Green, C. & Hummel, J.E. (2004). Relational perception and cognition: Implications for cognitive architecture and the perceptual-cognitive interface. In B.H. Ross (Ed.): *The Psychology of Learning and Motivation, Vol 44*, (pp. 201-226). San Diego, CA: Academic Press.

RELATIONAL PERCEPTION AND COGNITION: IMPLICATIONS FOR COGNITIVE ARCHITECTURE AND THE PERCEPTUAL-COGNITIVE INTERFACE

Collin Green & John E. Hummel

I. Introduction

A fundamental aspect of human intelligence is the ability to represent and reason about relations. Examples of relational thinking include our ability to appreciate analogies between different objects or events (Gentner, 1983; Holyoak & Thagard, 1995), our ability to apply abstract rules in novel situations (e.g., Smith, Langston & Nisbett, 1992), our ability to understand and learn language (e.g., Kim, Pinker, Prince & Prasada, 1991), our ability to learn and use categories (Ross, 1987), and even our ability to appreciate perceptual similarities (e.g., Palmer, 1978; Goldstone, Medin & Gentner, 1991; Hummel, 2000a; Hummel & Stankiewicz, 1996a).

Relational inferences and generalizations are so commonplace that it is tempting to assume that the psychological mechanisms underlying them are relatively simple. But this would be a mistake. The capacity to form and manipulate explicit relational (i.e., symbolic) representations appears to be a late evolutionary development (Robin & Holyoak, 1995), closely tied to the substantial increase in the size and complexity of the frontal cortex in the brains of higher primates, most notably humans (Stuss & Benson, 1987).

A review of computational models of perception and cognition also suggests that the question of how we represent and reason about relations is nontrivial (see Hummel & Holyoak, 1997, 2003): Traditional symbolic models of cognition (e.g., Anderson, Libiere, Lovett, & Reder, 1998; Anderson, 1990; Falkenhainer, Forbus & Gentner, 1989) simply *assume* relations as a given, making no attempt to understand the origins or detailed nature of these representations in the neural substrate; and traditional connectionist/neural networks models (e.g., Edelman & Intrator, 2003; St. John & McClelland, 1992; O'Reilly & Rudy, 2001; Riesenhuber & Poggio, 1999) fail to represent relations at all. Indeed, the proponents of such models typically reject the idea that the human cognitive apparatus is capable of representing relations explicitly (see Hummel, 2000; Hummel & Holyoak, 1997, 2003, for reviews). Comparatively few models have attempted to address the question of how a neural architecture can represent and process relational structures, or the related question of how early, non-relational representations and processes, for example in early vision, make contact with later, more explicitly relational/symbolic representations (e.g., as underlie reasoning; see Gasser & Colunga, 2001; Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2002; Shastri & Ajjenagadde, 1993; Strong & Whitehead, 1989).

II. Bridging the Gaps: Relating Symbols to Neurons and Cognition to Perception

This trend reflects, in large part, a tendency for researchers to focus on their domain of interest to the exclusion of related domains: As a practical matter, it is simply not possible to take the entire cognitive architecture into account in the attempt to understand, say, reasoning or object recognition. In the domain of reasoning, and higher cognition generally, this tendency often manifests itself in the starting assumption, "given that knowledge is represented in a symbolic format that is roughly isomorphic to propositional notation" (e.g., Anderson, 1990; Falkenhainer et al., 1989; Newell & Simon, 1976); rarely do such models pose the question

of where the proposed representations come from, or how they relate to the outputs of basic perceptual processes. In the domain of visual processing the starting assumption is typically the opposite (for a review and critique, see Hummel, 2000). For example, the goal of most models of object recognition is strictly *recognition* (e.g., Edelman, 1998; Edelman & Intrator, 2002, 2003; Poggio & Edelman, 1990; Tarr & Bülthoff, 1995; Ullman & Basri, 1989). Often, a second goal is to describe the resulting models as much as possible in terms of the properties of visual neurons (e.g., Edelman & Intrator, 2003; Reisenhuber & Poggio, 1999). Nowhere in the vast majority of these models is there any attempt to specify how the visual system might deliver *descriptions* of object shape or arrangements of objects in a scene that might be useful to later cognitive processes; indeed, such representations are often explicitly eschewed as unnecessary (see, e.g., Edelman & Intrator, 2003): It is as though once an object has been recognized, there is nothing else left to do (cf. Hummel, 2000; 2003).

The result is a sharp divide between researchers who assume symbolic representations as a given, and researchers who assume symbolic representations are a fiction. One seeming exception to this divide appears in the form of distributed connectionist models of cognition (e.g., Elman, 1990; Kruschke, 1992, 2001; McClelland, et al., 1995; McClelland & Rumelhart, 1981) and other models that represent concepts as vectors of features (e.g., Nosofsky, 1987; Shiffrin, & Styvers,1997). To the extent that (a) such vectors are reasonable approximations of symbolic representations and (b) basic perceptual processes can be viewed as delivering them as output, these models could serve both as an account of the interface between perception and cognition, and as a bridge between neural and symbolic accounts of knowledge representation. And it is convenient—and tempting—to view them as such. However, as detailed below, the outputs of perceptual processes are not well modeled as lists of features (i.e., (b) is false; Hummel & Biederman, 1992; Hummel, 2000, 2003); and even if they were, list of features are entirely inadequate as approximations of symbolic representations (i.e., (a) is false; Hummel & Holyoak, 1997; 2003). It is therefore necessary to look elsewhere for principles that can serve as an interface between perception and cognition on the one hand, and between neural and symbolic accounts of mental representation on the other.

This chapter reviews our recent and ongoing work toward understanding this interface. It is organized as follows. We begin by reviewing the role of relational processing in perception and higher cognition, with an emphasis on the implications of relational processing for mental representation and cognitive architecture more broadly. Next, we consider how the visual system might deliver such representations to cognition. Finally, we discuss how relational representations may be used as a basis for scene recognition and comprehension—a process that lies squarely at the interface of perception and cognition.

III. Relational Perception and Thinking

Imagine finding yourself in need of a hammer, and discovering that your children have placed your hammer in the configuration illustrated in Figure 1. Rather than simply grabbing the hammer, you would first remove the wine glasses from the top of the box, then lift the box out of the way, supporting the hammer with the other hand. This response to the situation in Figure 1, obvious as it seems, illustrates several important facts about our ability to comprehend novel visual scenes.

First, the inference that you should not simply grab the hammer depends on the ability to relate general knowledge (e.g., an understanding of support relations, of what happens to wine glasses that fall, etc.) to specific knowledge about the situation at hand. Second, most of the relevant knowledge, both the background knowledge and the understanding of the situation at hand, is specifically *relational*: It is not particularly relevant that the objects involved are wine glasses, boxes and a hammer; what matters is that a desired object is supporting a second object that is supporting something fragile. Third, these relations are delivered by the visual system: Without the ability to perceive the relations among the objects, it would be impossible to reason about them, or even to notice that there was anything that needed reasoning about. Finally, these abilities must be generic enough to work in situations that are completely novel (as the scene in Figure 1 presumably is). If the hammer were replaced by an unfamiliar widget, the widget's novelty would not render the scene incomprehensible.

