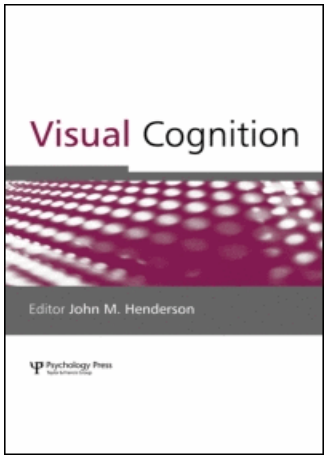

Visual Cognition
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713683696

Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition

John E. Hummel

Online Publication Date: 30 June 2001
To cite this Article: Hummel, John E. (2001) 'Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition', Visual Cognition, 8:3, 489 - 517
To link to this article: DOI: 10.1080/13506280143000214
URL: http://dx.doi.org/10.1080/13506280143000214

VISUAL COGNITION, 2001, *8* (3/4/5), 489–517

Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition

John E. Hummel

Department of Psychology, University of California, Los Angeles, USA

Behavioural, neural, and computational considerations suggest that the visual system may use (at least) two approaches to binding an object's features and/or parts into a coherent representation of shape: Dynamically bound (e.g., by synchrony of firing) representations of part attributes and spatial relations form a structural description of an object's shape, while units representing shape attributes at specific locations (i.e., a static binding of attributes to locations) form an analogue (image-like) representation of that shape. I will present a computational model of object recognition based on this proposal and empirical tests of the model. The model accounts for a large body of findings in human object recognition, and makes several novel and counter intuitive predictions. In brief, it predicts that visual priming for attended objects will be invariant with translation, scale, and left–right reflection, whereas priming for unattended objects will be invariant with translation and scale, but sensitive to left–right reflection. Five experiments demonstrated the predicted relationships between visual attention and patterns of visual priming as a function of variations in viewpoint. The implications of these findings for theories of visual binding and shape perception will be discussed.

HUMAN OBJECT RECOGNITION

The most important property of the human capacity for object recognition is our ability to recognize objects despite variations in the image presented to the retina. This ability takes two forms. The most commonly studied is recognition despite variations in viewpoint. We can recognize objects in a wide variety of views even though different views can present radically different images to the

Please address all correspondence to J.E. Hummel, Department of Psychology, University of California, 405 Hilgard Ave., Los Angeles, CA 90095-1563, USA.
Email: jhummel@lifesci.ucla.edu

This research was supported by NSF Grant 9709023 and by grants from the UCLA Academic Senate.

http://www.tandf.co.uk/journals/pp/13506285.html DOI:10.1080/13506280143000214

retina. This capacity is particularly challenging to understand because human object recognition is robust to some but not all variations in viewpoint. Recognition is invariant with the location of the image on the retina, left–right (mirror) reflection (Biederman & Cooper, 1991a), scale (Biederman & Cooper, 1992), and, to a lesser degree, rotation in depth (see Lawson, 1999, for a thorough review). However, it is sensitive to rotation in the picture plane (as when an object is upside-down; Jolicoeur, 1985, 1990; Lawson, 1999; Tarr & Pinker, 1989, 1990). The second form of object constancy is our ability to generalize recognition over variations in object shape. This capacity has at least two familiar and important manifestations. First, we are good at recognizing objects as members of a class, such as "chair" or "car", rather than just as specific instances, such as "my office chair" or "Toyota Camry". And second, we easily recognize novel members of known object classes: The first time we see a Dodge Viper, it is easy to recognize it as a car, even if we have never before seen a car with exactly that shape.

Together, these properties are challenging to explain because they defy explanation in terms of simple geometric laws. A system based strictly on the laws of projective geometry—e.g., that used those laws to match the information in an object's two-dimensional (2-D) image to a 3-D model the object's shape (e.g., Lower, 1987; Ullman, 1989, 1996)—would be equally able to accommodate all variations in viewpoint (which the human is not) but would not tolerate variations in an object's shape (which the human does).

These and other properties of human object recognition have led some researchers to postulate that we recognize objects on the basis of *structural descriptions* specifying an object's parts (or features) in terms of their spatial relations to one another (Biederman, 1987; Clowes, 1967; Marr & Nishihara, 1978; Palmer, 1977; Sutherland, 1968; Winston, 1975). The most explicit such theory to date is Biederman's (1987) *recognition by components* and its variants (Bergevin & Levine, 1993; Dickinson, Pentland, & Rosenfeld, 1992; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a, 1998). According to this theory, objects are recognized as collections of simple volumes (*geons*; Biederman, 1987) in particular categorical relations. For example, a coffee mug would be represented as a curved cylinder (the handle) side-attached to a straight vertical cylinder (the body). The relations are critical: If the curved cylinder were attached to the top of the straight cylinder, then the object would be a bucket rather than a mug. This type of representation provides a natural account of many properties of human object recognition. Note that it will not change if the mug is translated across the visual field, moved closer to or farther from the viewer, or left–right (mirror) reflected. But rotating the mug 90° about the line of sight (so that the body is horizontal and the handle is on top) will change the description. Like human object recognition, this description is sensitive to rotations about the line of sight, but insensitive to translation, scale, left–right reflection, and some rotations in depth. It is also

insensitive to things such as the exact length of the handle or the exact width of the body, making it a suitable basis for recognizing many different mugs as members of the same general class (Biederman, 1987; Hummel & Stankiewicz, 1998).

Consistent with this proposal, there is evidence that the visual system explicitly represents an object's parts (Biederman, 1987; Biederman & Cooper, 1991b; Tversky & Hemenway, 1984) in terms of their spatial relations (Hummel & Stankiewicz, 1996b; Palmer, 1978; Saiki & Hummel, 1996, 1998a, b), and that these representations are used in the service of object recognition (Biederman & Cooper, 1991b; Hummel & Stankiewicz, 1996b; Saiki & Hummel, 1996, 1998b). If these properties were all that were true of human object recognition, then we could conclude that recognition is based on generating and matching structural descriptions, as postulated by Clowes (1967) and Sutherland (1968), and later by Palmer (1977), Marr (1982; Marr & Nishihara, 1978), and Biederman (1987). However, several additional findings indicate that structural descriptions cannot provide a complete account of our ability to visually recognize objects. To understand why, it is necessary to consider the computational problem of generating and representing explicit structural descriptions.

Consider generating the description *curved cylinder side-attached to straight cylinder* from the image of a coffee mug (Hummel & Biederman, 1992). First, it is necessary to segment the object's local features (e.g., contours and vertices) into parts-based groups so that the features of one part do not interfere with the interpretation of the other. Likewise, any higher-level interpretations of the parts' attributes (e.g., the shapes of their cross sections and axes, their aspect ratio, etc.) must also be bound into sets. Next, the representation of *curved cylinder* must be bound to the agent role of *side-attached*, and *straight cylinder* to the patient role. An important problem for structural description concerns the nature of these bindings. Bindings can be either *dynamic* or *static*. A dynamic binding is one in which a single representational unit can be used in many different combinations. For example, one unit (or collection of units) might represent cylinders and another might represent the side-attached relation; a cylinder side-attached to another part would be represented by explicitly tagging these units as bound together (e.g., by synchrony of firing; Gray & Singer, 1989; Hummel & Biederman, 1992; von der Malsburg, 1981/1994). Because tags are assigned to units dynamically, the same units can enter into different conjunctions at different times. A static binding is one in which a separate unit is pre-dedicated for each conjunction. For example, one unit might respond to cylinders side-attached to other parts, another might respond to cylinders above other parts, and so forth. Structural description requires dynamic binding (Hummel & Biederman, 1992). The number of units required to pre-code all possible part–relation conjunctions would be prohibitive (growing exponentially with the number of relations).

More importantly, static binding sacrifices the independence of the bound attributes: The fact that a cylinder side-attached to something is more similar to a cylinder above something than to a slab above something is completely lost in a representation where each part–relation binding is coded by a separate unit. This loss of similarity structure is a fundamental property of static binding that cannot be overcome even with sophisticated static codes, such as Smolensky's (1990) tensor products (Holyoak & Hummel, 2000; Hummel & Biederman, 1992). Dynamic binding is thus a necessary prerequisite to structural description.

The problem is that dynamic binding imposes a bottleneck on processing: It is necessarily time consuming and capacity limited, and there is substantial evidence that it requires visual attention (Hummel & Stankiewicz, 1996a; Luck & Beach, 1998; Luck & Vogel, 1997). The limitations of dynamic binding are problematic for structural description theories of object recognition. Structural description cannot be faster or more automatic than dynamic binding, but object recognition apparently is. Face recognition in the macaque is accomplished to a high degree of certainty based on the *first* set of spikes to reach inferotemporal cortex (at least for overlearned stimuli; Oram & Perrett, 1992). Clearly, the macaque visual system recognizes faces without waiting around for several sets of desynchronized spikes. People also recognize objects very rapidly. Intraub (1981) showed that people can recognize common objects presented at the rate of 10 per second (see also Potter, 1976). These findings suggest that object recognition is much too fast to depend on dynamic binding for structural description. Similarly, although dynamic binding—and therefore structural description—requires visual attention, object recognition apparently does not, as shown by findings of both negative priming (e.g., Tipper, 1985; Treisman & DeSchepper, 1996) and positive priming (Stankiewicz, Hummel, & Cooper, 1998) for ignored objects.

