but rather involves a number of different subsystems (see Lu & Sperling 2001).

Chen (1985) showed subjects two frames in succession. In frame 1, a single stimulus occupied the center of the display, and in frame 2, the stimulus was replaced by two stimuli—one to the left of center and the other the same distance to the right of center. Subjects were asked to choose whether they saw motion from center to left or from center to right. For a wide variety of stimuli (and, again, with other differences between the stimuli controlled), subjects were significantly more likely to see motion from the central stimulus to a topologically equivalent stimulus than to a topologically distinct one. Thus, for example, if frame 1 contained a square with a square-shaped hole and frame 2 contained both a solid square and a ring, subjects were significantly more likely to see motion to the ring than to the solid square.

The proposal that topology is used in determining object identity over time has also been verified by a recent multiple object tracking study (Zhou et al. 2010). Subjects were asked to keep track of four stimuli as they moved about the screen in the presence of a set of distractors. The stimuli could undergo various sorts of feature changes during a trial. The critical measure was how such changes impacted subjects' ability to keep track of the stimuli. It was found that changes in topology—though not other feature changes (e.g., changes in color or metric shape)—significantly impaired the ability to track an object over time, indicating that the visual system relies heavily on topology in order to determine whether an object has remained the *same* object from one moment to the next.

Each of these studies provides evidence that the visual system treats objects (or time slices of an object) that are very different in metric properties as nevertheless having certain features in common—namely, topological features. And moreover, the visual system apparently *uses* this information in certain processing tasks, such as, e.g., motion computations.

I turn now to the perception of affine shape properties.

Much of the work on affine shape perception has been motivated by a large group of experimental findings indicating that perceivers' judgments of the metric properties (e.g., length and surface orientation) of objects are quite inaccurate under many conditions (see, e.g., Norman, Todd, & Phillips 1995; Norman et al. 1996).[26] The fact that metric perception is so inaccurate has led some researchers to hypothesize that perhaps the visual system is primarily in the business of producing estimates of more abstract shape properties—namely, affine properties. But how can this hypothesis be tested? One way is to place observers in restricted conditions in which *only* affine structure can be extracted (at least initially), and see whether it is indeed extracted.[27]

[26] Norman, Todd, and Phillips (1995) have found that judgments of surface orientation are inaccurate by an average of 14.5° even when subjects are given very reliable depth cues (e.g., binocular disparity, shading, texture, and motion). And Norman et al. (1996) found that subjects are highly inaccurate when asked to compare the lengths of line segments presented at random orientations in depth, though they are fairly accurate when asked to compare the lengths of nearby parallel lines. In particular, perceived length in depth tends to become progressively more compressed as a function of viewing distance, while length in the frontoparallel plane does not undergo this distortion. Incidentally, the finding that subjects are fairly accurate in comparing the lengths of parallel lines, but not non-parallel lines, lends some support to the view that perceivers represent affine shape. This is because the relative lengths of parallel line segments are preserved under affine transformations while the relative lengths of nonparallel line segments are not.

[27] There is another type of evidence that some have marshaled in support of affine shape perception. In a number of studies, Jan Koenderink and colleagues (e.g., Koenderink et al. 1996; Koenderink et al. 2001) have systematically investigated subjects' perceptual judgments of surface orientation. They have found that while perceivers' metric judgments are quite inaccurate (see note 26), perceived surface geometry nevertheless tends to be affine equivalent with real surface geometry. Specifically, perceived surface geometry tends to correspond to real surface geometry modulo a stretching or shearing in depth.

However, while some have taken these results to indicate that the visual system represents the affine properties of surfaces, I believe that this conclusion is too hasty. Rather, these findings can also be explained on the view that subjects *only* visually represent metric shape properties, but such representations simply tend to be non-veridical in systematic ways. That is, the visual system might non-veridically represent metric shapes that are distinct from, but affine equivalent to, metric shapes in the environment.

Researchers have explored this possibility quite extensively in the structure-from-motion paradigm.[28] In this paradigm, the observer is shown a display of dots or line segments that, when viewed statically, looks like a random 2-D configuration. The elements of the configuration, however, are generated by orthographic projection from elements of a (real or computer-generated) 3-D object. When the elements of the pattern begin to move in a way consistent with the movement of the 3-D object from which they were projected, the viewer spontaneously undergoes a percept of 3-D structure.

How does this happen? Ullman (1979) proved that it is possible to recover metric structure from a rigidly moving 3-D object on the basis of three distinct views (under orthographic projection) of four non-coplanar points of the object. For several years after Ullman's proof, it was assumed that the visual system solves the structure-from-motion problem by analyzing three views of the object and thus extracting precise metric shape.

However, while three views are necessary (and sufficient, assuming rigid movement) for extracting *metric* structure, it has been shown that with only two orthographic views, it is possible to recover the structure of an object modulo a uniform stretching in depth (Todd & Bressan 1990; Ullman 1983). In other words, one can recover the *X* and *Y* coordinates of each object point, but *Z* coordinates can only be recovered up to multiplication by a constant but unknown stretch factor *k*. As such, two views are sufficient to specify *relative* depth and *ratios* of distances along the depth axis, but are insufficient to specify *absolute* distances.

Recall that affine properties are, roughly, those that are preserved under stretching or shearing along an arbitrary direction. These transformations preserve ratios of

[28] For more—and critical—discussion of the evidence in favor of affine shape perception in structure-from-motion, see Bennett (2012).

distances along parallel lines, but disrupt absolute distances. Accordingly, the types of geometrical properties recoverable on the basis of two views in a structure-from-motion display are, roughly (though not exactly), affine properties.[29] Thus, an interesting test case for the proposal that the visual system extracts affine shape is to present an observer with an apparent motion sequence involving just two orthographic projections of elements of a 3-D object (with other depth cues removed), and see whether this produces a percept of 3-D structure. If so, then a reasonable interpretation is that the visual system generated this percept by recovering the affine structure (more specifically, structure modulo an unknown stretch factor along the depth axis) specified by the two views. And indeed, a number of studies have suggested that subjects *do* undergo percepts of 3-D structure under these restricted conditions, and that they can accurately identify aspects of the affine structure of the object. For instance, subjects are able, on the basis of just two views, to discriminate curved from planar surfaces (Norman & Lappin 1992), or determine whether two line segments are coplanar (Todd & Bressan 1990). As such, it is reasonable to conclude that the visual system *can* recover affine shape properties.[30]

But are affine shape properties extracted in the *general* case, when more precise metric information *is* available (at least in principle)? There are theoretical reasons to think that they are. Critically, many affine properties/relations of line segments

[29] The type of structure recovered here is actually slightly more determinate than affine shape. Certain objects that are related by affine transformation *can* be distinguished on the basis of two orthographic views, if they differ by more than a uniform stretching in depth (e.g., a shear or a stretching along either the horizontal or vertical axis). See Todd and Bressan (1990: 421).

[30] For now, I leave aside the issue of whether the structure-from-motion algorithms implemented by the visual system extract *only* affine shape. Todd and his colleagues have argued that the visual system is *incapable* of using more than two views in an apparent motion sequence to extract 3-D structure. This has, however, been challenged (Hogervorst & Eagle 2000; Bennett et al. 2012). Moreover, it is also possible that, on the basis of 2 views, subjects do perceive metric structure, but such percepts are generated on the basis of background heuristics rather than image data. However, even if this is right, it still seems plausible that such percepts are produced by way of prior representation of the affine structure determined by the 2 views (essentially, velocities of image elements).

(collinearity, parallelism, straightness, and curvedness) are more readily computable on the basis of retinal images than metric properties. The reason is that they tend to be preserved under projection to the retina,[31] and moreover they are highly unlikely to arise at the retina by accident of viewpoint. As such, they are often called *nonaccidental properties* (e.g., Biederman 1987). For instance, two collinear line segments in the world will always (discounting noise) project to collinear segments on the retinal plane—and the probability that two non-collinear segments in the world will project to collinear segments on the retinal plane is vanishingly small (Albert & Hoffman 1995). Similar remarks hold for the other properties listed above. As such, detection of such properties at the retina is sufficient for inferring that they are present in the world.[32] By contrast, since metric properties (e.g., lengths and angles) are not preserved under projection, the visual system must do an incredible amount of computational work to recover them.

6. Against Metric Views of Visual Shape Representation

The evidence discussed in section 5 indicates that information about abstract shape properties is likely extracted by early visual processes and used in a number of ways. I now argue that this raises a serious difficulty for metric views of shape representation at the level of subpersonal visual processing.

[31] Parallelism is preserved under parallel or orthographic projection, but is not in general preserved under perspective projection. As such, it is not strictly speaking the case that parallelism is preserved under projection to the retina. However, when the ratio of an object's extension in depth to its distance from the viewer (known as the perspective ratio) is very small—as is often the case—the effects of perspective are negligible, and the projection process approximates parallel projection (see, e.g., Todd 1995).

[32] This is not to suggest, however, that their detection at the retina is trivial. The detection of luminance edges—let alone geometrical relations between them—is an incredibly difficult computational task that has yet to be completely solved. The point, rather, is that such properties are in general *more easily* computable than metric features like depth and surface orientation.

The challenge for the metric view is simply to explain *how it is* that information about particular abstract shape properties is brought to bear in visual processing if visual shape representations do not make such information explicit. For as we saw earlier, 2½-D sketch-style (and other "view-based") representations make explicit only metric information, usually pertaining to individual surface points and edges (e.g., their numerical coordinates in a viewer-centered reference frame). Given just a representation of this sort, how can the visual system make use of the information that a certain object is, e.g., a triangle or a solid figure?

To make the problem more concrete, let's consider again the role that perception of topological properties plays in apparent motion perception.[33] Suppose that frame 1 contains a solid square, and that frame 2 contains both a square with a square-shaped hole and a solid triangle. If shown these two frames in succession, the subject is likely to see motion from the solid square to the solid triangle. And moreover, we have good reason to believe that it is sameness of topology (i.e., the property of being a solid figure) that accounts for this tendency. But for the visual system to compute motion paths in a manner that is selectively sensitive to topological similarities, respects of topological similarity must be *selected* or *highlighted* in contrast to other respects of geometrical similarity. That is, the respect in which the solid square is more similar to the solid triangle (both are solid figures) must be selected over the respects in which the solid square is more similar to the square with a hole (both have square-shaped bounding

[33] The emphasis on topology here is deliberate. Affine structure (or at least relations of affine equivalence across images) is often relatively easy to compute on the basis of image coordinates (see, e.g., Ullman 1996: 208-13). As such, the representation of affine structure perhaps need not require drastic revisions to extant view-based models. Topology, however, is another story. It is well-known that topological properties (e.g., connectedness) are quite difficult to extract because (as can be proved) such properties in general cannot be computed by any set of local procedures that each depend only on a fixed set of points (Minsky & Papert 1969: 12-14; Todd 2005). As such, the perception of topology may require large changes to current models of shape processing (see also Chen 2005).

contours). The problem is that metric representations don’t do this. Rather, in the 2½-D sketch, for instance, information about topology is *implicit alongside* information about, e.g., the lengths and angles of a shape’s bounding contour. As such, this type of representation alone cannot provide the basis for mapping the solid square to the solid triangle rather than the square with a hole.

But what is it to *select* information about a certain abstract shape property? Plausibly, it is just to construct a representation that explicitly encodes the property. As such, it seems likely that visual shape representations explicitly encode information about abstract shape properties, such as topological properties and affine shape properties.[34]

This weighs in favor of a *layered* view of visual shape representation. On this approach, visual representations of geometrical properties are layered in a hierarchy roughly in accordance with the stability of those properties. Thus, when you see a triangular surface of an object, your visual system constructs *numerous* representations arranged in a multi-level hierarchy: The object is represented at one level as having a quite specific metric shape (e.g., a surface composed of points such-and-such a distance away with such-and-such orientation relative to the line of sight), but it is also represented at another level as a triangle, and at a third level as a solid (filled) figure. The explicit representation of such abstract properties can enable them to exert an influence on other visual or vision-based processes, such as motion perception, object tracking, and shape discrimination.

The hierarchical aspect of the proposal is critical. When the visual system represents an object as both square and quadrilateral, these two representations are almost

[34] Or, failing this, visual shape representations must at least explicitly encode *relations* of affine or topological equivalence among objects.

certainly more functionally integrated with one another than either is with, e.g., representing the object as maroon. Plausibly, the former two representations will need to be deployed in many of the same computational processes. Thus, representations of various shape properties should be appropriately related, and their relation arguably should reflect the asymmetric entailments among the properties represented. Hierarchical structure is the appropriate framework for accomplishing this (figure 2.5). Directed edges linking property representations at distinct levels of the hierarchy encode entailment relations between those properties.

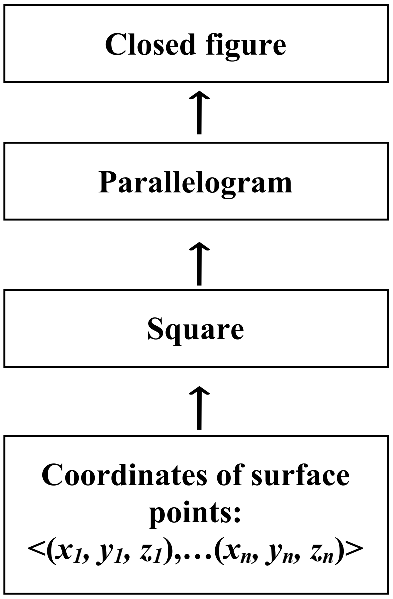

Closed figure
Parallelogram
Square
Coordinates of surface points:
$<(x_1, y_1, z_1),\ldots(x_n, y_n, z_n)>$

Figure 2.5. Hierarchy in accordance with geometrical stability. At the lowest level, highly unstable location features are represented. At the next level the property of being a square (invariant under similarity transformations) is represented. Next, the property of being a parallelogram (invariant under affine transformations). Finally, the property of being a closed figure (invariant under topological transformations).

Before moving on, I should forestall some potential misconceptions.

First, I should emphasize that I am *not* claiming that the visual system fails to construct a representation of metric features, such as the 2½-D sketch (though I remain agnostic about whether this is the best way to represent metric structure). Purely metric changes among affine equivalent objects are certainly registered by the visual system, and can be used to discriminate objects.[35] Indeed, if the visual system did not represent anything more specific than affine shape, then, as Li et al. (2013) note, a pizza box and a shoebox would be visually indistinguishable by shape. I am only claiming here that metric representations cannot *exhaust* the visual representation of shape. Shape representations must *also* explicitly encode information about more abstract shape properties.

Moreover, I am not claiming anything about the *order* in which the visual system extracts geometrical information on the basis of retinal input. Thus, it is possible that metric properties are extracted *prior* to abstract shape properties. This issue lies beyond my scope here. (However, there is reason to think that this is not the actual order of processing—see Chen 2005.)

Finally, the layered view should not be confused with the proposal that metric shape is represented in vision only at coarser levels of *precision* (which is almost certainly correct). For, just as an imprecise representation of being a poodle does not constitute a representation of being a dog, an imprecise representation of the sides and angular measurements of a particular triangle does not constitute a representation of the abstract property of being triangular.

7. Neural Underpinnings of Abstract Shape Perception

Given the strong psychophysical and theoretical support for visual representations of abstract shape, a natural next step is to inquire into their neural underpinnings. Vision

[35] Lee et al. (2012) have shown that a novel target object can be distinguished fairly reliably from a metrically distinct (but affine equivalent) object so long as the viewer sees the target from a variety of perspectives.

scientists have started to take this step, and have so far met with promising results (for a more comprehensive review of this research, see Biederman [2013]).

Studies of both humans and nonhuman primates have produced compelling evidence that higher-level ventral stream neurons are more sensitive to changes in abstract shape than to mere metric changes. Thus, Kayaert, Biederman, and Vogels (2003) recorded the responses of single neurons in the anterior inferotemporal cortex (IT) of rhesus monkeys as they were shown a base stimulus, followed by a set of variations of the base stimulus. The variations included pairs of changes equated in their low-level image differences from the base, such that one change involved a variation in affine shape (e.g., a change from straight sides to curved, or from parallel to nonparallel), while the other involved mere metric change (e.g., a change in aspect ratio or degree of curvature).[36] In approximately 65% of the cases studied, neural responses were altered significantly more by a difference in affine shape than by an equated difference in metric properties (see also Vogels et al. 2001 and Kayaert et al. 2005).

As regards the perception of topology, a recent fMRI study of the lateral occipital cortex (likely the human homologue of IT) has documented increased sensitivity to changes that disrupt the topological structure of a display (e.g., attaching two figures that were previously unattached) in comparison to equated changes that do not disrupt topology (Kim & Biederman 2012). Another fMRI study has revealed that when right-handed subjects perform a shape discrimination task, they exhibit greater activation in a region of the left inferior temporal gyrus when the figures to be discriminated are topologically distinct than when they are topologically equivalent (Wang et al. 2007),

[36] As noted above, “equating” low-level image differences is nontrivial. Kayaert et al. took the Euclidean distances between the gray levels of each pixel in the base stimulus and that pixel’s counterpart in the variant stimulus.

suggesting that there may be some lateralization in the ventral stream processing of topological properties.

Thus, the existing neurophysiological data corroborates psychophysical findings, indicating increased sensitivity to changes in abstract shape, at least in certain cortical regions. The evidence also implicates ventral stream areas already known to be involved in visual shape processing (see Denys et al. 2004).

8. Implications

Kulvicki (2007) has recently introduced the notion of *vertical articulateness* and argued that it is a highly general characteristic of perceptual content. On Kulvicki's characterization, a state has vertically articulate content "when for some property *P* that it represents, it also represents some *Q*, which is an abstraction from *P*" (Kulvicki 2007: 359). And roughly, one property is an abstraction from another only if there is an asymmetric entailment between the two: "*Q* is an abstraction from *P* only if being *P* entails being *Q* but the converse fails" (Kulvicki 2007: 359). While Kulvicki suggests that this view holds for a wide variety of types of perceptual content (e.g. color, shape, texture, etc.), he primarily motivates it in the case of color experience. Here I have mounted a sustained defense of the view that the content of visual states vis-à-vis geometrical properties is vertically articulate as well. This holds both for subpersonal representational states of the visual system, and for states of visual experience.

Kulvicki points out that vertically articulate perceptual content may have a crucial role to play in guiding concept acquisition. Most of our concepts concern fairly general categories. Thus, while most of us have the concept of red, few (if any) of us have concepts for maximally determinate shades of red. If perception presents us with the property redness (assuming there is such a property) in addition to presenting us with maximally determinate shades of red, then we have a much clearer picture of how we might acquire the concept of this general category. For, otherwise, generalization across specific shades would be left entirely up to post-perceptual cognition. Thus, it is quite possible that perceptual content *needs* to be vertically articulate if it is to form an adequate basis for learning.

Arguably, the need for vertical articulateness is even more pressing in the case of shape perception than in the case of color. It is well known that many of the earliest concepts children acquire reside at the so-called "basic" level, where perceptual "similarity" among members of the category is most salient (see Rosch 1978). At this level, cars may be grouped together and distinguished from buses, but no general concept MOTOR VEHICLE is yet available. But how is such "similarity" to be characterized? A number of authors have contended that *shape* serves as the most important respect of perceptual resemblance during concept learning (see Rosch et al. 1976; Landau et al. 1988; Margolis 1998).[37]

Nevertheless, while it is generally accepted that children must be sensitive to shape similarities during concept learning, members of the same basic category almost never share a common *metric* shape—e.g., some cars are longer or wider than others (and often drastically so). Rather, to locate the pertinent respect of shape similarity, arguably we must turn to abstract shape properties. And indeed, it appears that, at least in many cases, members of the same basic-level category (e.g., bottles or bowls) are at least

[37] However, it should be stressed that common shape is merely taken as a *guide* to common category membership. When information about "hidden" or "internal" features is available to children, it will often override shape information when making category judgments (see Gelman & Wellman 1991).

roughly affine equivalent: the shape of one member can roughly be obtained from the shape of another by some combination of scaling, stretching, and shearing (see, e.g., Ons & Wagemans 2011).

But how are children sensitive to similarities in abstract shape during the early stages of concept acquisition? Metric views of shape representation seem to have a difficult time answering this question, since abstract shape properties are left implicit in visual representation.[38] On the view outlined here, vision takes over much of the work that would otherwise have been left to cognition. That is, generalization across metrically distinct individuals occurs within vision. As such, on the layered view of visual shape perception I have offered, we gain a clearer conception of how visual shape perception may furnish a partial basis for early concept learning.

In closing, we should reject the idea that the representation of abstract shape properties belongs solely to the domain of post-perceptual cognition. Such properties are represented during vision proper, exert an influence on other perceptual processes, and are represented in visual experience.

[38] This difficulty is not new. View-based proposals have often been criticized for lacking a good account of basic level categorization (e.g., Palmer 1999: 452).

Chapter 3

On the Visual Experience of Structure

1. Introduction

Imagine that you are meeting a friend for coffee, and you see her walking toward your table. As she walks, her arms and legs turn about their joints. Moreover, her forearms turn slightly about her elbows, and her tibias move about her knees. I suggest that, despite these changes, her *overall structure* phenomenologically seems to remain *stable*. Call this experiential phenomenon *structure constancy*. Structure constancy is ubiquitous in our visual experiences of objects. In this chapter I'll offer an account of structure constancy. Then I'll argue that the phenomenon has important consequences for a viable understanding of the subpersonal underpinnings of visual spatial experience.

I'll begin in section 2 with a general discussion of perceptual constancy, and then I'll identify a critical respect in which structure constancy differs from the more familiar geometrical constancies. In section 3, I'll offer a characterization of *compositional structure*, and propose that structure constancy involves experientially representing an object as retaining compositional structure across certain geometrical changes. In section 4, I'll argue that the representation of compositional structure is plausibly involved in the perception of biological motion. In section 5, I'll argue that the phenomenology of structure constancy cannot be underpinned by a representational format that fails to make part structure explicit, and that this has implications for identifying the locus of visual shape phenomenology within visual system processing. In section 6, I'll argue that structure constancy raises a problem for views on which the visual representation that underlies our experience of spatial/geometrical properties is wholly viewer-centered. I suggest that our visual experience of geometrical properties plausibly reflects the simultaneous deployment of multiple spatial reference frames.

2. Perceptual Constancy

In this section I'll first offer some remarks about the nature of perceptual constancy in general, and then turn to the nature of *geometrical* constancy in particular. This will set up the rest of the chapter by clarifying why structure constancy differs from other geometrical constancies (e.g., size and shape constancy).

2.1. What is perceptual constancy?

Most theorists agree that perceptual constancy involves a type of *stability* in one's perceptual response across certain *changes* (cf. Cohen forthcoming). Thus, Tyler Burge (2010) writes: "Perceptual constancies are capacities systematically to represent a particular or an attribute as the same despite significant variations in registration of proximal stimulation" (408). Similarly, Stephen Palmer characterizes (visual) perceptual constancy as "the ability to perceive the properties of environmental objects, which are largely constant over different viewing conditions, rather than the properties of their projected retinal images, which vary greatly with viewing conditions" (Palmer 1999: 125).[1]

Under these characterizations, to exercise perceptual constancy with respect to a property *P*, one must at minimum perceptually represent *P* across changes in the way

[1] Roughly this notion also appears in Rock (1983: 24), Smith (2002), Pizlo (2008), and Hatfield (2014). Other notions of constancy instead focus on the stability of one's perceptual representation across changes in a property's *appearance* (e.g., Shoemaker 2000; Hill 2014).

one's sensory organs are stimulated. In the case of vision, this would be to perceptually represent *P* across changes in the stimulation of retinal cells.

While Burge's definition provides a useful starting point, it has a significant drawback. Burge does not say what is involved in representing a particular or attribute "as the same" across variations in proximal stimulation. On one reading, this would require that a subject (or a perceptual system) represent that something perceived under one condition of proximal stimulation *is the same*—or, at least, the same in respect of a particular attribute, such as color—as something perceived under a different condition of proximal stimulation. On another reading, it would require only that one perceptually attribute the same property *P* to individuals encountered under different conditions of proximal stimulation.

The first notion is arguably more demanding. To represent that things perceived under different conditions are the same in respect of a particular property, one must be able to perceptually represent *comparisons* or *relations* between those things. This might involve either representing that some property *P* is shared by things perceived in different conditions, or retained by a single thing perceived in different conditions. There is no such requirement in order to simply represent the same property *P* under two different conditions. We can call the first notion the *strong* type of constancy, and the latter the *weak* type.[2] I'll suggest below that structure constancy is generally of the strong type.

[2] Whether the "strong" notion of constancy is actually more demanding depends in part on whether the capacity to perceptually represent properties of individual objects is more or less basic than the capacity to represent comparisons or relations between objects. Most have assumed that the latter capacity relies on the former, although some (e.g., Morrison ms.) have recently suggested that the order of explanation is actually the reverse. Nothing will hang on this dispute for present purposes, and if the reader prefers to reverse my labeling of the two types of perceptual constancy, I have no objection.

2.2. Geometrical constancies

To set up the rest of the chapter, I want to briefly apply this account to geometrical constancies in particular. I'll understand an object's "geometrical properties" to include its size, shape, and location. Moreover, I'll henceforth focus on the strong type of perceptual constancy, where one not only recovers a property under two different conditions, but also represents that the property is shared or retained across changes in proximal stimulation.

To delineate the nature of geometrical constancy under this characterization, we need to know what features—or "cues"—within proximal stimulation are relevant to recovering distal geometrical properties. Research indicates that in the case of shape and size perception, there are a number of such cues—e.g., 2-D retinal shape, context, shading, texture, and motion, among others (Palmer 1999: ch. 5). For the sake of simplicity, however, let's just focus on 2-D retinal shape. Accordingly, our paradigm case of geometrical constancy in what follows will be one in which a subject perceptually represents an object as retaining a geometrical property (distal shape or size) across changes in the shape or size of its retinal projection.

Changes in the shape or size of an object's retinal projection result from *transformations* of the distal object within a viewer-centered frame of reference.[3] For instance, if the object undergoes a rotation transformation whereby it is slanted in depth relative to the line of sight, this issues in a change in the shape of its projection on the retina. A circular object presents a circular image when seen straight on, but an elliptical image when seen at a slant.

[3] More specifically, such changes result from transformations of the object within a viewer-centered reference frame where the directions of 'left-right' and 'up-down' are defined in accordance with the intrinsic structure of the retina.

Transformations are classified as *rigid* or *non-rigid*. Rigid transformations are ones that don't involve any changes to an object's intrinsic metric properties. By the "metric properties" of an object, I have in mind, roughly, those properties of the object that depend essentially on its constituent edge lengths, angles, and curvature. For instance, metric properties of a square surface include the property of having four angles of 90°. Rigid transformations include translation (simple change of position), rotation, and reflection (change in "handedness"). Such transformations do not alter the distances or angles between points of the transformed object.

Non-rigid transformations, on the other hand, do involve changes to an object's intrinsic metric properties. The simplest kind of non-rigid transformation is uniform scaling, in which an object changes in size but its constituent angles stay the same. Other non-rigid transformations include stretching, shearing, skewing, and bending, which disrupt both lengths and angles.[4] Both rigid and non-rigid transformations can result in changes to an object's 2-D retinal shape. For example, if a square is stretched into an oblong rectangle, this can result in a change in the shape of its projection on the retina.

Size constancy involves seeing things as sharing/retaining a property across rigid transformations in a viewer-centered reference frame (since non-rigid transformations usually change an object's size). For example, one might perceptually represent something as retaining a particular size property despite viewing it at different distances. Shape constancy, as it is normally introduced, involves seeing things as sharing/retaining a property despite either a rigid transformation (e.g., rotation or translation with respect to

[4] In geometry, transformations are arranged into groups, such as affine transformations and projective transformations. The group consisting only of rigid transformations and uniform scaling is the *similarity group*. Often, the similarity group is taken to be definitive of what we mean when we say that two objects have the "same shape." We mean that one can be brought into precise register with the other by some composition of similarity transformations (Palmer 1999: 364-365).

the retina and the line of sight), or uniform scaling. For example, one might see an object as retaining a particular distal shape despite viewing it at different orientations (slants) in depth.

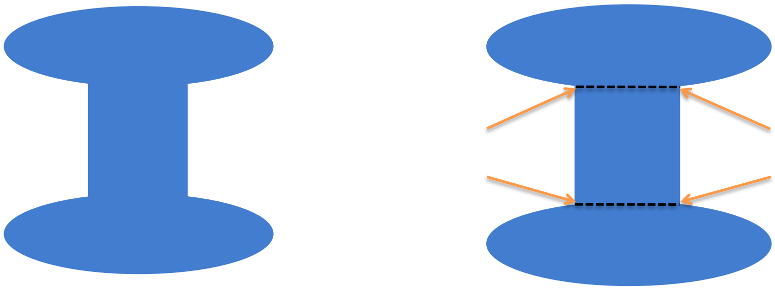
Structure constancy cannot be reduced to size or shape constancy. The reason is that structure constancy involves seeing an object as retaining a property (which I'll label "compositional structure") across certain non-rigid transformations that (unlike uniform scaling) disrupt both the distances and angles between parts of the object. As such, structure constancy is distinctive insofar as the transformations relevant to exercising structure constancy are different from (and, as we'll see, more geometrically complicated than) the transformations relevant to exercising the other geometrical constancies.