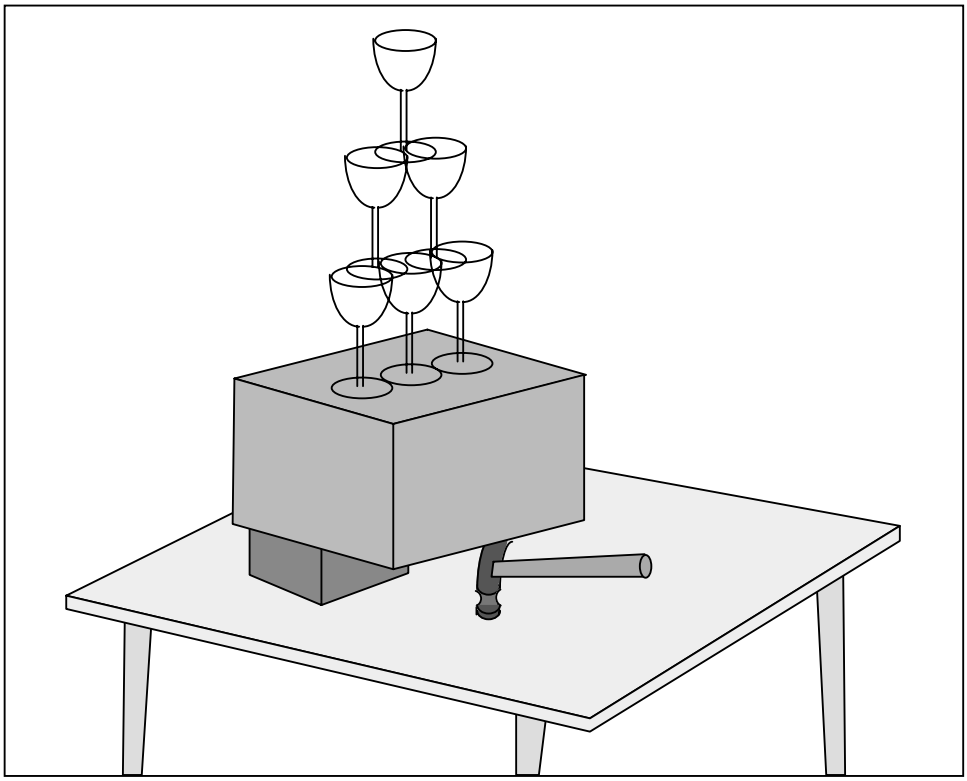

Figure 1. Illustration of a hammer that it is best not to move.

The kind of reasoning that the hammer scene invokes is both commonplace and illustrative of relational reasoning more generally. Relational inferences are inferences that are constrained by the relational roles that objects play, rather than by the identities or features of the objects themselves: It is not the hammer's identity as a hammer that prevents you from moving it, but its role as *object that supports the object that supports the fragile objects*. The capacity to make relational inferences depends on the ability to represent relational roles explicitly, as entities in their own right. In turn, doing so means representing those roles independently of their arguments (Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003): If the representation of the *supports* relation varied as a function of what was supporting what, then there would be little or no basis for generalizing anything learned about support relations in one context (e.g., the context of a pillow supporting a fishbowl) to novel contexts (e.g., a hammer supporting a box supporting wine glasses). It would be as though the two situations were simply *different*, with little or nothing in common.

Representing roles and fillers independently means having one set of units (e.g., neurons) represent relational roles, and a separate set represent the objects that can be bound to those roles. (The units do not need to be physically segregated in the network; they only need to be different units.) By keeping the units separate, any learning that pertains to a relation can be instantiated as connections to and from the units representing the relational roles. Since the connections refer to the roles only, whatever learning they embody (e.g., "if *supports* (*x*, *y*) and *fragile* (*y*), then *must- precede* (*remove-from* (*y*, *x*), *move* (*x*))") will generalize automatically to any new fillers of those roles (Hummel & Holyoak, 1997; 2003).

Representing relational roles independently of their fillers makes it necessary to specify which fillers happen to be bound to which roles at a given time—i.e., to *dynamically* bind roles to their fillers. One way to bind roles to their fillers dynamically is to exploit synchrony of firing (e.g., Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003; Shastri & Ajjenagadde, 1993; Strong & Whitehead, 1989; von der Malsburg, 1981/1994). The basic idea is that units representing bound roles and fillers fire

in synchrony with one another and *out* of synchrony with other role-filler bindings. For example, to represent *supports* (box, wine-glasses), units representing *supporter* would fire in synchrony with units representing the box, while units representing *supported* fire in synchrony with units representing the wine glasses; the *supporter*+box set would fire out of synchrony with the *supported*+glasses set. It is possible to imagine other dynamic binding codes. What is critical is that the binding code, whatever it is, must be independent of the signal that codes a unit's degree of certainty in the hypothesis to which it corresponds (e.g., its activation; Hummel & Biederman, 1992). At present, however, synchrony is the only proposed dynamic binding code with neurophysiological support (see Singer & Gray, 1995, for a review).

Dynamic role-filler binding is crucial for binding roles to fillers in working memory (WM), but it is also necessary to (a) store specific role-filler conjunctions (e.g., describing previously encountered relations between specific objects) in long-term memory (LTM), and (b) form localist tokens of role-filler conjunctions in WM. For both these purposes, conjunctive binding by localist units—i.e., units dedicated to specific role-filler conjunctions—is necessary (Hummel & Holyoak, 2003; see also Hummel & Biederman, 1992; Page, 2000).

In summary, representing relational (i.e., symbolic) structures in a neural architecture is not trivial, and requires a neural/cognitive architecture that is capable of meeting some very specific requirements. In the very least, it must: (1) represent roles independently of their fillers; (2) be able to bind these representations together dynamically in WM; and (3) bind them conjunctively as tokens in both WM and LTM (see Hummel & Holyoak, 2003, for a more complete list of requirements).

Scene comprehension is a problem at the interface of perception and cognition. To the extent that scene comprehension is a case of relational reasoning, it must rest on independent representations of objects and their relational roles; and to the extent that it depends on the outputs of perceptual processing, it also depends on the kinds of objects and relations perception is capable of delivering as output. Together, these considerations suggest that perception delivers to cognition (minimally) a representation of the objects in a scene in terms of their spatial relations to one another (there is evidence that it delivers a great deal more, including specification, for each object, of the relations among the object's parts; see Hummel, 2000). But the perceptual system starts with a representation that does not even specify the identities or locations of objects, much less their relations to one another (namely, the retinal image; and even the representation of local contour elements and "features" in V1 and V2). What does it take to go from a representation of the local features such as lines and vertices in an image to a specification of the objects in the scene in terms of their spatial relations to one another?

IV. From Images to Objects in Relations

Deriving an explicit description of the relations among the objects in a scene from the information in an early visual representation of that scene (e.g., as available in visual area V1) entails solving several problems, some of which, such as image segmentation, still elude satisfactory solutions in the computational literature (which is not to say they are unsolvable; it's just that we do not yet fully understand how the mind solves them; see Hummel, 2000): Starting with a representation of the locations of various "features" such as edges and vertices in an image, the visual system must (a) segment the image into discrete objects, (b) recognize those objects, (c) compute the spatial relations among those objects, (d) form tokens of the objects, their locations, interrelations, etc., as elaborated shortly, and (e) make inferences from the objects and their interrelations to likely interpretations of the meaning of the scene.

These requirements are complicated by the fact that an analogous set of operations characterizes the recognition of individual objects (at least attended objects; Hummel, 2001; Stankiewicz et al., 1998): The visual system must (a) segment the object's image into parts (e.g., geons; Biederman, 1987), (b) characterize those parts in terms of their abstract shape attributes, (c) calculate the spatial relations among the parts, and (d) match the resulting descriptions to long-term memory (Hummel & Biederman, 1992). Moreover, as elaborated later, objects within a scene may be organized into *functional groups*—groups of

objects that function together in the service of a goal (such as a table and chairs in a dining room) but which do not typically constitute an entire scene in themselves. That is, scenes are deeply hierarchical, making it necessary both to represent the various levels of the hierarchy and to relate the levels to one another (e.g., parts to objects, objects to functional groups, and functional groups to scenes). In the following, we describe how Hummel and Biederman's (1992) JIM model of object recognition solves some of these problems—and how it fails to solve others—in order to clarify the problems involved in mapping from an unstructured representation of a visual image (e.g., as in V1) to a structured—i.e., explicitly relational—representation of a visual scene.