COMPLEMENTARY SOLUTIONS TO THE BINDING PROBLEM

Both the strengths of the structural description account of shape perception (its ability to account for the flexibility of human object recognition, as well as the role explicit parts and relations in shape perception) and its weaknesses (its inability to account for the speed and automaticity of object recognition) stem from its ability to represent parts independently of their relations, and the resulting need for a dynamic solution to the binding problem. The question is how the visual system profits from the strengths of this approach without suffering its limitations.

Hummel and Stankiewicz (1996a) hypothesized that the visual system solves this problem by adopting a hybrid solution to the binding problem. Their model, *JIM.2* (so named because it is the successor to Hummel & Biederman's,

1992, *JIM* model), uses dynamic binding to generate structural descriptions of object shape when an object is attended, and uses static binding to maintain the separation of an object's parts when an object is ignored (or when dynamic binding otherwise fails). The basic idea starts with the observation that static binding—in which separate units responding to separate conjunctions of properties—is not capacity limited in the way that dynamic binding is. The theory predicts that when the visual system succeeds in segmenting an object into its parts, shape perception will have the characteristics of a structural description: Recognition will be largely invariant with variations in viewpoint, and part attributes will be represented independently of one another, and of the parts' interrelations. When the visual system fails to segment an image into its parts (e.g., due to inattention or insufficient processing time), shape perception will have the characteristics of the statically bound representation: It will be more sensitive to variations in viewpoint, and part attributes will not be represented independently of their spatial relations.

This paper presents the most recent version of this theory—a model I shall refer to as *JIM.3*, as it is the successor of *JIM.2*—and reviews several experiments that have been conducted to test its predictions (Stankiewicz & Hummel, 2000; Stankiewicz et al., 1998). In addition to accounting for a wide variety of findings in human shape perception and object recognition, *JIM.2* and *JIM.3* predict a number of counterintuitive relationships between visual attention and patterns of visual priming. As elaborated later, attended images should prime themselves, scaled and translated versions of themselves, and left–right reflections of themselves; by contrast, ignored images should prime themselves, scaled and translated versions of themselves, but not left–right reflections of themselves. Five experiments tested these (and other) predictions of the model, and all the predictions were supported by the empirical results.

THE MODEL

Space limitations preclude describing the model in detail, so I will describe it only in broad strokes, and note important departures from the *JIM.2* model of Hummel and Stankiewicz (1996a). The model is an eight-layer artificial neural network that takes a representation of the contours in an object's image as input, and activates a representation of the object's identity as output (Figure 1). Units in the first three layers represent local image features, including contours (layer 1), vertices and axes of symmetry (layer 2), and the shape properties of surfaces (e.g., layer 3). The explicit coding of object surfaces represents a significant departure from the models of Hummel and Biederman (1992) and Hummel and Stankiewicz (1996a). Each surface is represented in terms of five categorical properties of its shape: whether it is *elliptical* (i.e., is bounded by a single contour without sharp discontinuities) vs. *non-elliptical*

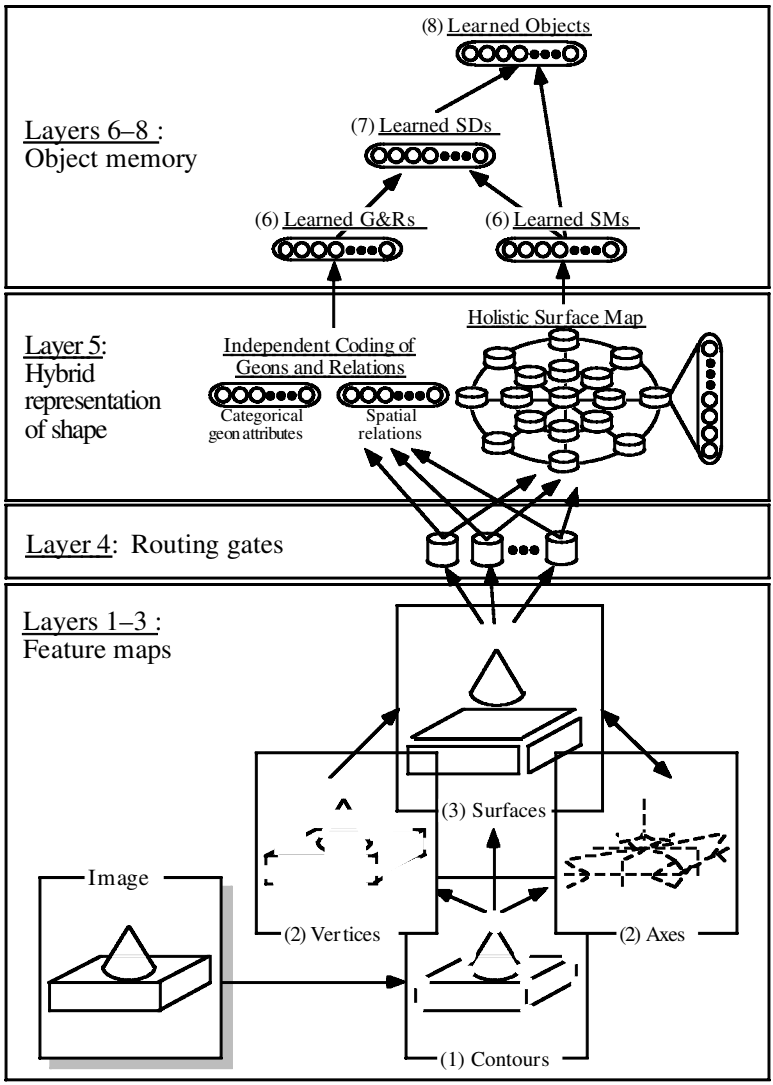

Figure 1. Illustration of *JIM.3*'s overall architecture. Units in layer 1 represent the contours in an object's image. Units in layer 2 represent vertices where contours coterminate and axes of symmetry between contours belonging to the same surface. Units in layer 3 represent the shape properties of object surfaces. Units in layer 4 gate the output of layer 3 to the independent geon shape units and the surface map in layer 5. Although the layer 4 gating units are depicted as a simple one-dimensional array in the figure, in the model they are distributed spatially over the visual field, like the local feature units in layers 1–3. Layer 5 is divided into two components: The independent units represent the shape attributes of an object's geons, and the units in the surface map represent shape attributes of surfaces at each of 17 location in a circular reference frame. Units in layers 6–8 encode patterns of activation in layer 5 (i.e., structural descriptions of objects) into long-term memory for recognition. See text for details.

(i.e., bounded by multiple contours that meet at vertices); the degree to which its axes of symmetry are *parallel*, *expanding* (i.e., wider at one end than the other), *concave* (i.e., wider at both ends than in the middle), or *convex* (i.e., wider in the middle than at either end); whether its major axis is *curved* or *straight*; whether it is *truncated* (meaning that some of its axes of symmetry terminate in the midsegments of its bounding contour or contours), or *pointed* (meaning that all its axes terminate at vertices where its contours meet); and whether it is *planar* (i.e., exists in a single 2-D plane in 3-D space) or *non-planar* (i.e., is curved in 3-D space). These properties are inferred from the properties of the vertices and axes of symmetry within a surface, and are used in subsequent layers of the model to infer the shape attributes of the geons to which the surfaces belong.

The local features coded in the model's first three layers group themselves into sets corresponding to geons by synchrony of firing: Lateral excitatory and inhibitory interactions cause the units in layers 1–3 to fire in synchrony when they represent features of the same geon, and out of synchrony when they belong to separate geons (Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a). The units in layer 3 activate units in layer 5 that represent an object's geons in terms of their shape attributes (e.g., whether a geon has a straight cross section or a curved cross section, whether its sides are parallel or non-parallel, etc.) and the configuration of their surfaces; interactions among the units in layer 5 compute the spatial relations among geons (e.g., whether a given geon is above or below another, larger than or smaller than another, etc.). Units in layer 6 learn to respond to specific patterns of activation in layer 5 (i.e., specific geons in specific relations), and units in layers 7 and 8 learn to respond to combinations of these patterns (i.e., collections of geons in particular relations). Together, the units in layers 6–8 constitute the model's long-term memory for known objects.