3. The Visual Phenomenology of Structure Constancy

Many of the most ecologically significant objects with which we interact are *biological objects*—especially animals and other humans. Many biological objects have an important characteristic: When they move, they *change shape*. This happens when, for instance, a person walks across a room. Even though the person's precise metric properties are constantly changing, intuitively we are able to see her body as retaining some important aspects of structure as she moves. In this section I'll first introduce the notion of *compositional structure*. Then I'll argue that structure constancy is most plausibly explained by the view that visual experience represents compositional structure.

3.1. Compositional structure introduced

Objects often seem to decompose into parts. For example, the object in figure 3.1a seems to have three natural parts, as shown in figure 3.1b.



Figures 3.1a (left) and 3.1b (right). Example of part decomposition.

In addition to being intuitively compelling, judgments about an object's decomposition into parts are remarkably consistent across observers (e.g., De Winter & Wagemans 2006). This, in addition to its role in several well-known theories of object recognition (Marr & Nishihara 1978; Biederman 1987), has led part decomposition to become a topic of extensive research in perceptual psychology.[5]

Critically, there are *rules* by which the visual system parses objects into parts. An important rule for our purposes is called the *minima rule*, first formulated by Hoffman and Richards (1984). The minima rule states that the boundaries between the perceived parts of an object tend to be found at extrema of negative curvature—roughly, places at which the surface of the object is locally most concave. Concave regions are, intuitively, regions where the object's surface curves "inward." Figure 3.2 illustrates two further applications of the minima rule in specifying part boundaries.

[5] For general discussion, see Singh and Hoffman (2001).

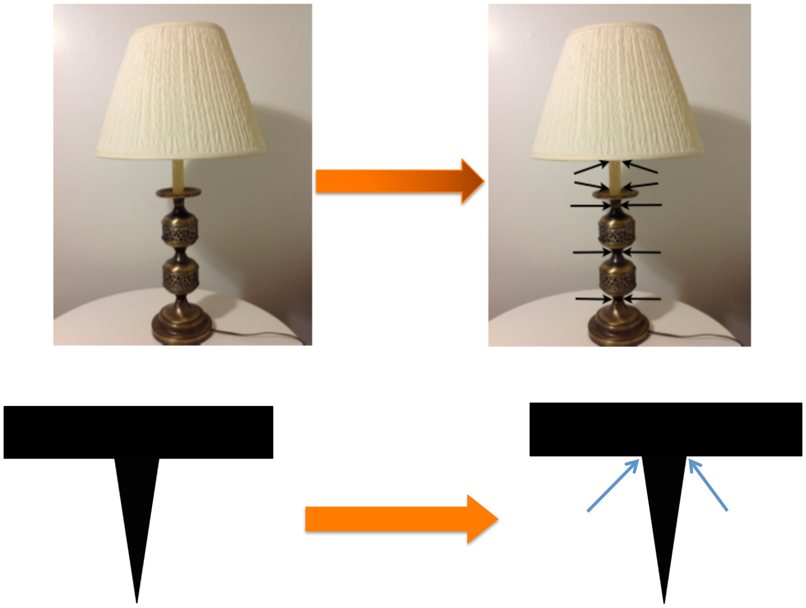


Figure 3.2. Further examples of part decomposition. Part boundaries are indicated by arrows.

While the minima rule tells us where to find boundaries between parts, it does not tell us precisely how to "slice" an object. That is, it does not specify how to make part *cuts*. Fortunately, this problem has also been studied extensively. Other things being equal, part cuts tend to obey the *short-cut rule* (Singh, Seyranian, & Hoffman 1999), which states that the visual system prefers part cuts that link negative minima of curvature, and generally opts for the shortest such links possible. The part cuts in figure 3.1b conform to the short-cut rule, as would the most obvious cuts of figures 3.2a and 3.2b.[6]

The representation of part decomposition has incredible psychological utility (e.g., Ling & Jacobs 2007). For example, many objects that move non-rigidly nevertheless change shape in a systematic manner. Roughly, their parts retain their intrinsic shapes, though the spatial relations between parts may change. The moving human body, as we saw, is an instance of this generalization, but so are the moving

[6] However, these rules have exceptions. See Singh and Hoffman (2001) for discussion.

bodies of most other animals, along with many manufactured devices (such as, e.g., a stapler or a reclining chair). By decomposing such objects into parts one can predict the ways they are likely to transform over time. They are disposed to move in ways that alter the spatial relations between parts, but unlikely to move in ways that either alter the intrinsic shapes of parts or displace the joints about which the parts rotate.

We are now ready to introduce the notion of compositional structure. A compositional structure of an object *O* consists of the following:

1. A decomposition of *O* into a pairwise disjoint set of (proper) parts $P_1...P_n$,
2. The approximate part-centered locations of boundaries between connected pairs of parts in $P_1...P_n$,
3. The approximate intrinsic shapes of $P_1...P_n$.

Structure constancy amounts, I suggest, to the ability to perceptually represent an object as retaining a particular compositional structure across proximal cue variations (e.g., changes in retinal shape) that result from non-rigid transformations of the object.

A terminological note: A set of parts $P_1...P_n$ will be called “pairwise disjoint” if and only if for all pairs (P_i, P_j) drawn from $P_1...P_n$, P_i and P_j do not overlap. Now, three substantive remarks on the visual representation of compositional structure:

First, according to my characterization of compositional structure, an object will have at least as many compositional structures as it has decompositions into parts. This may give rise to some initial concerns. For decompositions are *cheap*. An object can be decomposed in any number of ways, and it certainly does not seem as though we perceptually experience *all* of these decompositions, much less perceive them all as remaining stable as an object moves. However, the explanation of structure constancy offered here does not rely on this claim. Rather, the idea is that a *particular* compositional structure is perceptually represented, while the others are not.

Second, note that I take only the *approximate* intrinsic shapes of parts to figure in compositional structure. Due to, say, the deformation of muscle tissue, a person's upper arm does not retain its metric properties precisely as the arm rotates. So it is likely that to perceive an object as retaining compositional structure over time, the object's parts need only retain their shapes up to some more coarse-grained standards of precision.

Third, note that the locations of part boundaries must be specified in part-centered reference frames. This means that the locations of part boundaries are represented via their spatial relations to certain points on the connected parts themselves. The reason is this: If, say, the location of a perceived person's elbow (a boundary between forearm and upper arm) is specified in a viewer-centered reference frame, then its location *does* change as the person moves. Likewise, if its location is specified in a simple object-centered reference frame (e.g., with an origin at the center of gravity of the person's body), then its location changes as a result of rotation of the upper arm about the shoulder joint. Only when the elbow's location is specified in a frame of reference centered on either the forearm or upper arm (according to their intrinsic axes) does its location remain approximately stable across non-rigid movement of the body. Like the representation of metric part shapes, the representation of part boundaries should be somewhat coarse-grained. Even in a part-centered reference frame, part boundaries do not remain *perfectly* stable across non-rigid movement.

3.2. Representing compositional structure in experience

I've proposed that the compositional structure of an object is represented in experience, and that this accounts for the experience of structure constancy. But this claim requires further defense. As in the previous chapter, I'll defend it using a modification of Susanna Siegel's method of phenomenal contrast (Siegel 2010).

Siegel's method is introduced as a procedure for determining whether visual experiences represent a given property *F*. It requires us to examine two overall experiences that differ phenomenally, and determine whether the best explanation of their phenomenal contrast is that one of the overall experiences contains a visual experience that represents *F*, while the other does not.

Unfortunately, Siegel's method of phenomenal contrast cannot be straightforwardly applied in the current case. Consider any two experiences *A* and *B* that phenomenally differ, and are plausible candidates for differing vis-à-vis the compositional structures they represent. The method asks us to determine whether the phenomenal contrast between *A* and *B* is *best* explained by the hypothesis that they indeed differ with respect to the visual experiential representation of compositional structure. However, for any two such experiences, there will plausibly be *numerous other* differences in their visual experiential content, and some of these other differences would also seem to plausibly explain the phenomenal contrast.

Notice that if an object loses a particular compositional structure, it must cease to occupy precisely the same spatial region. For example, any change in the intrinsic shape of an object *O*'s part *P* necessitates a change in *O*'s compositional structure, but it also necessitates a change in the precise spatial region that *O* occupies. Thus, if we only consider this individual change, the difference in phenomenology that accompanies successive experiences of *O* (before and after the change) may seem to be explained just as well by the hypothesis that visual experience only represents the precise spatial region that *O* occupies, rather than *O*'s compositional structure.

How should we evaluate the hypothesis that visual experiences represent compositional structure? Rather than examining two *individual* experiences, I suggest that we examine *pairs of changes* in experience. We begin with an experience of a *base stimulus*, and a hypothesis about the particular compositional structure *C* of the base stimulus represented in experience. Next, we consider the experiences of two *test stimuli*. Test stimulus 1 shares compositional structure *C* with the base stimulus, while test stimulus 2 does not. However, *both* test stimuli differ from the base stimulus in their precise metric structure. If visual experiences represent compositional structure, then one might expect the difference between one's experiences of the base stimulus and test stimulus 2 to be *more salient* than the difference between one's experiences of the base stimulus and test stimulus 1. By "salience," I mean, intuitively, how perceptually *striking* a change or difference is. It is a fact about our phenomenology that some changes are more perceptually striking than others. For instance, a change from scarlet to aquamarine is (other things being equal) more perceptually striking than a change from scarlet to maroon. Similarly, a change from square to elliptical is more perceptually striking than a change from square to rectangular. It is this notion of salience that I have in mind in what follows.

However, for this to be a fair test, we need to ensure that, as regards factors *besides* compositional structure, the change from the base stimulus to test stimulus 1 is either roughly comparable to, or else greater than, the change from the base stimulus to test stimulus 2. In particular, we want to ensure that the increase in salience accompanying the change between the base stimulus and test stimulus 2 is not due to a greater difference in local features of the stimuli, or to a greater "overlap" in their spatial regions.[7]

There are a variety of ways to measure the amount of local point or feature difference between two figures (see, e.g., Kayaert et al. 2003; Veltkamp & Latecki 2006). Perhaps the most straightforward measure is "Hamming distance" (Ullman 1996: 5). To find this distance, we first specify the two figures within a coordinate system. Each is then represented by a binary vector indicating, for each point *p* within the coordinate system, whether *p* "belongs" to the figure ("1" if it belongs, "0" if it does not). Given this, we measure the distance between the two figures by normalizing the figures to a standard position and orientation, then summing the places in which the vectors for the two figures differ.

In what follows I'll only consider cases in which the Hamming distance between the base stimulus and test stimulus 1 is clearly either greater than, or roughly comparable to, the Hamming distance between the base stimulus and test stimulus 2.[8] Of course, differences in other, non-geometric features like color, luminance, and texture should be controlled as well. The argument is that if the difference between the base and test stimulus 2 is more phenomenologically salient under these conditions, then the *best explanation* is that visual experience represents the base stimulus as sharing a property

[7] For example, suppose that one change disrupts compositional structure while another change does not, and suppose further that the former change is more salient than the latter. However, suppose that the former change *also* involves rescaling the object by a factor of 3, while the latter change does not result in such rescaling. In this case, there would be no reason to attribute the difference in salience to the representation of compositional structure, rather than, e.g., the representation of size.

[8] I don't propose that this restriction must *always* apply when using the modified method of phenomenal contrast proposed here. The modified method of phenomenal contrast simply involves the idea of examining pairs of changes in visual experience, rather than pairs of individual experiences. Further modifications to the method are needed, I believe, for the specific phenomenological hypothesis that is being investigated.

with test stimulus 1, but doesn't attribute this property to test stimulus 2. My proposal is that the former two are visually experienced as sharing a compositional structure.

Consider figure 3.3. The compositional structure *C* of the base stimulus plausibly consists of the following: a decomposition into head, torso, arms, and legs; the approximate intrinsic shapes of these parts; and the joints at which they are connected to one another. Test stimulus 1 shares *C* with the base stimulus. Test stimulus 2 does *not* share *C* with the base stimulus (joint locations are changed). Phenomenologically, I find these changes to be qualitatively different. The transformation to test stimulus 2 seems much more perceptually striking, even though local feature differences have for all intents and purposes been held constant.[9] The proposal that visual experience represents compositional structure explains this. In the first case, the two objects are visually experienced as sharing a property (a particular compositional structure), while in the second case, they are not.

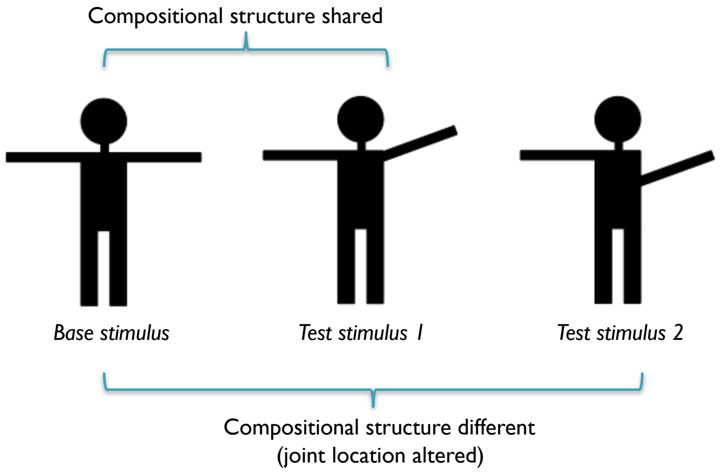
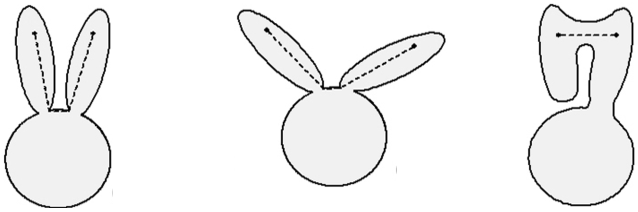


Figure 3.3. Two transformations of a base stimulus. The change that alters compositional structure (in this case, joint location) is intuitively more salient than the change that does not.

Consider another example, due to Ling and Jacobs (2007). The base stimulus is shown in figure 3.4a, while test stimuli 1 and 2 are shown in figures 3.4b and 3.4c,

[9] The arm has undergone the same amount of rotation in both cases. The only difference is whether its *axis* of rotation is its joint with the torso (test stimulus 1) or instead its endpoint (test stimulus 2).

respectively. Again, the Hamming distance between the base and test stimulus 1 is greater (i.e., the two have less overlap in local features), but the transition between the two arguably is less salient than the transition from the base to test stimulus 2. Once again, test stimulus 1 preserves compositional structure, while test stimulus 2 does not.



Figures 3.4a-4c (left to right). Figures from Ling and Jacobs (2007). © 2007 IEEE.

3.3. A post-perceptual explanation?

There are two potential worries with examples involving human bodies, bunny ears, and the like. First, it is unclear whether the contrast in salience here is due to *visual* experience, or rather to postperceptual expectations given familiarity with such objects and the ways they move. Second, even if the examples do reveal the representation of compositional structure in visual experience, it is unclear how general their implications are. Perhaps compositional structure is represented in visual experience only for highly familiar figures, and not for decomposable figures in general. For these reasons, it would be more persuasive if such contrasts in salience could be demonstrated using novel shapes.

Evidence suggests that compositional structure is extracted for novel shapes. Barenholtz and Tarr (2008) showed subjects a novel base shape, along with two transformations of the base shape. Only one of these transformations—which I'll again label test stimulus 1—preserved compositional structure under the minima and short-cut rules. The shape that failed to preserve compositional structure—test stimulus 2—could involve either a change in location of a boundary between parts, or a change in a part's intrinsic shape. Figure 3.5 shows a case in which test stimulus 2 involves a change of the former type. The differences between the base stimulus and test stimuli 1 and 2 are essentially equated in their low-level feature changes, because the narrower part on the right of the figure was rotated the same amount in both cases. The only difference was whether the part's axis of rotation was its joint with the rest of the object (preserving compositional structure) or its endpoint (altering compositional structure).

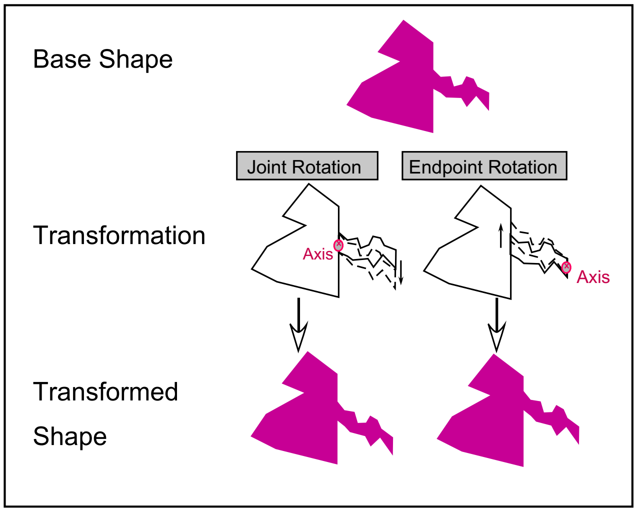


Figure 3.5. Stimuli from Barenholtz and Tarr (2008). The transformed shape at bottom left (test stimulus 1) preserves the compositional structure of the base (intrinsic part shapes, and locations of part boundaries). The transformed shape at bottom right (test stimulus 2) alters compositional structure, because the part boundary shifts upward. Reproduced from Barenholtz and Tarr (2008) with kind permission from Elsevier.

Participants saw the three shapes, and were simply asked to indicate which of the transformed shapes was more similar to the base. Barenholtz and Tarr found that subjects were significantly more likely to indicate that the shape that preserved compositional structure was more similar. The same pattern of results was obtained with other triples of shapes where the change that disrupted compositional structure instead altered the intrinsic shape of the base stimulus's part, rather than its joint location. Thus, there is evidence that the ability to extract compositional structure is highly general and not limited to particular classes of familiar objects.[10]

Nevertheless, how do we know that compositional structure isn't recovered post-perceptually, even in the case of novel objects? If this were the case, then structure constancy wouldn't really deserve to be labeled a *perceptual* constancy at all. In what follows, I will argue that compositional structure is recovered by the *visual system*. This forms the basis of an inference to the best explanation: (i) Compositional structure is represented in experience. (ii) Compositional structure is recovered during visual processing. Therefore, (iii) the best explanation is that compositional structure is represented in visual experience.

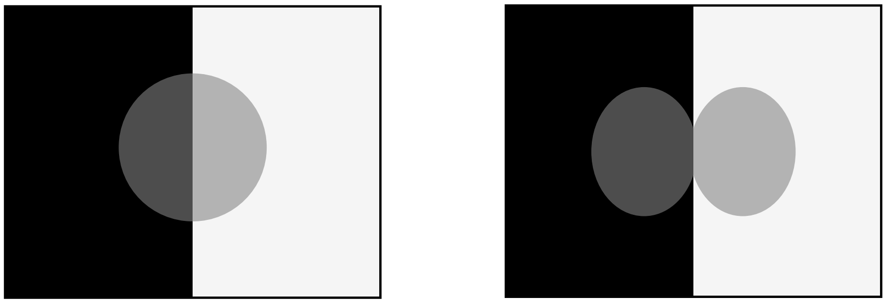
I've held that the explanation of structure constancy resides in our ability to decompose objects into parts and represent their boundaries and shapes independently. To show that structure constancy is represented during visual processing, I need to show (i) that parts are differentiated by the visual system, (ii) that the visual system processes part shapes independently of one another, and (iii) that the visual system utilizes part-centered reference frames. I will discuss evidence for (i)-(iii) in order.

[10] I should flag that it is not entirely clear how subjects in Barenholtz and Tarr's experiment understood the idea of "similarity" (or, indeed, whether all subjects understood this term in the same way). But given that the shapes were novel and had no particular cognitive significance for the observers, it seems reasonable to infer that subjects at least interpreted "similarity" to mean "perceptual similarity." It is also unclear, however, whether perceptual similarity bears any straightforward relationship to what I have called perceptual salience. That is, is it the case that when two objects are more perceptually similar, the change between them is always less perceptually salient? One way in which this would be true is if both perceptual similarity and perceptual salience (or strikingness) are determined by how close two perceptible properties are in the perceiver's underlying "quality space" (e.g., Rosenthal 2010). Roughly, when two properties are further apart in the perceiver's quality space, they will be less perceptually similar, and the change between them will also be more salient. However, I won't attempt here to defend quality space theory in general, or to show that it necessarily applies to the visual experience of compositional structure.

With respect to (i), perhaps the strongest evidence that parts are extracted during visual processing is that part decomposition influences other paradigmatically perceptual processes.

First, there is evidence that part structure influences the spread of visual attention throughout a scene. It is now fairly uncontroversial that attention can be allocated not just to locations (as in the "spotlight" model), but also to occupants of locations—particularly objects (see Scholl 2001; Chen 2013). This has been demonstrated convincingly using the "feature comparison" paradigm, in which subjects are asked to make a judgment pertaining to two visible features. Such judgments are both faster and more accurate when the features to be compared belong to the same object than when they belong to different objects, even if the spatial distance between them is held constant across the two cases. Recent work has shown that a similar pattern of results holds for *parts* of objects. Specifically, feature comparisons are faster (Barenholtz & Feldman 2003) and more accurate (Vecera et al. 2000) when the features to be compared belong to the same part of an object than when they belong to different parts, even when both the spatial distance and degree of contour curvature between the features (see Barenholtz & Feldman 2003) are held constant (see Chapter 4 for more discussion).

Another example of the role of part decomposition in visual processing involves the perception of transparency. Compare figures 3.6a and 3.6b.



Figures 3.6a (left) and 3.6b (right). Stimuli from Singh and Hoffman (1998). Reproduced with kind permission from SAGE Publications.

While figure 6a appears to depict a transparent gray filter in front of a half-dark, half-light background, in figure 6b the percept of transparency is greatly diminished (Singh & Hoffman 1998). Rather, the occluding object is perceived as an opaque figure with two differently shaded regions. The received explanation for this is that the visual system expects regions of a single part of an object to have the same reflectance, but it does not expect regions of different parts of an object to have the same reflectance (or at least it expects this less strongly). Since the object in figure 6b can be broken down into two natural parts, it can be interpreted as an opaque figure whose parts have different reflectances.

If part decomposition interacts with other perceptual processes, we have strong evidence that it is a perceptual process as well.[11] For, while it is possible to advert to a cognitive penetration account in these cases, I can think of no motivation for doing so. Moreover, it is worth noting that the tendency to parse objects into parts also strikes me

[11] Part perception has also been argued *inter alia* to influence figure-ground organization (Hoffman & Singh 1997) and pop-out effects in visual search (Xu & Singh 2002).

as involuntary—I cannot *help* seeing many objects as decomposed into natural parts. This is another hallmark feature of a perceptual process (e.g., Fodor 1983; Pylyshyn 1999).[12]

With respect to (ii), there are good reasons to believe that the visual system encodes the shapes of different parts separately from one another, and independently of their spatial relations. Though this hypothesis was initially put forth on computational and theoretical grounds (Biederman 1987; Marr & Nishihara 1978; Palmer 1978), there is now compelling experimental evidence for it. Consider a recent study of the subject S.M., an individual with integrative agnosia. Integrative agnosia is a visual disorder that affects processes involving the integration of local visual information into a global percept. Behrmann et al. (2006) found that S.M. was capable of correctly discriminating sequentially presented objects from one another when the objects differed in the intrinsic shape of a single part (e.g., a cube-shaped part versus an ellipsoid-shaped part), but, unlike "normal" participants, he could not discriminate objects when they differed purely in their parts' spatial configuration (e.g., a cube to the left of a cylinder versus a cube on top of a cylinder). In line with (ii), this suggests that there are visual processes that extract the shapes of individual parts, and these processes can remain intact despite an inability to extract the global configuration of an object (see also Davidoff & Roberson 2002).

A recent study also indicates that when the parts of a familiar object (e.g., a lamp) are rearranged into a novel object, the novel object visually *primes* its familiar counterpart (see Cacciamani et al. 2014). This occurred even though subjects were unaware of (or at least unable to report) the fact that the novel prime figures were rearranged versions of familiar objects. This transfer of priming suggests both that the

[12] Further evidence for the automaticity of part decomposition is provided by studies of human infants. Using a dishabituation paradigm, Bhatt et al. (2010) have provided compelling evidence that 6 ½ month-old infants are sensitive to the minima and short-cut rules.

visual system recovers the shapes of the parts of novel objects *and* that stored representations of familiar objects make explicit their component part shapes. Again, absent defeating evidence, I conclude that the processing of individual part shapes happens within perception.

The claim, (iii), that part boundaries are represented in part-centered reference frames is the hardest to establish. Before covering empirical support for this claim, we need to get clearer on what part-centered reference frames are, and how they have been developed in the vision science literature.

Constructing a reference frame involves choosing a set of parameters that permit the position of any point to be uniquely determined by specifying its values on these parameters (Klatzky 1998). When a reference frame is *centered* on an object *O*, this means that the positions of points are coded at least partly in terms of their spatial relations (e.g., distance and direction) to a point, or set of points, on *O*. For example, to construct a polar coordinate system, we first stipulate an origin *o* and an axis *A* through *o*, and then specify the location of any given point *p* in terms of two parameters: its distance from *o*, and the angle between *A* and the line from *o* to *p*.

Many shape representation theorists have proposed that the visual system recovers, roughly, the *medial axis structure* of an object (e.g., Blum & Nagel 1978; Rosenfeld 1986; Kimia 2003; Feldman & Singh 2006). The medial axis of a figure is composed of the set of points that have two or more closest points on the boundary of the figure. A figure's medial axis generally looks like a "skeleton" from which the figure is "grown." Medial axis representation schemes are centered on the points that compose the axis. Roughly, they represent the positions of points on the boundary of the shape by specifying their distances and directions from corresponding points on the axis (see the Appendix).

Importantly, the medial axis structure of an object often bears a close relation to its decomposition into parts under the minima and short-cut rules.[13] This is because different parts of the object tend to be associated with distinct axis branches (see figure 3.7). Thus, if the visual system extracts the medial axis structures of objects, and distinct parts are associated with distinct axis branches, then these distinct axis branches can be used to construct separate reference frames each centered on a distinct part. Accordingly, evidence for the visual representation of medial axis structure also counts as evidence that the visual system uses part-centered frames of reference.

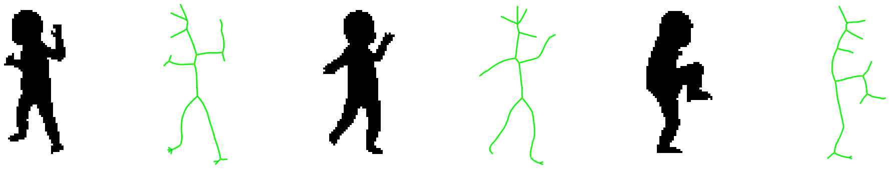


Figure 3.7. The medial axis structure of three human silhouettes. Note that in most cases the intuitive parts (arms, torso, and legs) correspond to distinct axis branches. Reproduced from Kimia (2003) with kind permission from Elsevier.

The prediction that vision extracts medial axis structure has recently been confirmed using a very simple paradigm. Firestone and Scholl (2014) showed subjects a novel shape, asked them to tap the shape wherever they liked, and recorded the locations of subjects' taps. If "tapping" behavior is guided by a visual shape representation that specifies the intrinsic (perhaps medial) axes of object parts, one might expect the

[13] In practice, however, the correspondence is not perfect. In standard models (e.g., Blum & Nagel 1978), small perturbations of a shape's contour give rise to "spurious" axis branches that do not intuitively correspond to distinct parts of the shape. Feldman and Singh (2006) have recently developed a Bayesian approach to axial description that "cleans up" the medial axis representation. The axes returned by their model are not medial axes, although for smooth shapes without many perturbations they closely resemble medial axes.

locations of subjects' taps to be influenced by these axes. Sure enough, Firestone and Scholl found that the recorded taps (when aggregated) corresponded closely to the medial axes of the shapes presented. That is, subjects were much more likely to tap an object somewhere along its medial axis than they were to tap other regions of the shape. This provides compelling evidence that medial axis structure is automatically extracted by vision, since the task did not require subjects to attempt to extract these axes.

Moreover, if the visual system represents spatial properties and relations by using intrinsic part axes, then it should encode the parts of an object as retaining their spatial relations to one another across transformations in viewer-centered position and orientation. For as long as these transformations are rigid, the *part-centered* relations between the constituents of the configuration will not change.