A. Image Segmentation

Segmenting an image means figuring out "what goes with what": which features belong to the same object part, which parts belong to the same object, and which objects belong to the same functional group. Simple perceptual properties such as collinearity and cotermination of visual image features, many of which were noted by the Gestaltists, serve to inform the grouping of basic features into objects or object parts (e.g., geons). JIM exploits these principles to group local image features into sets corresponding to geons. Specifically, it uses them to get features of the same geon firing in synchrony with one another, and out of synchrony with the features of other geons.

The Gestaltist principles JIM uses to group image features into geons are useful but by no means complete (see Hummel and Biederman, 1992; Hummel & Stankiewicz, 1996b, for discussions of several situations in which they fail). Many of the failures stem from the fact that Gestalt principles are all *local*, in the sense that they refer to the relations between individual (local) image elements (e.g., the collinearity of individual line segments) without regard for the figure to which the features belong as a whole. Conspicuously absent are top-down constraints based on knowledge of the various global shapes individual elements can form. By contrast, all other things being equal, the human visual system seems to prefer perceptual groupings that result in familiar objects over those that do not (see, e.g., Peterson & Gibson, 1994). For the same reason, Gestalt principles—and local grouping cues more generally—are poorly equipped to address the grouping of parts into objects or objects into functional groups. One exception may be the local cue of connectedness, which plays a role in determining whether the visual system interprets separate parts as belonging to the same object (Palmer & Rock, 1994; Saiki & Hummel, 1996, 1998a, 1998b).

B. Interpretation of Visual Properties

Whatever the complete set of cues to perceptual grouping turns out to be, the *results* of perceptual grouping are of paramount importance. The result of JIM's grouping of image features into parts is that JIM knows which image features refer to the same part, and which refer to different parts (subject to the limitations of its grouping algorithm). This knowledge is tremendously valuable because it allows the model to ignore details such as *where* individual features are located in the visual field for the purposes of figuring out what geon they collectively form (based on their identities) and to ignore *what* the features are for the purposes of figuring out where the geon is located in the visual field, how big it is, etc. In other words, JIM's ability to solve the binding problem at the level of image features allows it to selectively ignore various properties of those features in order to make inferences based on their *other* properties (Hummel & Biederman, 1992): Solving the binding problem makes it possible to treat different sources of information *independently* (see also Hummel & Holyoak, 1997, 2003).

The general principle is that any system that can bind information together dynamically is free to tear it apart—i.e., treat it independently—at will. The ability to do so is perhaps the single most important difference between symbolic cognition and the cognitive capabilities of purely associationist systems, such as traditional connectionist systems (see Hummel & Holyoak, 2003) and view-based approaches to shape perception (see Hummel, 2000, 2003): Such systems are unable to bind information together dynamically, so they are never at liberty to tear it apart when necessary.

1. The Importance of Keeping Separate Things Separate

This point bears elaborating, as it is both a central theme of this chapter and important to understanding mental representations generally. A local feature detector is a unit (e.g., neuron, symbol, etc.) that responds to a particular visual feature at a particular location in the visual field, a particular size, orientation, etc. In other words, it represents a *conjunction* of several different visual properties. Neurons in visual areas V1 and V2 are examples of local feature detectors (or, equivalently for the purposes of the current discussion, local *filters*). It is possible to build detectors for geons, or even complete objects, simply by connecting geon or object units directly to collections of local feature detectors, and this is exactly how view- (a.k.a. "appearance-") based models of object recognition operate: a "view" is an object detector that is connected directly to a set of local feature detectors. The resulting unit can detect its preferred geon or object (or "fragment"; Edelman & Intrator, 2003) when the corresponding local features are present in the image. But because such a unit takes its input from a specific set of local feature detectors, it, like the feature detectors, is only able to recognize its preferred object at a particular location, size and orientation in the image (with the right algorithm for matching features to stored views, such detectors are capable of modest generalization across rotation in depth; see e.g., Tarr & Bülthoff, 1995).

At the other extreme, it is possible to imagine a model that simply lists all the features in an image without any regard for their locations (e.g., Mel, 1997; Mel & Fiser 2000). The advantage of this approach is that it permits recognition of the geon (or object) regardless of where it appears in the image. The limitation is that, although the model knows *which* features are in the image, it has no idea where they are relative to one another: If the features of an object are all scattered about the image but not in the right relations to form the object, then the model will spuriously "recognize" the object even though it is not present. This problem can be alleviated somewhat by positing detectors for conjunctions of features in the right relations (e.g., A-connected-to-B, A-connected-to-C, etc.; Mel & Fiser, 2000). However, this approach only pushes the problem back one level: Now it is possible to fool the model with an A connected to a B in one place, an A connected to a C in another, etc.

The deep problem with both these approaches is that they either use all the information in the image in a conjunctive fashion (as in the case of view-based models), or simply discard information about feature location altogether (as in the feature-based approach) (see Figure 2). A better approach is simply to *separate* the feature ("what") information from the location ("where") information and use each for the tasks to which it is relevant (see Hummel & Biederman, 1992). There is evidence for this kind of separation of information at a gross level in mammalian visual systems (i.e., in the functions of the ventral ["what"] and dorsal ["where" or "how"] cortical visual processing streams; see, e.g., Goodale et al, 1991; Mishkin & Ungerleider, 1982).

This kind of separation of "what" from "where" is also essential within the "what" stream (and probably within the "where/how" stream as well). Like a feature-based model, JIM ignores the locations of a geon's features' for the purposes of inferring the shape attributes of the geon they form. That is, it uses the "what", ignoring the "where". But in contrast to a feature-based model, JIM is not fooled by a collection of unrelated geon features. The reason is that, due to the perceptual grouping of features into geon-based sets, features will only fire in synchrony with one another, and therefore be interpreted as belonging to the same geon, if they belong to the same geon. And while one processing stream in JIM is busy inferring the shape of a geon from its local features, a *separate* processing stream is using the features' locations to infer the geon's location, size, orientation, etc. This metric information, which is represented independently of (i.e., on separate units than) the geon's shape, is then used by routines that compute the relations between geons.

