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.

Please note: I am not confident this document is the absolutely final (i.e., in print) version of the paper. But it is at least close.

Two Roles for Attention in Shape Perception: A Structural Description Model of Visual Scrutiny

John E. Hummel Brian J. Stankiewicz
University of California, Los Angeles

Address correspondence to:
John E. Hummel
Department of Psychology
University of California, Los Angeles
405 Hilgard Ave.
Los Angeles, CA 90095-1563
USA

Abstract

This paper presents *MetriCat*, a model of the human capacity to recognize objects both as members of a general class (e.g., "chair") and as specific instances ("my office chair"), and of the role of visual attention in this capacity. MetriCat represents the attributes of an object's parts and their relations in a nonlinear fashion that provides a natural basis for recognition at both the class and instance levels (Stankiewicz & Hummel, 1996). Like previous structural description models (e.g., Hummel & Biederman, 1992), MetriCat represents part attributes and relations independently, dynamically binding them into structural descriptions. The resulting representation suggests two roles for visual attention in shape recognition: attention for binding and attention for signal-to-noise control. MetriCat implements both functions as special cases of a single mechanism for controlling the synchrony relations among units representing separate object parts. The model accounts for the time course of class- and instance-level classification, and makes several predictions about the relationships between attention, time, and levels of classification.

Two Roles for Attention in Shape Perception:
A Structural Description Model of Visual Scrutiny

Human object recognition is remarkably flexible. Upon viewing a novel instance of a known object class, it is usually possible to recognize the object immediately as a member of that class. For example, in a furniture store we effortlessly recognize the chairs as chairs, the tables as tables, and so forth, even if we have never seen those specific chairs or tables before. At the same time, it remains possible to tell the difference between one chair and another. Although these observations conform closely to common sense, it is challenging to understand how the visual representation of shape makes this kind of multi-level classification possible (Marr, 1980). Computational models of object recognition are divided with respect to their capacity to account for classification at different levels of abstraction. Models based on categorical structural descriptions (Biederman, 1987; Dickenson, Pentland & Rosenfeld, 1992; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a) provide a natural account of our ability to recognize objects as members of general classes (e.g., "chairs"), but do not provide a clear basis for distinguishing similar members of the same class. View-based models (Bülthoff & Edelman, 1992; Bülthoff, Edelman & Tarr, 1995; Edelman, Cutzu & Duvdevani-Bar, 1996; Poggio & Edelman, 1990; Tarr, 1995) are adept at classifying objects as individual instances, but are not well-suited to account for our ability to recognize novel instances of known object classes. A clear computational account of how both instance- and class-level recognition obtain simultaneously has yet to be developed.

A related question concerns the role of visual attention in shape perception and object recognition. A growing body of evidence suggests that we can recognize objects without attending to them (Tipper, 1985; Tipper & Driver, 1988; Treisman & DeShepper, 1996). However, this does not imply that attention plays no role in object recognition. Attention is known to play an important role in many basic visual tasks, including feature selection (see Bundeson, this issue) and visual binding (e.g., Treisman, 1982; Treisman & Gelade, 1980; but see Kanwisher & Wojculik, this volume). Using a priming paradigm, Stankiewicz, Hummel and Cooper (1997) recently showed that attention also plays a role in the visual representation of object shape. Specifically, we found that attended images visually prime both themselves and their left-right reflections, whereas ignored images prime only themselves (see also Stankiewicz & Hummel, 1997). Thus, although it does not determine *whether* an object will be recognized, attention seems to affect the qualitative *form* of the visual representations mediating recognition. This fact suggests the beginnings of an answer to the question of how we recognize objects at multiple levels of abstraction: Perhaps attention serves to "tune" the representation of object shape, alternating between coarse, categorical representations (e.g., for class-level recognition) and more precise representations (e.g., for instance-level classification). This idea is intuitive, and similar ideas have been proposed before (e.g., Biederman, 1987; Hummel & Stankiewicz, 1996a; Marr, 1980; Olshausen, Anderson & Van Essen, 1993). But to turn this intuitive idea into a falsifiable theory, it is necessary to specify both the form of the visual representations that are subjected to this attention-based "tuning", and the nature of the attentional mechanisms that perform the tuning.