There is intriguing evidence that areas of the visual system code for medial axis structure independently of viewer-centered position. In a recent fMRI study, Lescroart and Biederman (2013) presented subjects with figures that differed in either their medial axis configuration, their component part shapes, or their viewer-centered orientation. Figure 3.8 displays some of these figures: Shapes in the same row share the same medial axis structure, though their orientations and intrinsic part shapes vary. Stimuli in the same column share the same intrinsic part shapes, but differ in medial axis structure. The line segments next to the figures indicate viewer-centered orientation. Lescroart and Biederman found that by area V3, patterns of BOLD activity could classify stimuli according to shared medial axis structure at a rate significantly better than chance (even though such stimuli differed in the shapes of their component parts), and classification of medial axis structure was significantly more accurate than classification of orientation.

This is precisely what would be expected on the proposal that extrastriate areas of the visual system represent configurations according to spatial arrangements of part axes.[14]

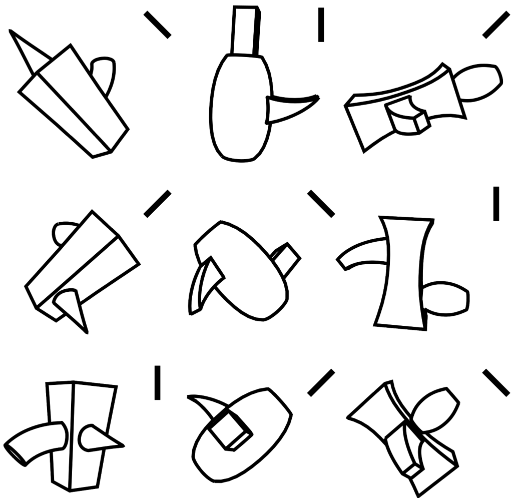


Figure 3.8. Stimuli from Lescroart and Biederman (2013). Reproduced from Lescroart and Biederman (2013) with kind permission from Oxford University Press.

Thus, there is evidence that each of the characteristics that figure in compositional structure is recovered during vision proper. I have also already argued that compositional structure is represented in phenomenal experience. As such, I contend that the proposal that compositional structure is represented in *visual experience* offers the best account in light of all the evidence at our disposal, including both the patterns of phenomenological salience associated with shape changes, and the empirical data on visual processing of geometrical structure.

[14] The proposal that shape is represented via part-centered reference frames is also consistent with work on view-invariance in vision. For example, several studies have found transfers of visual priming across changes in viewer-centered location, and sometimes across mirror reflection (Biederman & Cooper 1991; Stankiewicz et al. 1998; cf. Biederman & Bar 1999).

4. The Perception of Human Motion in Point-Light Displays

I have argued so far that structure constancy is underpinned by our ability to recover an object's compositional structure. I've also argued that compositional structure is represented both in visual experience and subpersonal visual processing. In this section I want to focus specifically on the perception of human motion, since this is the example with which we started, and the topic has received much psychophysical and neurophysiological investigation.

A large portion of the research on biological motion perception has utilized "point-light" motion displays. In a seminal study, Johansson (1973) fixed small lights onto the head and joints of an actor, and showed subjects movie clips that (by adjustment of contrast) depicted only the light markers, and not the rest of the actor's body. While subjects simply perceived a random array of dots while the actor was static, once the actor began moving they spontaneously reported perceiving a human performing certain actions. The phenomenal experience of viewing such displays is extraordinary—one has the inescapable feeling of perceiving dots attached to a moving person. Subsequent work has shown that, from point-light displays, subjects can fairly reliably identify numerous other characteristics, such as the gender of the actor (Kozlowski & Cutting 1977), the type of action being performed (Dittrich 1993), and certain emotional expressions of actors (Clarke et al. 2005; see Blake & Shiffrar 2007 for review). When the actor attempts to lift an item, subjects can even accurately guess the weight of the item he or she is attempting to lift (Bingham 1993).

Given that the characteristic experience of seeing a human body in motion seems also to be evoked by point-light displays, such displays are relevant to the current discussion. For if the standard experience of seeing human motion involves (at least in part) the perceptual representation of the person's body as retaining its compositional structure, then compositional structure should also be extracted when viewing point-light displays. However, there are at least two reasons one might be skeptical of this proposal, even if one agrees that the visual system extracts compositional structure in general.

First, it might be that human body motion is associated with characteristic *local* motion signals (i.e., the motion trajectories of small points or patches of the human body) that distinguish it from other types of motion, and that detecting these local motion trajectories alone (e.g., the motions of the individual dots) underlies the remarkable experience of bodily motion in point-light stimuli. Second, it might be that even though perceiving bodily motion in point-light stimuli involves extracting some kind of global shape information, it does not involve representing part decomposition. Rather, perhaps the visual system computes a familiar sequence of whole-body forms from the sparse data provided in point-light movie displays, without ever differentiating the body's subparts.

In what follows, I'll consider these two possibilities in turn. I will argue that the perception of human motion in point-light stimuli plausibly involves representing global shape properties, and, more specifically, the part structure of the human body.

Recent research has addressed the question of whether the identification of human motion in point-light displays could be facilitated purely by the detection of local motion signals. One suggestive line of evidence against this view derives from neurophysiology. Some patients who have lesions to early visual motion processing areas are nevertheless relatively unimpaired in detecting biological motion in point-light displays, distinguishing human motion from other kinds of point-light motion (e.g., the motion of a ball or a puppet), and identifying the type of action an actor is performing (Vaina et al. 1990; McLeod et al. 1996). This suggests that the detection of local motion signals may not be essential to perceiving human motion in point-light displays (although of course they may be used when they are available).

Second, and more convincingly, Beintema and Lappe (2002) have recently developed a novel kind of point-light display, which they call the "sequential position walker." In this computer-generated display, the lights on the walker's limbs shift from one part of the limb to another from each frame to the next. For instance, while the actor's arm moves forward, the light on her arm may nevertheless move backward. This display lacks the appropriate local motion signals for human body motion, because the motions of the individual lights are no longer informative for computing the overall motion of the body. However, the display still includes (highly degraded) form information within the spatial array of light markers in each frame. Several studies have now documented that subjects are nearly as efficient at detecting human motion (including both the direction of motion and whether the actor is walking forward or backward) in the sequential position walker display as they are in standard point-light displays. As such, it is unlikely that the experience of human motion in point-light displays depends essentially on the detection of a certain set of local motion signals (though, again, such signals may be used when they are informative). The visual system seems able to compute information about the form and global motion of the human body from the sequence of sparse spatial arrays of dots visible in the movie frames, even without appropriate motion of the individual dots (see Beintema & Lappe 2002; Lange et al. 2006 for discussion).

Nevertheless, even if the extraction of shape information is involved in perceiving human motion in point-light displays, this does not yet show that part decomposition is involved. Indeed, some have offered models on which the identification of human motion instead relies on matching the sparse form information contained in the series of dot arrays to learned sequences of “whole-body” forms, without extracting the body’s subparts (e.g., Lange et al. 2006). However, recent studies indicate that the differentiation of individual parts likely plays an important role in the perception of biological motion.

First, there is reason to think that individual limbs are recovered in the process of extracting biological motion from a point-light display. Thus, Pinto and Shiffrar (1999) found that when point-light walker displays are scrambled so that the motion of individual limbs is preserved, but limbs are presented in random locations of the screen, observers remain above chance in detecting human motion (although performance is significantly reduced relative to normal displays).

Neri (2009) has recently provided more convincing evidence that limbs are represented individually during the perception of point-light walkers. In each trial of the experiment, participants were shown two movie sequences depicting martial arts fighters, and were asked to indicate which sequence more closely resembled a real fight between human actors. Critically, one of the movie sequences had been temporally “scrambled,” while the other had not. In the scrambled sequence, the actors’ markers were divided into three triplets, and the triplets were shifted slightly out of temporal phase with one another. The crucial manipulation was whether the members each triplet in the scrambled sequence belonged to a *single* limb (e.g., three markers from the right arm, three from the left arm, and three from the left leg), or to *different* limbs. In the former case, the temporal relations between markers *within* each limb were preserved (though phase relations across limbs were altered), while in the latter case they were not. Neri reasoned that if the recovery of bodily motion from point-light stimuli relies on differentiating individual limbs, then processing should be altered more by the latter kind of scrambling, where relationships within limbs were disrupted. Consistent with this proposal, it was found that for a given phase shift, subjects were more likely to (mistakenly) select a scrambled sequence that preserved phase relations within individual limbs as more similar to a real fight between actors.

Furthermore, it has also been found that whether or not observers are able to learn to identify the motion of a novel object in a point-light display is heavily determined by whether the object is decomposable into parts that move in a piece-wise rigid fashion, as determined by an underlying skeleton. Jastorff et al. (2006) showed participants three types of point-light stimuli. One stimulus was consistent with an underlying human skeleton, while the other two were generated from novel objects. Critically, one of the novel stimuli was consistent with having been generated by an object with an underlying skeleton, so that it deformed in a systematic manner, preserving the shapes of individual "limbs." The other novel stimulus was inconsistent with an underlying skeleton. However, local motion signals were roughly controlled across the three types of stimuli. It was found that subjects could learn to reliably identify either of the first two types of stimuli after a relatively short number of training trials (20 repetitions), but could not learn to reliably identify the third type of stimulus. This comports with the idea that the process of extracting an object’s structure and motion from a point-light stimulus exploits the organization of the object into component parts whose shapes stay relatively stable over time.

Nonetheless, while it is plausible that the perception of human bodily motion relies in part on a general perceptual capacity to extract compositional structure, other processes contributing to the perception of human motion are plausibly specialized. For example, the ability to recognize the gender of an actor in a point-light display seems to rely on domain-specific information about the dynamics of male vs. female walking patterns, such as the type of body sway (Mather & Murdoch 1994). The proposal I recommend, then, is that the perception of human motion relies both on the ability to recover compositional structure, and also on a set of domain-specific capacities specialized for biological (or even human) movement.

5. Mereological Structure and Shape Representation Schemes

What does structure constancy tell us about the subpersonal underpinnings of our visual experience of spatial/geometrical properties? I believe it has at least two important consequences for a viable understanding of these underpinnings. In this section, I’ll argue that structure constancy has the consequence that certain aspects of shape experience must be underpinned by a representation scheme that is *mereologically structured*. In the next, I’ll argue that because structure constancy must recruit non-viewer-centered reference frames, it raises difficulties for approaches on which spatial phenomenology is subserved by some enrichment of Marr’s 2½-D sketch.

Call a representation *R* mereologically structured iff:

(1) *R* purports to introduce individuals *O* and *O** independently, and
(2) *R* represents that *O* is a proper part of *O**.

To purport to introduce *n* individuals *independently* is to deploy *n* distinct representational items that each purport to introduce distinct individuals.[15] For instance, the phrases "John's cat" and "John's dog" purport to introduce two individuals independently. The central idea, then, is that if a representation *R* is mereologically structured, then distinct constituents of *R* purport to pick out distinct entities that are related through mereological composition, and *R* represents the parthood relations that those entities stand in to one another.

Some shape representation schemes are not mereologically structured. Consider, for instance, schemes found in the *view-based* approach to object recognition (see, e.g., Ullman and Basri 1991; Ullman 1996; Edelman 1999; Riesenhuber & Poggio 2002). On several of these models, the representation of shape just amounts to the representation of a vector composed of the viewer-centered coordinates of some of the object's "critical features"—e.g., vertices, inflection points, and curvature maxima.[16] This type of scheme does not incorporate the representation of parthood at all, and proponents of the view-based approach have often downplayed the role of part decomposition in visual processing (e.g., Edelman 1999: 89-94).

Perhaps the most popular mereologically structured scheme is *hierarchical description* (see, e.g., Palmer 1977; Marr & Nishihara 1978; Feldman 2003; Leek et al.

[15] The notion of introducing an individual is left deliberately vague. For present purposes, it does not matter whether the relevant individuals are introduced by description or by singular reference. I will return to this issue in chapter 4.

[16] Such views have been offered primarily in order to account for findings indicating that object recognition is sensitive to viewpoint. Such results have sometimes been believed problematic for hierarchical approaches to shape representation, which commonly invoke non-viewer-centered reference frames. However, for an argument that hierarchical models can accommodate viewpoint effects on recognition, see Bar (2001).

2009; Hummel 2013). A hierarchical description (figure 3.9) is a representational structure that contains distinct nodes corresponding to each individual introduced, encodes either mereological or spatial relations between nodes, and associates monadic featural information with each node. It is usually depicted as a tree.

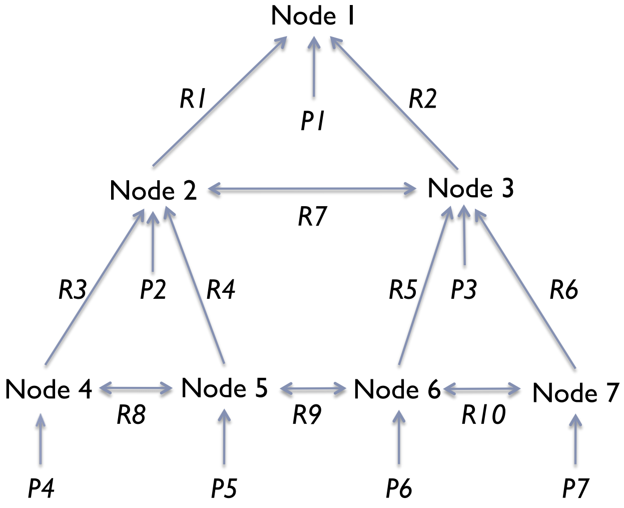


Figure 3.9. The format of a hierarchical description. *P*s represent monadic properties, while *R*s represent dyadic relations. *R*s linking nodes at the same level of the description represent spatial relations, while *R*s linking nodes at different levels represent mereological relations.

Edges traversing levels of a hierarchical description represent parthood. For present purposes, I'll assume that the visual system's representation of parthood is transitive: If a hierarchical description represents O_1 as part of O_2 and O_2 as part of O_3, then it also represents O_1 as part of O_3. Edges linking nodes at the same level of a description represent spatial relations between parts. Let's call edges representing parthood *M-edges* (for "mereological edges"), and edges representing spatial relations *S-edges*. A subset of the S-edges will describe the locations of *boundaries* between parts: They will represent, for a pair of connecting parts, the points where those parts meet (in part-centered coordinates). Call these *B-edges*.

Several influential models of shape processing invoke both an earlier, view-based stage and a later, hierarchical stage (Marr 1982; Hummel & Biederman 1992; Hummel 2001, 2013). If this is right, it is natural to ask which (if either) of these stages underpins shape phenomenology. I argue that structure constancy provides strong reason to locate at least certain aspects of shape phenomenology at the hierarchical stage.

It is hard to see how a view-based scheme could underpin structure constancy. Because view-based schemes do not introduce the parts of objects as distinct individuals, such models do not *prioritize* any particular part decomposition over others. Each of the many possible decompositions of an object into parts is compatible with, say, the same arrangement of vertices and curvature extrema along the object's bounding contour. Because view-based schemes fail to prioritize a specific part decomposition, they lack the resources for distinguishing changes that leave intrinsic part shapes intact while altering the global shape of the object from changes that alter the intrinsic shapes of parts. Indeed, any given change could—relative to *some* decomposition—be considered a change in the intrinsic shapes of an object's parts. Thus, without a specification of which decomposition is the *relevant* one, we cannot determine whether a particular change does or does not deform intrinsic part shapes.

Hierarchical description, on the other hand, can be readily applied to the explanation of structure constancy. Let's spell this out using the human body as an example. Given a hierarchical description that introduces a human body *O*, a *compositional structure* of *O* is encoded in (i) the intrinsic shape information associated with nodes at some level of the description lower than the level at which *O* is introduced, such as a level introducing the head, torso, arms, and legs, (ii) the M-edges linking these nodes to the node introducing *O*, and (iii) the B-edges linking these nodes to one another—e.g., torso-centered locations of the shoulders (where the arms intersect the torso), and torso-centered locations of the hip joints (where the legs intersect the torso). By distinguishing this information from the information encoded by the remaining S-edges (such as the angle formed between an arm and the torso) and information about *O*'s global metric structure, the representation enables the visual system to distinguish transformations that leave a given compositional structure intact from those that do not. As such, hierarchical descriptions may underpin structure constancy.[17]

6. Metric vs. Categorical Representation of Boundaries

I believe that structure constancy requires that vision represent the locations of certain elements (viz., part boundaries) within reference frames defined according to the intrinsic axes of individual parts. However, even among representation schemes that utilize part-centered reference frames, there are significant differences. Perhaps the most important difference concerns whether the relations between parts are represented *metrically* or *categorically*.

In a metric framework, boundaries between parts are specified using numerical coordinates—e.g., part *A* intersects part *B* at coordinates <*x*, *y*, *z*> relative to an origin on *B*. In a categorical framework, boundaries between parts are instead specified using qualitative categories—e.g., part *A* is on top of part *B*; part *A* is to the left of part *B*; or part *A* intersects part *B* somewhere above the midpoint of *B*'s axis. Marr's 3-D model is a

[17] The view that shape representation is hierarchical should not be confused with the view that shape representation is *volumetric* (i.e., the view that vision decomposes objects into 3-D components—Marr & Nishihara 1978; Biederman 1987). Hierarchical representation (and part decomposition) is compatible with either a volumetric or surface-based representation scheme (Leek et al. 2009).

version of the metric approach (Marr & Nishihara 1978; Marr 1982, ch. 5), while Biederman's recognition-by-components model is a version of the categorical approach (Biederman 1987; Hummel & Biederman 1992).

Structure constancy is compatible with either a metric or a categorical encoding of part boundaries. However, the issue has implications for the precise *pattern* structure constancy should be expected to take. Thus, if the categorical approach is right, then one should perceptually represent two objects as having the same compositional structure, so long as they agree in the metric shapes and *categorically specified* boundaries between parts (e.g., whether one part intersects another above, or below, its midpoint). If the metric approach is right, on the other hand, then there should be finer distinctions than this—certain objects that agree in the *categorical* relations among their parts may be perceptually represented as differing in compositional structure, so long as the numerical coordinates of their boundaries are different.

An experiment by Hummel and Stankiewicz (1996) bears on the issue of metric versus categorical encoding of part relations. The experiment utilized a variety of triples consisting of a base shape and two variants of the base shape. The shapes in each triple were unfamiliar configurations of lines (see figure 3.10). In the first variant—test stimulus 1—only metric relations between parts were altered. In the second variant—test stimulus 2— a categorical relation between parts (viz., whether a part intersects another above or below the former's midpoint) was altered. And, critically, in each case the overall pixel difference from the base was *greater* for test stimulus 1 than for test stimulus 2. When given a discrimination task, subjects were better at differentiating test stimulus 2 from the base. Moreover, when asked to indicate which stimulus was more similar to the base, they reliably judged test stimulus 1 to be more similar. For instance, in one experiment using a variety of shape triples, the base was never correctly discriminated from test stimulus 2 *less* than 80% of the time, while the base was never correctly discriminated from test stimulus 1 *more* than 20% of the time.

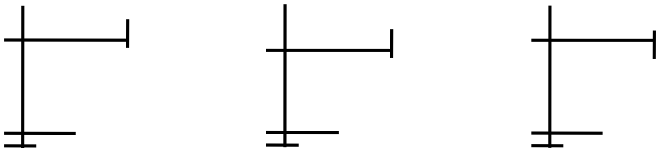


Figure 3.10. Figures similar to those used by Hummel and Stankiewicz (1996). From left to right: Base stimulus, test stimulus 1, and test stimulus 2. See the main text for explanation.

This study demonstrates increased perceptual sensitivity to changes in categorical relations between parts. However, it does not really show that the visual system is completely *insensitive* to purely metric changes in part boundaries (and Hummel and Stankiewicz acknowledge this). Indeed, the fact that we *can* clearly distinguish the base stimulus from test stimulus 1 lends some intuitive support to the view that such metric information is encoded, at least coarsely.[18] The question, it seems, is how to account for the *contrast* in sensitivity between the two types of changes.

In light of this, I believe that an intermediate view is most reasonable. The view involves two proposals. First, part boundaries are explicitly encoded *both* numerically *and* categorically. Because categorical relations are made explicit, this plausibly enables such categories to be *highlighted* in shape experience. Second, I propose that the *precision* with which boundaries are numerically encoded is a function of how near those boundaries are to certain "critical" locations along a part's axis, such as its midpoint or

[18] It could be, however, that the difference between the base stimulus and test stimulus 1 is not encoded within a hierarchical shape description, but rather by some distinct visual representation, such as, e.g., a Marrian 2½-D sketch. I leave this open as a possibility, though it is not the view I'll develop here.

endpoint. These critical locations plausibly mark the borders of qualitative categories. Thus, I suggest that a representation scheme adequate to subserve shape phenomenology should encode the coordinates of a part boundary very finely when the boundary is near either a midpoint or an endpoint of a part's axis, but more coarsely (i.e., within a broader interval) when the boundary is farther away from such critical points. I am confident that a version of this approach would predict the results obtained by Hummel and Stankiewicz (1996). For it implies that perceivers should be *somewhat* sensitive to purely metric changes in part boundary location, but *more* sensitive to metric changes that also cross the border of a qualitative category.

This proposal also gives substance to the idea, offered earlier, that part boundaries are specified only "coarsely" or "approximately" within visual phenomenology. If a subpersonal representation of the type suggested here underlies important aspects of shape phenomenology, then the degree of coarseness in one's experience of part boundaries will be a function of where those boundaries are. Accordingly, the precise amount of local feature change necessary to disrupt the experienced compositional structure of an object *O* will vary depending on where an *O*'s part boundaries lie.

Thus, we arrive at the following view. A representation scheme adequate to underlie shape phenomenology (and structure constancy) should: (i) be hierarchical, (ii) encode boundaries between parts in part-centered reference frames, and (iii) encode boundaries both categorically and metrically, although the metric encoding should be more precise near certain critical points along object parts, such as their midpoints and endpoints.

7. Comparison with Other Approaches to Spatial Experience

In this section I contrast the approach developed above with other existing theories of the subpersonal underpinnings of visual spatial experience. I argue that several views on offer do not comport well with the phenomenon of structure constancy, because they cast the subpersonal underpinnings of spatial experience as entirely *viewer-centered*.

Many have been attracted to the idea that visual phenomenology seems to present us with an array of facing surfaces, rather than, say, the 2-D retinal image or the volumetric structure of objects (e.g., Jackendoff 1987; Tye 1991, 1995; Prinz 2012). In light of this, theorists influenced by Marr's (1982) pioneering tripartite theory of vision have sought to locate the underpinnings of visual consciousness at the "intermediate" level of processing, which describes the geometry of surfaces. In Marr's framework, the intermediate level is occupied by the 2½-D sketch, so theorists have often appealed to the 2½-D sketch, though usually with some alterations or enrichments, which I'll discuss below.

The 2½-D sketch is an array specifying the viewer-centered distance, direction, and local orientation at each point (up to a certain resolution) for all visible surfaces in the scene (see Marr 1982: 275-279). It can be construed as a type of "depth map" representing certain spatial properties of thousands of very small surface patches within one's field of vision. The important thing to note is that the 2½-D sketch lacks two features that I have argued are central to explaining structure constancy. First, the scheme is mereologically unstructured. This is because the 2½-D sketch only attributes geometrical features to very small surface patches in one's field of vision, and it does not represent the composition of such surface patches into larger individuals. Second, the scheme is *wholly viewer-centered*. That is, all locations in the visual field are represented relative to an origin centered on the viewer. So the locations of part boundaries are not represented in part-centered coordinates.

Jackendoff (1987) calls on the 2½-D sketch in his account of the subpersonal underpinnings of visual experience, but recognizes that Marr's representational structure has important defects (e.g., lack of explicit surface segmentation, perceptual grouping, etc.). As such, he develops an *enriched* 2½-D sketch, which he calls the 2½-D structural description (see Jackendoff 1987: 331-338). More recently, Prinz (2012) has appealed to Jackendoff's theory in his "intermediate view" of the subpersonal basis of visual consciousness.

Jackendoff enriches Marr's depth map with the primitive elements *boundary* and *region*, and the predicates *directed*, *abutting*, *overflow*, and *occlusion*. Boundaries and regions are obtained by appropriately segmenting the initially undifferentiated 2½-D sketch. The predicates represent properties and relations of these boundaries and regions. For example, the 2½-D structural description has the resources to encode (via the *directedness* predicate) figure-ground relations, and can encode (via the *overflow* predicate) that a region extends outside one's field of vision. Moreover, Jackendoff also incorporates parthood into his 2½-D structural description. He proposes that boundaries are identified not only where one finds luminance edges in the retinal image, but also in accordance with Hoffman and Richards' minima rule.

For our purposes, the important point is this. Jackendoff's model organizes the visual array into objects and parts, but it does not alter the basic reference frame of the 2½-D sketch. The depth map is segmented, and certain properties of segmented regions are represented, but the underlying coordinate frame remains wholly viewer-centered. Likewise, although Prinz (2012) offers some revisions to Jackendoff's model, he agrees that the representation underlying visual consciousness is wholly viewer-centered (Prinz 2012: 50-57). See also Tye (1991: 90-97; 1995: 140-141) for a similar view.

For the reasons canvassed above, viewer-centered representational schemes cannot plausibly underpin structure constancy. Whenever an object moves relative to the perceiver, the viewer-centered locations of its part boundaries change. But to explain the patterns of phenomenological salience associated with such transformations, we need a representation that treats part boundaries as remaining *stable* across such changes in viewer-centered location, so long as they don't shift their positions relative to the connected parts themselves. A part-centered scheme does this, while a viewer-centered scheme does not.

As such, the view I have offered importantly departs from these approaches on a critical dimension of shape representation (viz., its reference frame), though it does have a feature in common with them (viz., incorporating part-based organization).

I should underscore, however, that the view that the locations of certain things are experienced in part-centered reference frames does not imply that we *fail* to also experience things in a viewer-centered reference frame. Indeed, it is an undeniable aspect of our phenomenology that we perceive from a perspective—e.g., that objects are seen to have certain spatial relations to our point of view (e.g., Peacocke 1992; Schellenberg 2008; Bennett 2009).

I think that on the most plausible analysis of our phenomenal experience of spatial properties, vision utilizes *multiple* reference frames simultaneously (cf. Briscoe 2009;

Humphreys et al. 2013). Indeed, the view that perception uses multiple reference frames seems to comport best with the *overall* phenomenology of watching a non-rigid object move. When a person walks, for example, there is a sense in which her joint locations seem to stay stationary, but also a sense in which they seem to move relative to your viewpoint. As such, our visual experience of geometrical properties is at least twofold. We are aware of the spatial relations that things in the world (e.g., objects and their parts) bear to our current viewpoint, but we are also aware of the spatial relations that parts of an object bear to one another, independently of their relations to our current perspective.

7. Conclusion

In this chapter I have offered an account of a novel type of perceptual constancy, which I've called *structure constancy*. I've argued that we visually experience objects as retaining their compositional structure despite certain changes that alter their intrinsic metric properties. Moreover, I've drawn out implications of structure constancy for both the representational content and the subpersonal underpinnings of visual spatial experience.

Appendix

Most part-centered models of shape description are *axial*. They involve specifying intrinsic axes for the overall shape such that each part is associated with a separate axis "branch" (e.g., Blum & Nagel 1978; Rosenfeld 1986; Kimia 2003; Feldman & Singh 2006; Feldman et al. 2013).[19] The boundary of the shape is generated or "grown" from its component axes. Here is a broad outline of how such representation schemes work. I will only consider the 2-D case, in which axes are planar curves.

Representation of part axes: We can describe a part O_1's axis M by the following discrete approximation: (i) the length L of M; (ii) a sequence $(p_1,\ldots, p_n)$ of points, such that for each p_i and p_{i+1}, p_i and p_{i+1} are separated by an interval of length L/n; and (iii) a sequence of *turning angles* $(\alpha_{1, 2},\ldots, \alpha_{n - 1, n})$, where each $\alpha_{i, i+1}$ is the angle between the tangent vectors to M at points p_i and p_{i+1}. A representation of (i)-(iii) specifies the length and curvature of M in a way that is invariant to translations and rotations of O_1 with respect to the viewer.

Representation of part shapes: If M is O_1's medial axis, then M is the set of points such that each $m \in M$ is the center of a "maximally inscribed disc"—i.e., a disc that is wholly contained in O_1 and tangent to the boundary of O_1 at two or more points. The "growth" of O_1 from M is represented by specifying the *radius function* r, where r maps each point on M to the radius of its associated maximally inscribed disc. Thus, by specifying $(r(p_1),\ldots,$

[19] As noted above, however, the correspondence between intuitive parts and medial axes is generally imperfect. See Feldman and Singh (2006) for a Bayesian approach to axial representation that may better correspond to intuitive part decomposition.

$r(p_n)$), we can describe the positions of points on O_1's outer contour in relation to corresponding points on M.