The synchrony relations established in layers 1–3 are preserved in layers 4–6, where they serve to bind together the various attributes of a geon's shape, and to bind geons to their spatial relations. However, the lateral interactions that establish synchrony and asynchrony take time, so initially (i.e., in the first several iterations [tens of ms] after an image is presented) all the features in an image will tend to fire at once, whether they belong to the same geon or not. The inhibitory interactions that cause the features of separate geons to fire out of synchrony with one another also require visual attention, so if an object is never attended, its features will never group themselves into parts-based sets (Hummel & Stankiewicz, 1996a, 1998). The model's fourth layer is a collection of inhibitory gates that project the instantaneous outputs of layer 3 (surface properties) to layer 5 (geons and relations) (Hummel & Stankiewicz, 1996a). The interactions between layers 3, 4, and 5 allow the model to capitalize on the dynamic binding of features into parts-based sets when synchrony can be established, and prevent catastrophic failure when it cannot.

Layer 5: The representation of object shape

Layer 5 is divided into two components: A collection of units that represent geon shape attributes independently of one another and of the geons' interrelations (the left-hand side of layer 5 in Figure 1), and a holistic *surface map*, which represents the shape attributes of an object's surfaces (the same shape attributes coded in layer 3) separately at each of several locations in a circular reference frame (right-hand side of layer 5 in Figure 1).[1] The independent shape units represent the shape of a geon in terms of five categorical attributes (cf. Biederman, 1987): whether its cross-section is *straight* (like that of a brick) or *curved* (like that of a cylinder); whether its major axis is *straight* or *curved*; whether its sides are *parallel* (like those of a brick), *expanding* (like a cone or wedge), *convex* (like a football) or *concave* (narrower in the middle than at the ends); and whether the geon is *pointed* (like a cone) or *truncated* (like a cylinder or truncated cone). These attribute units are capable of distinguishing 31 different kinds of geons.[2] For example, a brick has a straight cross-section, a straight major axis, parallel sides, and is truncated; a curved cone has a round cross-section, a curved major axis, expanding sides, and is pointed. Additional units code whether a geon's aspect ratio is *flat* (like a disk), *intermediate* (like a cube), or *elongated* (like a pipe). Other units code the spatial relations between geons (specifically, whether a given geon is *above*, *below*, *beside*, *larger-than*, and/or *smaller-than* other geons in an object).

These units represent geons attributes and relations independently in the sense that a unit that responds to a given property will respond to that property in the same way regardless of the geon's other properties. That is, separate units

---

[1] In *JIM.2*, the units in the holistic map represent, not surface properties, but geon attributes (the same attributes represented on the independent units). The shift to surface properties in the current version of the model is more a matter of convenience (i.e., simply copy surface features rather than having to use them to compute geon properties) than a strong theoretical claim. I am not theoretically committed to any particular vocabulary of features in the surface map; there is simply no empirical or theoretical basis for choosing a specific vocabulary at this time. The only strong theoretical claims about the surface map are that it is (a) holistic, (b) invariant with translation and scale, and (c) able to be activated automatically (i.e., without attention or other provisions for dynamic binding).

[2] This is a substantial improvement over the eight different kinds of geons the Hummel and Biederman (1992) and Hummel and Stankiewicz (1996a) models are capable of distinguishing. The previous models are incapable of distinguishing pointed geons from truncated geons, and are incapable of distinguishing different kinds of non-parallelism in a geon's sides. The difference between the current model and the previous models stems from the fact that the current model uses the vertices and axes in an object's image to infer the properties of the object's surfaces, and uses the surface properties to infer the properties of the geons. That is, the mapping from vertices and axes to geon properties is a two-stage mapping in the current model. By contrast, the previous models infer geon properties directly from vertex and axis properties (i.e., a one-stage mapping). It is tempting to speculate that geon properties are not linearly separable in the space of vertex and axis properties, and therefore cannot be unambiguously computed in a one-stage mapping.

are responsible for representing separate attributes and relations. This independence has two important consequences. First, it makes the representation completely invariant with translation, scale, and left–right reflection (the relations *left-of* and *right-of* are both coded simply as *beside*; Hummel & Biederman, 1992), and relatively insensitive to rotation in depth; it also permits the model to respond to the shapes of an object's parts independently of their relations and vice versa. The second consequence of the independence is that it makes the representation heavily dependent on synchrony of firing to bind geon attributes and relations into geon-based sets. For example, if the local features of the cone in Figure 1 fire in synchrony with one another and out of synchrony with the local features of the brick, then the properties of the cone will fire out of synchrony with the properties of the brick; together, resulting activation vectors will unambiguously specify that a cone is on top of a larger brick. However, if the local features of the cone happen to fire in synchrony with those of the brick, then the properties of the two geons will be superimposed on the independent units. The resulting pattern of activation cannot distinguish a cone and a brick from a cylinder and a wedge, or even from a single geon with a combination of cone and brick properties. That is, when dynamic binding fails, the representation formed on the independent units of layer 5 is effectively useless (cf. Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a).

The holistic surface map is much less sensitive to such binding errors than are the independent geon attribute and relation units. This component of layer 5 is a collection of units that represent the shape attributes of an object's surfaces at each of 17 locations in a circular reference frame (see the right-hand side of layer 5 in Figure 1). The map is holistic in the sense that each unit codes a static binding of one property to one location in the map (cf. Hummel, 2000; Hummel & Stankiewicz, 1996b). The units in layer 4 gate the instantaneous outputs of layer 3 (surface shape properties) to the independent units of layer 5 and to specific locations in the map: Surface properties are projected to the map in a way that preserves their topological relations (i.e., surfaces in adjacent locations in layer 3 project to adjacent locations in the surface map; see Hummel & Stankiewicz, 1996a, for details).[3] As a result, the units in the surface

[3]Units in layer 4 have circular receptive fields in layer 1. Different units have receptive fields of different sizes and locations, so that, collectively, they cover the entire visual field. Layer 4 units gate the connections between layer 3 units (surface features) and units in layer 5 (both the independent units on the left-hand side [layer 5i], and the units in the holistic surface map on the right-hand side [layer 5s]). A layer 4 gating unit, *i*, will become active at time *t* if all the layer 1 contour units with non-zero outputs at time *t* are: (1) contained entirely with its receptive field (i.e., if the gating unit can "see" all the contour units that are currently generating outputs), and (2) are not contained entirely within the receptive field of any other unit, *j*, whose receptive field is contained wholly within *i*'s receptive field. When *i* becomes active, it enables the connections of layer 3 surface units to send their outputs to: (1) the independent units in layer 5, and (2) those units in the surface map whose locations *relative to the surface map* are the same as the features' locations

(Continued overleaf)

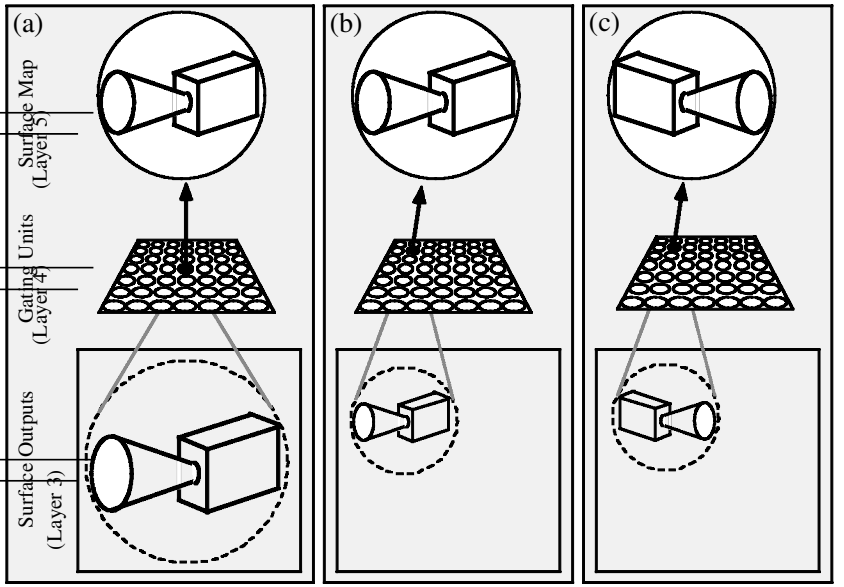

Figure 2. Illustration of the mapping of layer 3 (surface feature unit) outputs to the surface map of layer 5 via the gating units of layer 4. Adjacent surfaces in layer 3 project to adjacent locations in the map in a way that discards their absolute locations and sizes in the image. (a) and (b) illustrate that the same image in different locations and sizes in layer 3 produce exactly the same representation on the surface map. The image in (c) is a left–right reflection of the image in (b), so the representation of (c) is a left–right reflection of the representation of (b) on the surface map.

map maintain the separation of geon attributes even when multiple geons fire at the same time. (Different geons, occupying different locations in the image, will project to different locations in the map.) Although the mapping from layer 3 to the surface map preserves the topological relations of the surfaces, it discards their absolute locations and sizes in the image. Thus, the representation formed on the surface map is sensitive to rotation (both in depth and in the picture plane) and left–right reflection (compare Figures 2b and 2c), but it is invariant with translation and scale (compare Figures 2a and 2b).