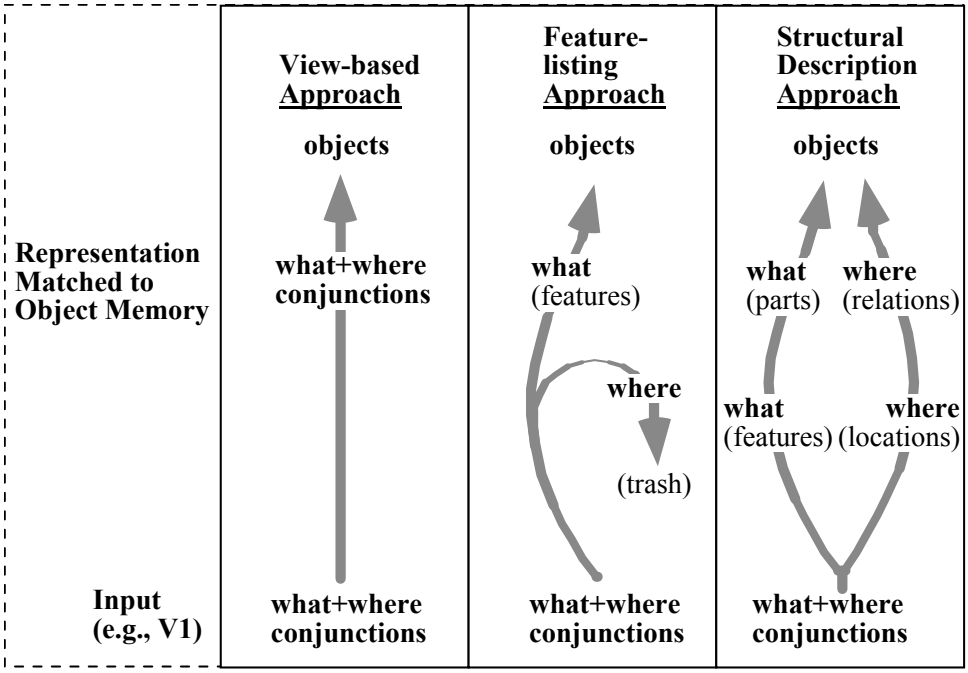

Figure 2. Three approaches to the use of feature and location information in object recognition. Only structural description models separate feature and location information while preserving both.

C. Computing Relations

The "heavy lifting" of generating a structural description is separating the "what" from the "where", a function made possible by the dynamic binding of features into parts-based sets. Once this separation is accomplished, using the "what" information to compute geon shape attributes and the "where" information to compute relations (such as relative location, relative size and relative orientation) is relatively straightforward (see Hummel & Biederman, 1992; Hummel, 2001). The latter can be accomplished by simple comparitor circuits composed of neural-like units. For example, consider a comparitor for *relative location in the vertical dimension* that receives a signal indicating *vertical location 3* (i.e., input from a unit that responds whenever a geon is located at coordinate 3 on the vertical axis) at time $t = 1$, and a signal indicating *vertical location 5* as input at $t = 2$. These inputs indicate that whatever fired at time 1 is below (lower in the visual field than) whatever fired at time 2, and serve as a natural basis for computing that relation (e.g., by serving as inputs to a matrix of units that effectively perform subtraction). The next time it gets *vertical location 3* as input (say, at $t = 3$), the comparitor need only activate a unit for *below* (i.e., the result of the "subtraction" 3 - 5) as output; and the next time it gets *vertical location 5* as input (at $t = 4$), it need only activate a unit for *above* as output (the result of the subtraction 5 - 3). As long as the units representing the shape of the geon at location 3 are firing in synchrony with the representation of location 3 (which, in JIM, they will be), activating *below* in synchrony with *location 3* not only specifies that the geon at location 3 is below something, but it also binds the representation of the relational role to the representation of the geon's shape. That is, the same simple operations both calculate the relations and, as a natural side effect, bind them to the appropriate geons (see Hummel & Biederman, 1992).

Analogous operations can be used to compute the spatial relations between whole objects for the purposes of scene perception and comprehension. The primary difference is that the arguments of the relations (and thus the inputs to the relation-computing machinery) are descriptions of complete objects rather than object parts.

D. TOKEN FORMATION

The geons and relations comprising an object fire in geon-based sets with separate geons firing out of synchrony with one another, so the final stages of object recognition in JIM are performed by two layers of units (see Figure 3). In the first of these layers, which happens to be the 6th layer of units in JIM, units learn to respond in a localist fashion (i.e., with one unit per pattern) to specific geons in specific relations. For example, a coffee mug would be represented by two such units, one that responds to curved cylinders beside and smaller than other parts (the handle), and one that responds to straight vertical cylinders beside and larger than other parts (the body of the mug). (Simply coding the relations as "beside" and "larger" or "smaller" is admittedly simplified. In particular, specification of their connectedness relations is conspicuously absent.) In the second recognition layer (JIM's 7th layer), units integrate their inputs over time (i.e., in order to pool the outputs of multiple Layer 6 units, which are all firing out of synchrony with one another) to learn to respond to specific combinations of part-relation conjunctions—i.e., to whole objects. Like the Layer 6 units, the units in Layer 7 respond to their preferred patterns in a localist fashion, with one unit for each object.

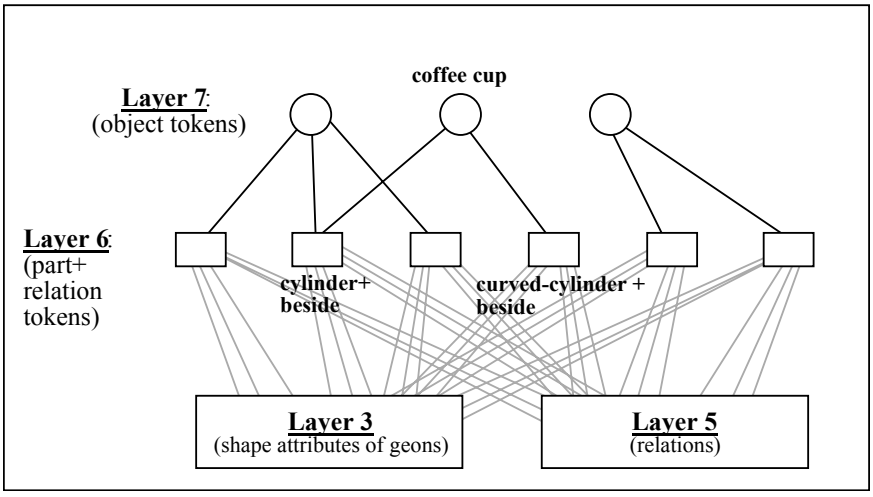

Figure 3. The upper two layers of the JIM model. Units in Layer 6 respond to specific conjunctions of part attributes and relations. Units in Layer 7 integrate their inputs over time to respond to collections of Layer 6 units, i.e., to complete objects.

The localist nature of these representations is no accident. In addition to making the representation unambiguous with regard to which part(s) and object(s) are present in the image (see Page, 2000, for a discussion of the utility of localist representations), the localist nature of these units also allows them to act as explicit *tokens* of (or "pointers to") the things they represent. As elaborated in the next section, tokens play an essential role in relational reasoning (Hummel & Holyoak, 2003), including scene comprehension.

1. A Case Study in the Importance of Tokens

In the context of object recognition, the importance of tokens for specific part-relation conjunctions—and conjunctions of part-relations conjunctions, and conjunctions of *those* conjunctions—is illustrated by one of the most important and severe limitations of Hummel and Biederman's 1992 version of JIM. Recall that each unit in JIM's Layer 6 represents a conjunction of (a) the shape attributes defining a single geon (e.g., with units for *curved cross section*, *curved major axis* and *parallel sides* together representing a curved cylinder) and (b) that geon's relations to the other geon(s) in the object. What is not specified by this representation is the other geon(s) to which the relations refer: The curved cylinder is beside and

smaller than *something*, but what? When an object has only two parts, as in the case of the mug, the resulting representation is unambiguous. But when an object has more than two parts, the representation can quickly become ambiguous.

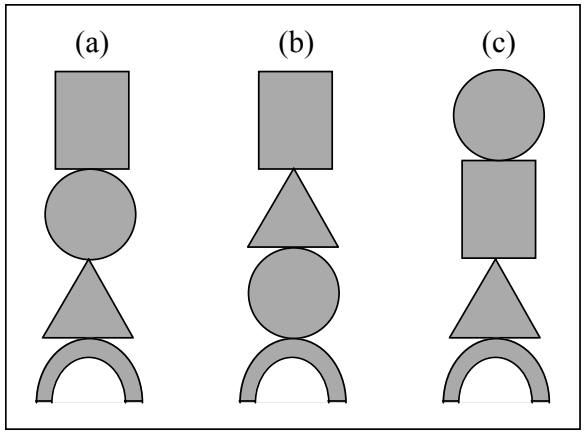

Figure 4. Three "totem pole" objects. Hummel and Biederman's (1992) JIM model of object recognition predicts that (a) and (b) should be more confusable than either is with (c).