Representation of part boundaries: Axial models aim to identify intrinsic axes in such a way that the axial representation of an overall shape includes distinct branches for each individual part. On such models, part boundaries are associated with points at which the shape's axis branches. Thus, in a hierarchical description, the *B*-edge representing the boundary between parts O_1 and O_2 might (for example) specify either the point p_i or an interval $[p_i, p_j]$ along O_1's axis M at which the endpoint of O_2's axis intersects. This representation would be genuinely part-centered, because the representation of axes described above is invariant to translations and rotations with respect to the perceiver. The representation would also stay approximately stable over rotations of O_2 with respect to O_1 that change the relative orientation of the two parts' axes, but not their location of intersection along M.

Chapter 4

Attentive Visual Reference[*]

1. Introduction

One of the most important advances in the study of visual attention over the last few decades has been the finding that visual attention can be allocated to objects, and not just to the locations they occupy. Since this discovery, there have been various proposals concerning the subpersonal mechanisms and cognitive role of attention to objects. Almost all of them agree on this much: When a person visually attends to an object, her visual system deploys a representation that picks out the object.[1] Call such subpersonal level visual system representations *attentive visual object representations* (hereafter AVORs),[2] and call the representational relation between them and the objects that they designate *attentive visual reference*. This chapter concerns the nature of attentive visual reference.

There is strong support for the view that visual attention can be allocated to individuals of a special category known as 'visual objects' or, better, 'objects of vision.' Very roughly, these are individuals that can be discriminated by early visual processes according to perceptual organization criteria (e.g., perceptual grouping and figure-ground segregation).[3] For present purposes, I'll just call them 'objects.' Thus, a central finding of research into visual attention is the *same-object advantage*: When we are asked to compare two targets, we are both faster and more accurate when the targets appear on the

[*] This chapter is adapted from Green (forthcoming).

[1] See, e.g., Burge (2010), Carey (2009), Dickie (2011), Fodor and Pylyshyn (2015), Kahneman, Treisman, and Gibbs (1992), Levine (2010), Matthen (2005), Pylyshyn (2007), Recanati (2012), and Scholl (2001).

[2] Theorists have given various names to the object representations involved in attention. Kahneman et al. (1992) call them 'object files,' Pylyshyn (2003, 2007) calls them 'FINSTs,' and Recanati (2012) calls them 'perceptual files.'

[3] See, e.g., Brovold and Grush (2012), Feldman (2003), Kimchi (2009), and Yantis (1992).

same object than when they appear on different objects, even if the spatial distance between the targets is held constant.[4] It seems, then, that attention often *selects* or *adheres to* objects. A now standard explanation of this data is that the processes underlying attention deploy representations—AVORs—that pick out certain objects in the scene before our eyes, marking them as selected.

But what is it for an AVOR to 'visually refer' to an object? Attentive visual reference (hereafter 'visual reference'[5]) no doubt requires a certain kind of causal relation between an object (or a state or event in which the object participates) and an AVOR.[6] I won't, however, be concerned with the kind of causal relation in question. My primary interest is in whether *descriptive content* figures in the mechanism of visual reference determination, and if so, what kind of descriptive content so figures. I will assume, as is commonplace in vision science, that the visual system deploys some representations that have descriptive contents (Biederman 1987; Frisby & Stone 2010; Marr 1982; Palmer 1977, 1999). The issue is whether such description-like representations play a reference-fixing role vis-à-vis AVORs.

The plan is as follows. Section 2 motivates the view that AVORs *directly* refer to objects in the scene. In other words, the content of an AVOR just is the object it designates. Section 3 considers two views about the mechanism by which direct visual reference is secured to an object in a given context. The *pure causal view* holds, roughly,

[4] See, e.g., Behrmann et al. (1998) and Marino and Scholl (2005). Feature comparison is one of two paradigms commonly used to demonstrate the same-object advantage. The other is a version of the spatial cuing paradigm (see, e.g., Egly et al. 1994).

[5] This is just a convenient shorthand. I do not claim here that attentive visual reference is the *only* type of singular reference secured by the visual system. It is possible that reference to particulars occurs at a variety of levels of visual processing, and that some of these levels are attentive while others are preattentive (cf. Burge 2010: 451).

[6] Even those who reject a causal requirement on singular reference in general usually concede that reference within *perceptual systems* is likely subject to a causal requirement (see, e.g., Hawthorne & Manley 2012: 26).

that an AVOR refers to the object to which it stands in an appropriate causal or information-carrying relation. The *location-based view* holds, roughly, that an AVOR comes to refer to an object partly by being associated with descriptive information about the spatial location of that object. I argue that both of these views face serious difficulties. Section 4 develops my alternative proposal—the *structure-based view*—according to which an AVOR comes to refer to an object partly by being associated with descriptive information about some of the *geometrical* and *mereological* features of that object. I thus claim that descriptive information plays a role in determining the referent of an AVOR. Section 5 considers two challenges to the structure-based account.

2. Attentive Visual Reference is Direct Reference

2.1. Content-giving vs. Reference-fixing

It is common to distinguish two roles that a description may play with respect to a singular referring expression (such as a name or indexical). First, a description *D* may give the *content* of a referring expression *E*. If so, then by specifying the content of *D* we thereby specify the content of *E*. When *D* plays this type of role with respect to *E*, call *D* a *content-giving description* with respect to *E*. Second, a description *D* may play a *reference-fixing* role with respect to a referring expression *E*. If so, then *E* secures reference to an object *o* partly because *E* is (or has in the past been) associated with *D*, and *o* satisfies (or satisfied) the content of *D*. If a description *D* plays this type of role with respect to an expression *E*, call *D* a *reference-fixing description* with respect to *E*.

If a referring expression *E* lacks a content-giving description, it is standard to hold that *E* refers *directly*. That is, the content of *E* just is the individual it picks out, rather than a descriptive condition that the individual satisfies. Thus, Kripke (1980) makes the case that proper names refer directly primarily by attacking the proposal that their contents can be given descriptively, and Kaplan (1989) makes a similar argument in the case of indexicals. Here I'll just assume that such arguments succeed—I'll assume that if an expression *E* refers to an individual *o* but lacks a content-giving description, then *E* refers to *o* directly.

It is also fairly commonplace to hold that a description can be a reference-fixing description with respect to a given expression without being a content-giving description with respect to that expression. For example, it has been proposed that the names 'Jack the Ripper' and 'Neptune' refer directly, but nevertheless have (or had) their references fixed in part by description (e.g., Kripke 1980; Jeshion 2002).

Accordingly, there are two distinct questions about the role that descriptions play with respect to AVORs. First, do AVORs have content-giving descriptions? Second, do AVORs have reference-fixing descriptions? In what follows, I'll argue that the answer to the first question is 'no,' while the answer to the second question is 'yes.' AVORs directly refer, but they have associated reference-fixing descriptions.

2.2. An Argument for Direct Visual Reference

In this subsection I'll offer a brief argument for the view that AVORs refer to objects directly. This view has recently been endorsed by a number of theorists (see Dickie 2011; Levine 2010; Pylyshyn 2007; Recanati 2012), though the position is by no means universal.[7] A primary argument for the view—which I'll call the *temporal argument*—appeals to our ability to track objects through feature changes.

Consider, for example, the *apparent motion* paradigm: When one stimulus is flashed followed by another flashed in a different location, then with a suitable spatiotemporal gap between the flashes, the viewer will have a visual experience as of a single object moving continuously from one location to the other. In a key variation on the design, Navon (1976) showed that apparent motion occurs even when the two successive stimuli differ in both color and shape (e.g., a blue square followed by a red circle). One will, in similarly compelling fashion, have an experience as of a single enduring object moving from one location to the other. However, the object appears to change both shape and color along the way.

Similar findings are supplied by the *multiple-object tracking* (MOT) paradigm. In this paradigm, subjects are asked to keep track of several (3-5) targets that move around a screen in the presence of a set of distractors. Incredibly, people are highly adept at performing this task, routinely displaying accuracy greater than 90% (Pylyshyn & Storm 1988; vanMarle & Scholl 2003). But most crucially, it seems that people can successfully track targets despite *changes* in their shapes or colors.[8, 9] The mechanisms responsible for

[7] For instance, some, such as Searle (1983), hold that the perceptual representation of objects is wholly descriptivist. A similar position is suggested by views on which perceptual content is entirely general or existentially quantified (see, e.g., Davies 1992; Pautz 2009).

[8] See Bahrami (2003), vanMarle and Scholl (2003), and Zhou et al. (2010). The situation is complicated somewhat by the latter study, which indicates that topological changes (e.g., the addition of a hole) do slightly disrupt tracking. However, even in this case overall performance is still quite good (85-90% accuracy).

[9] There is disagreement concerning the processes that enable performance on the multiple-object tracking task. While Pylyshyn and Storm (1988; Pylyshyn 2007) interpret MOT as involving a set of independent object representations that each pick out a distinct object and track their referents in parallel, others have proposed that the visual system perceptually groups the targets as vertices of a single 'virtual polygon' (e.g., Yantis 1992).

visual tracking thus appear to treat objects as enduring through these changes, at least within the context of the MOT task.[10]

These findings demonstrate what I shall call *object-selective sensitivity*—visual attention that is selective for an object over time, but not selective for any of that object's features (e.g., its color, size, or shape). This in turn suggests that AVORs are *tolerant of changes* in the properties of the objects to which they refer—they can continue to refer despite such changes.

This fact about AVORs is in tension with the view that they have content-giving descriptions. To see the problem, suppose that you attend to a red ball at time t_0, and your visual system deploys an AVOR that designates the ball. And suppose that this AVOR were equivalent to a description of some of the ball's features—say, 'the red ball at location *l*.' With respect to any given time, this description denotes whatever is a unique red ball at *l*. However, the existing evidence warrants the prediction that (at least within certain contexts) your AVOR could *continue* to designate the object even if, between t_0 and t_2, it should shift its location and morph into a green cone. Since the description doesn't do this, your AVOR differs from that description in what it picks out with respect to various times. Therefore, the two cannot be equivalent in content. This is the temporal argument for direct visual reference.[11]

I should emphasize that the temporal argument does not rely on the claim that features of an object are not *registered* or *used* during tracking.[12] Rather, it relies only on

[10] A similar phenomenon occurs in the *tunnel effect* (Flombaum et al. 2004).

[11] Kahneman et al. (1992) and Pylyshyn (2007), among others, seem to endorse some version of the temporal argument. One will notice the analogy between the temporal argument and Kripke's (1980) modal argument for the direct reference of proper names.

[12] Indeed, it is plausible that when the visual system tracks an object over time, it often takes into account some of the object's features, including its recently stored location and motion trajectory (Tripathy et al. 2011), and sometimes its surface properties such as color, size, and shape (Feldman & Tremoulet 2006;

the claim that the visual system can treat objects as retaining identity over time across *changes* in their features. The latter claim does not entail the former. Thus, suppose (for illustration) that the visual system tracked objects by applying a very simple rule: For successive times t_0 and t_1, an object *o* at t_0 is identical to an object *o** at t_1 just in case *o*'s represented location at t_0 is closer to *o**'s represented location at t_1 than it is to the represented location of any other object discriminated at t_1. To apply this rule in a given case, the visual system may well need to register a property of *o* at t_0 (its location), and use this property in determining how *o* persists over time. Still, the visual system would be prepared to track *o* through a change in that very property, since the rule obviously permits *o*'s successor at t_1 to differ from *o* in its location. As such, even if the visual system *uses* location (or other properties, such as color or shape) while tracking an object, a description that specifies the object's location at a particular time would still be inadequate to designate the object at later times.

The main strike against descriptivism about visual reference, then, is that the view fails to capture our ability to track—and hence maintain representations of—changing objects over time. So it seems reasonable to conclude that AVORs lack content-giving descriptions. They refer directly, in a manner akin to indexicals like 'this' or 'that.' This

Hollingworth & Franconeri 2009). As an anonymous reviewer has pointed out, different features of an object will likely be used in computing object persistence in different contexts, and the nature of the MOT task may lead the visual system to prioritize an object's represented location over other features, such as its shape and color. However, other tasks facilitate the use of surface features. One such task is the 'bouncing/streaming' paradigm employed by Feldman and Tremoulet (2006), in which information about spatial location is made deliberately ambiguous, so the visual system must rely on surface features. Furthermore, evidence from the *visual marking* paradigm indicates that a set of items can be inhibited as a group according to their surface features (e.g., color or shape), and that changes in these features abolish inhibition (Watson & Humphreys 1997). As such, there is considerable evidence that the visual system uses features in a number of object-based processes, such as computing object persistence over time and maintaining attentional inhibition.

is an important result, because it means that visual object representation cannot be analyzed into a type of feature ascription.

3. Pure Causal Views, Location-based Views, and Their Discontents

I turn now to the primary task of this chapter: offering an account of how visual reference is secured in context. That is, I am interested in the facts that make it the case that, relative to a context, an AVOR secures reference to an object. We can distinguish between two positions. On one type of view, an AVOR refers to a given object simply by virtue of the object standing in a certain causal or informational relation to the AVOR. Call this the *pure causal view*. On another type of view (a 'causal-descriptivist' view), causal and informational relations must be supplemented with descriptive content that the object satisfies in order for an AVOR to secure reference to the object. Within this latter camp, there are many specific positions possible, depending on the type of descriptive content that the theory imports. However, the most popular variant has held that visual reference to an object depends critically on the possession of information about the *spatial location* of the object in question. Call this the *location-based view*. In this section, I'll argue against the pure causal view and the location-based view in turn.

3.1. Pure Causal Accounts

We can state the pure causal view more rigorously as the claim that for a token AVOR *R* to directly refer to an object *o* in a context *C*, it is sufficient that, in *C*, the event of *R*'s being deployed is appropriately caused by some event or state in which *o* participates.[13]

[13] The 'appropriateness' condition will be needed in order to circumvent, for example, the well-known problem of deviant causal chains (e.g., Coates 2000).

Views of this sort have been defended by Pylyshyn, Fodor, and Dretske, among others.[14] For present purposes I will concentrate on the version of the view found in Pylyshyn's work.

Pylyshyn (2003, 2007) posits a set of 4-5 *FINST*s, which are visual symbols deployed to designate objects. The theory of FINSTs offers one way of developing the notion of an AVOR. On Pylyshyn's view, FINSTs are context-sensitive, they refer to external objects directly, and they are assigned as part of the process of *selection* (an attention-like process whereby the visual system devotes additional processing to some subset of the information to which it has access). A crucial feature of Pylyshyn's account is that, relative to a context, the reference of a FINST is fixed in a purely causal manner. Pylyshyn frequently depicts the process of reference-fixing as consisting in an object's 'grabbing' or 'capturing' a FINST. By this, he means that an object (or rather, an event involving the object) appropriately causes a FINST to be deployed, and because of this, the FINST refers to the object in question. Pylyshyn is adamant that the mechanism for determining the reference of a FINST does not involve descriptive information—securing reference to an object does not, on this view, depend on having represented any of its properties.[15] So reference-fixing descriptions are ruled out wholesale.

Though there is much that I find attractive in Pylyshyn's account, I will argue that in virtue of its purely causal character, the theory of FINSTs encounters two significant difficulties. First, it faces problems of referential indeterminacy due to what I call the

[14] See Dretske (1995), Fodor (2008), Pylyshyn (2007), and Fodor and Pylyshyn (2015). Recanati (2012) joins these theorists in the view that the mechanism of visual reference determination is entirely non-descriptive.

[15] For example: 'There are specific properties that cause a FINST index to be assigned and that enable it to keep track of the indexed individuals—but these properties are not encoded, and a representation of these properties is not used in carrying out those functions' (Pylyshyn 2007, p. 206).

circumscription problem. Second, it appears to conflict with findings in computational vision science.

One of the most pressing issues facing any purely causal account of reference determination is to avoid pervasive referential indeterminacy. Proponents of pure causal theories need to offer an account of which link in a (non-deviant) causal chain of events preceding the tokening of a symbol counts as the *referent* of that symbol. I want now to focus on a particular variety of which-link problem that is particularly difficult for the pure causal theorist to solve. I suggest that any purely causal account of reference determination for AVORs gives rise to a quite worrisome indeterminacy between *parts* and *wholes*.

To draw out the difficulty, I will first motivate the claim that states of visual attention can refer to parts of larger individuals. Thus, consider the lamp in figure 4.1, and do the following. First fix your attention on the entire lamp. Now shift your attention to focus only on the lampshade. Finally, shift your attention once again, and focus only on the lamp base. The result should be clear: it is possible to attend selectively to the parts of an individual, either at the expense of, or in addition to, attending to the individual of which they are parts. (Which of these alternatives is the case will not matter in what follows—indeed, both may be true.)

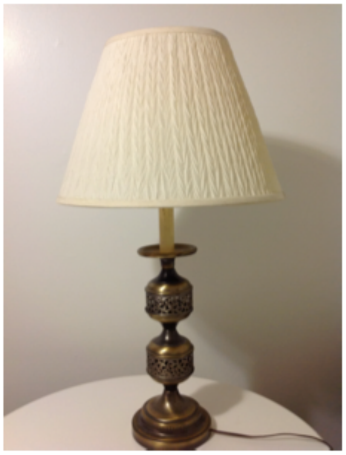
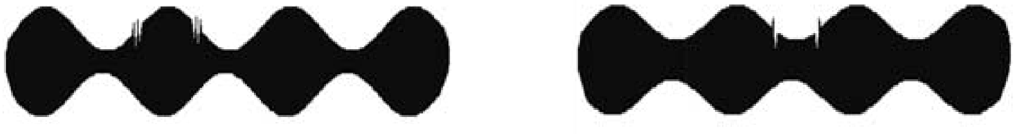


Figure 4.1. A multi-part lamp

While I find the foregoing phenomenological demonstration sufficiently compelling, the claim that attention can be selective for parts also finds empirical confirmation. As discussed earlier, there is considerable evidence that visual attention often adheres to objects. But research also indicates that visual attention can be *part-based*: Judgments that pertain to features of the same part of an object are faster and more accurate than judgments that pertain to features of different parts.

Barenholtz and Feldman (2003) presented subjects with one of the stimuli shown in figures 2a and 2b and asked them to indicate whether the two marks on the contour of the object were the same or different. Each figure seemed to decompose naturally into parts—most perceivers see the 'humps' as separate parts of the object. And critically, the marks to be compared could appear either on the *same* hump (figure 4.2a) or on *different* humps (figure 4.2b). It was found that even though the distance between the marks was held constant, judgments were significantly faster in the former case. This provides compelling evidence that the distribution of attention is sensitive to part boundaries within objects (see also Vecera et al. 2000; Watson and Kramer 1999).



Figures 4.2a (left) and 4.2b (right). Stimuli used by Barenholtz and Feldman (2003). Reproduced from Barenholtz and Feldman (2003) with kind permission from Elsevier.

But if, as I have argued, visual attention to a thing is sufficient for direct visual reference to that thing, then it must be the case that AVORs can refer either to objects or to their parts. When you attend to the whole lamp in figure 1, your visual system directly refers to the whole lamp. But when you attend only to a part, your visual system directly refers to that part. Thus, I contend that visual attention can incorporate reference to the lampshade, the lamp base, or the whole lamp, depending on how it is distributed.[16]

How does this bear on the pure causal view of visual reference determination? Recall that according to the pure causal view, for a token AVOR *R* to directly refer to an object *o* in a context *C*, it is sufficient that, in *C*, *o* appropriately cause *R* to be deployed. Now consider a case in which a subject *S*'s visual system deploys an AVOR *R* that refers to the lampshade in figure 1. The pure causal theorist holds that *R* comes to refer to the lampshade because some event in which the lampshade participates appropriately causes the deployment of *R*. But now the problem should come into focus: The deployment of *R* is *also* caused—and, in all likelihood, appropriately caused—by an event in which the *whole lamp* participates. So how can the pure causal theorist ensure that the lamp is not the referent of *R*? Call this the *circumscription problem*: Whenever visual reference is

[16] However, it is clear that some parts of objects are more natural targets for attention than others. For example, one cannot (without difficulty) attend only to the left half of the lampshade. This is to be explained by appeal to the rules according to which early vision parses objects into parts. I discuss these rules in section 4.

fixed to a particular thing, it is necessary to resolve the *spatial boundaries* separating the thing from its background and from other nearby objects.[17]

As noted earlier, the circumscription problem is an instance of the more general ‘which-link’ problem for causal theories of content determination. Since a problem of this sort is generally acknowledged, it is also generally acknowledged (even by those I have labeled pure causal theorists) that something beyond a *bare* causal relation must be called on for determining reference. Nevertheless, existing approaches to solving the which-link problem will not solve the circumscription problem, so the circumscription problem is an especially pressing instance of the which-link problem. Here I’ll consider two such approaches.

Fodor (2008: ch. 7) has recently appealed to a method of *triangulation* for solving the which-link problem. Triangulation works as follows. Imagine that a perceiver *S* is attentively viewing an object *o*, and deploys an AVOR *R*. What makes it the case that *R* refers to *o* rather than another entity in the causal chain leading from *o* to *R*, such as a pattern of retinal stimulation? Fodor suggests that we consider a *counterfactual duplicate* of *S* standing, say, three feet to the right of *S*, who views *o* and deploys an AVOR of her own. Call this duplicate *S** and call her AVOR *R**. Now trace back the causal chains preceding the deployments of both *R* and *R**. If the two chains intersect at a link, then, Fodor suggests, *R* and *R** each has that link as its referent. The proposal is that the *only* link these chains will have in common is the object *o*.

However, despite its ingenuity, Fodor’s suggestion does not solve the circumscription problem. For if *S* and *S** both view the lampshade in figure 1, then the causal chains preceding *R* and *R** will share at least *two* links: one involving the lamp,

[17] I borrow the term ‘circumscription’ for this process from Keane (2009).

and one involving the lampshade. So the perceivers triangulate on at least two objects. Thus, we still cannot determine which of these objects counts as the referent of *R*, and the circumscription problem remains.

An earlier suggestion is due to Dretske (1981), who attempts to solve the which-link problem by appeal to information-carrying relations. His account is, roughly, this: A perceptual representation *R* (e.g., an AVOR) can carry information about a distal object *o* (e.g., a lamp) rather than a more proximal cause *c* (e.g., a pattern of retinal firing) because, given constancy mechanisms, *o* would have led to *R* via a causal chain that did *not* involve *c* had there been, say, a slight difference in viewpoint. Thus, the probability of *o* given *R* is higher than the probability of *c* given *R*, so *R* carries more information about *o* than about *c*.

Irrespective of the merits of this approach for ruling out *proximal* links along the causal chain, it does not seem to address the circumscription problem. For it seems clearly possible that the following two facts might hold: (i) Whenever the lamp causes *R*, it has the lampshade as a part, and (ii) whenever the lampshade causes *R*, it is part of the lamp. In such a case, *R* carries the same amount of information about the two objects, so information-carrying relations alone cannot recommend one over the other as the referent of *R*.

The circumscription problem thus seems to resist the solutions that Fodor and Dretske have offered to the which-link problem. I'll now consider two more potential responses on behalf of the pure causal view.

First, one might propose that we can solve the problem of referential indeterminacy while maintaining the non-descriptive character of visual reference-fixing if we supplement the facts about causation with brute facts about attention allocation. Thus, when *S* allocates attention to the lampshade rather than the whole lamp, she is indeed in causal contact with both, but since she only attends to one of them, this fact about attention can resolve the indeterminacy.

However, this response leaves unexplained how visual attention *itself* gets allocated to the lampshade rather than the whole lamp. And the pure causal theorist needs to explain this without appealing to descriptive information that distinguishes the two objects, lest the view collapse into a type of causal-descriptivist position. Here the pure causal theorist faces a dilemma. She can either hold that causal/informational relations on their own are sufficient to explain how attention is allocated to the lampshade, or she can hold that they are not. The first option does not appear promising, because the same indeterminacy problems will resurface. Since *S*'s attentional systems are in causal contact with both the lamp and the lampshade, we require an explanation of how she manages to attend *only* to the lampshade. But if the pure causal theorist adopts the second option, she owes us an account of what *does* determine how attention gets allocated to the lampshade rather than the lamp. Either way, we are back where we started—we need to resolve the lamp/lampshade indeterminacy. The only difference is that the indeterminacy now concerns attention allocation itself, rather than visual reference.

Another response would be to claim that the evidence for part-based attention does not compel us toward the view that AVORs can refer to parts of objects directly. Perhaps the visual system refers to 'whole' objects directly (using AVORs), but picks out parts of objects only by description. The pure causal theorist could then resolve the indeterminacy between parts and wholes by claiming that an AVOR directly refers to the *whole object* that causes its deployment.

But why should we accept that parts are picked out only by description? It might be suggested that, unlike attending to a whole object, attending to a part does not enable one to track the part through feature changes. However, this seems wrong. Imagine, say, tracking a person's arm as it turns about its joint and gradually contracts and changes color.[18] If the pure causal theorist wishes to pursue this sort of response, we are owed an account of why attention involves direct reference when allocated to whole objects, but not when allocated to their parts.

It should be noted that Pylyshyn's view also faces another serious problem. Pylyshyn claims not only that FINSTs have their reference determined without the help of descriptive content; he also sometimes suggests that FINSTs are assigned *prior* to the visual representation of features such as shape, color, texture, etc. (e.g., Pylyshyn 2003: 217-219). But that claim appears mistaken. To take a single example: Assigning a FINST to an individual object arguably requires, among other things, first segregating that object from its background. Indeed, Pylyshyn accepts this requirement (see Pylyshyn 2007: 31). But essentially all theoretical models of figure-ground segregation invoke the representation of geometrical features such as convexity, symmetry, and closure (Vecera & O'Reilly 1998; Peterson et al. 2000). Very briefly, to determine which of two neighboring visible surfaces is figure and which is background, the visual system must

[18] It should be noted that when subjects are asked to track 'arbitrary' parts of objects (e.g., the endpoints of lines) rather than whole objects as the objects move about the screen, tracking is highly impaired. But tracking under these conditions is not nearly as impaired for intuitively 'good' parts of objects (e.g., square ends of dumbbells) as for arbitrary parts (see Scholl et al. 2001; Howe et al. 2012). (I'll return to this in chapter 5.) Moreover, a full assessment of the ability to track parts must also take into account conditions in which the rest of the object is stationary, and only a part changes location (e.g., rotation of an arm).

represent and compare features of those two surface regions: the surface region that is, e.g., more convex or symmetric is more likely to be figure, while the region that is, e.g., less convex or symmetric is more likely to be background. It is unclear how Pylyshyn's view can avoid straightforward inconsistency with these models.

I maintain that the most promising way to avoid the foregoing difficulties is to incorporate descriptive information into the mechanism of visual reference determination. An AVOR can refer to the lampshade rather than the whole lamp because, *inter alia*, it is associated with descriptive content that distinguishes the lampshade from the lamp. But what type of descriptive content will do? The most common answer philosophers have given is 'location.'

3.2. Location-based Accounts

Location-based views claim that securing reference to an object through perception requires locating it in space (in either egocentric or allocentric coordinates).[19] One familiar location-based model of reference-fixing is due to Evans (1982), who holds that to have the ability to demonstrate an object directly on the basis of perception, it is necessary to meet two conditions. First, one must enjoy an 'ongoing information-link' with the object. This consists roughly in the disposition to update beliefs that involve one's concept of the object in response to information gained through a causal link with the object. Second, one must possess 'a conception of [the object] as the occupant of such-and-such a position (at such-and-such a time)' (1982: 149). This condition seems to require that one be in possession of a description that specifies the spatial location of the

[19] For such views, see Strawson (1963: ch. 1), Evans (1982: ch. 6), and Clark (2000: ch. 4).

object to be referred to.[20] This description is a reference-fixing description on my construal, since its being satisfied by the object is necessary for securing reference to that object.

One can naturally apply this type of account to visual reference as follows. For an AVOR *R* deployed by a subject *S*'s visual system to directly refer to an object *o*, (i) there must be an ongoing information-link from *o* to *R* through which representations involving *R* are updated, and (ii) *S*'s visual system must identify *o* via a description of the form 'the thing at region *l*,' and associate *R* with that description. An advantage of this view over the pure causal position is that it enables us to distinguish reference to whole objects from reference to their parts, because location information would be sufficient to solve the circumscription problem. The lamp occupies a different spatial region from the lampshade (although the two overlap), and so only one of the two candidate referents can satisfy the descriptive condition on visual reference-fixing. Unfortunately, however, (ii) is dubious.

Consider, first, a mundane example: Suppose that you view an object *o* through a mirror of which you are unaware, so that *o* appears to be at a location it is not.[21] It seems plausible that despite this you may be in a position to track *o* as it moves about and undergoes feature changes. As such, it seems plausible that your visual system deploys an AVOR that directly refers to *o*. But if so, then your AVOR refers to *o*, and it is not associated with an accurate description of *o*'s location. So condition (ii) is false.

[20] There is another possible reading of Evans's second condition, according to which having a conception of an object's location is not to have a suitable descriptive *thought* about the object, but instead to have certain *behavioral dispositions* with respect to the object, such as being able to locate it through grasping movements. On this version of the location-based view, visual reference to an object requires having the appropriate dispositions to locate it in behavior. Although I won't consider this type of view in detail here, the discussion below is sufficient to cast doubt on it as well.