(Footnote 3–continued)
relative to i's receptive field. For example, surface features in the upper left of *i*'s receptive field will project their outputs to the upper-left of the surface map (i.e., the corresponding layer 3-to-layer 5 connections will be enabled); features in the centre of the receptive field will project their outputs to the centre of the surface map, etc. These gating operations, which can all be performed in a strictly feed-forward fashion, cause the representation generated on the surface map to be invariant with translation and scale. That is, due to the operation of the layer 4 gating units, a given shape will have the same representation in the surface map regardless of its size or location in the image. See Hummel and Stankiewicz (1996a) for details.

The independent units and the surface map work together to permit the model to generate structural descriptions of object shape when dynamic binding succeeds, and to permit recognition of objects in familiar views even when dynamic binding fails. When an object image is first presented for recognition, all the local features in the image (contours, vertices, axes, and surfaces) will tend to fire at once, whether they belong to the same geon or not. In layer 5, the resulting pattern of activation on the independent units blends all properties of all the geons in the image, but the pattern on the surface map keeps the geons spatially separate (see Figure 3a). Although the blended representation on the independent units is useless for specifying the object's identity, the holistic representation on the surface map can specify the object's identity, provided the object is depicted in a familiar view. The initial (globally synchronized) burst of activation in layers 1–3 also serves as the units' first opportunity to exchange excitatory and (when the stimulus is attended) inhibitory signals. The excitatory signals encourage features belonging to the same geon to continue to fire in synchrony with one another, and the inhibitory signals encourage the features of separate geons to fire out of synchrony with one another (see Hummel & Stankiewicz, 1996a, 1998), so as processing proceeds, the object's geons

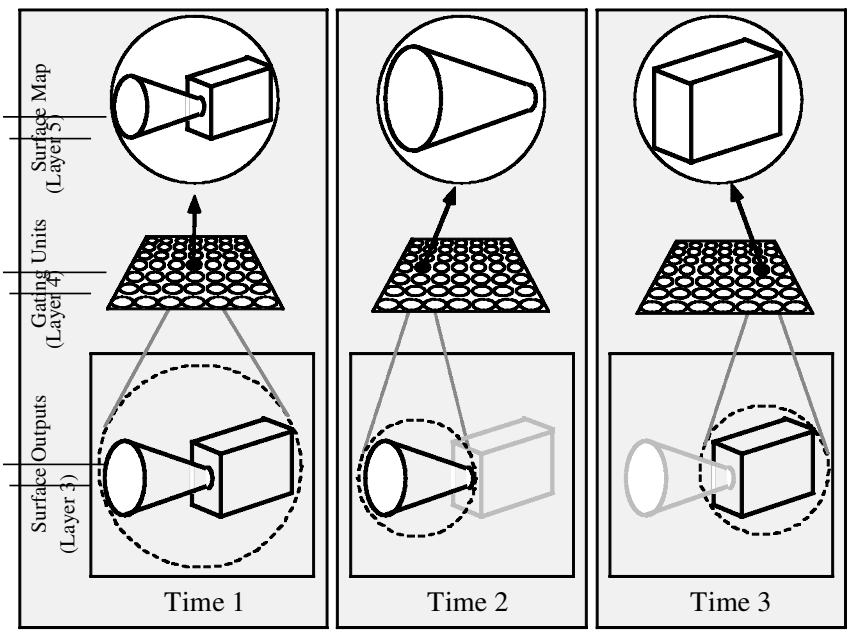

Figure 3. Illustration of the representation formed on the surface map as a function of the synchrony relations among the features of an object's geons. When all the features of an object fire at once (time 1), the separation of the object's surfaces is preserved in their separate locations in the surface map. After the objects' geons have desynchronized their outputs (times 2 and 3), individual geons are projected to the surface map one at a time.

come to fire out of synchrony with one another. In layer 5, the resulting series of patterns of activation constitute a structural description specifying the object's geons in terms of their shape attributes, spatial relations, and the topological relations of their constituent surfaces. This structural description can specify the object's identity even if it is depicted in a novel view, and even if the object is a novel member of a known object category (i.e., even if its shape differs somewhat from the shape of familiar members of the category).

Layers 6–8: Encoding shapes into long-term memory

The patterns of activation generated in layer 5 are encoded into the model's long-term memory by a simple kind of unsupervised Hebbian learning (see Hummel & Saiki, 1993). Patterns of activation generated on the independent units are learned by individual units in layer 6 (one unit per pattern; layer 6i in Figure 1), as are patterns of activation generated on the surface map (layer 6s in Figure 1). That is, each unit in layer 6 learns to respond to the shape attributes of one geon (or collection of geons) and its relations to other geons in the object (layer 6i), or to the arrangement of surfaces in a geon (or collection of geons) (layer 6s). Units in layer 7 sum their input from units in layers 6i and 6s over time to reconstruct the desynchronized patterns representing an object's constituent geons into a complete structural description of the object as a whole (i.e., a complete *object model*; see Biederman & Cooper, 1991a). Units in layer 7 activate object identity units in layer 8. These units are assumed to correspond to non-visual neurons representing the object's identity (and other aspects of its semantic content). The activation of the object identity units is taken as the measure of object recognition. The pattern generated on the surface map when an image is first presented (i.e., before the local features in layers 1–3 come to fire in geon-based sets) will be a holistic representation of the entire object. The unit in layer 6s that encodes this pattern is allowed to connect directly to the corresponding object identity unit in layer 8, providing a fast holistic route to recognition for objects in familiar views.

SIMULATIONS

Simulations of existing findings

JIM.3 evolved from Hummel and Biederman's (1992) *JIM* model in response to findings suggesting that dynamic binding is not strictly necessary for object recognition (as did Hummel & Stankiewicz's, 1996a, *JIM.2*). The resulting model is substantially more complicated than *JIM*. It is therefore important to show that it can still account for the findings against which *JIM* was tested—namely, the effects of various transformations in viewpoint on recognition performance. To this end, the model was trained on one view of each of 20 simple

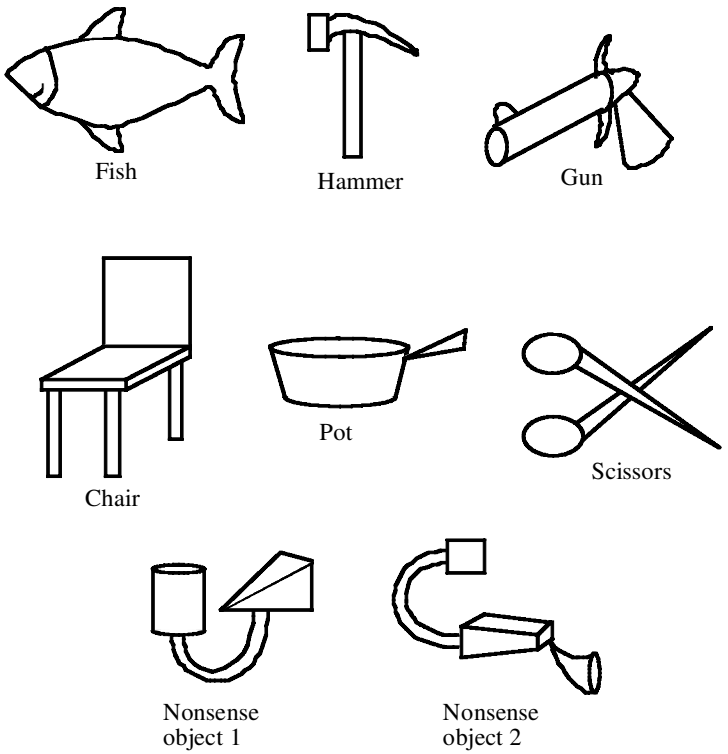

Figure 4. Eight of the twenty object images on which *JIM.3* was trained.

objects,[4] and tested for its ability to recognize those objects in new (untrained) views. A subset of the images on which *JIM.3* was trained is depicted in Figure 4. The objects were designed to be structurally complex: Most have several parts, many (such as the fish) contain ambiguous segmentation cues, and most are left–right asymmetrical. The model was trained by presenting each view once, and allowing the units in layers 6–8 to encode its representation on layer 5 (i.e., the object' s structural description) into long-term memory (see Hummel & Stankiewicz, 1996a). It was then tested for its ability to recognize the images on which it was trained, translated versions of those images (i.e., with the same image translated to a new location in the visual field), scaled versions of the images, left–right reflections of the images, and images rotated 45°, 90°, 135°, and 180° in the picture plane. In all the simulations described here, I allowed the

[4]All objects had at least two parts (most had three or more). They comprised: a car, a cat, a chair, a fish, a revolver, a hammer, a house, the cone-on-a-bric k object in Figure 1, a telephone, a cooking pot, a sailboat, a pair of scissors, a pair of eyeglasses, a teacup, a teapot, a sliding board, and four nonsense objects.

model to run until one object identity unit in layer 8 achieved an activation of at least 0.5, and was at least 0.2 more active than its next-highest competitor (activations range from 0 to 1). The "winning" object identity unit was taken as the model's response. I recorded the model's recognition time (RT: the number of iterations it had to run to satisfy these criteria) and accuracy (i.e., whether the winning unit corresponded to the object actually depicted in the image).