This paper presents *MetriCat* (for *metric* and *categorical* properties), a model of shape recognition addressed to the question of how we recognize objects at multiple levels of abstraction (Stankiewicz & Hummel, 1996), and to the role of visual attention in this capacity. MetriCat is an extension of our previous work on structural descriptions of object shape (Hummel & Biederman's, 1992, *JIM*, and Hummel & Stankiewicz's, 1996a, *JIM.2*). Like JIM and JIM.2, MetriCat represents object shape in terms of the qualitative properties of volumetric parts (i.e., geons; Biederman, 1987), and the qualitative relations among them. Also like its predecessors, it represents geon attributes and relations *independently* of one another. These properties give MetriCat the capacity for class-level recognition enjoyed by models based on categorical structural descriptions. However, MetriCat differs from these models in that its representation of shape attributes and relations is not strictly categorical. Rather, part attributes and relations are represented in a nonlinear form that emphasizes differences across categorical boundaries without discarding all metric information within those boundaries. This property permits the model to represent and match object shape at

multiple levels of numerical specificity, and provides a natural basis for classifying objects at multiple levels of abstraction (Stankiewicz & Hummel, 1996).

This approach to the representation of object shape suggests two separate but related roles for attention in shape perception. Coding part attributes independently of their relations makes it necessary to bind them *dynamically* (i.e., actively) into part-based groups (Hummel & Biederman, 1992). Attention is known to play a central role in binding independent visual attributes (see Schneider, 1995 for a thorough review), and therefore figures prominently in the generation of structural descriptions from object images (Hummel & Biederman, 1991, 1992; Hummel & Stankiewicz, 1996a; Stankiewicz, et al., 1997; see also Logan, 1994). Like its predecessors, MetriCat requires attention to bind attributes into parts-based structural descriptions (although it implements attention more explicitly than either JIM or JIM.2). It also uses attention to generate the metrically-detailed representations that permit classification at progressively finer levels of specificity (e.g., from *chair*, to *office chair*, to *my office chair*). Both these operations—attribute binding, and the generation of detailed representations of shape—are performed by a single mechanism: namely, by enhancing the ability of object parts to inhibit one another in the competition for binding and processing resources. The model's attention mechanism thus unifies the binding and selective processing functions performed by visual attention (see Schneider, 1995). (However, we should note that the model is not intended as a general theory of visual attention. Among other things, it is not explicitly addressed to phenomena such as location-based visual search, visual neglect following brain damage, etc. Rather, our current focus is strictly on the role of attention in the representation and classification of object shape.) The resulting model makes testable predictions about the roles of time and attention in recognition at different levels of visual specificity, suggests a specific basis for intelligent search for diagnostic features for subordinate- and instance-level classification, and generates testable predictions about the relationship between levels of classification and the effects of viewpoint on recognition performance.

Theories of Object Recognition

Recent theories of object recognition fall into two general classes with respect to their ability to account for shape classification at different levels of abstraction. One class of theories assumes that objects are represented as *structural descriptions* specifying an object's parts in terms of their categorical attributes and relations to one another (Biederman, 1987; Dickenson, et al., 1992; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a). For example, a table might be represented as a horizontal slab on top of four vertical posts; a coffee mug might be represented as a squat vertical cylinder with a curved cylinder end-attached to its side (see Biederman, 1987). Categorical representations of this type have several desirable properties, including robustness to noise and variations in viewpoint, and a natural capacity to generalize over variations in object shape (see Biederman, 1987; Pinker, 1984). As a consequence, they provide a natural account of the human ability to recognize objects in novel viewpoints (Biederman & Cooper, 1991a, 1992; Biederman & Gerhardstein, 1993; Cooper, Biederman & Hummel, 1992), our ability to recognize objects in the presence of noise or occlusion, and our ability to recognize novel instances of known object classes (Biederman, 1987; Clowes, 1967). Some structural description models also account for a variety of more subtle properties of human shape perception, including the role of convex parts in shape perception (Biederman, 1987; Biederman & Cooper, 1991b; Hoffman & Richards, 1985; Tversky & Hemenway, 1985), the role of categorical relations in shape perception and object recognition (Hummel & Stankiewicz, 1996b; Saiki & Hummel, 1996), and the role of time and attention in shape perception and object recognition (Hummel & Stankiewicz, 1996a; Stankiewicz et al., 1997).