[21] Campbell (2002: 111) raises this type of example as a problem for Evans.

One might respond that in this case your AVOR does not refer to *o*, but rather to *o*'s reflection in the mirror. However, this does not work, because the object you track appears to be located *not* where *o*'s reflection is located, but rather somewhere *behind* the mirror.

Another response is that, intuitions notwithstanding, in fact visual reference is not secured in this case—your AVOR doesn't refer to anything. Although it risks being question-begging, this move is not entirely unreasonable. Perhaps we are misled simply because the case is an isolated instance, occurring against the backdrop of accurate localization abilities. As such, I believe that the mirror argument on its own is not sufficient to refute the location-based view. It would thus be more instructive to consider a scenario where a subject suffers from *sustained* location illusions. If the location-based view is right, then this disability should clearly preclude securing visual reference.

As it happens, such disabilities do exist, although they are very rare. The subject A.H., studied extensively by McCloskey and colleagues,[22] suffers from a deficit in which stimuli are normally seen as reflected across either the central horizontal or central vertical axis of her visual field. Thus, a stimulus presented 10º to the left of the central vertical axis of her visual field will often be seen as located 10º to the right. This leads A.H. to exhibit systematic errors in reaching for objects. Nevertheless, A.H. is completely unimpaired in classifying objects according to shape or color, which suggests that she represents the properties of objects perfectly well, save for location. Moreover, there is good reason to hold that A.H. has a genuinely *visual* impairment, rather a deficit of location coding in the action system. First, A.H. is also error-prone when simply asked to *report* the location of an object, rather than to reach for it. Second, when A.H. is required

[22] See McCloskey et al. (1995), McCloskey and Rapp (2000). This case is discussed in Ayob (2008).

to act toward objects *without* the guidance of vision (e.g., pointing to the source of a sound heard with her eyes closed) she exhibits normal performance. This makes it unlikely that A.H. has a deficit of location coding in the action system, because if she did, then the deficit would be expected to carry over to auditorily guided tasks as well.

While A.H.'s deficit appears to be congenital, a different patient, P.R., suffers from a similar mirror-reversal of visual localization brought on by cerebral hypoxia. Like A.H., P.R. displays mirror-reversal when asked to copy a line drawing (Pflugshaupt et al. 2007). Pflugshaupt et al. (2007) also found that when a target appeared to the right of her fixation point, P.R. would consistently make saccades in the opposite direction. Furthermore, a letter-reading task revealed that P.R. is significantly faster at reading mirror-reversed characters relative to normal characters, and is significantly faster when asked to read a string of characters from right to left than when asked to read the string from left to right.

Location-based views entail that A.H.'s and P.R.'s attentional states do not directly refer to objects, since they generally fail to satisfy condition (ii): their AVORs are associated with locational descriptions that are not satisfied by the objects to which they bear information links.[23]

Contra this, I argue that these subjects retain the ability to directly refer to objects through attention. My argument is the following: (1) A.H.'s and P.R.'s AVORs designate objects in their fields of vision. (2) If A.H.'s and P.R.'s AVORs designate objects in their fields of vision, then as long as they retain the ability to display object-selective

[23] The same goes for alternate location-based views where visual reference to an object is partially secured by having the appropriate behavioral dispositions toward that object (see note 20). A.H.'s immediate behavioral dispositions (e.g., reaching behavior and even eye movements) are misaligned with the true locations of objects.

sensitivity, their AVORs directly refer to objects in their fields of vision. (3) A.H. and P.R. plausibly retain this ability. Therefore, their AVORs directly refer to objects.[24]

I endorse the first premise of the argument because I find it to be the best explanation of the capacities A.H. and P.R. possess—their visual capacities are, like ours, specific to particular objects. For instance, when a yellow banana causally affects A.H.'s visual system, it will seem to her that she is seeing and attending to a particular object that is yellow and banana-shaped. As such, she appears clear able to discriminate the banana from its surroundings, and hence to solve the circumscription problem. Moreover, since she is able to attend and respond selectively to changes in objects, if the banana changes color, shape, or texture, her visual system will respond by revising its representation of the object. Since the capacities to discriminate and respond selectively to changes in objects are capacities specific to particular objects, the most plausible view is that the contents of the representations that underlie these capacities are specific to the objects in question. Thus, A.H.'s and P.R.'s AVORs designate objects.

Note that the first premise on its own is not enough to raise problems for the location-based position. Condition (ii) only places a requirement on *directly* referring to objects through vision, and it may be that while (e.g.) A.H.'s attentional states designate objects, they do not directly refer to objects. For both definite descriptions and singular terms designate. But, turning to the second premise, as long as A.H. and P.R. retain the

[24] Impairments of visual localization also occur in Balint's syndrome, though they are generally accompanied by simultanagnosia (the inability to perceptually represent more than one object at a time), and impairments in feature binding. The latter deficit leads to an increased rate of illusory conjunctions, in which, say, a display containing a green triangle and yellow circle is misperceived as containing a green circle and yellow triangle (Robertson et al. 1997). Whether Balint's syndrome cases pose a problem for the location-based view depends, I believe, on whether some form of feature binding is a necessary condition on securing visual reference (and moreover, on the precise nature of the binding deficits exhibited by Balint's patients). I lack the space to consider these issues here, but see the recent disagreement between Campbell (2007) and Schwenkler (2012).

capacity for object-selective sensitivity (i.e., the capacity to attend selectively to an object over time through changes in its features), we can call on the argument given in section 2.2 to show that their AVORs directly refer to objects.

Do localization deficits provide any reason for predicting that a subject's AVORs are intolerant of feature changes? It appears not. For example, a deficit in localization needn't affect the ability to compute mappings between stimuli in apparent motion, since this depends primarily on encoding the stimuli as having the appropriate spatial distance from each other (e.g., Dawson 1991). But A.H.'s and P.R.'s visual systems retain the ability to encode distances, since reflecting the visual field about a central axis preserves distances. As such, our best conjecture is that they retain the ability to perceive apparent motion. And if this ability is unimpaired, then they will likely be able to treat two successive stimuli as the same object even when the stimuli differ in their sensory features.

If A.H. and P.R. retain the capacity for object-selective sensitivity, then we have the fuel to construct a temporal argument for direct reference. Since they have the ability to maintain representations of objects through changes in sensory features, their AVORs cannot be equivalent in content to (present-tensed) descriptions that cite sensory features. Their AVORs will differ from such descriptions with respect to their temporal profiles.[25] We should conclude, I contend, that A.H.'s and P.R.'s attentional states refer to objects directly. Given that, it follows that (ii) is false: It is possible to be in a visual state that directly refers to an object without correctly locating it.[26]

[25] The 'temporal profile' associated with a representation consists of those things that the representation picks out with respect to various times. More specifically, it is a function from times to extensions.

[26] As I remarked earlier, A.H. and P.R.'s deficit is very rare. As such, there is perhaps a greater concern than usual that the deficits they exhibit have not yet been correctly characterized. However, I believe these

Note, however, that while I have relied on the claim that A.H. and P.R. have *inaccurate* visual representations of location (and thus that their location representations cannot serve as reference-fixing descriptions), I am *not* claiming that their visual representations of location fail to be *used* in a variety of visual processing tasks.[27] Indeed, perceptual location representations might be computationally useful despite being inaccurate. Imagine, for example, that the representation of location is critical for binding together features encoded in separate visual feature maps as features of the same individual (e.g., Treisman & Gelade 1980). On this view, it is because two features (e.g., color and texture) are represented as occupying the same location that they are represented as features of the same object. If the misrepresentation of location is *systematic* across feature maps (e.g., if all feature maps are mirror-reversed relative to the distal environment), then location representation could still adequately perform its feature-binding function despite being inaccurate (cf. Campbell 2002: 90-96). For instance, if both color and texture feature maps are mirror reversed, then color and texture features that belong to the same individual will still be *represented as* having the same location, even though both are represented as having the *wrong* location. But the former is all that is needed for feature binding to succeed.

The upshot is this. To avoid referential indeterminacies such as the indeterminacy between the lamp and the lampshade, AVORs must be associated with descriptive information sufficient to solve the circumscription problem (i.e., the problem of resolving an object's spatial boundaries). But while veridical information about an object's location

concerns are somewhat ameliorated by the fact that both patients have been studied in great detail, using a variety of experimental paradigms. These various paradigms provide converging evidence for selective deficits in visual localization.

[27] Thanks to an anonymous reviewer for pressing me to clarify this point.

would arguably be *sufficient* to solve the circumscription problem, the cases of A.H. and P.R. indicate that it is not *necessary*. I now argue that there is a better option. While visual reference to an object can plausibly be secured in the absence of accurate information about its location, I contend that the abilities to circumscribe and so visually refer to an object require the visual system to possess at least roughly accurate information about its *shape*.[28] As such, visual representations of shape plausibly play a reference-fixing role vis-à-vis AVORs.

4. The Structure-based View of Visual Reference Determination

This section develops a view of visual reference determination that avoids the problems with pure causal and location-based views and, moreover, enjoys empirical support in vision science. I propose that direct visual reference to an object is fixed partly on the basis of a *hierarchical description* that specifies (i) the object's geometrical properties, and (ii) its mereological relations to both its parts and objects of which it is a part. Moreover, to secure visual reference to an object, the object must at least roughly satisfy a description of this type. First, I'll explicate the idea of hierarchical shape description and present the structure-based view of visual reference determination. Then, in 4.2 and 4.3, I'll discuss some empirical evidence in favor of this view.

4.1. The Structure-based View

Consider once again the lamp in figure 4.1. Most observers, when asked to indicate the intuitive parts of the object, will judge that it is naturally divided at roughly the points indicated by the arrows in figure 4.3 (cf. DeWinter & Wagemans 2006). These points

[28] The view that a type of shape representation is critical in establishing perceptual links to particular objects can also be found in Schwenkler (2012).

correspond to *negative minima of curvature*. That is, the part boundaries are located at points where the contour of the shape is locally most concave. This is called the *minima rule*, first proposed by psychologists Hoffman and Richards (1984).[29] The idea that shapes are visually parsed into components at concavities has been widespread among perceptual psychologists for decades,[30] and a cursory examination of one's immediate environment reveals that the shapes of most complex objects are decomposable in this manner.

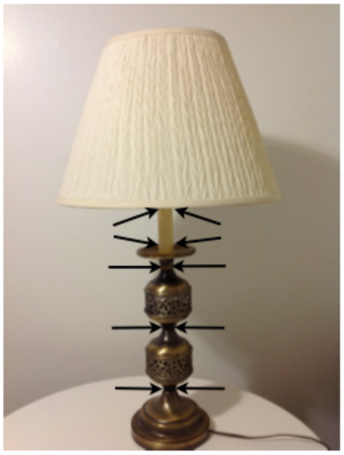


Figure 4.3. Lamp containing part boundaries marked with arrows.

To *parse* a shape is to represent that shape in terms of its parts and the spatial relations among them (i.e., the ways in which they are connected). The fact that decompositions such as the one shown in figure 4.3 are so phenomenologically natural indicates that the visual system indeed engages in shape parsing. Geometrical structure is

[29] More formally, the minima rule for parsing 3-D shapes says to 'divide a surface into parts at loci of negative minima of each principal curvature along its associated family of lines of curvature' (Hoffman & Richards 1984). Roughly, the principal curvatures of a surface at a point are the maximum and minimum 'bends' of the surface at that point. The directions of principal curvature of a surface at any given point are always orthogonal to one another.

[30] See, e.g., Biederman (1987), Palmer and Rock (1994), Singh and Hoffman (2001), DeWinter and Wagemans (2006), and Barenholtz and Tarr (2008). There are, of course, others who reject the approach (e.g., Edelman 1997).

processed at a variety of spatial scales, with structure at lower levels nested within structure at higher levels.

Hierarchical description is a representational format that can be applied to the problem of shape parsing, and more generally to representing both geometrical and mereological structure.[31] Hierarchical descriptions are networks composed of nodes, relations among nodes, and monadic predicates associated with individual nodes. They are usually depicted as trees (see figure 4.4). The uppermost node in a tree is called the *root node*, and is associated with monadic predicates that characterize the shape in question at the most global level. The root node is in turn associated with children, or subordinate nodes, that represent parts of the global shape. The tree will represent the spatial relations that these parts stand in to one another (i.e., how they are connected), along with their monadic geometrical properties (e.g., shape, size, and orientation). Subordinate nodes may in turn have children of their own, and the decomposition may iterate until some perceptual shape primitives are reached.[32]

[31] See Palmer (1977) and Feldman (2003) for this formalism.

[32] Most hierarchical models of shape description are also *part-centered* (e.g., Marr and Nishihara 1978; Feldman and Singh, 2006). What this means is that: (i) the shape of an individual part is represented in terms of axes centered on the part itself, and (ii) the relations between parts are represented in terms of the intrinsic axes of those parts (e.g., the angle formed between the parts' axes). Thus, hierarchical models of shape description are intended to represent an object's shape in a manner that is invariant to the object's viewer-centered location.

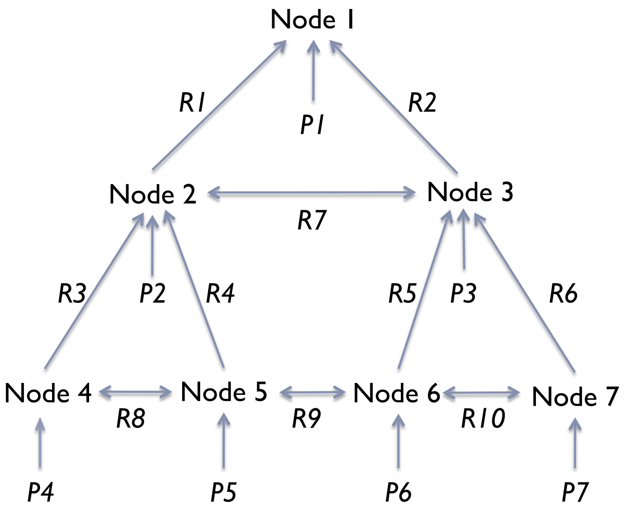


Figure 4.4. The format of a hierarchical description. *P*s are monadic predicates and *R*s are dyadic relations

This much is standard. But it is not always clear how we should conceive of the structural units within hierarchical descriptions. Here I propose that, in the general case, each node is an existentially bound variable. Thus, the content of a hierarchical description can be given (less transparently) by an existentially quantified conjunction. As such, hierarchical descriptions have the logical form of indefinite descriptions—they pick out objects (and parts of objects) via satisfaction.

To make this idea more concrete, consider the double-headed arrow shown in figure 4.5a. A coarse-grained and rather idealized hierarchical description for this shape (according to the minima rule) is shown in 4.5b with quantifiers omitted, and a picturesque decomposition is shown in 4.5c. The content of the hierarchical description in 4.5b is given by the sentence:

> $\exists w \exists x \exists y \exists z(\textit{Double-headed-arrow}(w) \wedge \textit{Triangle}(z) \wedge \textit{Rectangular-Bar}(y) \wedge \textit{Triangle}(x) \wedge \textit{Part-of}(z\ w) \wedge \textit{Part-of}(y\ w) \wedge \textit{Part-of}(x\ w) \wedge \textit{Abutting}(z\ y) \wedge \textit{Abutting}(y\ x))$.

The hierarchical description in 4.5b represents the double-headed arrow, but it does not refer to it directly. The object is only represented *qua* witness to the existential statement above.

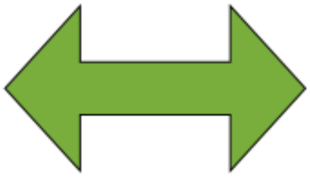


Figure 4.5a. A double-headed arrow

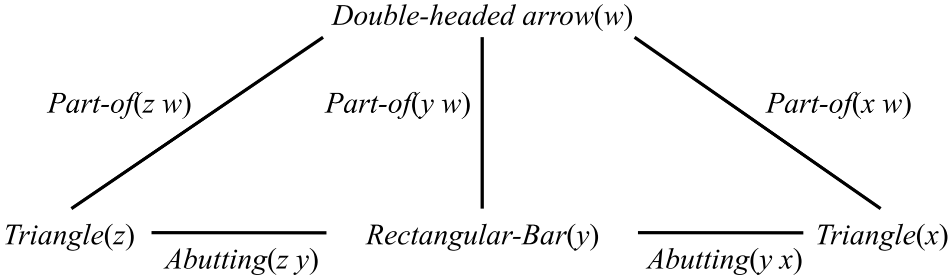


Figure 4.5b. Hierarchical description of the arrow in 4.5a

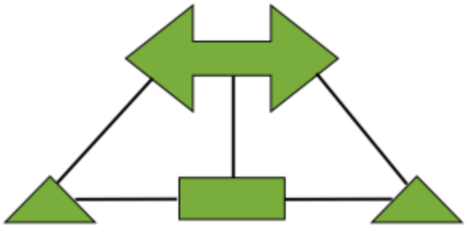


Figure 4.5c. A pictorial rendering of the description in 4.5b

Figure 4.6 shows how the lamp's shape would be decomposed into parts according to the minima rule. Of course, in addition to part *boundaries*, in figures 4.5 and 4.6 I have had to make choices regarding part *cuts*. That is, I have had to choose precisely how to divide the shape, since this is not determined by the minima rule alone. But this problem has also been studied extensively. The visual system appears to prefer part cuts that link two negative minima of curvature, and it usually prefers the shortest such links possible.[33] So, according to these rules, the lamp's shape can be parsed (in the first instance) into six subshapes.

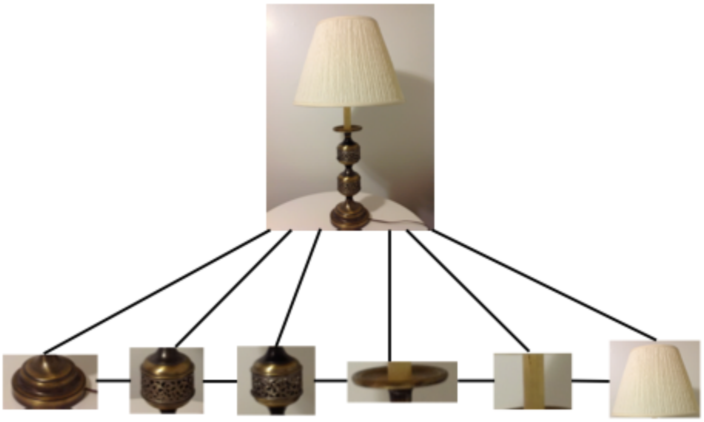


Figure 4.6. Pictorial rendering of a hierarchical description of the lamp in figure 4.1

To a first approximation, then, my proposal is this: Attentive visual reference is fixed partly on the basis of hierarchical shape descriptions generated by early vision. More rigorously, the structure-based model of visual reference determination can be outlined as follows:

1. Preattentive visual processes generate hierarchical shape descriptions for some set of objects in the scene. Each such description contains a set of nodes, and each node is an existentially bound variable. Nodes are associated both with monadic geometrical predicates and with spatial and compositional (part-whole) relations to other nodes.

2. When allocating attention to an object, a mechanism of AVOR deployment operates on a hierarchical description *H*. When this happens, an AVOR is paired with a particular node in *H*.

[33] See, e.g., Singh and Hoffman (2001). However, as these authors note, there are exceptions to this rule.

3. When an AVOR *R* is deployed, it takes an object *o* as its referent only if (i) *o* takes part in an event that appropriately causes the deployment of *R*,[34] (ii) *o* at least roughly satisfies the geometrical and mereological descriptive content with which *R* is associated, and (iii) while the causal connection to *o* remains appropriate, visual representations containing *R* are reliably updated in response to information received from *o*.

4. When an AVOR *R* is paired with a particular node, the predicates and relations previously associated with that node in the hierarchical description take *R* as an argument, substituting it in place of the node.

5. AVORs are akin to indexicals. They are tolerant of changes in the properties of the objects to which they refer. In particular, they are tolerant of changes in the shapes of their referents, despite the fact that geometrical representations are used descriptively during reference-fixing.

I should make three remarks about this model. First, (3) gives only *necessary* conditions for securing visual reference to an object.[35] I leave open whether further conditions may ultimately need to be included. Second, a gloss on the qualification 'roughly' in the second condition of (3): I do not claim that it is impossible to visually refer to an object if one's visual system gets its shape slightly wrong (think of seeing something in a distorting funhouse mirror), but I do think there are limits to how extreme such errors can be. I return to this issue in section 5.2. Third, it is a consequence of (5) that AVORs will differ from their reference-fixing descriptions in temporal profile, so the view is tailored to accommodate object-selective sensitivity. Thus, although the reference of an AVOR is fixed partly on the basis of a hierarchical description, these descriptions do not give the contents of AVORs.

The structure-based view avoids the problems that beset both location-based views and pure causal views. It avoids the problems with the former because hierarchical

[34] I am here assuming a property exemplification account of events (Kim 1976), on which an event is construed as the exemplification of a property by an object at a time. Events, on this view, are individuated by their constituent objects, properties, and times. For an object to *participate* in an event, then, is for it to be a constituent of the event.

[35] By "necessary," I mean "nomologically necessary."

descriptions specify *shape*, rather than location, so accurate localization is not a requirement for securing visual reference. As such, A.H. and P.R. are in a position to satisfy the conditions for visual reference determination.

The structure-based view also sidesteps the problems facing pure causal views, since, in addition to causal relations, the structure-based view also avails itself of descriptive information during the process of visual reference-fixing. As such, visual reference is not indeterminate between objects and their parts. When an AVOR *R* refers to the lampshade in figure 4.1, it is paired with a particular node in the hierarchical description constructed for the lamp, and the referent of *R* must satisfy the descriptive content associated with that node. Specifically, the referent of *R* must, *inter alia*, be (at least roughly) a lampshade-shaped thing that is part of a whole lamp. The lampshade—but not the whole lamp—fits this description, and takes part in an event that causes *R* to be deployed. So *R* refers to the lampshade, not to the whole lamp.

Finally, the causal component of the structure-based model (the first condition of (3)) enables the view to deal with another type of case. Suppose that a perceiver views a scene containing two distinct but qualitatively identical objects, *o* and *o**. The perceiver's visual system constructs two separate shape descriptions, *D* and *D**, upon viewing the scene. Plausibly, if we were to trace the causal chains leading from both *o* and *o** all the way through the perceiver's early visual system, we would find that *one* of the two objects—say, *o*—was appropriately causally responsible for her visual system's generating *D*, while the *other* was appropriately causally responsible for her visual system's generating *D**.[36] Now suppose that a mechanism of AVOR deployment operates

[36] Thus, note that the visual system constructs a shape description as a causal consequence of registering a collection of cues present in proximal retinal stimulation. Given that the visual system constructs *two* such

on the root node of description *D*. Under these conditions, the AVOR refers to *o*, rather than *o**, because it bears an appropriate causal connection to *o*, and not to *o**.[37]

4.2. Support for the Structure-based View: Visual search

If the structure-based view is correct, then two predictions should hold. First, hierarchical descriptions should be constructed *prior* to allocating attention to objects. Second, when the ability to visually process an object's shape is impaired, the ability to selectively attend and so visually refer to that object should be impaired as well. In this subsection I discuss evidence supporting the first prediction. In the next subsection I discuss evidence supporting the second.

There are several lines of empirical support for the claim that shape parsing occurs preattentively. As it happens, we have already encountered one line of support: The research on part-based attention covered in 2.1 indicates that visual attention is naturally distributed in a manner that exhibits sensitivity to part boundaries. The most plausible conclusion to draw from this is that shapes are parsed prior to attention allocation (or, more conservatively, prior to *object-based* attention allocation), and hence that hierarchical shape descriptions are constructed before fixing visual reference. Here I discuss a visual search experiment that also supports this conclusion.

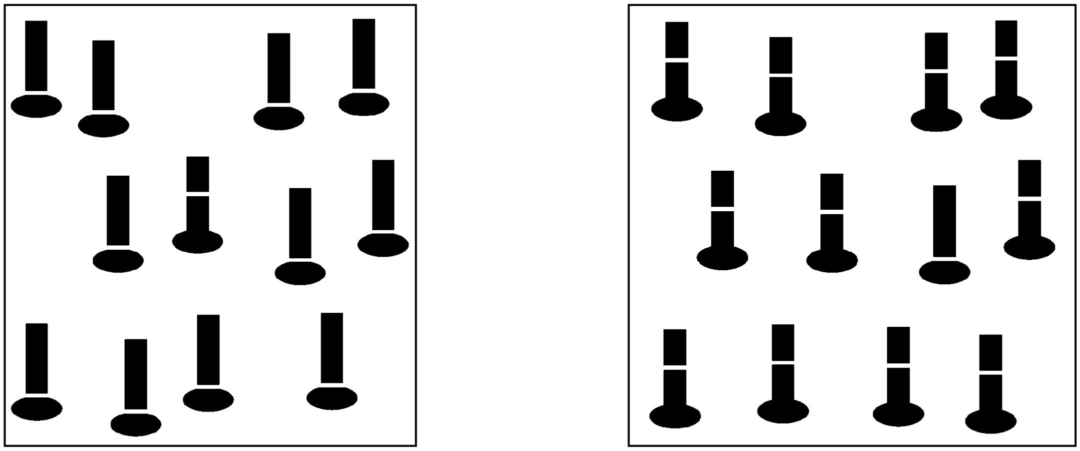
descriptions (and is operating normally in other respects), there must be *two* such collections of proximal cues. One of these collections of cues will have been supplied by *o*, while the other will have been supplied by *o**. Thus, the causal chain leading from *o* to *D* (leading through a particular collection of proximal cues) will be distinct from the causal chain leading from *o** to *D**.

[37] This approach also appears to generalize to cases in which the qualitatively identical objects are incorrectly localized. Thus, suppose that A.H. views *o* and *o**, and mislocates both objects. Because she successfully discriminates the two objects, she will construct two separate shape descriptions. However, one of these shape descriptions will trace its causal ancestry back to *o*, while the other will trace its ancestry back to *o**. When an AVOR is paired with one of these descriptions, then, it takes as its referent the object that is causally responsible for the construction of that description. Thus, the distinction between the two causal chains preceding the two shape descriptions enables the resolution of some of the referential indeterminacy that is left unresolved by the contents of the shape descriptions alone.

In a visual search study, a display of items is presented and the subject is asked to report the presence or absence of a target item. These studies are often used to determine whether a feature is represented preattentively. The idea is that if a display is sufficiently complex, features which are encoded preattentively will be represented in parallel across the display, whereas features whose encoding requires attention will not, since attention is limited at any time to at most a few items in the display.

A common finding in visual search studies is that an item that is discriminable from the rest of the display by the *presence* of a single feature (e.g., a square with a triangular protrusion viewed in a field of normal squares) tends to 'pop out.' This means that the time taken to indicate the item's presence is short and independent of how many other items are in the display. However, an item discriminable only by the *absence* of a single feature (e.g., a normal square presented in a field of squares with triangular protrusions) does not pop out (Treisman & Gelade 1980). If a target fails to pop out, then search is slow and dependent on how many other items are in the display. So it appears that single features, but not absences of features, are generally perceived preattentively. This is called a *search asymmetry*.

Using a sophisticated design, Xu and Singh (2002) sought to determine whether there are search asymmetries associated with preattentive shape parsing. They reasoned that if shapes are parsed preattentively, then, when all items in a display share the same shape, the following two predictions should hold: First, an item that includes a physical gap at an 'unnatural' part boundary (according to the minima rule) should pop out if the rest of the items include physical gaps corresponding to their 'natural' part boundaries (see figure 4.7a). Second, an item that includes a physical gap corresponding to its ‘natural’ part boundary should *not* pop out if the rest of the items include physical gaps at ‘unnatural’ boundaries (see figure 4.7b).



Figures 4.7a (left) and 4.7b (right). Search displays used by Xu and Singh (2002). Reproduced from Xu and Singh (2002) with kind permission from Springer Science and Business Media.

The rationale for these predictions is as follows. Assuming that shapes are preattentively parsed at curvature minima, the presence of a *physical* gap at an object’s curvature minima—a ‘natural’ gap—plausibly does not create an *additional* feature relative to an object that is not physically parsed at its curvature minima. This is because (by hypothesis) the visual system will automatically parse the latter object at precisely that location anyway. However, the presence of a physical gap at some location other than an object’s curvature minima—an ‘unnatural’ gap—*does* create an additional feature, because the object would not have been parsed in that location otherwise. Thus, in figure 4.7a, the target object is distinguished from the rest by the *presence* of a feature (an unnatural gap), while in figure 4.7b, the target object is distinguished from the rest only by the *absence* of that feature (again, an unnatural gap).