Figure 5 shows the model's performance on the trained, translated, scaled and reflected images (the RTs shown are means over 10 runs), and Figure 6 shows its performance on the rotated images (mean RT and error rate over 10 runs). Errors are not reported in Figure 5 because the model made no errors in these simulations. Like *JIM* and *JIM.2*, *JIM.3* accounts both for the invariance of human object recognition with translation, scale, and left–right reflection, and for the detrimental effects (on both response time and accuracy) of rotation in the picture plane. Importantly, it also accounts for the "cusp" in the rotation function at 180°: *JIM.3*, like people (see Jolicoeur, 1985) and like *JIM* (see Hummel & Biederman, 1992), is faster and more accurate to recognize images that are completely upside-down (i.e., 180° off upright) than images that are slightly less than perfectly upside-down (e.g., 135° off upright; see Figure 6).

As shown in Figure 5, the model was just as fast to recognize translated and scaled versions of the images on which it was trained as to recognize the original images themselves. It was somewhat slower to recognize the left–right reflected images than to recognize the original images (although it was perfectly accurate in its recognition of the reflected images). This advantage for the original images over the left–right reflected images in the model's RT

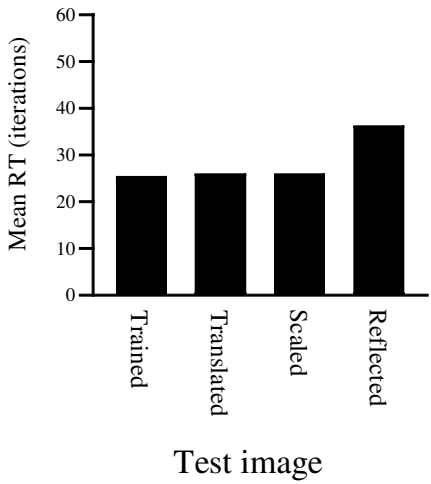

Figure 5. *JIM.3*'s recognition performance (recognition time, RT) with the images on which it was trained, and translated, scaled, and left–right reflected versions of those images. RTs reflect means over 10 runs of all 20 objects.

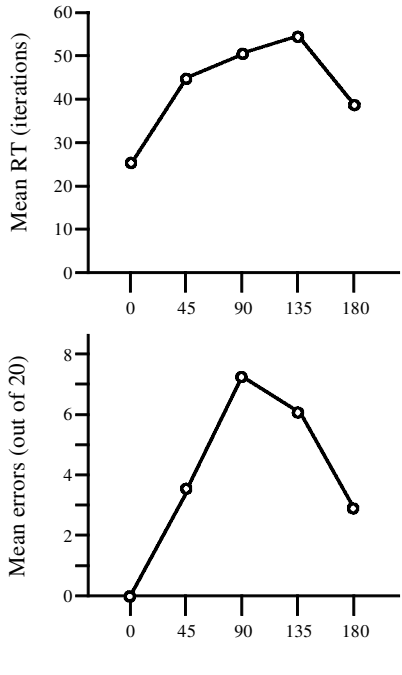

Figure 6. *JIM.3*' s recognition performance (RT and errors) with the images on which it was trained (0°), and 45°, 90°, 135°, and 180° rotated versions of those images. Rotations were in the picture plane.

performance reflects the operation of the holistic surface map: The stored holistic object representation (in the direct connection from layer 6s to layer 8) allows the model to recognize an object in a familiar view based solely on the first pattern of activation generated on the surface map. Because the surface map is invariant with translation and scale, this makes recognition of translated and scaled images just as fast as recognition of familiar images. However, the surface map is not invariant with left–right reflection, so recognition of reflected images requires the model to decompose the image into parts and generate a structural description. The resulting structural description is invariant with left–right reflection, but the process of generating it takes time. Hence, recognition of left–right reflected images is slower than recognition of the original images. At first sight, this result seems to be inconsistent with the fact that Biederman and Cooper (1991a) showed that visual priming is completely invariant with left–right reflection (in the sense that images visually primed their left–right reflections just as much as they primed themselves). However, Biederman and Cooper measured visual priming over long prime-probe delays (on the order of minutes), and they did not report subject' s response times recognizing familiar objects in unfamiliar left–right orientations. As discussed in

detail shortly, *JIM.3* predicts this result. However it also predicts that short-term priming (i.e., on the order of seconds rather than minutes) will be sensitive to left–right reflection, and that it will take slightly longer to recognize familiar objects in unusual left–right orientations (although accuracy should be high). The former prediction (about short-term priming) turns out to be correct, as discussed shortly (Stankiewicz et al., 1998); to the best of my knowledge, the latter prediction remains untested (although there are well-known effects of non-canonical viewpoints on the time required for recognition; see Lawson, 1999).

Novel predictions

The fundamental theoretical tenet underlying *JIM.3* (and *JIM.2*) is that the visual system uses two solutions to the binding problem for the representation of object shape, and that these solutions have complementary properties: Dynamic binding is "expensive", requiring both visual attention and time to establish, but results in structural descriptions that specify object properties independently, and is therefore highly flexible (e.g., among other things, it is robust to variations in viewpoint and the metric properties of an object' s shape); by contrast, static binding is "inexpensive", requiring neither attention nor time to establish, but results in a representation that lacks the flexibility of a structural description. This tenet leads to the general prediction that the visual representation of an attended image should differ *qualitatively* from the visual representation of an unattended image. One specific manifestation of this general prediction is that attended images should visually prime both themselves and their left–right reflections (by virtue of the structural description generated on the independent units), whereas ignored images should prime themselves (by virtue of the holistic surface map), but not their left–right reflections.[5]

To illustrate the implications of this relationship between attention and priming, I tested the model for its ability to recognize the trained images after either the objects' independent representations had been primed (layer 6i), or their holistic representations had been primed (layer 6s), or both had been primed. (The simulation results in Figure 5 serve as a baseline condition, in

[5]One very important (and very hard) problem the model does not solve is the figure–ground segmentation problem. Like the vast majority of all computational models of object recognition, *JIM.3* can "view" only one object at a time. We assume that when an object is ignored, its features occasionally fire, but when they do, they all fire in synchrony with one another (i.e., all ignored objects are forced to share a single "time slice" in the oscillatory firing; see Hummel & Stankiewicz, 1998). This account predicts that the relationship between attention and patterns of priming should vary as a function of how many ignored objects there are in the visual field: The more ignored objects, the smaller the probability that any given one of them will have the opportunity to fire by itself within any fixed amount of time. In addition, with multiple ignored objects, there may be the opportunity for binding errors, even in the surface map, if two or more objects happen to fire at the same time. We have yet to investigate these possibilities empirically.

which neither representation had been primed.) I primed these representations by turning up the gain (i.e., the growth rate in the activation function) on the units in layer 6i and/or 6s (respectively), which allows them to become active faster in response to input from layer 5.[6] The assumptions underlying this manipulation are the following: (1) If an image is attended, then the representation of that image on both the independent units and the surface map will become active and therefore be primed; hence, to simulate priming for attended images, I primed both layer 6i and layer 6s. (2) If an image is ignored, then its (holistic) representation in the surface map will become active, but no useful representation will be activated on the independent units; to simulate priming for ignored images, I therefore primed layer 6s, but not layer 6i. (3) Priming for the surface map is less enduring than priming for the independent units. (This latter was less an a priori assumption than a hypothesis suggested by the Stankiewicz et al. 1998, data summarized later. However, by assuming it, the model provides a straightforward account of a subtle and counterintuitive aspect of those data.) To simulate the effects of long-term priming of attended images, I primed layer 6i but not layer 6s. I then tested the model with the trained images and recorded its RT with each. I operationalized priming as the difference between the mean RT (over objects and runs) in the unprimed condition (Figure 5) minus the mean RT in each primed condition (layer 6i only, layer 6s only, or both 6i and 6s).

Figure 7 shows the simulation results (i.e., magnitude of priming in iterations) in the three priming conditions (both layer 6i and 6s primed, only 6i primed, and only 6s primed). Based on the characteristics of the independent and holistic representations in layer 5, it is possible to turn these simulation results into specific behavioural predictions about the relationship between visual attention and priming for various kinds of object images. First consider the model's predictions regarding short-term priming—i.e., when the probe trial (i.e., the second presentation of an image) follows immediately after the prime trial (Figures 8 and 9). Recall that the independent representation, but not the holistic representation, is (1) invariant with left–right reflection but (2) requires visual attention. Therefore, attended images should visually prime both themselves and their left–right reflections, whereas ignored images should prime themselves but not their left–right reflections. There are four components to this prediction (see Figure 8). (1) A person who attends to an image on one trial (the prime), and then immediately sees the very same image on the next (probe) trial should profit from priming in both the independent representation (because the image was attended) and the holistic representation

[6] It is arguably more realistic to assume the priming resides, not in the units, but in the mapping between representations (i.e., in the connections between units; see Cooper, Biederman, & Hummel, 1992). However, for the purposes of the current simulations, the two produce equivalent results.