For the same reason that they support recognition at the level of object classes, categorical structural descriptions are insufficient to make some within-class distinctions. For example, a categorical structural description of a robin would be indistinguishable from a categorical structural description of a blue jay: On the basis of such a description alone, both objects would simply be recognized as birds. In such cases, people search for distinguishing features (such as the red breast

on the robin, or the blue crest on the jay) as a basis for classifying these objects at the subordinate-levels of "robin" and "blue jay" (Biederman, 1987; Biederman & Schiffrar, 1987). But while such features play a critical role in subordinate-level recognition, it is unlikely that they constitute our only basis for distinguishing structurally similar objects. Although it is difficult, we can discriminate objects that differ only in their metric properties (Bülthoff & Edelman, 1992; Edelman et al., 1996; Hummel & Stankiewicz, 1996b).

In contrast to structural description models, *view-based* models represent objects as holistic "views" specifying the 2-D coordinates of their features as they appear in specific views (Bülthoff & Edelman, 1992; Bülthoff at al., 1995; Edelman et al., 1996; Poggio & Edelman, 1990; Tarr, 1995; Vetter, Poggio & Bülthoff, 1994). This approach differs from the structural description approach in that (a) there is no explicit decomposition of an object into its parts, and (b) an object's features are represented in terms of their numerical coordinates in the view, rather than their categorical (e.g., "above/below") relations to one another. This kind of representation is metrically precise and information-rich, making it suitable for distinguishing structurally similar objects (Bülthoff et al., 1995; Tarr, 1995). There is substantial support for the idea that some instance-level classification tasks—most notably face recognition—are mediated by this type of holistic representation of object shape (e.g., Farah, 1992; Tanaka & Farah, 1993). In addition, there is evidence for the role of view-like representations in our capacity to recognize objects without attending to them (Stankiewicz et al., 1997), although the views implicated in automatic recognition may differ substantially from the metrically precise, information-rich representations postulated in current view-based models (see Hummel & Stankiewicz, 1996a).

If the view-based approach provides a general account of instance-level classification (i.e., beyond the case of faces and similar objects), then our capacity for multi-level classification may simply reflect the simultaneous operation of both views (for instance recognition) and structural descriptions (for class recognition). However, not all subordinate- or instance-level recognition is performed on the basis of holistic views. For example, Biederman and Schiffrar (1987) showed that categorical features play the central role in chick-sexing[1], a decidedly subordinate-level classification task, and Hummel and Stankiewicz (1996b) report evidence against the role of holistic views in the recognition of structurally-similar objects. Thus, although holistic views likely play an important role in face recognition, and although diagnostic categorical features likely play an important role in subordinate- and instance-level recognition, neither proposal provides a complete account of our ability to recognize objects at multiple levels of abstraction. As such, we believe it is worth considering alternative explanations for this capacity.

A Unified Model of Instance- and Class-Level Recognition

As noted previously, categorical representations (as postulated in structural description theories) naturally support recognition at the entry- or class-level, while metrically-rich representations (as postulated in view-based theories) naturally support instance- and subordinate-level recognition. MetriCat is based on a single representation that captures both these properties. The basis of MetriCat is a simple extension of the representations postulated in categorical structural description models. A categorical representation entails a nonlinear mapping from a stimulus domain to the perceptual representation of that domain. For example, consider an image containing two figures, A and B. The relation *Above(A, B)* is strictly categorical if it evaluates to *true* (or 1) for all locations of A and B such that Y_A (the location of A on the vertical axis, Y) is greater than Y_B, and to *false* (or 0) for all locations such that $Y_A \leq Y_B$. In this case, the mapping from the stimulus domain (coordinates in the image) to the relation is a step function (i.e., Above(A, B) is 1 for all $(Y_A - Y_B) > 0$, and 0 for all $(Y_A - Y_B) \leq 0$). A step function is strictly categorical in the sense that its value (0 or 1) changes only at the categorical boundary (i.e., the derivative is infinite at the categorical boundary and zero everywhere else), discarding all information on either side of the boundary. However, a step is a special case of a more general class of nonlinear