These predictions were confirmed. The unnaturally segmented target in figure 4.7a popped out from the display—search time was independent of the number of distractors. However, search for the naturally segmented target in figure 4.7b was slow and inefficient. This indicates that in the display shown in figure 4.7b, the unnaturally segmented distractor items were preattentively parsed at their curvature minima. Because such parsing corresponded precisely with the physical gap in the target, the target's physical gap could not induce pop out.

There is thus sound empirical support for the claim that shape parsing is an early, preattentive process. So we may conclude that hierarchical shape representations are likely available early enough to aid in fixing visual reference, just as the structure-based view requires.

4.3. Support for the Structure-based View: Visual Form Agnosia

According to the structure-based view, a particular type of shape representation (namely, hierarchical description) is a prerequisite for selectively attending to an object, and hence visually referring to that object. If this is true, then visual shape processing impairments should also lead to impairments in selecting objects through visual attention. In fact, we now have evidence that this is the case.

Lesions to the lateral occipital area of the visual system are known to result in a condition called 'visual form agnosia.' Visual form agnosics lack the ability to perceive or recognize the shapes of objects, but are relatively unimpaired in discriminating non-geometrical features such as color and motion (Benson & Greenberg 1969). Because of this selective deficit, visual form agnosics provide an interesting test of the structure-based account. If the account is right, then visual form agnosics should exhibit deficits in selecting objects through visual attention.

de-Wit, Kentridge, and Milner (2009) have recently tested the patient D.F., a well-known visual form agnosic, on two of the standard paradigms for assessing object-based attention (viz., feature comparison and spatial cueing paradigms). For instance, she was asked to make discrimination (same/different) judgments about two probes that appeared either on the same object or on different objects, with distance between the probes held constant. Crucially, D.F. did not display the usual pattern of results in either of these tasks. For example, unlike healthy participants, she was *not* faster at making within-object comparisons than between-object comparisons.[38] Describing these results, de-Wit et al. write: "We have found, using two different object-based attention paradigms…, that there was no evidence that the deployment of [D.F.'s] attention was sensitive to the presentation of objects. In fact her performance was so insensitive to the presence of objects in the display that it matched that produced by healthy participants performing the task when no objects were present at all" (1488).

The foregoing findings comport with the proposal that a type of shape representation is indeed a prerequisite for attentive visual reference. On my analysis, D.F.'s deficit is twofold. She is impaired at visually processing and representing the shapes of objects, and as a result, her processes of visual attention cannot secure reference to individual objects. Her AVORs cannot secure reference, and this is because no visual descriptions of shape are available to guide them.

[38] However, while D.F.'s attention allocation patterns did not exhibit sensitivity to *objects*, they were sensitive to spatial *location*, as indicated by her normal performance on the spatial cueing paradigm. Thus, if she was cued to attend to a particular location, then she was faster to detect a target when it appeared at the cued location than when it appeared at a different location (de-Wit et al. 2009: Experiment 3). Thus, one reasonable hypothesis is that visual form agnosics retain the ability to attentionally select *locations* but lose the ability to attentionally select *objects*. (Mole 2008 has endorsed a similar view in the case of blindsight.)

At this point it is worth highlighting the contrast between the case of D.F. and the cases of A.H. and P.R. discussed earlier. While D.F. has arguably lost the ability to attentively select particular objects through vision, there is no evidence of such a deficit in either A.H. or P.R. Rather, the latter subjects *can* attentively select objects through vision. They are simply inaccurate in representing the locations of those objects. Severe localization impairments, then, can coexist with relatively normal processes of attentional object selection. However, the case of D.F. suggests that severe shape processing impairments disrupt attentional object selection.

5. Challenges to the Structure-based View

This section addresses two important challenges facing the structure-based view. The first alleges that visual reference, as construed by the structure-based account, cannot be determinate, and thus cannot be singular. The second alleges that the view cannot accommodate the possibility of shape illusions.

5.1. Is Visual Reference Singular Reference?

While the structure-based view arguably has the resources to resolve referential indeterminacy between an object and its proper parts, there are other types of indeterminacy that it is not obviously equipped to resolve. Suppose that a statue is suspended in midair by steel cables and you attentively track it as it is moved about in your field of vision. Assuming your visual system deploys an AVOR (as seems plausible), then it would seem to have at least two candidate referents: It might refer to the statue, or it might refer to the lump of clay that constitutes the statue. Given that these are different objects, singular reference can be secured to at most one of them. But how can we determine which, if either, your AVOR refers to?[39]

Note that geometrical representation on its own cannot resolve this indeterminacy, since the statue and the lump have the *same* geometrical properties (at those times when both exist).[40] Nor is it plausible to appeal to the subject's application of sortal concepts (e.g., STATUE), since such concepts are unlikely to be available to the subpersonal processes that deploy AVORs.

There are two responses to this difficulty. First, one might simply concede that visual reference is indeterminate (and thus nonsingular), but maintain that this does not destroy the analogy between AVORs and indexicals, since certain uses of indexicals are plausibly indeterminate in much the same way. Imagine, for instance, a young child who deploys the bare demonstrative concept 'that' while looking at a statue, but lacks the sortal concepts necessary for identifying either a statue or a lump of clay. Does her use of the concept refer to the statue or the lump? Perhaps there is simply no fact of the matter.

Nevertheless, it seems unpalatable to countenance pervasive indeterminacy of this sort, so this is not the response I favor. While I won't attempt to settle whether, in the case under consideration, your AVOR refers to the statue, the lump of clay, or instead some distinct but spatially coincident object, I do claim that there is plausibly a fact of the matter, and moreover that the issue is open to empirical investigation. The determinacy of

[39] It should be noted that indeterminacy problems are quite pervasive in both thought and natural language, so they are by no means specific to visual reference. For instance, there would appear to be myriad equally good candidates for being the referent of 'Wyoming,' since the practices of language users fail to single out precisely one body of land in the relevant vicinity (see, e.g., McGee & McLaughlin 2000). Doubtless this type of problem arises in connection with AVORs too, but the same can be said for almost every singular term.

[40] Whether they have the same *mereological* properties at a given time is more controversial. For example, one might hold that the lump of marble that constitutes the nose of the statue of David is a part of the lump of marble that constitutes the statue, but not a part of the statue itself. I won't address such issues here.

visual reference may derive from the principles internalized by visual mechanisms of object individuation. Because of these principles, the visual system is keyed to certain kinds of individuals and not to others (see also Campbell 2006). The technical term 'visual object' has sometimes been used for the things that the visual system individuates, though here I've just called them 'objects.'

In more detail, my proposal is as follows. When the visual system selects an object through visual attention, at least two kinds of mechanisms of object individuation are operative: mechanisms of *discrimination* and mechanisms of *tracking*. The former mechanisms incorporate criteria for circumscribing an object at a time—i.e., specifying the boundaries that separate it from its background and from other contemporaneous things in the scene. Such criteria include, I suggest, standard rules of perceptual organization such as Gestalt grouping rules (see Chapter 5, and also Brovold & Grush 2012; Wagemans et al. 2012), but also the rules of shape parsing discussed earlier. In contrast, mechanisms of tracking incorporate *persistence* criteria—criteria specifying when objects encountered at different times count as temporal stages of the same persisting individual. In other words, the persistence criteria specify how to establish *correspondences* between perceived objects at distinct times.

As regards persistence criteria, several have recently proposed that the visual system incorporates a principle of *spatiotemporal priority* (see Mitroff & Alvarez 2007; Scholl 2007), according to which two object-stages count as stages of the same persisting individual so long as they are linked by a spatiotemporally continuous path—even if they differ in surface features such as color, size, or shape.[41] However, surface features likely

[41] It is compatible with the structure-based view that the visual system opts to preserve spatiotemporal continuity over, say, geometrical continuity in computing object persistence. The structure-based view is a

play an important role in certain conditions, especially when the available spatiotemporal information is ambiguous (Feldman & Tremoulet 2006; Hollingworth & Franconeri 2009). Moreover, the persistence criteria recruited by the visual system will plausibly vary according to context and task requirements (cf. note 12).

The point I wish to emphasize is that such persistence criteria—whatever they are—may play a critical role in resolving referential indeterminacy between spatially coincident objects. This is because distinct spatially coincident objects generally differ in their persistence conditions. Thus, in the familiar example, the lump of clay can persist despite being squashed, while the statue cannot.

Very roughly, then, the sequence may work as follows. Whenever an AVOR is deployed, the visual system simultaneously recruits tracking mechanisms that incorporate a set of persistence criteria. These criteria specify the types of changes that the selected object can undergo while remaining the same individual. Furthermore, these persistence criteria aid in singling out a particular referent for the AVOR out of the available candidates (more specifically, out of the candidates that are not already ruled out via the causal and descriptive conditions introduced in section 4.1). When multiple spatially coincident objects are candidate referents for a given AVOR, the AVOR comes to refer to the object whose persistence conditions match—or perhaps 'best match'—the conditions laid down by the recruited persistence criteria. So if one of the candidate referents *can* persist through all of the changes that are permitted under the recruited criteria, while the other cannot, then the first matches those criteria better than the second, and so is a better candidate referent for the AVOR.

theory of how the requirements for securing reference to an object in the first place, not a theory of how vision establishes correspondences over time.

While this account is still rather preliminary, it appears plausible that visual reference may be determinate despite the fact that visual reference-fixing cannot rely on sortal concepts. Further investigation of the persistence criteria recruited by the visual system in various task environments will likely shed further light on this issue.

5.2. Shape Illusions

I now turn to the second challenge facing the structure-based view. To discuss the challenge, let me first introduce two items of terminology. First, I'll characterize a *shape illusion* as a case in which a visual state refers to an object, yet at least partially misrepresents its shape. Second, a requirement that calls for certain aspects of visual representational content to be veridical (or at least not to be nonveridical) in order to secure visual reference to a given object will be called a *veridicality requirement* for securing visual reference to that object.

If the structure-based view is correct, then there are veridicality requirements for visual reference associated with shape representation. This is because an AVOR *R* takes an object *o* as its referent only if *o* at least roughly satisfies the geometrical and mereological descriptive content with which *R* is associated. As such, shape illusions pose a *prima facie* challenge for the structure-based view. For how could the structure-based view be right if it is possible for a state to visually refer to an object while misrepresenting its shape?

In response, I agree that certain shape illusions are compatible with successful visual reference. However, as I suggested above, I think that there are limits to how extreme such illusions can be. In what follows, I'll support this claim in two steps. I'll propose, first, that securing visual reference to an object requires successfully *circumscribing* the object. Second, I'll argue that while certain errors about an object's shape are compatible with successfully circumscribing the object, *dramatic* errors about the shape of an object plausibly preclude circumscription. Accordingly, such errors also preclude visually referring to the object.

Recall that the circumscription problem is the problem of resolving the spatial boundaries that separate an object from its background and from other nearby objects. It is highly plausible that circumscription is a necessary condition for securing visual reference. To visually refer to an object, your visual system must have discriminated that object both from spatially distinct objects in the environment and from objects that stand in compositional relations to it.[42] This enables the visual system to, among other things, package information received from that object separately from information received from other objects. I propose, then, that an error of visual representation precludes securing visual reference to an object provided that the error is so extreme that it is incompatible with having successfully discriminated the object from its surroundings. I'll argue in what follows that extreme errors of shape representation are of this type.

Certain errors about geometrical structure seem to rule out even detecting an object's spatial boundaries, and so to preclude visually discriminating the object. Arguably the most extreme kind of geometrical illusion is one that involves misrepresenting *topological* relations (more specifically, relations like connectedness and disconnectedness). Suppose, for instance, that you monocularly view a scene containing

[42] A potential exception: It may be possible to visually refer to a single object that takes up one's entire field of vision, such as a uniform wall viewed very close up (see Dretske, 1969, p. 26). In this case, the object is not circumscribed because its boundaries are not visible. However, this case is unusual precisely because there are no other objects in one's field of vision from which the object *needs* to be discriminated. Thus, even if this type of case shows that discrimination is not *always* required for visual reference, a weaker claim still appears true: viz., to visually refer to an object that does *not* occupy one's entire field of vision, it is necessary to circumscribe that object.

the following: a half-disc, and a somewhat larger whole-disc that is directly behind the half-disc. Suppose further that the objects are lined up perfectly, so that it appears to you that there is a single bounded disc in the environment. This issues in an inaccurate representation of the scene's topological structure (the scene appears to contain one bounded figure, when in fact it contains two).

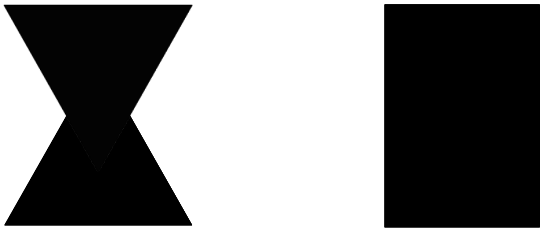
It is implausible in this case that you succeed in circumscribing either of the two objects, because you cannot visually discriminate them from one another, and information acquired from the two cannot be separately packaged. As such, if your visual system deploys an AVOR, then it simply fails to refer. Thus, if an error involves failing to detect the boundaries between two objects, then the error plausibly precludes securing visual reference to either one of them.

It is also reasonable to hold that a perceiver cannot successfully discriminate an object if she is too *wrong* about the geometry of its boundaries. For example, it is difficult to see how a perceiver could successfully circumscribe a normal coffee mug while representing it as having the shape and size of a minivan. And it is similarly implausible that a perceiver could circumscribe a six-foot-tall triangular object despite visually representing it as a circle with one-foot diameter. Such errors preclude even approximately demarcating the borders separating the object from its surroundings, so they also plausibly preclude securing visual reference to the objects in question.

These examples contrast with much less extreme (and more familiar) cases, in which the actual shape of an object is related to its visually represented shape by a fairly minor geometrical transformation (e.g., slight scaling or stretching). For instance, suppose you view a tomato through a drinking glass, and the glass slightly compresses the tomato's appearance. This kind of error does not seem to rule out discriminating the tomato, because you can still at least approximately discern the boundaries separating the tomato from its background and from other objects in the scene. But there are still other cases where our judgments are less clear. For example, might it be possible to successfully circumscribe a one-foot-tall triangular object while representing it as a one-foot-tall square? There are likely to be borderline instances where it is simply unclear whether the shape error is extreme enough to preclude successful discrimination. In such cases, it may even be indeterminate whether visual reference has been secured.[43]

What about the veridicality requirements for visually referring to *parts* of objects? I've argued so far that a plausible condition on successfully discriminating (and so visually referring to) an object is that one not be *too* wrong about the geometry of its boundaries. I believe that this condition also holds for parts. Recall that the boundaries between parts of an object are normally characterized by surface regions of high concavity. As such, if a perceiver's visual system fails to register the concavities that mark a part's boundaries, then the perceiver also fails to discriminate that part. For example, suppose that you view the object shown in figure 4.8a, but, due to distance or poor visual acuity, you visually represent it as having the shape of the object in 4.8b. Because the concavities of the object are not detected, your visual system cannot even approximately demarcate the boundaries separating the top and bottom parts. Accordingly, this prevents you from visually discriminating the parts, and so precludes securing visual reference to either of them.

[43] Notice that this indeterminacy concerns whether or not a given AVOR has secured reference to an object *at all*. It does not concern *which* object is the referent of an AVOR. As such, the foregoing type of indeterminacy is different from the type of indeterminacy discussed in 4.1.



Figures 4.8a (left) and 4.8b (right). A shape illusion. You view the object in 4.8a, but visually represent it as having the shape of the object in 4.8b.

Thus, there plausibly *are* veridicality requirements for discriminating (and so visually referring to) an object, and some of these requirements plausibly involve how one perceptually represents the geometrical properties of the object. One cannot succeed in visually referring to an object when one is too wrong about its geometrical properties. This is reflected in (3) above: The object must at least roughly satisfy the shape representation that the perceiver's visual system constructs.

But what is it for an object to 'roughly' satisfy a shape representation? To make this idea more precise, I propose that hierarchical descriptions are enriched not only with scales of representation, but also with *layers* of representation at each scale. Such layers have varying degrees of specificity with respect to the precise metric structure of the shape. Thus, recall from Chapter 2 that while a given tomato may be represented as having very specific metric features (e.g., precise curvature, exact size, etc.), it may also be represented simply as an ellipsoid, or even simply as a closed figure with some very approximate size. Similarly, while a given coffee-mug has a very precise metric structure, it is also describable simply as cylindrical.[44] Thus, for an object to 'at least roughly' satisfy a given shape representation is, to a first approximation, for it to satisfy at least

[44] See Marr and Nishihara (1978) and Biederman (1987) for shape representation schemes that incorporate roughly this level of abstraction. There is strong evidence that such abstract geometrical properties are indeed extracted by the visual system (see Chapter 2; Bennett 2012).

some of the predicates contained within that representation. So when you see a tomato through a drinking glass that compresses its appearance, your visual system constructs a shape description that the tomato roughly satisfies (it represents that there is something ellipsoidal), and so succeeds in referring to it. Similar remarks hold, I believe, for other familiar shape illusions.[45]

Moreover, this strategy may also enable us to handle a related prima facie difficulty for the structure-based view. Suppose that you view an object from a distance, and due to poor visual resolution you are unable to make out the precise sizes, angles, and degrees of curvature involved in the shape. Still, despite this imprecision, it seems plausible that your visual system is able to refer to the object. This, however, is compatible with the current approach. For although your visual state is quite noncommittal with respect to the object's metric properties, arguably it does accurately specify the object's shape at a higher level of abstraction. For instance, when you view a human body from afar, the body still appears to consist of an ellipsoid (the head) and a set of broadly cylindrical components (torso, arms, and legs). Because these more qualitative contents are satisfied by the object, the object at least roughly satisfies the hierarchical description your visual system constructs, and thus it can be the referent of an AVOR.

[45] Take, for instance, the case of perceiving your body through a funhouse mirror. While the precise form of your body is deformed, its constituent shapes and their relations are preserved at a more qualitative level. (One's head remains seen as ellipsoidal, one's torso remains seen as broadly cylindrical, etc.) How exactly to characterize this qualitative level of shape representation lies outside my scope here, but one attractive possibility is to take qualitative shape properties to be (roughly) affine invariants (e.g., Todd et al. 1998). Such properties include, e.g., ellipticality, triangularity, and being a parallelogram.

6. Conclusion

This chapter has offered an account of the type of referential link secured to objects during visual attention. The account includes two components: a theory of the contents of visual object representations (AVORs), and a theory of how visual reference is secured to an object. The first maintains that visual reference is direct reference: The content of an AVOR (relative to a context) is simply the object it picks out. The second—which I have labeled the structure-based account—holds that visual reference to objects is determined in a manner that is both causal and descriptive. Securing visual reference to an object requires both that the object cause the deployment of an indexical-like object representation (an AVOR) and that the AVOR be associated with descriptive content that the object satisfies. Since the descriptive content in question specifies the geometrical structure of the object to be referred to, it aids in discriminating the object from other nearby things in the environment. The requirement that such content be satisfied thus yields a highly plausible constraint on successful visual reference.

Chapter 5

Objects, Object Files, and Object Principles

1. Introduction

As discussed in the previous chapter, a host of work within cognitive science has established that processes of visual attention and tracking recruit a system for representing individual objects. This system is often called the “object file” system. Moreover, many have held that by virtue of internalizing certain principles about objects, the object file system is *tuned* or *keyed* to a particular kind of thing out in the world. In this chapter, I’ll consider the kinds of things to which the object file system is tuned.

I’ll focus mainly on a view recently defended by Tyler Burge, Susan Carey, and others on which the object file system is tuned, roughly, to *Spelke-objects*. These are a class of entities that obey certain geometrical, topological, and kinematic constraints. In section 2, I’ll introduce this view and then contrast it with a more permissive view on which the object file system individuates and selects objects in accordance with familiar principles of perceptual organization. In sections 3 and 4, I’ll argue that the available evidence is consistent with—and in fact favors—the more permissive view.

2. Spelke-objects vs. Perceptual Units

2.1. Spelke-objects

There is extensive evidence that object representations are recruited by a number of mid-level visual processes, generally involving the allocation of visual attention. For example, attention often seems to “spread throughout” an object, enabling speeded comparison of features belonging to the same individual (Chen 2012; Scholl 2001). Moreover, the reidentification of a target (e.g., a letter) is faster when the target reappears on the same object on which it initially appeared, even if the object changes its location in the interim (Kahneman et al. 1992). Finally, perceivers have an impressive capacity to keep track of a small set of objects over time, even as they move randomly amidst a field of distractors (Pylyshyn & Storm 1988; Pylyshyn 2007).

To account for such experimental phenomena, researchers have proposed the existence of an "object file" system, which is dedicated to selecting and tracking individual objects. The object file system consists of 3-5 representations called object files. Object files secure reference to individuals in the scene and "stick" to their referents over time.[1] Moreover, they appear to incorporate temporary memory stores (or files) in which the current and recent features of the represented object can be recorded. The object file system has also been proposed to be inborn, operative from infancy through adulthood (Scholl & Leslie 1999; Carey & Xu 2001).

Recently, several theorists have proposed that the object file system (hereafter the OF-system) selects objects in accordance with a set of *object principles*, which specify conditions under which something counts as an object.[2] If this is right, then plausibly these principles characterize the kind of thing to which the OF-system is tuned.

In a series of well-known papers, Elizabeth Spelke proposed that certain principles are used when singling out and tracking objects over time (e.g., Spelke 1990,

[1] See Chapter 4 for an account of how the object representations deployed during visual attention and tracking secure reference to individuals in the scene.

[2] What it means for object principles to be "internalized" is a matter of dispute. Some have seemed to think that object principles are part of the infant's explicitly represented knowledge (e.g., Spelke 1990). Others suggest that the principles are instead akin to Marrian natural constraints—they are built into the architecture of the visual system, but they are not explicitly represented (e.g., Bernal 2005).

1994). Following Spelke, subsequent authors—notably Tyler Burge and Susan Carey—have proposed that the processes of visual selection and tracking are governed by these principles (Burge 2010; Carey 2009; Carey & Xu 2001; Rosenberg & Carey 2009). Burge writes:

> [T]o represent something as a body, the individual's perceptual system must segment a three-dimensional whole from a surround by either synchronic or diachronic means. Its doing so is governed by principles for identifying cohesiveness and boundedness of three-dimensional volume shapes. And it must be able to track the wholes over time, either in motion or at rest. Tracking depends on attribution of maintenance of cohesiveness and boundedness of volume shapes. (Burge 2010: 464)

In this passage Burge identifies a set of principles purportedly used when picking out and tracking particular objects (which he calls "bodies"), including the principles of *three-dimensionality* (hereafter "3-D"), *cohesion*, and *boundedness*. It is clear from the surrounding context that Burge also believes that these principles are internalized by the OF-system (e.g., Burge 2010: 453-454). Carey adopts a similar view, although she includes some further criteria as well: "[O]bject files symbolize physical objects, by which I mean bounded, coherent, 3-D, separable, spatio-temporally continuous wholes" (Carey 2009: 97).

In what follows, I'll evaluate the view that the OF-system internalizes the 3-D, cohesion, and boundedness principles. By this, I have in mind the idea that the OF-system is tuned to such entities *alone*, rather than (say) to some wider class of entities that includes them. Before evaluating this view, however, we need to unpack what the principles mean.

The 3-D principle, as Burge introduces it, requires that objects are *volumetric*. They have volume, and so are distinct from either 2-D regions of the retina or 2-D surface patches.

As Burge and Carey acknowledge, the cohesion and boundedness principles are due to Spelke (1990). Spelke's rules supply topological conditions on objecthood. The cohesion principle is stated as follows:

> *Cohesion*: "Two surface points lie on the same object only if the points are linked by a path of connected surface points" (Spelke 1990: 49).

The cohesion principle, then, entails that objects are material, topologically connected figures. For any two surface points *x* and *y*, if there is a single object *O* such that *x* and *y* both belong to *O*, then *y* can be reached from *x* by following a continuous path *P* where each point along *P* is a surface point.[3] For example, my cell phone satisfies the cohesion constraint because any two surface points on it can be reached by following a connected path of surface points.

The boundedness principle is stated as follows:

> *Boundedness*: "Two surface points lie on distinct objects only if no path of connected surface points links them" (Spelke 1990: 49).

The boundedness principle entails that for any two points *x* and *y*, if there are distinct objects *O* and *O** such that *x* belongs to *O* and *y* belongs to *O**, then *y cannot* be reached from *x* by following a continuous path *P* such that each point along *P* is a surface point. Thus, my cell phone and my toaster count as separate objects by the boundedness constraint, because one cannot reach a point on the toaster from a point on the phone by

[3] Burge (2010: 446) prefers to formulate the cohesion principles in terms of small surface patches or edges rather than in terms of individual surface points. This difference will not matter for present purposes, so I will work with Spelke's original formulation. However, if the reader prefers a formulation in terms of local surface patches rather than surface points, she should feel free to make the necessary substitutions throughout.

following a connected path of surface points. Furthermore, the boundedness constraint entails that, e.g., the left and right halves of my phone *cannot* count as separate objects, because one can reach a point on the left half from a point on the right half by following a connected path of surface points.

2.2. Perceptual units

Before evaluating whether the evidence really supports the 3-D, cohesion, and boundedness principles, I want to contrast Burge and Carey's approach with a different view found in a number of vision scientists. This view characterizes "visual objects" by appeal to criteria of perceptual organization.

Kimchi (2009) characterizes objects as "elements in the visual scene organized by Gestalt factors into a coherent unit" (25). Likewise, Chen (2012) characterizes them as "the elements in the visual scene organized by one or more Gestalt grouping principles and/or uniform connectedness" (785).[4] The idea these authors share is that the principles guiding visual object individuation are simply the rules of perceptual organization.

We can divide principles of perceptual organization into *grouping* principles and *parsing* principles. The former specify rules by which the visual system composes smaller units into larger units, while the latter specify rules by which the visual system decomposes larger units into smaller units.

The traditional grouping principles include proximity, similarity, good continuation, and common fate (see Wagemans et al. (2012) for a review). For example, the principle of proximity states that items that are close together tend to be grouped, while the common fate principle states that items that move along similar motion paths

[4] Similar views can be found in Brovold and Grush (2012), Driver et al. (2001), Feldman (2007), Kahneman et al. (1992), Xu (2002), and Yantis (1992).

tend to be grouped. Newer grouping principles include *element connectedness* and *uniform connectedness* (Palmer & Rock 1994; Palmer 1999: ch. 6). The former states that topologically connected elements tend to be grouped together. The latter states that regions of the visual field that have some uniform property (e.g., color or texture) tend to be treated as units. Some grouping phenomena are shown in figure 5.1.

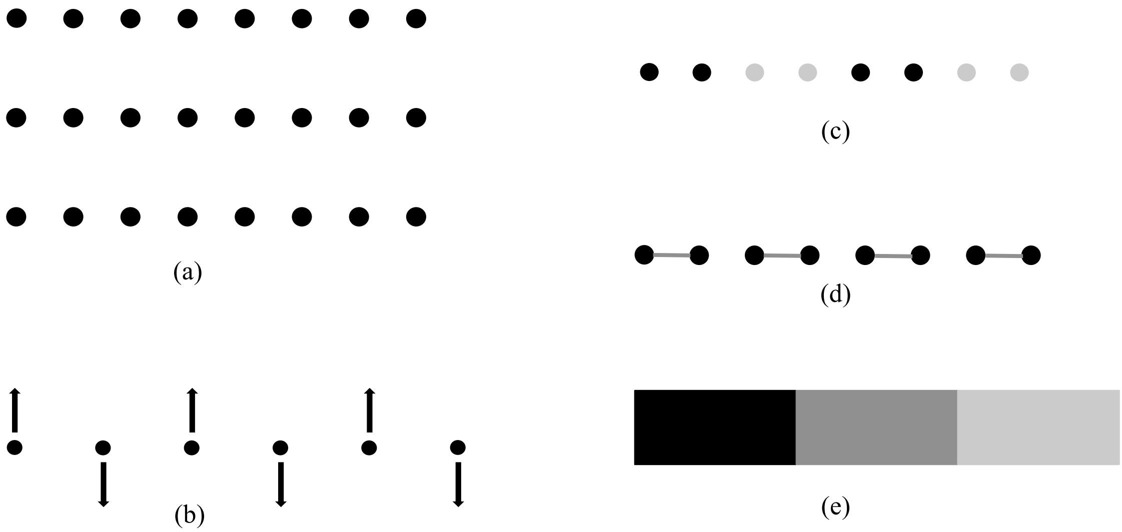
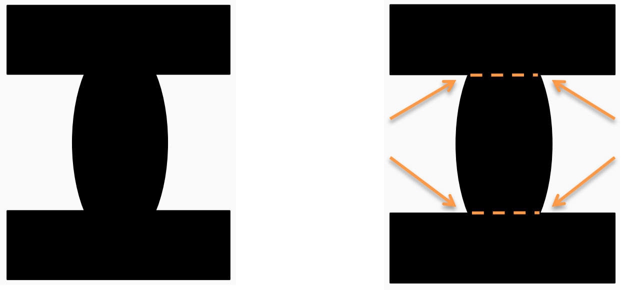


Figure 5.1. Perceptual grouping phenomena. (a) Proximity, (b) Common fate, (c) Similarity, (d) Element connectedness, and (e) Uniform connectedness.