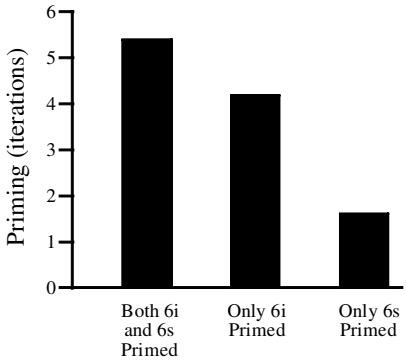

Figure 7. *JIM.3*'s recognition performance in the priming conditions expressed as magnitude of priming: RT on unprimed simulation runs (Figure 5) minus RT on primed simulation runs (both 6i and 6s primed, only 6i primed or only 6s primed).

(because the prime and probe images were identical, and the prime-probe delay was short); that is, the magnitude of priming in this case should be equivalent to the magnitude of priming in the *both 6i and 6s* condition (Figure 8, attended/identical). (2) A person who attends to an image on the prime trial and then immediately sees its left–right reflection on the probe trial should profit from priming in the independent representation (because the image was attended, and the independent representation is invariant with left–right reflection), but should not profit from priming in the holistic representation (because the holistic representation is sensitive to left–right reflection); the magnitude of priming in this case should be equivalent to the magnitude of priming in the *6i only*

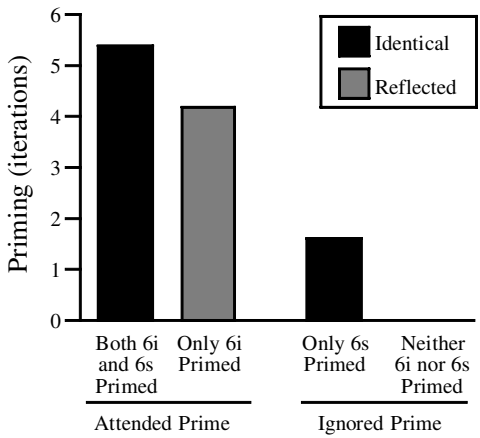

Figure 8. *JIM.3*'s predictions regarding the relationship between attention (i.e., whether the prime image is attended or ignored) and visual priming for identical images and left–right reflections. See text for details.

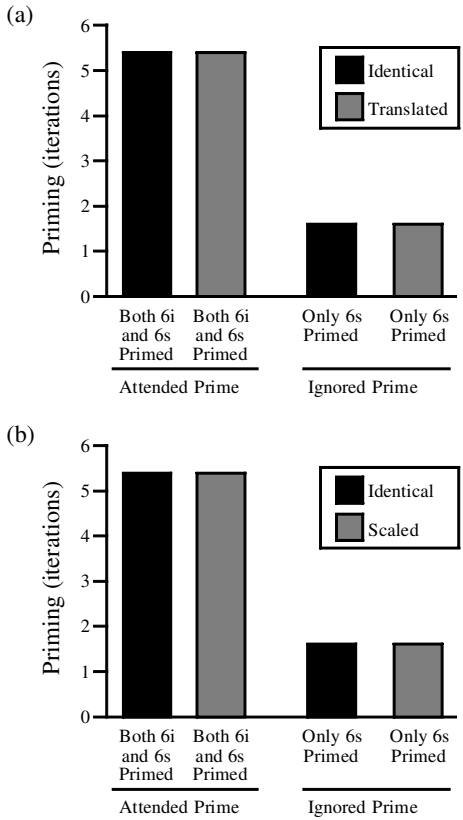

Figure 9. *JIM.3*'s predictions regarding the relationship between attention (i.e., whether the prime image is attended or ignored) and visual priming for identical images (a and b), translated images (a), and scaled images (b). See text for details.

condition (Figure 8, attended/reflected). (3) A person who ignores an image on the prime trial and then immediately sees the same image on the probe trial should not profit from priming in the independent representation (because the image was not attended), but should profit from priming in the holistic representation (because the prime and probe images were identical, and the prime-probe delay was short); that is, the magnitude of priming in this case should be equivalent to the magnitude of priming in the *6s only* condition (Figure 8, ignored/identical). (4) Finally, a person who ignores an image on the prime trial and then immediately sees its left–right reflection on the probe trial should profit neither from priming in the independent representation (because the image was not attended), nor from priming in the holistic representation (because the probe image was the left–right reflection of the prime); that is, there should be no priming at all (Figure 8, ignored/reflected).

Although the holistic representation is sensitive to left–right reflection, it is invariant with translation and scale (recall Figure 3). The pattern of visual priming effects should therefore be substantially simpler if the left–right reflected probe images are replaced with probes that are either translated relative to the prime (i.e., so that the prime and probe images are presented in different parts of the visual field) or scaled relative to the prime (i.e., so that the prime and probe images are presented in different sizes). In this case, the only variable that should matter is whether the prime is attended or ignored: Attended images should profit from priming in both the independent and holistic representations, regardless of whether the probe is identical to the prime (Figure 9a and b, attended/identical), translated relative to the prime (Figure 9a, attended/translated), or scaled relative to the prime (Figure 9b, attended/scaled); ignored images should profit from priming in the holistic representation but not the independent representation, regardless of whether the probe is identical to the prime (Figure 9a and b, ignored/identical), translated relative to the prime (Figure 9a, ignored/translated) or scaled relative to the prime (Figure 9b, ignored/scaled).

Finally, consider the predicted effects of attention and viewpoint for long prime-probe delays (i.e., on the order of several minutes, as in the experiments of Biederman & Cooper, 1991a). If priming in the holistic representation (layer 6s) is short lived (i.e., persists for seconds but not minutes; Stankiewicz et al., 1998), then with long prime-probe delays, all holistic priming will disappear, resulting in the pattern depicted in Figure 10. A few properties of this pattern are notable. First, the magnitude of priming in all conditions is lower than in the corresponding conditions with short prime-probe delays (the one exception being the ignored/reflected condition, which was already at zero even for short prime-probe delays). Second, priming in all the ignored conditions goes to

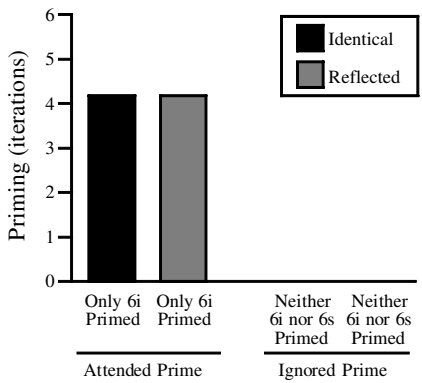

Figure 10. *JIM.3*'s predictions regarding the relationship between attention (i.e., whether the prime image is attended or ignored) and visual priming for identical images and left–right reflections over long prime-probe delays. See text for details.

zero. And third, priming in the attended/reflected condition is equivalent to priming in the attended/identical condition: That is, over long prime-probe delays, priming is completely invariant with left–right reflection, as reported by Biederman and Cooper (1991a). The predicted advantage for attended/identical images over attended/reflected images in the short prime-probe delay simulations is due to the holistic (layer 6s) priming in the identical condition. With long prime-probe delays, this advantage is predicted to disappear, with the result that images prime their left–right reflections just as much as they prime themselves.

TESTS OF THE MODEL'S PREDICTIONS

Brian Stankiewicz and his colleagues (Stankiewicz & Hummel, 2001; Stankiewicz et al., 1998; see also Stankiewicz, 1997) ran five experiments to test these predictions. In the basic paradigm with short prime-probe delays, trials were grouped into prime/probe pairs. Each prime trial began with a fixation cue (a cross in the centre of the screen), followed by an empty box either left or right of fixation. The box was followed by two line drawings of objects: One appeared inside the box, and the other outside the box (on the other side of fixation). The subject's task was to name the object that appeared inside the box, ignoring the other object. The precue box served both as an endogenous attentional cue (subjects knew to attend to the image in the box) and as an exogenous attentional cue (its abrupt onset automatically attracts attention). The images remained on the computer screen for 165 ms, and were then masked. The entire prime trial (from precue to mask) lasted only 195 ms, which is too brief to permit a saccade to the attended image. After a 2 s pause, the prime trial was followed by the corresponding probe trial, which presented a fixation cue in the centre of the screen, followed by a single line drawing either in the centre of the screen (Stankiewicz et al.; Stankiewicz & Hummel, Exp. 2), or in a new location off fixation (i.e., a location occupied by neither the attended nor the ignored image during the prime trial; Stankiewicz & Hummel, Exp. 1). The subject's task was to name the object depicted in the line drawing. In addition to the paired trials producing short prime-probe delays, one experiment (Stankiewicz et al., Exp. 3) investigated the effects of priming for attended and ignored images over delays lasting several minutes (as in Biederman & Cooper, 1991a). In this experiment, any prime object that was not probed on the trial immediately following the prime was probed in a separate block of trials at the very end of the experiment: If the probe trial presented the ignored object from the immediately preceding prime trial, then the (un-probed) attended object was presented in the probe block at the end of the experiment; and if the probe trial presented the attended image from the immediately preceding prime trial, then the (un-probed) ignored object was presented at the end of the experiment. This long-delay probe condition permitted Stankiewicz et al. to compare the

effects of attention (and viewpoint) on short-delay priming (i.e., delays lasting a few seconds) to their effects on long-delay priming (i.e., delays lasting several minutes).