Consider, for example, the "totem pole" objects in Figure 4, along with the schematic depictions of JIM's representation of them in its Layers 6 and 7 in Figure 5A. To JIM, objects (a) and (b) are identical: They both consist of a crescent that is below something, a circle and a triangle that are both above and below something, and a square that is above something. Hummel and Biederman (1992) noted that this limitation constitutes a novel prediction: JIM predicts that (a) and (b) ought to be more confusable with one another than either is with (c), in which the rectangle changes places with the circle (changing which relations are bound to which parts). Logan (1994) tested and falsified this and related predictions, thereby falsifying Hummel & Biederman's (1992) original version of JIM. (see Hummel & Stankiewicz, 1996b, Hummel, 2001, and Stankiewicz, Hummel & Cooper, 1998, for reviews of empirical phenomena that reveal other limitations of the original formulation of JIM.).

Logan's (1994) falsification of Hummel and Biederman's (1992) JIM suggests that the architecture of JIM's Layers 6 and 7 does not adequately capture an object's structure. But as pointed out by Logan, they do not falsify its general approach to structural description, based on dynamic binding of independent shape attributes and relations. Indeed, there is substantial empirical support for the general approach (see Hummel, 1994, 2000, 2001, Hummel & Stankiewicz, 1996a, 1996b; Kurbat, 1994; Logan, 1994; Saiki & Hummel, 1998a, 1998b; Stankiewicz & Hummel, 2002; Stankiewicz et al., 1998). Instead, Logan's findings underscore the hierarchical nature of object shape, and the importance of representing each level of this hierarchy explicitly—a task for which JIM's 6th and 7th layers are inadequate.

To elaborate, consider the augmented version of JIM's upper layers depicted in Figure 5B. This representation codes the relations between *pairs* of geons: Rather than coding a geon's relations to all other geons in the same unit (as in JIM's 6th layer), units in Layer 6a code a geon's shape attributes and its relations to *one* other geon. Although not shown in the figure, these units may code multiple relations—e.g., relative location, relative size, etc.—as long as all those relations refer to the same geon. Units in Layer 6b code for pairs of units in Layer 6a (e.g. *triangle+below* and *circle+above*) and represent complete propositions expressing the relations between two (and only two) geons (e.g., *above* (circle, triangle), or *above-and-smaller* (circle, triangle)). Layer 7 takes its input from Layer 6b, and responds to complete objects (like Layer 7 in JIM). The resulting augmented representation, like Logan's (1994) subjects but unlike the original JIM, easily distinguishes objects (a) and (b). Importantly, its ability to do so stems from the fact that it forms explicit tokens for elements at every level of part-relation hierarchy.

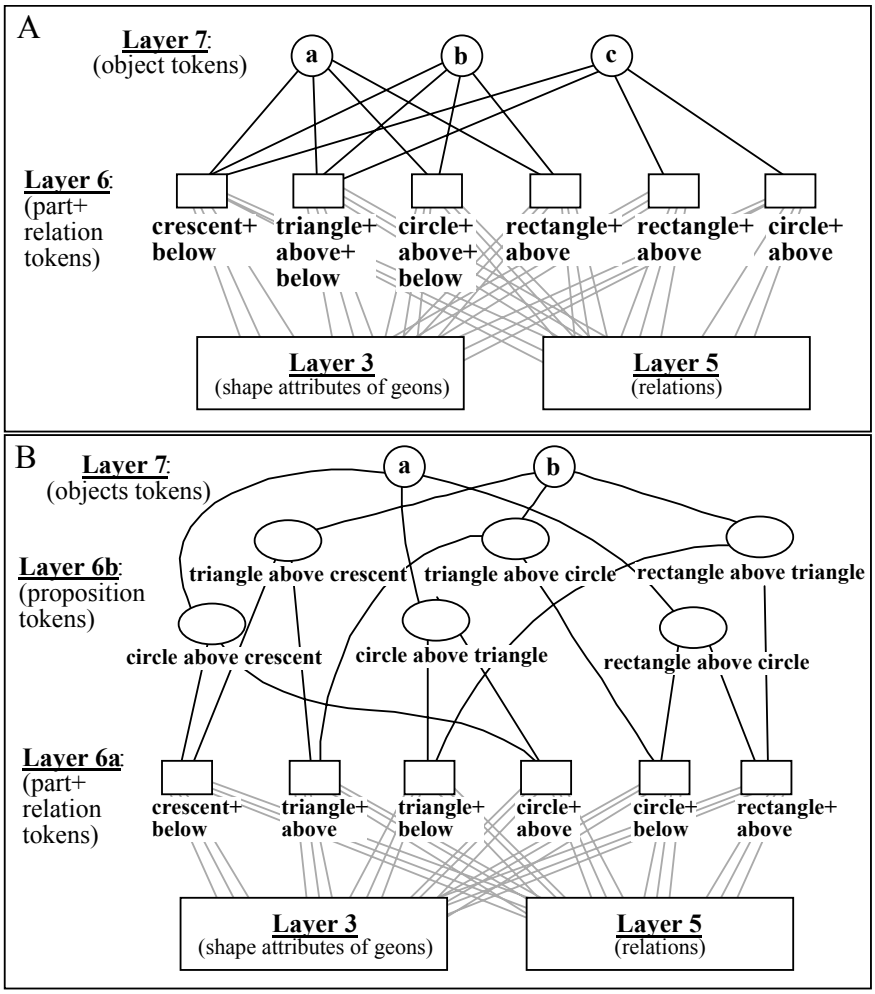

Figure 5. Structural description of objects without (A) and with (B) proper token formation. JIM's (Hummel & Biederman, 1992) representations are like those in (A), but the representations in (B) are better suited to general object recognition.

One final extension of the Hummel & Biederman (1992) representation is worth noting. In the original JIM, the outputs of units representing a geon's shape and relations to other geons fed directly into units in Layer 6 representing part-relation conjunctions (rectangles in Figures 5A and 5B). If, instead, we allow the outputs of units representing aspects of a geon's shape to feed into one set of units (circles in Layer 5b of Figure 6A), and aspects of its relation(s) to other geons feed into a *separate* set of units (triangles in Layer 5b of Figure 6A) before shape and relation information is combined in Layer 6a, then we get the situation depicted in Figure 6A: a geon's shape is represented by one token in Layer 5b and its relations to one other geon are represented by a separate token.

The resulting representation is isomorphic with the representational scheme Hummel and Holyoak's (1997, 2003) LISA model uses to code propositions for relational reasoning (Figure 6B): Although Figure 6A illustrates a hierarchy of tokens for representing structural descriptions of object shape, the very same hierarchy can be used to represent and reason about propositions describing *any* relational structure. For example, Figure 6B illustrates how LISA uses this hierarchy to represent the proposition "the café sells coffee". At the bottom of the hierarchy, relational roles (e.g., *seller* and *sold*) and their arguments (café and coffee) are represented as patterns of activation distributed over units that code their semantic content. Localist tokens for individual roles (triangles in Figure 6) share bidirectional excitatory connections with the semantic features of those roles, and tokens for objects (circles in Layer 5b of Figure 6) share connections with the semantic features of those objects. *Sub-proposition* (*SP*) units serve as

tokens for specific roll-filler bindings (such as the binding of café to the *seller* role; rectangles in Figure 6), and share bidirectional excitatory connections with the corresponding role and filler units. SPs are connected to *P* units, which serve as tokens for complete propositions. Not shown in Figure 6B are units that code for *groups* of related propositions (analogous to the object units in Layer 7 of Figure 6A), which share excitatory connections with the corresponding P units. The important point is that the representation depicted in Figure 6A—which is a straightforward extension of the representations JIM generates in response to an object's image—has been shown to serve a basis for relational reasoning, and is thus a suitable starting point for a model of scene comprehension.