The study of parsing principles is more recent, but there is now good evidence that the visual system engages in parsing. Objects often seem to have a privileged decomposition into parts, and part decomposition seems to exert an effect on other perceptual processes (Singh & Hoffman 2001). Two important parsing principles are the *minima rule* (Hoffman & Richards 1984) and the *short-cut rule* (Singh et al. 1999). The minima rule states that the boundaries between separate parts of an object tend to be found at negative minima of curvature—i.e., places at which the bounding contour of a shape is most concave. The short-cut rule states that part divisions tend to be made by linking two minima of curvature along the shortest paths possible. Thus, observe that the object in figure 5.2a seems to break down naturally into parts, as shown in figure 5.2b. This decomposition follows the minima and short-cut rules. Let's call those parts of objects returned by parsing principles *parsable parts*.



Figures 5.2a (left) and 5.2b (right). An example of part decomposition

Grouping and parsing principles both intuitively involve the visual representation of *units* or *individuals*. Phenomenally, one experiences the elements of a row as "going together" as part of a single unit. Such experiences arise, of course, in natural scenes as well. Flocks of geese, swarms of bees, and trails of ants are phenomenally experienced as units that can be tracked over time. Likewise, one often experiences the separate parts of an object as separate units.[5] For instance, we readily differentiate a person's arm as a different perceptual unit from her torso, and track it as it swings about her shoulder. But do such units count as objects for the OF-system?

If they do, this raises difficulties for Burge and Carey. For instance, if perceptual groups count as objects for the OF-system, then the OF-system does not impose cohesion as a necessary condition on objecthood. Consider a pair of points on separate elements

[5] Note that this is consistent with holding that the object of which they are parts also is experienced as a unit. Indeed, visible scenes often phenomenally appear to be *hierarchically organized*, containing units at various spatial scales (cf. Marr & Nishihara 1978; Saiki & Hummel 1998).

(e.g., dots) in a perceptual group. Because the two elements are disconnected, one cannot reach one point from the other by following a continuous path of surface points.

Likewise, if parsable parts count as objects for the OF-system, then the OF-system does not impose boundedness as a constitutive requirement on objecthood. Consider, for instance, surface points belonging to distinct parts in figure 5.2b. The boundedness principle states that if these parts count as distinct objects, then one surface point *cannot* be reached from the other by following a continuous path of surface points. But clearly the two points can be linked in this way, since the parts to which they belong are connected. Indeed, this point is not lost on proponents of the Spelke criteria. Fei Xu writes: "[P]art of an object is not an object so long as that part does not fall off and start to move independently on its own" (Xu 1997: 387).[6]

Note also that perceptual organization principles apply to both volumetric and non-volumetric things alike. There is no reason why (e.g.) a group of planar dots cannot be organized by rules like proximity, similarity, or common fate. As such, things that do not satisfy the 3-D principle can nevertheless satisfy perceptual organization principles.

Thus, the principles of perceptual organization are distinct from Spelke's principles. The latter are in general far more restrictive. While most things satisfying 3-D, cohesion, and boundedness can be grouped according to perceptual organization principles (e.g., according to the principles of uniform or element connectedness), many units individuable by perceptual organization criteria do not satisfy 3-D and/or cohesion. Thus, in what follows I'll call the view on which the principles governing visual object selection and tracking include perceptual organization criteria the *permissive view*:

[6] Contrast this view with the one found in Kahneman et al. (1992): "Visual objects are hierarchically organized; a group of dancers can be a visual object, as can an individual dancer, or her right hand. At any instant one of these levels may be dominant in the parsing of the scene" (178).

> *Permissive view*: The principles governing visual object selection and tracking include the traditional criteria of perceptual organization, such as the principles of perceptual grouping and perceptual parsing.

Correspondingly, I'll call the view on which the OF-system selects and tracks in accordance with the 3-D, cohesion, and boundedness principles the *restrictive view*.

3. Reevaluating the 3-D, Cohesion, and Boundedness Principles

In this section, I will argue that the available evidence—including the data standardly evinced in support of the restrictive view—is consistent with, and may in fact support, the permissive view.

3.1. Evidence for 3-D?

While Burge and Carey both propose 3-D as an object principle, many of the experimental tasks used to study the OF-system instead employ 2-D figures on a computer screen. This is true, for instance, of most experiments on both multiple-object tracking[7] and object-based attention[8] (e.g., Pylyshyn & Storm 1988; Behrmann et al. 1998; Scholl & Pylyshyn 1999). Such stimuli do not have volume. Thus, if 3-D is an object principle internalized by the OF-system, then they do not satisfy the OF-system's object principles. *Prima facie*, this presents a problem, since the OF-system treats these things in much the same way as it treats volumetric figures.

Carey recognizes this issue, and responds as follows:

> Does the fact that 2-D bounded entities activate object-files mean that their content is more perceptual—perhaps *closed shape*? Should object-

[7] In a multiple object-tracking task, subjects are asked to keep track of several target objects as the objects move randomly about the screen in the presence of a set of (usually identical) distractors.

[8] There are a variety of paradigms used to study object-based attention. However, perhaps the most common involves showing that comparisons involving two features are faster when the features appear on the same object than when they appear on different objects, even if distance between the features is held constant.

> files be called "closed shape-files" or "perceptual individual-files"? No, they should not. For computer displays to work, we must present many of the cues for depth in 2-D arrays, and surfaces arranged in 3-D are routinely perceived in such displays. That the system can be fooled into accepting 2-D entities as objects does not mean that it is not representing the stimuli as real objects, just as the fact that the system can be fooled into seeing depth in 2-D displays…does not mean it is not representing the stimuli as arrayed in 3-D space. (2009: 98)

Carey's response rests on the plausible idea that a system can be *tuned* to a certain kind *K* even though it sometimes picks out not-*K*s. However, she recognizes that to defend the proposal that such a system is tuned to *K*s, one needs some explanation of why the system occasionally picks out not-*K*s. Her explanation is that in such cases, the not-*K*s are misrepresented as *K*s. Thus, in the current case, the explanation for why the OF-system occasionally picks out 2-D figures is that it misrepresents them as 3-D. And the evidence for this, she suggests, is that in order for 2-D figures to properly affect the OF-system, they must supply appropriate cues to depth.

To assess Carey's response, we must first observe there is an unfortunate ambiguity in the requirement that objects are "3-D." On a stronger reading—the one I have assumed so far—this means that objects must have volume. But on a weaker reading, it just means that objects are arrayed within 3-D space, and that they stand in depth relations both to the observer and to other things. Note that something can have the second characteristic without having the first. Distal surfaces, for instance, are often construed as 2-D entities, but they do stand in depth relations to the observer and to other things in the environment.

As such, even granting Carey's claim that the stimuli in 2-D computer displays are represented as having locations in *depth*, this does not entail that such stimuli are represented as *volumetric*. Thus, her response at most rescues the weaker version of the

3-D requirement, not the stronger one. As we saw above, Burge understands the 3-D requirement in the stronger way, so Carey's response is irrelevant to his version of the requirement. Carey is less explicit on the matter, but she may have in mind the weaker version. However, the weaker reading seems quite forced—it counts planar polygons like squares and circles as "3-D," as long as they are embedded in 3-D space.

Turning to the stronger version of the 3-D principle, despite Burge's claims, there is very little evidence that the OF-system is selectively tuned to entities with volumetric shapes. For example, while stimuli in standard MOT or object-based attention studies do indeed supply standard cues to depth, such as occlusion and figure-ground cues (e.g., Scholl & Pylyshyn 1999), they rarely supply standard cues to volume (such as differential shading, texture density, or surface orientation edges). Moreover, such stimuli *look* non-volumetric. They look like flat, 2-D figures. Thus, while the weaker version of the 3-D requirement may hold of the OF-system, evidence for the stronger version of this requirement is lacking.

3.2. Evidence for cohesion?

Studies of infant cognition and adult mid-level visual processing have been cited in support of the view that the OF-system incorporates the cohesion principle. In this subsection I'll argue that the available evidence in fact better supports the permissive view.

Huntley-Fenner et al. (2002) contrasted 8-month-old infants' ability to keep track of cohesive objects with their ability to keep track piles of sand that lost their cohesion during motion. Infants in the 'object' condition saw a rigid entity that had precisely the same shape, color, and texture as a pile of sand. The object was lowered onto a stage in view of the infants in a manner that preserved its cohesion. After this, two screens were placed on the stage, one of which occluded the object. Next, infants saw another sand-pile-shaped cohesive object lowered behind the second screen. Finally, both screens were removed, revealing either one object (unexpected) or two objects (expected), and infants' looking times were monitored. The 'sand' condition was the same as the object condition, except that the experimenter poured sand out of a cup into piles on the stage, rather than lowering cohesive, pile-shaped objects. (The pouring action disrupted the internal connectedness, and so the cohesion, of the pile of sand.)

In the 'object' condition, looking times were longer in response to the unexpected outcome of only one object after the screens were removed, while in the 'sand' condition, there was no significant difference in looking times for the two outcomes. This is consistent with the proposal that object files were maintained for the cohesive pile-shaped objects (enabling infants to keep track of how many such objects there were), but not for the non-cohesive piles of sand.

Mirroring the Huntley-Fenner et al. (2002) study, vanMarle and Scholl (2003) compared tracking performance in a standard multiple object tracking task with performance under conditions where each object (both the targets and the distractors) disintegrated into a number of small pieces and seemed to "pour" from one location to the next (a violation of cohesion). They found that tracking performance was significantly worse in the latter condition (89% accuracy versus 67% accuracy).

More recently, Cheries et al. (2008) have shown that simply splitting an object into two pieces (another loss of cohesion) disrupts infants' object representations. Infants were divided into a 'no-split' condition and a 'split' condition. Those in the no-split condition saw one graham cracker placed in a container, and two already disconnected graham crackers placed in a different container. In the split condition, they also saw one graham cracker placed in one container and two graham crackers placed in the other container. However, the two graham crackers resulted from breaking a single larger graham cracker into two pieces within view of the infants. It was found that a majority of infants crawled toward the 2-cracker container in the no-split condition, but crawling behavior was at chance in the split condition. Cheries et al. suggest that the lack of preferential crawling in the split condition was because infants represented the loss of cohesion involved in splitting as a violation of cohesion, and this led them to discard the object file for the larger cracker. Further, because infants did not have enough time to assign new files to the resulting pair of crackers before they were placed in the container, they could not keep track of how many crackers were in the 2-cracker container.[9]

I will suggest an alternative explanation of these findings that does not advert to cohesion per se, but before doing so, it is important to note that there is independent evidence that loss of cohesion may not be the key factor leading to failures of tracking or individuation in these studies. This is because (i) *failures* in these or similar paradigms are observed with cohesive stimuli, and (ii) *successes* in these or similar paradigms are observed with *non*-cohesive stimuli.

As regards (i), object tracking is impaired when stimuli retain their cohesion, but expand or contract in a manner similar to the way a pile of sand changes shape when it is poured from a cup. In a third condition of their experiment, vanMarle and Scholl (2003) found that when the items to be tracked expanded and contracted so that they appeared to move like “slinkies” (albeit while maintaining their internal connectedness), tracking was

[9] These findings with infants mirror earlier work on adult mid-level vision (Mitroff, Scholl, & Wynn 2004).

just as impaired as when the items seemed to “pour” from one location to the next. This suggests that the tracking impairments in both Huntley-Fenner et al. (2002) and the pouring condition of vanMarle and Scholl (2003) may be due to the expansion and contraction involved in pouring, rather than the loss of cohesion.[10]

There is also evidence that tracking is *unimpaired* in the face of non-cohesion, as long as the elements of a group move together as a cluster. In a fourth condition of vanMarle and Scholl’s (2003) experiment, elements that broke apart into several pieces (destroying cohesion) but still moved together as a tight cluster were tracked just as well as figures in the standard MOT task.

Similarly, Wynn, Bloom, and Chiang (2002) found that infants are capable of enumerating perceptual groups that move together as a cluster. Infants were first habituated to a display containing either two or four groups of three dots that moved across a computer screen as a cluster. During test trials, they saw displays containing either two groups of four dots or four groups of two dots. Critically, infants who were habituated to displays containing two groups looked longer at test displays containing four groups, while infants who were habituated to displays containing four groups looked longer at test displays containing two groups.[11] A natural conclusion is that in this case, infants assigned an object file to each group of dots on the basis of common fate and proximity grouping cues, thus treating each group as a distinct object.[12]

[10] Howe et al. (2013) have recently confirmed tracking impairments in response to expansion and contraction under a variety of different conditions.

[11] Note that these results cannot be explained on the hypothesis that infants merely enumerated the individual dots in the display, because the two test displays had the *same number* of dots, differing only in the number of *groups* of dots.

[12] Wynn et al. themselves reject this interpretation. They argue that it is unlikely that the groups were treated by the OF-system as individual objects, because the dots within a group moved somewhat independently of one another, and that it is known that when elements move independently, the visual system tends to treat them as distinct objects (e.g., Spelke 1990). My response is simple. While the dots in a

In light of this, I want to offer an alternative explanation of the results often taken to support cohesion. The reason for failures of tracking and individuation in these cases may simply be that the visual system is highly sensitive to correlated motion paths (i.e., common fate grouping). Elements are grouped on the basis of cues to correlated motion, but if the OF-system is supplied with cues to *independent* motion among elements, those elements are unlikely to be grouped (or, if they are already grouped, they are likely to become ungrouped).

Thus, consider again the “slinky” condition of vanMarle and Scholl (2003) in which subjects were asked to track entities that expanded and contracted along their direction of motion. The present hypothesis suggests that even though the entities maintained cohesion, the visual system was given strong cues that the elements comprising them (e.g., their front and back edges) moved independently. This led to a failure to group the edges as belonging to a single object, and hence led to a failure to maintain object files for the slinkies. A similar explanation applies to the “pouring” condition of vanMarle and Scholl’s study, and also to the Huntley-Fenner et al. (2002) findings. Moreover, a similar explanation is also available for Cheries et al. (2008). When the initially cohesive object split apart, the two resulting pieces initially followed very different motion trajectories. Because of this, they could no longer be perceptually grouped. Again, on this proposal it is not the violation of cohesion per se that explains tracking failures, but rather cues to independent motion.

group did not follow precisely parallel motion paths, their velocities had a sufficiently close relationship to enable common fate grouping. Similar remarks hold for the elements that moved as clusters in vanMarle and Scholl’s (2003) MOT experiment. Indeed, there is evidence that the neural pooling of motion signals is a fairly flexible process that may accommodate differences in both the speed and direction of motion among a collection of elements (e.g., Webb et al. 2011). It is possible that such flexible motion computations underlie common fate grouping.

Moreover, in the studies demonstrating tracking or individuation *success* despite lack of cohesion (Wynn et al. 2002; vanMarle & Scholl 2003), I conjecture that the key feature is that the elements that formed a non-cohesive group nevertheless followed similar motion paths. This led them to be grouped into a single unit and targeted by an object file.

3.3. Evidence for Boundedness?

Recall that the boundedness requirement precludes undetached parts from counting as objects for the OF-system. But is there evidence that the OF-system imposes such a requirement?

Although boundedness has received less attention relative to cohesion, some studies have been marshaled in its support. Thus, Fodor and Pylyshyn (2015) write:

> It turns out that subjects can track dumbbells but can't track their weights (unless we remove the rod that connects them). Connecting the parts of the dumbbell creates a single new object, not just an arrangement of the parts of one (see Scholl, Pylyshyn, and Feldman 2001). (…) The world is the totality of *things*, not undetached parts of things. (131)

Fodor and Pylyshyn believe that the cited experiment by Scholl, Pylyshyn, and Feldman (2001) indicates that visual tracking mechanisms are selectively keyed to bounded things, rather than to undetached parts. They also believe that this helps in framing a response to Quine's "gavagai" problem. I'm not concerned here with whether visual tracking experiments have any special relevance to traditional problems of referential indeterminacy. However, it's worth asking in any event whether the evidence that Fodor and Pylyshyn cite actually supports the claim that tracking mechanisms impose a boundedness constraint.

Fodor and Pylyshyn claim that subjects *can't track* the weights of dumbbells unless we remove the rod that connects them, and that this indicates that perceptual/attentional representations cannot target parts of objects. But closer inspection reveals that the study cited does not actually substantiate this claim. Scholl, Pylyshyn, and Feldman (2001) tested tracking under a number of conditions, but three are most critical. A baseline condition replicated the standard MOT paradigm described above. In another condition, subjects were asked to track the endpoints of lines as the lines moved about the screen in the presence of distractors. In a third condition, subjects were asked to track the square ends of dumbbells. (A single dumbbell in this experiment was composed of two squares linked by a line segment.) Importantly, dumbbell weights are likely to be segregated by perceptual parsing criteria, while the endpoints of lines are not. So if the OF-system selects things in accordance with parsing criteria, then dumbbell weights should count as candidate objects.

Consistent with Fodor and Pylyshyn's claim that we cannot track parts, subjects were significantly better in the first condition than in either of the other two, and performance in the third condition (tracking line endpoints) was only slightly better than chance (see also Howe et al. 2012). However, performance in the second condition (tracking dumbbell weights) was well above chance. Tracking accuracy with dumbbell weights was roughly 84%, versus 92% in the baseline condition (see Scholl, Pylyshyn, and Feldman 2001: 170, fig. 3). Thus, while there was *some* (statistically significant) decrement associated with tracking weights versus tracking individual boxes, this hardly warrants the claim that we can't track them.

Moreover, there is an unfortunate asymmetry involved in comparing tracking performance with “whole” objects versus parts of objects. For, suppose that both topologically bounded figures (e.g., dumbbells) and certain parts of such figures (e.g., dumbbell weights) count as candidate objects for the OF-system. Moreover, suppose (as seems intuitively plausible) that during tracking, the OF-system sometimes mixes up a target with one of its parts. If this were the case, then we would *expect* performance in MOT tasks to be slightly better for whole objects than for parts of objects, because of the following simple fact: Selectively tracking a part of an object is *sufficient for* reidentifying the *whole* object at the end of a MOT task, while tracking a whole object is *insufficient for* reidentifying one of its *parts* at the end of a MOT task. Thus, if a subject needs to track a whole object, but her OF-system mixes up that object with one of its parts, she will still be able to perform the task just as well. However, if a subject needs to track only a part of an object, but her OF-system mixes up that part with either the whole object or with another part of the same object, her performance will suffer. So we can explain why tracking should be slightly better for whole objects than for their parts by appeal to this performance limitation alone. We needn’t suppose that the OF-system incorporates boundedness as a constitutive requirement on objecthood.

Another study sometimes taken to support the boundedness criterion is due to Mitroff, Scholl, and Wynn (2005). Mitroff et al. studied the effect of object *merging* on the object-specific preview benefit (OSPB). They showed subjects an initial display containing three circular objects. Letters briefly appeared on each circle and then vanished. Next, the objects underwent one of two motion processes. In one condition, two of the circles merged to become a single circle. In another condition, two of the circles approached one another, but did not merge. Finally, after motion ended, a single letter appeared on one of the circles, and subjects had to indicate whether it was the same as a letter previously viewed. OSPBs were associated with faster “same” responses to a letter that reappeared on the same object on which it initially appeared. Mitroff et al. found that merging disrupted OSPBs, though, puzzlingly, only for the lower of the two objects that merged.

Merging involves a loss of boundedness. When two objects merge, it becomes possible to reach a surface point on one object from a surface point on another by following a connected path of surface points. As such, one might take this experiment to indicate that the OF-system imposes a boundedness condition on objecthood—when the objects ceased to display boundedness, the OF-system concluded that one of them (usually the lower one) went out of existence.

However, there are two problems with this inference. First, the evidence shows at most that the OF-system internalizes a *conditional* principle: *If* an object is bounded at one time, then it will *stay* bounded at later times. Second, and more importantly, this study cannot distinguish the view that the OF-system internalizes boundedness as a requirement on objecthood from the view that it accords with less stringent perceptual organization (grouping and parsing) criteria. This is because the type of merging employed in this study (two circles merging into a single circle) would also be expected to disrupt the visual system’s ability to organize the initially separate objects into distinct perceptual units. As such, the results have no implications concerning whether the OF-system can treat nonbounded entities that *are* decomposable according to well-established parsing criteria (e.g., the minima and short-cut rules) as objects.[13]

4. Further Support for the Permissive View: Holes

There is additional evidence to favor the permissive view over the restrictive view. Here I'll focus on just one particularly interesting line of research. *Nonmaterial* things—particularly *holes* in an object—can be visually attended and tracked over time, and, in many cases, they can be attended and tracked just as efficiently as solid, cohesive objects.

Note that the cohesion principle is standardly spelled out in terms of "surface points." This suggests that the things that count as objects for the OF-system must be composed of material surfaces. It is worth asking, however, whether even this is correct.

Researchers on perceptual organization have frequently observed that *holes* can be organized by Gestalt criteria into perceptual units, because they display (*inter alia*) the important grouping cues of closure and surroundedness (e.g., Palmer 1999: 285-287). But holes are not composed of material surfaces. Rather, they are (arguably) composed of empty space wholly surrounded by material surfaces (e.g., Casati & Varzi 1999).

Many have proposed that holes are visually processed differently from other regions of empty space. Indeed, the hole in a doughnut seems to be a kind of "thing" that moves from place to place along with its material host. In accordance with this, while perceivers are generally poor at remembering the shapes of background regions (e.g., the background of an unambiguous figure-ground display), they appear to remember the

[13] Note, moreover, that there is independent evidence that parsable parts elicit the same pattern of results as "whole" objects on a number of paradigms standardly used to study object files. Thus, there is evidence that the OF-system can maintain short-term memory stores for parsable parts, just as it does for whole objects (Xu 2002).

shapes of holes just as well as the shapes of solid objects (Palmer et al. 2008; Nelson et al. 2014; although see Bertamini 2006 for a different perspective).

There also exists more direct evidence that holes count as potential objects for the OF-system. Giralt and Bloom (2000) found that there were no significant differences between three-year-olds' ability to track and enumerate holes and their ability to track and enumerate individual solid objects. (This study is also notable because it utilized real physical objects and holes, rather than computer simulations.) Furthermore, holes are tracked just as efficiently as standard stimuli in the multiple object-tracking paradigm, regardless of whether they are defined through monocular depth cues or through stereo disparity (Horowitz & Kuzmova 2011). This evidence is readily accommodated by the permissive view, since on the latter view the OF-system needn't require that objects be composed of material surfaces.

One option for a defender of the restrictive view would be to suggest that holes are being *misrepresented* as 3-D, bounded, cohesive individuals in these studies. If so, however, then it is incumbent on her to support this claim, since the studies in question incorporated a large amount of depth information specifying the surface visible through the hole as behind its occluder (indeed, Giralt and Bloom (2000) used full-cue physical stimuli).

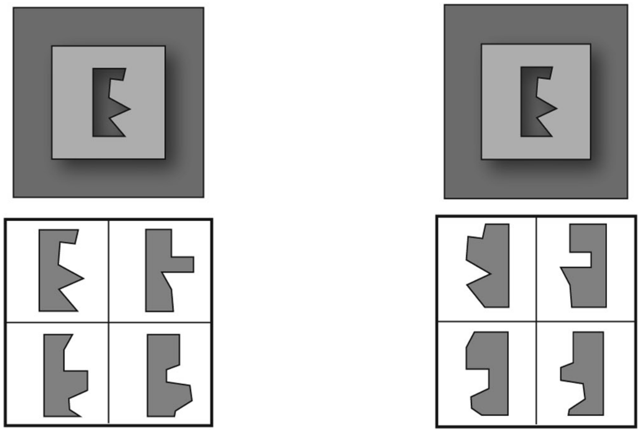
A more plausible response would be to claim that participants in these studies were not tracking nonmaterial entities *per se*, but were instead tracking the material bounds of those entities (e.g., the 2-D surface regions or 1-D contours that bounded the holes).[14] However, note first that this would nevertheless conflict with the assumption of a 3-D requirement on objecthood. Second, there seems to be little theory-independent

[14] Thanks to Chris Hill for this suggestion.

motivation for the idea that while the OF-system can pick out the totality of a cohesive object, it can pick out only the boundary of a hole. Indeed, holes seem phenomenologically to be *things* bounded by material hosts. Finally, this response fits badly with evidence on how the visual system processes the shapes of holes (Palmer et al. 2008; Nelson et al. 2014). Let me explain.

Note than when a contour divides two regions, it partially determines the shapes of both regions. Thus, the boundary of the hole of a doughnut partially determines (along with the doughnut's outer boundary) that the doughnut is a torus, and it also determines that the hole of the doughnut is circular. However, a well-known fact about visual shape processing is that the visual system does not always (or even usually) *encode* the shapes of both of the regions divided by a contour. Thus, in an unambiguous figure-ground display, subjects usually encode only the shape of the figural region, and not the shape of the ground region (e.g., Palmer 1999: 280-281). Note, further, that if the visual system encodes the shape of a region bounded by a contour, this suggests that it *picks out* that region, and attributes a property (viz., shape) to it. Thus, if (as on the current proposal) perceivers do *not* pick out the interiors of holes, and instead select only the interiors and boundaries of their material hosts, then the boundary of a hole should presumably be used to represent *only* the shape of the material host (the hole's *exterior*), and *not* the shape of the hole's interior.

However, studies specifically testing this possibility have found that the visual system likely uses the boundary of a hole to encode the shape of the hole's interior. For example, Nelson et al. (2014) showed participants an object with a hole and asked them to indicate which of a set of test figures had a boundary that partially matched the boundary of the hole. They found that subjects were more accurate when the target figure shared the shape of the hole's interior than when it shared the shape of a portion of the hole's exterior material host (see figures 5.3a and 5.3b). This strongly suggests that subjects indeed attributed shape properties to the *interior* of the hole, rather than merely to its exterior. Moreover, it corroborates (though does not yet conclusively establish) the proposal that when holes are selected and tracked, the visual system genuinely picks out the interior of the hole, rather than merely picking out its material bound.



Figures 5.3a (left) and 5.3b (right). Figures from Nelson et al. (2014). Figure 5.3a shows a case in which the target figure matched the shape of the hole's interior. Figure 5.3b shows a case in which the target figure matched a portion of the shape of the hole's exterior. In both cases, the target figure is located in the top-left quadrant of the bottom grid. Reproduced from Nelson et al. (2014) with kind permission from SAGE Publications.

I conclude that there is compelling evidence that the OF-system can select and track holes of objects. Moreover, it is unlikely that it does this simply by misrepresenting holes as bounded and cohesive, or by selecting only the material bounds of holes.

5. Conclusion

I have contrasted two views about the principles used by the OF- system when individuating and tracking objects. On one view, which I've called the restrictive view, the OF-system is selectively keyed to 3-D, bounded, and cohesive individuals. On a more permissive view, the OF-system selects objects in accordance with familiar criteria of perceptual organization. I have argued that the available evidence—including the evidence often cited in support of the former view—is consistent with the permissive view. Moreover, additional data may provide positive reason to favor the permissive view over its more restrictive competitor.

Bibliography

Albert, M., & Hoffman, D. (1995). Genericity in spatial vision. In D. Luce, K. Romney, D. Hoffman, & M. D'Zmura (eds.), *Geometric Representations of Perceptual Phenomena: Articles in Honor of Tarow Indow on his 70th Birthday*. New York: Erlbaum, pp. 95-112.

Amir, O., Biederman, I., Herald, S.B., Shah, M.P., & Mintz, T.H. (2014). Greater sensitivity to nonaccidental than metric shape properties in preschool children. *Vision Research*, *97*, 83-88.

Austin, J.L. (1962). *Sense and Sensibilia*. Oxford: Oxford University Press.

Ayob, G. (2008). Space and sense: The role of location in understanding demonstrative concepts. *Proceedings of the Aristotelian Society*, *108*, 347-354.

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

Bar, M. (2001). Viewpoint dependency in visual object recognition does not necessarily imply viewer-centered representation. *Journal of Cognitive Neuroscience*, *13*, 793-799.

Barenholtz, E., & Feldman, J. (2003). Visual comparisons within and between object parts: Evidence for a single-part superiority effect. *Vision Research*, *43*, 1655-1666.

Barenholtz, E., & Tarr, M.J. (2008). Visual judgment of similarity across shape transformations: Evidence for a compositional model of articulated objects. *Acta Psychologica*, *128*, 331-338.

Behrmann, M., Peterson, M.A., Moscovitch, M., & Suzuki, S. (2006). Independent representation of parts and the relations between them: Evidence from integrative agnosia. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1169-1184.

Behrmann, M., Zemel, R., & Mozer, M.C. (1998). Object-based attention and occlusion: Evidence from normal participants and a computational model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1011-1036.

Beintema, J.A., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences*, *99*, 5661-5663.