In all these experiments, the critical manipulation was the relationship between the probe image and the images presented on the corresponding prime trial (see Table 1). The probe either depicted the object that was attended on the prime trial (the *attended* condition), the object that was ignored on the prime trial (the *ignored* condition), or an object the subject had not previously seen in the experiment (the unprimed *baseline* condition). In the attended and ignored conditions, the probe image could either be: (1) identical to the corresponding prime image (the *identical* condition; Stankiewicz et al., 1998; Stankiewicz & Hummel, 2001), (2) a left–right reflection of the prime image (the *reflected* condition; Stankiewicz et al.), (3) identical to the prime image except for its location in the visual field (the *translated* condition; Stankiewicz & Hummel, Exp. 1), (4) identical to the prime except for its size (the *scaled* condition; Stankiewicz & Hummel, Exp. 1), or (5) an image of a different object with the same basic-level name as the object depicted in the prime trial (the *different exemplar* control condition; Stankiewicz et al., Exp. 2). For example, if the prime presented an image of a jumbo jet (basic level name "airplane"), the probe image in the different exemplar condition would depict a small private

TABLE 1
Summary of conditions in the Stankiewicz et al. (1998) and Stankiewicz and Hummel (2001) experiments

Prime-probe relationship	*Attention condition*		*Unprimed baseline*
	Attended	*Ignored*	
Identical	SHC: 1,2,3 SH: 1,2 LD	SHC: 1,2,3 SH: 1,2 LD	
Left–right reflected	SHC: 1,2,3 LD	SHC: 1,2,3 LD	
Translated	SH: 1	SH: 1	SHC: 1,2,3 SH: 1,2 LD
Scaled	SH: 2	SH: 2	
Different exemplar	SHC: 2	SHC: 2	

Table entries indicate which conditions appeared in which experiments. Letters refer to papers: "SHC" denotes Stankiewicz et al. (1998) and "SH" denotes Stankiewicz and Hummel (2001). Numbers refer to experiments in corresponding experiments (Experiments 1, 2, and 3 in SHC, and Experiments 1 and 2 in SH). "LD" indicates that the corresponding condition was run in both the *long prime-probe delay* and *short prime-probe delay* conditions (Stankiewicz et al., 1998, Exp. 3); all conditions not so marked were run in the *short prime-probe delay* condition only.

plane such as a Cessna (basic level name "airplane"). The different exemplar control condition serves as a basis for estimating what fraction of any observed priming is specifically visual (i.e., reflects priming in visual representations), as opposed to non-visual (e.g., priming for an object's name or concept; see Biederman & Cooper, 1991a, b, 1992). I will not discuss this condition further except to note that the majority of the priming observed in the experiments summarized here is specifically visual, and that none of the effects summarized here are attributable to the non-visual components of the priming (see Stankiewicz et al. for details). In all five of the experiments, priming was operationalized as the time it took subjects to name an object in the unprimed baseline condition minus the time it took them to name the objects in the corresponding primed condition.

The results of all five experiments are strikingly consistent with the behavioural predictions of the model. Consider first the role of attention in short-delay priming for images and their left–right reflections (Figure 11). As predicted (Figure 8), attended images primed both themselves and their left–right reflections, whereas ignored imaged primed themselves, but not their left–right reflections. Also as predicted, the advantage for identical images over their left–right reflections in the attended condition was the same as the advantage for identical images over their left–right reflections in the ignored condition (about 50 ms. in both cases). That is, the effects of attention (attended vs. ignored) and view (identical vs. reflected) were strictly additive. Next consider the predicted role of attention in priming for translated and scaled images, and recall that the model predicts that, although ignored images should not prime their left–right reflections, priming for ignored images should none the less be invariant with translation and scale (Figure 9). Consistent with this prediction, ignored images primed translated versions of themselves just as much as they primed themselves (Figure 12a), and primed scaled versions of themselves just as much as they primed themselves (Figure 12b). Also as predicted, although short-delay priming did not vary as a function of translation or scaling for either the attended or ignored images, priming for attended images was generally

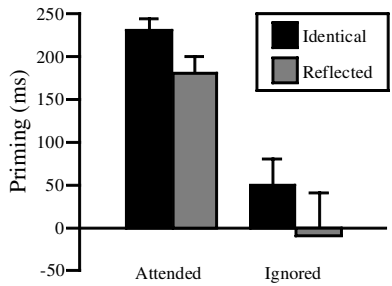

Figure 11. Patterns of visual priming for attended and ignored identical and reflected images (data from Stankiewicz et al., 1998, Exp. 1). See text for details.

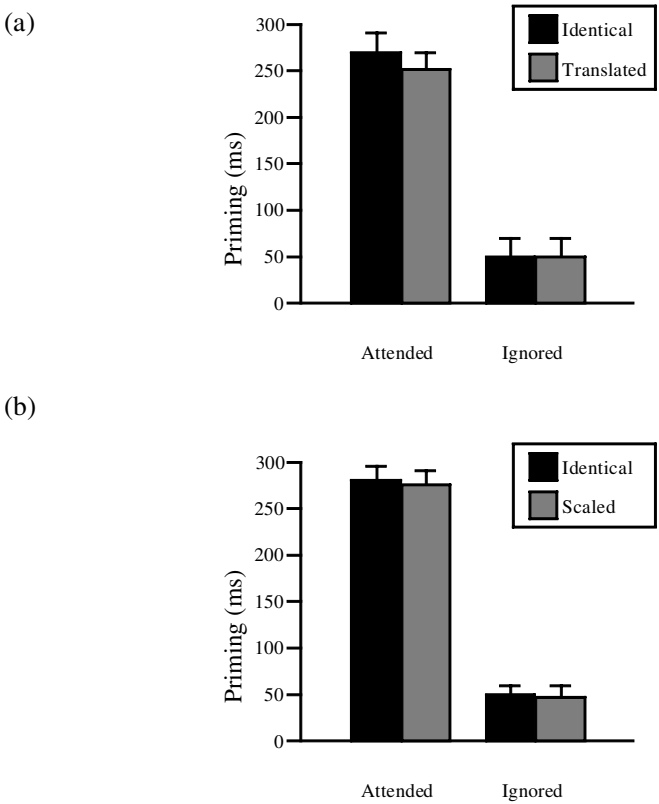

Figure 12. Patterns of visual priming for attended and ignored identical (a and b), translated (a), and scaled (b) images. The data in (a) are from Stankiewicz and Hummel (2001, Exp. 1); those in (b) are from Stankiewicz and Hummel (2001, Exp. 2). See text for details.

much larger than priming for ignored images. Finally, consider the effects of attention and view (identical vs. reflected) over long prime-probe delays (Figure 13). In contrast to short prime-probe delays, which show a priming advantage for identical images over their left–right reflections (as predicted by the model), long prime-probe delays are predicted to show no such advantage: The model predicts that long-delay priming for attended images will be invariant with left–right reflection, and that there will be no priming for ignored images (Figure 10). This pattern of effects is borne out exactly in the experimental data (Figure 13).

GENERAL DISCUSSION

Many aspects of the human capacity for shape perception and object recognition indicate that we represent an object's shape in terms its parts and their

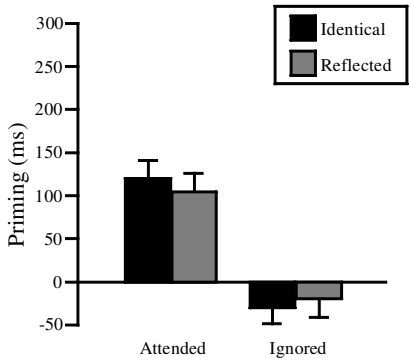

Figure 13. Patterns of visual priming for attended and ignored identical and reflected images over long prime-probe delays (data from Stankiewicz et al., 1998, Exp. 3). See text for details.

categorical spatial relations. The view-invariances and sensitivities characterizing human object recognition are captured precisely by the structural description model of Hummel and Biederman (1992; Hummel & Stankiewicz, 1996a). Structural descriptions also provide a natural account of both our ability to recognize objects as members of a general class (Biederman, 1987; Marr, 1982) and as specific instances (Hummel & Stankiewicz, 1998). More direct evidence for the role of structural descriptions in shape perception comes from studies showing that our visual systems explicitly represent objects in terms of their surfaces (Nakayama & He, 1995; Nakayama & Shimojo, 1992) and parts (Biederman, 1987; Biederman & Cooper, 1991b), and from studies showing that we represent the spatial relations among an object's parts both explicitly and independently of the parts they relate (Hummel & Stankiewicz, 1996a; Saiki & Hummel, 1996, 1998a, b; see also Palmer, 1978). The role of structural descriptions in shape perception is also supported by our ability to appreciate the relational similarity between shapes, independently of whether similar parts stand in corresponding relations (see Hummel, 2000, for a review). At the same time, however, the speed and automaticity of object recognition suggest that the visual system is not bound by the capacity limits imposed by this approach to representing object shape: Because of the computational demands of dynamic binding, generating a structural description from the information in an object's 2-D image is necessarily time consuming and attention demanding; by contrast, object recognition is both fast and automatic (see Hummel & Stankiewicz, 1996a, for a review).