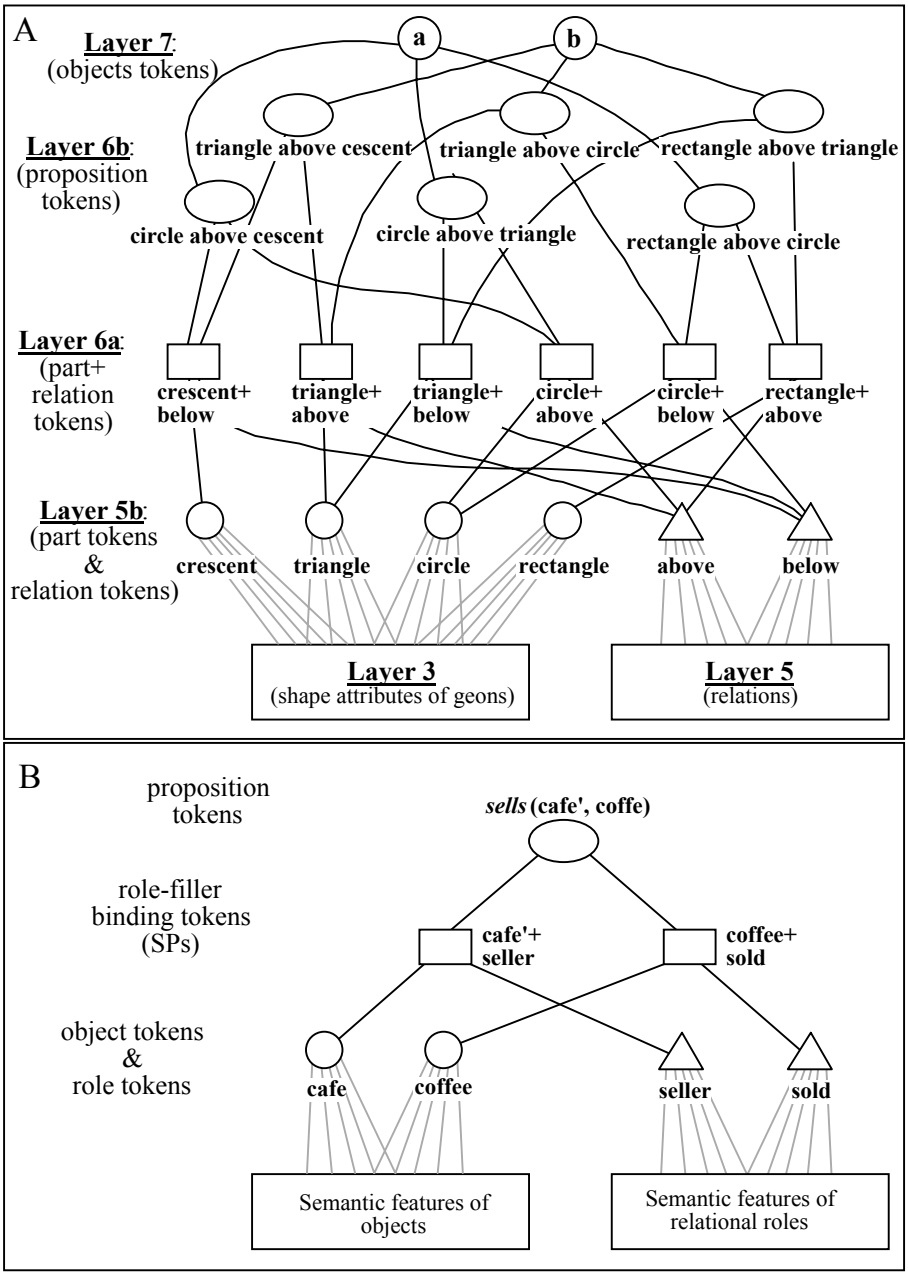

Figure 6. A representational scheme for recognizing multiple objects simultaneously (A) would serve as a basis for constructing abstract relational descriptions of visual scenes (B).

V. Toward a Model of Scene Comprehension

The representational scheme illustrated in Figure 6 is a "missing link" between models of perception on the one hand and models of cognition on the other: Given a retinotopically-mapped representation of a visual scene as input, it is possible to generate the representation in Figure 6 using visual routines such as those embodied in JIM (both the original 1992 version, and more recent versions; Hummel & Stankiewicz, 1996a, 1998; Hummel, 2001); the resulting representation serves as a natural basis for relational reasoning (Hummel & Holyoak, 1997; 2003). It is worthwhile to note that the construction of properly tokenized, abstract, structural representations of the environment is decidedly non-trivial. A number of difficult obstacles are yet to be overcome in solving this problem. However, JIM serves as an excellent starting point from which to address the construction of more complex and abstract structural descriptions. We next consider, in a bit more detail, how a representational scheme such as the one depicted in Figure 6 can serve as the basis for recognizing visual scenes and reasoning about their properties.

Object recognition is undoubtedly useful for scene recognition (e.g., seeing coffee makers, cash registers, tables and chairs suggests that one may be in a café), but it is neither necessary nor sufficient. Biederman (1987) demonstrated that it is possible to recognize otherwise ambiguous objects based strictly on their locations in a scene: In such cases, scene recognition precedes and supports object recognition, rather than the reverse. That object recognition is not sufficient for scene recognition is illustrated by the fact that identifying a collection of objects as tables, chairs, coffee makers, etc., is not sufficient to distinguish a café from a café supply warehouse. In order to distinguish a café from a café supply warehouse, it is necessary to understand the relations among the objects. In addition, a number of results suggest that top-down influences on scene perception are substantial. Change detection (Werner & Thies, 2000), eye movements (Hollingworth & Henderson, 2000; Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978), and object detection (Moores, Laiti, & Chelazzi, 2003) are all influenced by semantic knowledge about visual scenes and objects.

Scenes, composed of objects in particular relations, are thus analogous to objects, composed of parts in particular relations (Biederman, 1987). However, there is an important difference: The spatial relations among the parts of an object are fairly tightly constrained. For example, the handles of various mugs may vary in their exact size, shape and location, but they will almost always be attached to the sides of the mugs' bodies. By contrast, the spatial relations among the objects in a scene are free to vary widely. What is it about the spatial relations among the objects in a scene that determines whether they form a café or a café supply warehouse?

A. The Functional Relations Hypothesis

One intuitive hypothesis is that *functional* relations, rather than specific *spatial* relations, are what distinguish one category of scenes from another: A scene is a café if and only if the objects in the scene are arranged in a way that supports making, buying and drinking coffee. (This hypothesis is closely related to Gibson's, 1950, 1979, notion of *affordances*: a scene is a café if and only if it *affords* these functions.) Although this idea is intuitive, it underscores the abstract relational nature of visual scene recognition. It implies that, not only is it not good enough to be able to recognize the objects in a scene, it is also not good enough to know where the objects are located, or even to know where they are located relative to one another. Instead, it is necessary to be able to compute, from their spatial relations, their functional relations. Doing so requires knowledge of things such as goals and ways to satisfy those goals. It is in this sense that scene recognition is a task at the interface of perception and cognition.