Belhumeur, P.N., Kriegman, D.J., & Yuille, A.L. (1999). The bas-relief ambiguity. *International Journal of Computer Vision*, *35*, 33-44.

Bennett, D.J. (2009). Varieties of visual perspectives. *Philosophical Psychology*, *22*(3), 329-352.

Bennett, D.J. (2012). Seeing shape: Shape appearances and shape constancy. *The British Journal for the Philosophy of Science*, *63*, 487-518.

Bennett, D.J., Zhao, H., Vuong, Q.C., & Liu, Z. (2012). Humans can use information beyond 2 frames in structure from motion. Poster presented at VSS 2012.

Benson, D.F., & Greenberg, J.P. (1969). Visual form agnosia: A specific defect in visual discrimination. *Archives of Neurology*, *20*, 82-89.

Berkeley, G. (1710/1982). *A Treatise Concerning the Principles of Human Knowledge*. Indianapolis, IN: Hackett Publishing Company.

Bernal, S. (2005). Object lessons: Spelke principles and psychological explanation.

Philosophical Psychology, *18*(3), 289-312.

Bertamini, M. (2006). Who owns the contour of a visual hole? *Perception*, *35*, 883-894.

Bhatt, R.S., Hayden, A., Kangas, A., Zieber, N., & Joseph, J.E. (2010). Part perception in infancy: Sensitivity to the short-cut rule. *Attention, Perception, & Psychophysics*, *72*(4), 1070-1078.

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

Biederman, I. (2013). Psychophysical and neural correlates of the phenomenology of shape. In L. Albertazzi (*ed.*), *Handbook of Experimental Phenomenology: Visual Perception of Shape, Space, and Appearance*. Malden, MA: Wiley-Blackwell, pp. 417-436.

Biederman, I., & Bar, M. 1999). One-shot viewpoint invariance in matching novel objects. *Vision Research*, *39*, 2885-2889.

Biederman, I., & Cooper, E.E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, *20*(5), 585-593.

Bingham, G.P. (1993). Scaling judgments of lifted weight: Lifter size and the role of the standard. *Ecological Psychology*, *5*, 31-64.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, *58*, 47-73.

Block, N. (2014). Seeing-as in the light of vision science. *Philosophy and Phenomenological Research*, DOI: 10.1111/phpr.12135.

Blum, H. & Nagel, R. N. (1978). Shape description using weighted symmetric axis features. *Pattern Recognition*, *10*(3), 167-180.

Brannan, D.A., Esplen, M.F., & Gray, J.J. 2012). *Geometry*. Cambridge: Cambridge University Press (2nd edition).

Briscoe, R. (2008). Vision, action, and make-perceive. *Mind & Language*, *23*(4), 457-497.

Briscoe, R. (2009). Egocentric spatial representation in action and perception. *Philosophy and Phenomenological Research*, *79*, 423-460.

Brovold, A., & Grush, R. (2012). Towards an (improved) interdisciplinary investigation of demonstrative reference. In A. Raftopoulos and P. Machamer (eds), *Perception, Realism, and the Problem of Reference*. Cambridge: Cambridge University Press.

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Cacciamani, L., Ayars, A.A., & Peterson, M.A. (2014). Spatially rearranged object parts can facilitate perception of intact whole objects. *Frontiers in Psychology*, *5*, 1-11.

Campbell, J. (1996). Molyneux's question. *Philosophical Issues*, *7*, 301-318.

Campbell, J. (2002). *Reference and Consciousness*. Oxford: Oxford University Press.

Campbell, J. (2006). Does visual attention depend on sortal classification? Reply to Clark. *Philosophical Studies*, *127*, 221-237.

Campbell, J. (2007). What's the role of spatial awareness in visual perception of objects? *Mind & Language*, *22*, 548-562.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Carey, S., & Xu, F. (2001). Infant knowledge of objects: Beyond object files and object tracking. *Cognition*, *80*, 179-213.

Casati, R., & Varzi, A. (1999). *Parts and Places: The Structures of Spatial*

Representation. Cambridge, MA: MIT Press.

Chen, L. (1982). Topological structure in visual perception. *Science*, *218*, 699-700.

Chen, L. (1985). Topological structure in the perception of apparent motion. *Perception*, *14*, 197-208.

Chen, L. (1990). Holes and wholes: A reply to Rubin and Kanwisher. *Perception & Psychophysics*, *47*, 47-53.

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition*, *12*(4), 553-637.

Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, *74*, 784-802.

Cheries, E.W., Mitroff, S.R., Wynn, K., & Scholl, B.J. (2008). Cohesion as a constraint on object persistence in infancy. *Developmental Science*, *11*, 427-432.

Clark, A. (2000). *A Theory of Sentience*. Oxford: Oxford University Press.

Clarke, T.J., Bradshaw, M.F., Field, D.T., Hampson, S.E., & Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, *34*, 1171-1180.

Coates, P. (2000). Deviant causal chains and hallucinations: A problem for the anti-causalist. *The Philosophical Quarterly*, *50*, 320-331.

Cohen, J. (2015). Perceptual constancy. In M. Matthen (ed.), *The Oxford Handbook of Philosophy of Perception*. Oxford: Oxford University Press, pp. 621-639.

Cole, M.S., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., & Singh, M. (2009). How well do line drawings depict shape? *ACM Transactions on Graphics*, *28*, 1-9.

Davidoff, J., & Roberson, D. (2002). Development of animal recognition: A difference between parts and wholes. *Journal of Experimental Child Psychology*, *81*, 217-234.

Davies, M. (1992). Perceptual content and local supervenience. *Proceedings of the Aristotelian Society*, *92*, 21-45.

Dawson, M.R. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review*, *98*, 569-603.

Denys, K., Vanduffel, W., Fize, D., Nelissen, K., Peuskens, H., Van Essen, D., & Orban, G.A. (2004). The processing of visual shape in the cerebral cortex of human and nonhuman primates: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, *24*, 2551-65.

DeWinter, J., & Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale, integrative study. *Cognition*, *99*, 275–325.

De-Wit, L., Kentridge, R.W., & Milner, A.D. (2009). Object-based attention and visual area LO. *Neuropsychologia*, *47*, 1483-1490.

Dickie, I. (2011). Visual attention fixes demonstrative reference by eliminating referential luck. In C. Mole, D. Smithies, & W. Wu (eds.), *Attention: Philosophical and Psychological Essays*. Oxford: Oxford University Press.

Dittrich, W.H. (1993). Action categories and the perception of biological motion. *Perception*, *22*, 15-22.

Dretske, F. (1969). *Seeing and Knowing*. Chicago: The University of Chicago Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention, and phenomenal visual objects. *Cognition*, *80*, 61-95.

Edelman, S. (1997). Computational theories of object recognition. *Trends in Cognitive Sciences*, *1*(8), 296-304.

Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.

Egly, R., Driver, J., & Rafal, R.D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*, 161-177.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Evans, G. (1985). Molyneux's question. In A. Phillips (ed.), *Gareth Evans: Collected Papers*. Oxford: Clarendon Press. Reprinted in A. Noë & E. Thompson (eds.), *Vision and Mind: Selected Readings in the Philosophy of Perception*, pp. 319-349. Cambridge, MA: MIT Press.

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, *7*(6), 252-256.

Feldman, J. (2007). Formation of visual "objects" in the early computation of spatial relations. *Perception & Psychophysics*, *69*, 816-827.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014-18019.

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. (2013). An integrated Bayesian approach to shape representation and perceptual organization. In S.J. Dickinson & Z. Pizlo (eds.), *Shape Perception in Human and Computer Vision*. New York: Springer, pp. 55-70.

Feldman, J., & Tremoulet, P.D. (2006). Individuation of visual objects over time. *Cognition*, *99*, 131-165.

Firestone, C., & Scholl, B.J. (2014). "Please tap the shape, anywhere you like": Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, DOI: 10.1177/0956797613507584.

Flombaum, J.I., Kundey, S.M., Santos, L.R., & Scholl, B.J. (2004). Dynamic object individuation in rhesus macaques: A study of the tunnel effect. *Psychological Science*, *15*, 795-800.

Fodor, J.A. (1983). *Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. (2008). *LOT2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, J.A., & Pylyshyn, Z.W. (2015). *Minds without Meanings: An Essay on the Content of Concepts*. Cambridge, MA: MIT Press.

Frisby, J.P., & Stone, J.V. (2010). *Seeing: The Computational Approach to Biological Vision*. Cambridge, MA: MIT Press.

Gelman, S.A., & Wellman, H.M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*, 213-244.

Giralt, N., & Bloom, P. (2000). How special are objects? Children's reasoning about objects, parts, and holes. *Psychological Science*, *11*(6), 497-501.

Graf, M. (2006). Coordinate transformations in object recognition. *Psychological Bulletin*, *132*(6), 920-945.

Green, M. (1986). What determines correspondence strength in apparent motion? *Vision*

Research, *26*(4), 599-607.

Green, E.J. (2015). A layered view of shape perception. *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axv042.

Green, E.J. (forthcoming). Attentive visual reference. *Mind & Language*.

Hatfield, G. (2014). Psychological experiments and phenomenal experience in size and shape constancy. *Philosophy of Science*, *81*, 940-953.

Hawthorne, J., & Manley, D. (2012). *The Reference Book*. Oxford: Oxford University Press.

Hill, C.S. (2014). The content of visual experience. In *Meaning, Mind, and Knowledge*, pp. 218-238. Oxford: Oxford University Press.

Hoffman, D.D. (1998). *Visual Intelligence: How We Create What We See*. New York: Norton.

Hoffman, D.D., & Richards, W.A. (1984). Parts of recognition. *Cognition*, *18*, 65-96.

Hoffman, D.D., & Singh, M. (1997). Salience of visual parts. *Cognition*, *63*, 29-78.

Hogervorst, M.A., & Eagle, R.A. (2000). The role of perspective effects and accelerations in perceived three-dimensional structure-from-motion. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(3), 934-955.

Hollingworth, A., & Franconeri, S.L. (2009). Object correspondence across brief occlusion is established on the basis of both spatiotemporal and surface feature cues. *Cognition*, *113*, 150-166.

Horowitz, T.S., & Kuzmova, Y. (2011). Can we track holes? *Vision Research*, *51*, 1013-1021.

Howe, P.D., Holcombe, A.O., Lapierre, M.D., & Cropper, S.J. (2013). Visually tracking and localizing expanding and contracting objects. *Perception*, *42*, 1281-1300.

Howe, P.D., Incledon, N.C., & Little, D.R. (2012). Can attention be confined to just part of a moving object? Revisiting target-distractor merging in multiple object tracking. *PloS one*, 7, e41491.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, *8*, 489-517.

Hummel, J.E. (2013). Object recognition. In D. Reisburg (ed.), *Oxford Handbook of Cognitive Psychology*, pp. 32-46. Oxford: Oxford University Press.

Hummel, J.E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480-517.

Hummel, J.E., & Stankiewicz, B.J. (1996). Categorical relations in shape perception. *Spatial Vision*, *10*(3), 201-236.

Humphreys, G., Gillebert, C.R., Chechlacz, M., & Riddoch, M.J. (2013). Reference frames in visual selection. *Annals of the New York Academy of Sciences*, *1296*, 75-87.

Huntley-Fenner, G., Carey, S., & Solimondo, A. (2002). Objects are individuals but stuff doesn't count: Perceived rigidity and cohesiveness influence infants' representations of small groups of entities. *Cognition*, *85*, 203-221.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jastorff, J., Kourtzi, Z., & Giese, M.A. (2006). Learning to discriminate complex movements: Biological versus artificial trajectories. *Journal of Vision*, *6*, 791-804.

Jeshion, R. (2002). Acquaintanceless *de re* belief. In J. Campbell, M. O'Rourke, & D. Shier (eds.), *Meaning and Truth*. New York: Seven Bridges Press.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*, 195-204.

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

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, & H. Wettstein (eds.), *Themes From Kaplan*. Oxford: Oxford University Press.

Kayaert, G., Biederman, I., & Vogels, R. (2003). Shape tuning in macaque inferior temporal cortex. *The Journal of Neuroscience*, *23*, 3016-3027.

Kayaert, G., Biederman, I., & Vogels, R. (2005). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cerebral Cortex*, *15*, 1308-1321.

Kayaert, G., & Wagemans, J. (2010). Infants and toddlers show enlarged visual sensitivity to nonaccidental compared with metric shape changes. *i-Perception*, *1*, 149-158.

Keane, B. (2009). Visual objects as the referents of early vision: A response to A Theory of Sentience. In L. Trick, & D. Dedrick (eds.), *Computation, Cognition, & Pylyshyn*. Cambridge, MA: MIT Press.

Kellman, P.J. (2003). Visual perception of objects and boundaries: A four-dimensional Approach. In R. Kimchi, M. Behrmann, & C.R. Olson (eds.), *Perceptual Organization in Vision: Behavioral and Neural Perspectives*. Mahwah, NJ: Lawrence Erlbaum, pp. 155-201.

Kim, J. (1976). Events as property exemplifications. In M. Brand, & D. Walton (eds.), *Action Theory*. Dordrecht: Reidel.

Kim, J.G., & Biederman, I. (2012). Greater sensitivity to nonaccidental than metric changes in the relations between simple shapes in the lateral occipital cortex. *NeuroImage*, *63*, 1818-1826.

Kimchi, R. (2009). Perceptual organization and visual attention. *Progress in Brain Research*, *176*, 15-33.

Kimia, B.B. (2003). On the role of medial geometry in human vision. *Journal of Physiology-Paris*, *97*, 155-190.

Kirsh, D. (2003). Implicit and explicit representation. In L. Nadel (ed.), *Encyclopedia of Cognitive Science*, Vol. 2. London: Macmillan, pp. 478-481.

Klatzky, R.L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freksa, C. Habel, & K.F. Wender (eds.), *Spatial Cognition*. Berlin: Springer, pp. 1-17.

Koenderink, J.J., van Doorn A.J., & Kappers A.M.L. (1996). Pictorial surface attitude and local depth comparisons. *Perception & Psychophysics*, *58*, 163-173.

Koenderink, J.J., van Doorn, A.J., Kappers, A.M.L., & Todd, J.T. (2001). Ambiguity and the 'mental eye' in pictorial relief. *Perception*, *30*, 431-448.

Kozlowski, L.T., & Cutting, J.E. (1977). Recognizing the sex of a walker from a dynamic point light display. *Perception & Psychophysics*, *23*, 459.

Kramer, A.F., & Jacobson, A. (1991). Perceptual organization and focused attention: The role of objects and proximity in visual processing. *Perception & Psychophysics*, *50*(3), 267-284.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kulvicki, J. (2007). Perceptual content is vertically articulate. *American Philosophical Quarterly*, *44*(4), 357-369.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299-321.

Lange, J., Georg, K., & Lappe, M. (2006). Visual perception of biological motion by form: A template-matching analysis. *Journal of Vision*, *6*, 836-849.

Lee, Y-L., Lind, M., Bingham, N., & Bingham, G.P. (2012). Object recognition using metric shape. *Vision Research*, *69*, 23-31.

Leek, E.C., Reppa, I., Rodriguez, E., & Arguin, M. (2009). Surface but not volumetric part structure mediates three-dimensional shape representation: Evidence from part–whole priming. *The Quarterly Journal of Experimental Psychology*, *62*(4), 814-830.

Lescroart, M.D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex*, *23*, 629-637.

Levine, J. (2010). Demonstrative thought. *Mind & Language*, *25*, 169-195.

Li, Y., Sawada, T., Shi, Y., Steinman, R.M., & Pizlo, Z. (2013). Symmetry is the *sine qua non* of shape. In S.J. Dickinson, & Z. Pizlo (eds.), *Shape Perception in Human and Computer Vision*, pp. 21-40. New York: Springer.

Ling, H., & Jacobs, D. W. (2007). Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 286-299.

Lu, Z.-L., & Sperling, G. (2001). Three-systems theory of human visual motion perception: Review and update. *Journal of the Optical Society of America A*, *18*, 2331-2370.

Margolis, E. (1998). How to acquire a concept. *Mind & Language*, *13*(3), 347-369.

Marino, A.C., & Scholl, B.J. (2005). The role of closure in defining the "objects" of object-based attention. *Perception & Psychophysics*, *67*, 1140-1149.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Marr, D., & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London, B: Biological Sciences*, *200*, 269–294.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London B: Biological Sciences*, *258*(1353), 273-279.

Matthen, M. (2005). *Seeing, Doing, and Knowing*. Oxford: Oxford University Press.

McCloskey, M., Rapp, B., Yantis, S., Rubin, G., Bacon, W.F., Dagnelie, G., Gordon, B., Aliminosa, D., Boatman, D.F., Badecker, W., Johnson, D.N., Tusa, R.J., & Palmer, E. (1995). A developmental deficit in localizing objects from vision. *Psychological Science*, *6*, 112-117.

McCloskey, M., & Rapp, B. (2000). Attention-referenced visual representations: Evidence from impaired visual localization. *Journal of Experimental Psychology*, *26*, 917-933.

McGee, V., & McLaughlin, B.P. (2000). The lessons of the many. *Philosophical Topics*, *28*, 129-151.

McLeod, P., Dittrich, W., Driver, J., Perrett, D., & Zihl, J. (1996). Preserved and impaired detection of structure from motion by a "motion-blind" patient. *Visual*

Cognition, *3*(4), 363-391.

Minsky, M., & Papert, S.A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Mitroff, S.R., & Alvarez, G.A. (2007). Space and time, not surface features, guide object persistence. *Psychonomic Bulletin & Review*, *14*, 1199-1204.

Mitroff, S.R., Scholl, B.J., & Wynn, K. (2004). Divide and conquer: How object files adapt when a persisting object splits into two. *Psychological Science*, *15*, 420-425.

Mitroff, S.R., Scholl, B.J., & Wynn, K. (2005). One plus one equals one: The effects of merging on object files. Poster presented at the annual meeting of the Psychonomic Society, Toronto, ON, Canada.

Mole, C. (2008). Attention and consciousness. *Journal of Consciousness Studies*, *15*, 86-104.

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

Morrison, J. (ms). Perceptual structuralism.

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

Navon, D. (1976). Irrelevance of figural identity for resolving ambiguities in apparent motion. *Journal of Experimental Psychology: Human Perception & Performance*, *2*, 130-138.

Nelson, R., Reiss, J.E., Gong, X., Conklin, S., Parker, L., & Palmer, S.E. (2014). The shape of a hole is perceived as the shape of its interior. *Perception*, *43*, 1033-1048.

Neri, P. (2009). Wholes and subparts in visual processing of human agency. *Proceedings of the Royal Society B*, *276*, 861-869.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Norman, J.F., & Lappin, J.S. (1992). The detection of surface curvatures defined by optical motion. *Perception & Psychophysics*, *51*(4), 386-396.

Norman, J.F., Todd, J.T., & Phillips, F. (1995). The perception of surface orientation from multiple sources of optical information. *Perception & Psychophysics*, *57*(5), 629-636.

Norman, J.F., Todd, J.T., Perotti, V.J., & Tittle, J.S. (1996). The visual perception of three-dimensional length. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 173-186.

Ons, B., & Wagemans, J. (2011). Development of differential sensitivity for shape changes resulting from linear and non-linear planar transformations. *i-Perception*, *2*, 121-136.

Palmer, S.E. (1977). Hierarchical Structure in Perceptual Representation. *Cognitive Psychology*, *9*, 441-474.

Palmer, S.E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. Lloyd (eds.), *Cognition and Categorization*, pp. 261-304. Hillsdale, NJ: Lawrence Erlbaum.

Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Palmer, S.E. (2003). Perceptual organization and grouping. In R. Kimchi, M. Behrmann, & C.R. Olson (eds.), *Perceptual Organization in Vision: Behavioral and Neural*

Perspectives. Mahwah, NJ: Lawrence Erlbaum, pp. 3-43.

Palmer, S.E., Davis, J., Nelson, R., & Rock, I. (2008). Figure-ground effects on shape memory for objects versus holes. *Perception*, *37*, 1569-1586.

Palmer, S.E., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*, 29-55.

Pautz, A. (2009). What are the contents of experiences? *The Philosophical Quarterly*, *59*, 483-507.

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.

Peterson, M.A., de Gelder, B., Rapcsak, S.Z., Gerhardstein, P.C., & Bachoud-Levi, A.C. (2000). Object memory effects on figure assignment: Conscious object recognition is not necessary or sufficient. *Vision Research*, *40*, 1549-1567.

Pflugshaupt, T., Nyffeler, T., von Wartburg, R., Wurtz, P., Lüthi, M., Hubl, D., Gutbrod, K., Juengling, F.D., Hess, C.W., & Müri, R.M. (2007). When left becomes right and vice versa: Mirrored vision after cerebral hypoxia. *Neuropsychologia*, *45*, 2078-2091.

Pinker, S. (1997). *How the Mind Works*. New York: W.W. Norton & Company.

Pinto, J., & Shiffrar, M. (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica*, *102*, 293-318.

Pizlo, Z. (2008). *3D Shape: Its Unique Place in Visual Perception*. Cambridge, MA: MIT Press.

Prinz, J.J. (2012). *The Conscious Brain: How Attention Engenders Experience*. Oxford: Oxford University Press.

Pylyshyn, Z.W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341-423.

Pylyshyn, Z.W. (2003). *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.

Recanati (2012). *Mental Files*. Oxford: Oxford University Press.

Riesenhuber, M. & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*, 162-168.

Robertson, L., Treisman, A., Friedman-Hill, S., & Grabowecky, M. (1997). The interaction of spatial and object pathways: Evidence from Balint's syndrome. *Journal of Cognitive Neuroscience*, *9*, 295-317.

Rock, I. (1983). *The Logic of Perception*. Cambridge, MA: MIT Press.

Rosch, E.H. (1978). Principles of categorization. In E. Rosch, & B. Lloyd (eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Rosch, E.H., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Rosenberg, R.D., & Carey, S. (2009). Infants' representations of material entities. In B.M. Hood & L.R. Santos (eds.), *The Origins of Object Knowledge*. Oxford: Oxford University Press, pp. 165-188.

Rosenfeld, A. (1986). Axial representations of shape. *Computer Vision, Graphics, and Image Processing*, *33*(2), 156-173.

Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues*, *20*, 368-393.

Saiki, J., & Hummel, J.E. (1998). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 227-251.

Schellenberg, S. (2008). The situation-dependency of perception. *The Journal of Philosophy*, *105*, 55-84.

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

Scholl, B.J. (2007). Object persistence in philosophy and psychology. *Mind & Language*, *22*, 563-591.

Scholl, B.J., & Pylyshyn, Z.W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, *38*, 259-290.

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.

Schwenkler, J. (2012). Does visual spatial awareness require the visual awareness of space? *Mind & Language*, *27*, 308-329.

Schwenkler, J. (2013). Do things look the way they feel? *Analysis*, *73*(1), 86-96.

Schwitzgebel, E. (2011). *Perplexities of Consciousness*. Cambridge, MA: MIT Press.

Scimeca, J.M., & Franconeri, S.L. (2015). Selecting and tracking multiple objects. *WIREs Cognitive Science*, *6*, 109-118.

Searle, J.R. (1983). *Intentionality*. Cambridge: Cambridge University Press.

Sekuler, A., & Palmer, S. (1992). Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General*, *121*(1), 95-111.

Shoemaker, S. (2000). Introspection and phenomenal character. *Philosophical Topics*, *28*, 247-273.

Siegel, S. (2010). *The Contents of Visual Experience*. Oxford: Oxford University Press.

Singh, M. & Hoffman, D. D. (1998). Part boundaries alter the perception of transparency. *Psychological Science*, *9*, 370-378.

Singh, M., & Hoffman, D.D. (2001). Part-based Representations of Visual Shape and Implications for Visual Cognition. In T.F. Shipley & P.J. Kellman (eds.), *From Fragments to Objects: Segmentation and Grouping in Vision*. New York, NY: Elsevier Science.

Singh, M., Seyranian, G.D., & Hoffman, D.D. (1999). Parsing silhouettes: The short-cut rule. *Perception & Psychophysics*, *61*, 636-660.

Smith, A.D. (2002). *The Problem of Perception*. Cambridge, MA: Harvard University Press.

Spelke, E.S. (1990). Principles of object perception. *Cognitive Science*, *14*, 29-56.

Spelke, E.S. (1994). Initial knowledge: Six suggestions. *Cognition*, *50*, 443-447.

Stankiewicz, B.J., Hummel, J.E., & Cooper, E.E. (1998). The role of attention in priming for left-right reflections of object images: Evidence for a dual representation of object shape. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 732-744.

Strawson, P.F. (1963). *Individuals*. New York: Anchor Books.

Tarr, M.J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233-282.

Todd, J.T. (1995). The visual perception of 3-dimensional structure from motion. In W.

Epstein, & J. Rogers (eds.), *Handbook of Perception and Cognition, Volume 5: Perception of Space and Motion*. Orlando, FL: Academic Press, pp. 201-226.

Todd, J.T. (2004). The visual perception of 3D shape. *Trends in Cognitive Sciences*, *8*, 115-121.

Todd, J.T. (2005). Stability and change. *Visual Cognition*, *12*(4), 639-642.

Todd, J.T., & Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception & Psychophysics*, *50*, 509-523.

Todd, J.T., Chen, L., & Norman, J.F. (1998). On the relative salience of Euclidean, affine, and topological structure for 3D form discrimination. *Perception*, *27*, 273-282.

Todd, J.T., Weismantel, E., & Kallie, C.S. (2014). On the relative detectability of configural properties. *Journal of Vision*, *14*(1), 1-8.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Tripathy, S.P., Ogmen, H., & Narasimhan, S. (2011). Multiple-object tracking: A serial attentional process? In C. Mole, D. Smithies and W. Wu (eds), *Attention: Philosophical and Psychological Essays*. Oxford: Oxford University Press.

Tse, P.U. 1999). Volume completion. *Cognitive Psychology*, *39*, 37-68.

Tye, M. (1991). *The Imagery Debate*. Cambridge, MA: MIT Press.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *203*(1153), 405-426.

Ullman, S. (1983). Recent computational studies in the interpretation of structure from motion. In J. Beck, B. Hope, & A. Rosenfeld (eds.), *Human and Machine Vision*. New York: Academic Press, pp. 459-480.

Ullman, S. (1996). *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, *67*, 21-44.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *13*, 992-1006.

Vaina, L.M., Lemay, M., Bienfang, D.C., Choi, A.Y., & Nakayama, K. (1990). Intact “biological motion” and “structure from motion” perception in a patient with impaired motion mechanisms: A case study. *Visual Neuroscience*, *5*, 353-369.

vanMarle, K., & Scholl, B.J. (2003). Attentive tracking of objects versus substances. *Psychological Science*, *14*, 498-504.

Vecera, S.P., Behrmann, M., & McGoldrick, J. (2000). Selective attention to the parts of an object. *Psychonomic Bulletin and Review*, *7*, 301-308.

Vecera, S.P., & O'Reilly, R.C. (1998). Figure-ground organization and object recognition processes: An interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 441-462.

Veltkamp, R.C., & Latecki, L.J. (2006). Properties and performance of shape similarity measures. In *Proceedings of the 10th IFCS Conference on Data Science and Classification*, Slovenia, July 2006.

Vogel, E.K., Woodman, G.F., & Luck, S.J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92-114.

Vogels, R., Biederman, I., Bar, M., & Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *Journal of Cognitive Neuroscience*, *13*, 444-453.

Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin*, *138*, 1172-1217.

Wang, B., Zhou, T.G., Zhuo, Y., & Chen, L. (2007). Global topological dominance in the left hemisphere. *Proceedings of the National Academy of Sciences*, *104*, 21014-21019.

Watson, D.G., & Humphreys, G.W. (1997). Visual marking: Prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological Review*, *104*, 90-122.

Watson, S.E., & Kramer, A.F. (1999). Object-based visual selective attention and perceptual organization. *Perception & Psychophysics*, *61*, 31-49.

Webb, B.S., Ledgeway, T., & Rocchi, F. (2011). Neural computations governing spatiotemporal pooling of visual motion signals in humans. *The Journal of Neuroscience*, *31*(13), 4917-4925.

Williams, P., & Tarr, M.J. (1999). Orientation-specific possibility priming for novel three-dimensional objects. *Perception & Psychophysics*, *61*(5), 963-976.

Wynn, K., Bloom, P., & Chiang, W.-C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, *83*, B55-B62.

Xu, Y. (2002). Encoding color and shape from different parts of an object in visual short-term memory. *Perception & Psychophysics*, *64*(8), 1260-1280.

Xu, Y., & Singh, M. (2002). Early computation of part structure: Evidence from visual search. *Perception & Psychophysics*, *64*(7), 1039-1054.

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, *24*, 295-340.

Zhou, K., Luo, H., Zhou, T., Zhuo, Y., & Chen, L. (2010). Topological change disturbs object continuity in attentive tracking. *Proceedings of the National Academy of Sciences*, *107*(50), 21920-21924.