JIM.3 is a computational instantiation of a theory of how the human visual system exploits the flexibility and expressive power of explicit structural descriptions when it attends to an object's image, without suffering catastrophic failures of recognition when it does not. According to this theory (and its predecessor, *JIM.2*; Hummel & Stankiewicz, 1996a), the visual system solves this problem by adopting two complementary approaches to the problem

of visual feature binding: When an image is attended, dynamic binding of local features into parts-based sets, and of parts to their relations, results in an explicit structural description that supports recognition despite variations in the view in which an object is depicted, and even variations in the object's exact 3-D shape; when an image is ignored, static binding of features to locations in a semi-object-centred reference frame permits recognition provided the image depicts the object in a familiar view (although even in this case, recognition is invariant with translation and scale). The theory makes several novel predictions about the relationship between attention and shape perception, and all the predictions tested to date have been empirically confirmed (Stankiewicz & Hummel, 2001; Stankiewicz et al., 1998).

JIM.3 accounts for a very large number of findings in human shape perception and object recognition, and is entirely consistent with many others. Like its predecessors, *JIM* (Hummel & Biederman, 1992) and *JIM.2* (Hummel & Stankiewicz, 1996a), *JIM.3* provides a direct account of the major view-invariances and view-sensitivities characterizing human object recognition. It also provides an account of the role of categorical shape attributes in object recognition, and of the human capacity to recognize objects at multiple levels of abstraction (e.g., as "a car," or "a Honda Civic," or "my Honda Civic"; cf. Hummel & Stankiewicz, 1998). It is also consistent with our ability to represent an object's parts independently of one another, and of their spatial relations, and suggests a direct basis for appreciating the relational similarity of objects composed of different parts (see Hummel, 2000; Hummel & Holyoak, 1997). More generally, in its ability to account for the known properties of human shape perception, and to successfully predict previously unknown properties, the model illustrates how the strengths and limitations of dynamic binding map onto the strengths and limitations of human shape perception: Both are flexible, and permit the generation of sophisticated structured representations; and both are inherently capacity limited and therefore require finite working memory and attentional resources (cf. Hummel & Holyoak, 1997).

In recent years, models based on varieties of *view-matching*—in which objects are recognized by matching holistic representations of the precise locations of their 2-D features directly to memory (e.g., Edelman, 1998; Poggio & Edelman, 1990; Riesenhuber & Poggio, 1999; Tarr & Bülthoff, 1995)—have become the dominant account of human object recognition in the literature (for reviews, see Hummel, 2000; Lawson, 1999). According to these models, all of shape perception is based on representations akin to the holistic surface map of the current theory. Accordingly, they can only account for the properties of object recognition that stem directly from that kind of representation. Theories in this tradition account for only a fraction of the view-invariances of human object recognition, and are inconsistent with most other aspects of shape perception—most notably, all those phenomena that stem from our ability to represent features or parts independently of their configuration (i.e., virtually all

the interesting properties of human shape perception; see Hummel, 2000). *JIM.3*, like *JIM.2*, acknowledges the important role of holistic representations in our ability to recognize objects without the aid of visual attention. But by integrating these holistic representations into explicit relational descriptions, the current theory also provides a natural account of all the phenomena that depend on our ability to represent an object's parts and relations independently. In turn, the model's capacity to do so depends entirely on its ability to solve the dynamic binding problem.

REFERENCES

Bergevin, R., & Levine, M. D. (1993). Generic object recognition: Building and matching course descriptions from line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15,* 19–36.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

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

Biederman, I., & Cooper, E.E. (1991b). Priming contour deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology, 23,* 393–419.

Biederman, I., & Cooper, E.E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 121–133.

Clowes, M.B. (1967). Perception, picture processing and computers. In N.L. Collins & D. Michie (Eds.), *Machine intelligence* (Vol. 1, pp. 181–197). Edinburgh, UK: Oliver & Boyd.

Cooper, E.E., Biederman, I., & Hummel, J.E. (1992). Metric invariance in object recognition: A review and further evidence. *Canadian Journal of Psychology, 46,* 191–214.

Dickinson, S.J., Pentland, A.P., & Rosenfeld, A. (1992). 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14,* 174–198.

Edelman, S. (1998). Representation is representation of similarities. *Behavioural and Brain Sciences, 21,* 449–498.

Gray, C.M., & Singer, W. (1989). Stimulus specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National academy of Sciences, USA, 86,* 1698–1702.

Holyoak, K.J., & Hummel, J.E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–264). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hummel, J.E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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

Hummel, J.E., & Holyoak, K.J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466.

Hummel, J.E., & Saiki, J. (1993). Rapid unsupervised learning of object structural descriptions. In *Proceedings of the fifteenth annual conference of the Cognitive Science Society* (pp. 569–574). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hummel, J.E., & Stankiewicz, B.J. (1996a). An architecture for rapid, hierarchical structural description. In T. Inui & J. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 93–121). Cambridge, MA: MIT Press.

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

Hummel, J.E., & Stankiewicz, B.J. (1998). Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition, 5,* 49–79.

Intraub, H. (1981). Identification and processing of briefly glimpsed visual scenes. In D. Fisher, R.A. Monty, & J.W. Sender (Eds.), *Eye movements: Cognition and visual perception* (pp. 181–190). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory and Cognition, 13,* 289–303.

Jolicoeur, P. (1990). Identification of disoriented objects: A dual systems theory. *Mind and Language, 5,* 387–410.

Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica, 102,* 221–245.

Lower, D.G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision, 1,* 57–72.

Luck, S.J., & Beach, N.J. (1998). Visual attention and the binding problem: A neurophysiological perspective. In R.D. Wright (Ed.), *Visual attention* (pp. 455–478). New York: Oxford University Press.

Luck, S.J., & Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

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

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

Nakayama, K, & He, Z.J. (1995). Attention to surfaces: Beyond a Cartesian understanding of focal attention. In T.V. Papathomas, C. Chubb, A. Gorea, & E. Kowler (Eds.), *Early vision and beyond* (pp. 181–186). Cambridge, MA: MIT Press.

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

Oram, M.W., & Perrett, D.I. (1992). The time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology, 68,* 70–84.

Palmer, S.E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9,* 441–474.

Palmer, S.E. (1978). Structural aspects of similarity. *Memory and Cognition, 6,* 91–97.

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

Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509–522.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 11,* 1019–1025.

Saiki, J., & Hummel, J.E. (1996). Attribute conjunctions and the part configuration advantage in object category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1002–1019.

Saiki, J., & Hummel, J.E. (1998a). Connectedness and part-relation integration in shape category learning. *Memory and Cognition,* 26, 1138–1156.

Saiki, J., & Hummel, J.E. (1998b). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 227–251.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46,* 159–216.

Stankiewicz, B.S. (1997). *The role of attention in viewpoint-invariant object recognition*. Unpublished doctoral dissertation, University of California, Los Angeles, USA.

Stankiewicz, B.S., & Hummel, J.E. (2001). Automatic priming for translation- and scale-invariant representations of object shape. *Manuscript submitted for publication*.

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,* 732–744.

Sutherland, N.S. (1968). Outlines of a theory of visual pattern recognition in animals and man. *Proceedings of the Royal Society of London, Series B, 171,* 95–103.

Tarr, M.J., & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1494–1505.

Tarr, M.J., & Pinker, S. (1989). Mental rotation and orientation dependence in shape recognition. *Cognitive Psychology, 21,* 233–283.

Tarr, M.J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science, 1*(4), 253–256.

Tipper, S.P. (1985). The negative priming effect: Inhibitory effects of ignored primes. *Quarterly Journal of Experimental Psychology, 37A,* 571–590.

Treisman, A., & DeSchepper, B. (1996). Object tokens, attention, and visual memory. In T. Inui & J.L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 15–46). Cambridge, MA: MIT Press.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113,* 169–193.

Ullman, S. (1989). Aligning pictoral descriptions: An approach to object recognition. *Cognition, 32,* 193–254.

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. Cambridge, MA: MIT Press.

von der Malsburg, C. (1994). The correlation theory of brain function. In E. Domany, J.L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks II* (pp. 95–119). Berlin: Springer. (Original work published 1981.)

Winston, P. (1975). Learning structural descriptions from examples. In P. Winston (Ed.), *The psychology of computer vision* (pp. 157–209). New York: McGraw-Hill.