One implication of the functional relations hypothesis is that the meaning (semantics) of a scene is more than the sum of the semantic properties of its constituent objects: Scenes are defined by semantics that reflect the functional relations among their constituent objects. In the limit, the objects themselves can become nearly irrelevant: Along with some basis for making coffee and collecting money, a collection of appropriately arranged rocks or logs could form a perfectly fine café. Or, as illustrated by

Biederman (1987), an array of appropriately arranged abstract shapes can form a perfectly fine office, thereby disambiguating the identities of its constituent objects.

Another implication of the functional relations hypothesis is that *functional groups*—groups of objects in spatial relations that satisfy various functional relations—form an explicit intermediate level of representation between objects and complete scenes. A single scene will typically contain multiple functional groups. In the case of a café, for example, there would be groups for sitting and drinking coffee, groups for preparing coffee, and groups for purchasing coffee. Different scenes may share many of the same functional groups. Or example, a restaurant will share many functional groups with a café, and a wood shop may share many groups with a metal shop or a laboratory. A key prediction of the functional relations hypothesis is that scenes and scene categories should be confusable to the extent that they share functional groups, even controlling for the absolute number of shared objects and spatial relations.

1. Functional Groups in the Recognition of Familiar Scenes

Biederman (1987) proposed that scenes may be recognized on the basis of *geon clusters*—collections of abstract, coarsely-coded shapes corresponding to complete objects, but perceptually coded, at least initially, as simple geons, in particular relations. For example, a brick-like shape with a roughly vertical slab-like shape behind it would form a geon cluster for a desk and chair. The presence of such a cluster could provide a useful basis for recognizing the scene as an office.

We hypothesize that, based on the statistics of the arrangements of objects in familiar scenes, geon clusters are likely to correspond, not to whole scenes, but to functional groups within scenes. For example, across the population of various kitchens, refrigerators are unlikely to appear in any particular location relative to sinks, since a refrigerator and a sink are unlikely to form a functional group. By contrast, sinks, counters and dish drains do form a functional group (e.g., "dish washing station") and are therefore likely to appear in regular spatial relations to one another across scenes (e.g., *beside* (counter, sink), *beside* (dish drain, sink) and *on-top* (dish drain, counter)). If so, then experience with a few typical kitchens could provide ample opportunity to learn a geon cluster for the functional group "dish washing station". Similar statistical regularities would provide opportunities to learn geon clusters for functional groupings of cutting boards and knives, pots and stoves, etc. Once learned, such geon clusters could provide a rapid route to the recognition of scenes in familiar arrangements, even before the objects within the clusters/groups are visually recognized (as observed by Biederman).

2. Functional Groups in Scene Comprehension and Novel Scene Recognition

Geon clusters for familiar functional groups may provide a fast route for the recognition of scenes with objects in familiar configurations, and functional groups as a more general construct—i.e., as groups of objects in spatial relations that afford particular functions—may also provide a basis for scene recognition and comprehension even in the absence of familiar geon clusters. An extreme example was given in Figure 1: The spatial relations among of the hammer, boxes and wine glasses is unlikely to activate any familiar geon cluster, but based on one's knowledge of functional relations (such as support) it is straightforward to comprehend what the scene "means". A less extreme example would be a kitchen in which the table is adorned with beakers, test tubes, and a scale. This scene contains an unfamiliar geon cluster (the table and chairs with scientific instruments) that is nonetheless interpretable as "a kitchen table that is (probably temporarily) being used as some kind of laboratory". As in the hammer and wine glass example, it is the functional relations among the table and the instruments, rather than the familiarity of the particular geon cluster, that suggests the "kitchen laboratory" interpretation.

B. From Spatial Relations to Functional Inferences: The Cognitive Side of the Perceptual-Cognitive Interface

Central to the functional relations hypothesis is the idea that objects in spatial relations activate representations of the objects' functional relations. This hypothesis raises the question of how we know which spatial relations afford which functional relations. Although the question of which functional affordances/relations are learned vs. innate is well beyond the scope of this chapter, it seems uncontroversial that at least some functional affordances must be learned. For example, most people would probably agree that the fact that the shifter in a car affords changing the ratio of the rate of revolution of the engine to the rate of revolution of the wheels is most likely learned, rather than innate.

Everyday experience provides ample opportunity to observe patterns of covariation between spatial relations and functional relations, so it is tempting to assume that learning to map from one to the other is a simple matter of learning to associate them. And to a first approximation, this is probably correct. However, the learning is complicated by the fact that the "associations" in question are not simple associations between objects or features, but rather between abstract relations, which themselves can take variable arguments. As a result traditional connectionist learning algorithms (e.g., Rumelhart et al., 1986; O'Reilly and Rudy, 2002) are fundamentally ill-suited to the task: These architectures cannot represent relational structures (see Hummel and Holyoak, 1997, 2003; Marcus, 1998), so they are unable to learn associations between them; a model cannot learn to associate that which it cannot represent.

By contrast, Hummel and Holyoak's (1997, 2003) LISA model of relational learning and reasoning provides an ideal platform to simulate this kind of learning. As noted previously, LISA operates on representations of relations and their arguments (i.e., propositions) like those illustrated in Figure 6—representations that, at least in the domain of spatial relations, can be generated by visual routines embodied in a system such as JIM. Although LISA's operation is too complicated to describe here in detail, it is sufficient to note that LISA is able to learn abstract relations among propositions (e.g., that one relation causes or affords another), and to use its knowledge of familiar situations—both in the form of specific examples, and in the form of abstract schemas or rules—to infer new facts about analogous novel situations. For example, given a description of the spatial relations among some tables and chairs, e.g., as delivered by JIM, and given similar descriptions, along with descriptions of the functional relations among those objects in LTM, LISA can use its knowledge of the familiar situations to infer the (as yet unstated) functional relations among the tables and chairs in the new situation (see Hummel & Holyoak, 2003).

C. Open Questions and Future Directions

Many problems remain to be solved before the general ideas presented here can be turned into a working model of scene recognition and comprehension. We shall briefly mention only a few of the thorniest.

Some of the most difficult problems surround the hierarchical nature of visual scenes: Scenes are composed of functional groups, which are composed of objects in specific relations, and objects consist of parts in specific relations. The image segmentation routines described by Hummel and Biederman (1992) are designed to take an image of a single object and decompose that object into its constituent parts. The model does not address the problem of segmenting an object from a complex background (the familiar figure-ground segregation problem), or the related problem of knowing which object parts in a multi-object display belong to the same object and which belong to different objects (but see Saiki & Hummel, 1996, 1998a, 1998b, for some progress in this direction). A related problem is that, in a multi-object display, the number of separate object parts will quickly exceed the capacity of visual working memory (which capacity is approximately four discrete units, e.g., objects or object parts; Luck & Vogel, 1998). For example, a scene with four objects, each with four parts, contains 16 parts, for a total of 120 non-redundant sets of inter-part relations. Clearly, it is neither possible nor desirable for the visual system to compute all sets of pair-wise relations between all parts in an object image.

In order to deal with the hierarchical nature of visual scenes, a model of scene perception will need, among other things, intelligent routines for directing attention between levels of the hierarchy, and for relating elements at one level to elements at other levels. The representational format illustrated in Figure 6 is one step in the direction of specifying how elements at different levels of the visual hierarchy are related, but it is by no means sufficient. Among other limitations, this representational scheme assumes, at least tacitly, that every object in a scene is represented in terms of its complete parts structure. However, to the extent that Biederman's (1987) idea of geon clusters is correct, "objects" in a cluster may act more like geons (small circles in Layer 5b of Figure 6A) than like complete objects (Layer 7 of Figure 6A). Similarly, there is evidence that we can recognize objects in familiar views without first decomposing them into their parts (Stankiewicz et al., 1998; Stankiewicz & Hummel, 2002; see also Hummel & Stankiewicz, 1996b; Hummel, 2001). These facts could either simplify the problem of representing scenes hierarchically by obviating the need to represent every part of every object explicitly, or they could complicate it by making it unclear at which level of the hierarchy the representation of an object qua element in a cluster should reside: Is such an object an "object" that should reside at Layer 7, or a part that should reside at Layer 5b? It seems likely that an adequate solution to the hierarchical representation problem will make the latter part of this problem ("is this an object or a part?") simply "go away". But it is difficult to know for sure until we see what that solution looks like.

Implementing a model of scene comprehension will entail solving several other more minor problems as well. And although the general framework presented here arguably raises more questions about scene recognition and comprehension than it answers, we are encouraged that it provides a framework for posing the questions at all.

VI. Conclusion

Scene recognition and comprehension provide an excellent platform for thinking about problems at the perceptual-cognitive interface, as they depend jointly on perceptual input and existing functional and relational knowledge. The problem of scene comprehension—of making the connection between representations given by the visual system and the conceptual knowledge structures that underlie relational reasoning—underscores the importance of developing models of perception that can deliver representations that are useful to the rest of cognition on the one hand, and developing models of cognition whose basic representations and operations can be grounded in the outputs of perceptual processing on the other. Cognitive science is still far from being able to connect a camera to a computer and have the computer make intelligent inferences about the objects in a scene and the actions that can be performed there. Many technical and theoretical problems must be solved before we will be able to fully automate scene comprehension in this way. But one of the most important and basic of those problems is to elucidate the nature of the perceptual-cognitive interface. In turn, one of the most important abilities at the interface of perception and cognition is the ability to tear information apart—e.g., about the identities and locations of features, parts and objects—to put it back together as needed, and to form and manipulate tokens representing the resulting visual and cognitive entities and their relations.

ACKNOWLEDGEMENTS

The authors would like to thank Irving Biederman, Steve Engel, Keith Holyoak, Zili Liu, and the members of the LISA lab and the CogFog group for valuable comments and discussion related to the issues presented in this chapter. Preparation of this chapter was supported in part by NIH NRSA F31-NS43892-01. Address correspondence to: Collin Green, Department of Psychology, University of California - Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563 (email: cbgreen@ucla.edu).

REFERENCES

Anderson, J.R. (1990). *The adaptive character of thought.* Erlbaum.

Anderson, J.R., Libiere, C., Lovett, M.C., & Reder, L.M. (1998). ACT-R: A higher-level account of processing capacity. *Behavioral & Brain Sciences, 21*, 831-832.

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

Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. Pylyshyn, (Ed.) *Computational Processes in Human Vision*. pp. 370-428. New York: Ablex.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral & Brain Sciences, 21*, 449-498.

Edelman, S., & Intrator, N. (2000). (Coarse coding of shape fragments) + (retinotopy) approximates representation of structure. *Spatial Vision, 13*, 255-264.

Edelman, S. & Intrator, N. (2002). Models of perceptual learning. In Fahle, M. (Ed.) & Poggio, T. (Ed.). *Perceptual Learning*. MIT Press: Cambridge, MA. 337-353.

Edelman, S., & Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science, 27(1)*, 73-109.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Falkenhainer, B., Forbus, K.D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1-63.

Gasser, M., & Colunga, E. (2001). Learning relational correlations. In E.M. Altmann, (Ed.) & A. Cleeremans, (Ed), *Proceedings of the 2001 Fourth International Conference on Cognitive Modeling* (pp. 91-96). Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gibson, J.J. (1950). *The perception of the visual world*. Houghton Mifflin: Oxford, England.

Gibson, J. J (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology, 23*, 222-262.

Goodale, M. A., Milner, D. A., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature, 349*, 154-156.

Henderson, J.M., Weeks, P.A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception & Performance, 25(1)*, 210-228.

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition, 7(1/2/3)*, 213-235.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Hummel, J. E. (1994). Reference frames and relations in computational models of object recognition. *Current Directions in Psychological Science, 3*, 111-116.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman (Eds.). *Cognitive Dynamics: Conceptual Change in Humans and Machines* (pp. 157 - 185). Hillsdale, NJ: Erlbaum.

Hummel, J.E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition, 8(3-5)*. 489-517.

Hummel, J. E. (2003). "Effective systematicity" in, "effective systematicity" out: A reply to Edelman & Intrator (2003). *Cognitive Science*, *27*, 327-329.

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., & Holyoak, K.J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110(2)*. 220-264.

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

Hummel, J. E., & Stankiewicz, B. J. (1996b). An architecture for rapid, hierarchical structural description. In T. Inui and J. McClelland (Eds.). *Attention and Performance XVI: Information Integration in Perception and Communication* (pp. 93-121). Cambridge, MA: MIT Press.

Kim, J.J., Pinker, S., Prince, A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science, 15*. 173-218.

Kruschke, J. K. (1992). ALCOVE: An examplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J.K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45(6)*. 812-863.

Kurbat, M.A. (1994). Structural description theories: Is RBC/JIM a general-purpose theory of human entry-level object recognition? *Perception, 23(11)*. 1339-1368.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance, 20 (5)*, 1015-1036.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception & Performance, 4(4)*, 565-572.

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.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37*, 243-282.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-437.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

Mel, B. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation, 9*, 777-804.

Mel, B., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation, 12*, 247-278.

Mishkin, M. & Ungerleider, L.G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research, 6(1)*. 57-77.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience, 6(2)*, 182-189.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*, 113-126.

Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13(1)*, 87-108.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108*, 311-345.

Page, M. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioural & Brain Sciences, 23(4)*. 443-512.

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

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

Peterson, M. A., & Gibson, B. S. (1994). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science, 5*, 253-259.

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

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

Robin, N., & Holyoak, K. J. (1994). Relational complexity and the functions of prefrontal cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 987-997). Cambridge, MA: MIT Press.

Ross, B. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629-639.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1)* (pp. 318-362). Cambridge, MA: MIT Press.

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 the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 227-251.

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

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences, 16*, 417-494.

Shiffrin, R. M., & Styvers, M. (1997). A model for recognition memory: REM--retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience, 18*, 555-586.

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science, 16*, 1-40.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217-257.

Stankiewicz, B.J. & Hummel, J.E. (2002) The role of attention in scale- and translation-invariant object recognition. *Visual Cognition, 9*, 719-739.

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.

Strong, G. W., & Whitehead, B. A. (1989). A solution to the tag-assignment problem for neural networks. *Behavioral and Brain Sciences, 12*, 381-433.

Stuss, D.T. & Benson, D.F. (1987). The frontal lobes and control of cognition and memory. In Perecman, Ellen (Ed.). *The Frontal Lobes Revisited*. New York, NY, US: The IRBN Press. 141-158.

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.

Ullman, S. & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 992-1006.

von der Malsburg, C. (1981/1994). The correlation theory of brain function (1994 reprint of a report originally published in 1981). In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks II* (pp. 95-119). Berlin: Springer.

Werner, S., & Thies, B. (2000). Is "change blindness" attenuated by domain-specific experitise? An expert-novice comparison of change detection in football images. *Visual Cognition, 7(1/2/3)*, 163-173.