[1]Chick-sexing is the task of deciding, on the basis of the appearance of the genetalia, whether a given chick is male or female.

functions whose derivative is a maximum over some point (such as a categorical boundary), and reduced elsewhere. A common example is the logistic function:

$$y = \frac{1}{1 + e^{-x}} . \quad (1)$$

Like a step function, y changes fastest over a single point in x (specifically, $x = 0$), but unlike a step, it continues to change (although progressively less) as x deviates further from zero. This "softer" nonlinearity is meaningfully categorical in that, like a step function, it emphasizes differences across categorical boundaries (e.g., from Above to Below). (Indeed, in the literature of categorical perception, this kind of function is the behavioral manifestation of categorical perception; see, e.g., Foster & Ferraro, 1989). But the logistic function has the advantage that it does not discard all information on either side of the boundary.

We argue that extant structural description models are insufficient for instance-level recognition largely because they are based on categorical properties defined by step functions. MetriCat uses logistic (rather than step) nonlinearities to represent categorical properties (such as whether one part is above or below another, and whether a geon's major axis is straight or curved). In other respects, MetriCat is like its predecessors, JIM and JIM.2: It represents part attributes and relations explicitly and independently, and binds those properties into relational structures dynamically; that is, the representations in MetriCat are structured rather than holistic (see Hummel & Biederman, 1992). The resulting representations have all the advantages of a categorical structural description, but also permit fine metric discriminations when necessary. In combination with an appropriate set of routines for matching these representations to object memory, the result is unified account of both class- and instance-level recognition. Moreover, this approach to representing object shape suggests a unifying account of two seemingly different functions of visual attention. In MetriCat, a single mechanism controls attention for binding (e.g., as discussed by Treisman and others) and attention for signal-to-noise control (e.g., as discussed by LaBerge and others).

Assumptions and Relation to Prior Work

The focus of the MetriCat model is the representation of object shape and the operations that match those representations to memory. (Specifically, it is an extension of the upper layers of JIM and JIM.2, which are responsible for shape representation and object classification; the lower layers of these models are responsible for image segmentation, etc.) MetriCat takes a description of an object's parts and their spatial relations as input. Rules for segmenting images into part-based groups, and for recovering the parts' attributes and spatial relations are described elsewhere (Dickenson et al., 1992; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a), so for the purposes of the current model, we shall simply assume that these operations have taken place.

Overview

As input, MetriCat takes a numerical representation of a geon's shape attributes and relations to other geons (Figure 1, "Input"). One such pattern is given for each geon in an object. In the model's first layer (Figure 1, "Layer 1"), separate collections of units respond to: (1) the parallelism of a geon's sides, (2) the curvature of its major axis, (3) the curvature of its cross section, (4) its aspect ratio, the (5) pointedness of its major axis (i.e., is the major axis pointed, as in a cone, or truncated, as in a cylinder or funnel), (6) whether the geon is above or below any other geons, and (7) whether it is beside or centered on any other geons. (This is an admittedly simplified set of attributes. Among other potentially important properties, we are omitting the geons' connectedness relations. For a richer set of geon and relation descriptors, see Biederman, 1987.) Shape attributes and relations are represented independently, in the sense that separate collections of units represent separate attributes. Units are bound into geon-based sets by synchrony of firing (Hummel & Biederman, 1992): If two units "fire" (i.e., become active) in synchrony with one another, then they are treated as properties of the same geon; if they fire out of synchrony, then they are treated as belonging to separate geons. Synchrony and, more importantly, *asynchrony* are established by

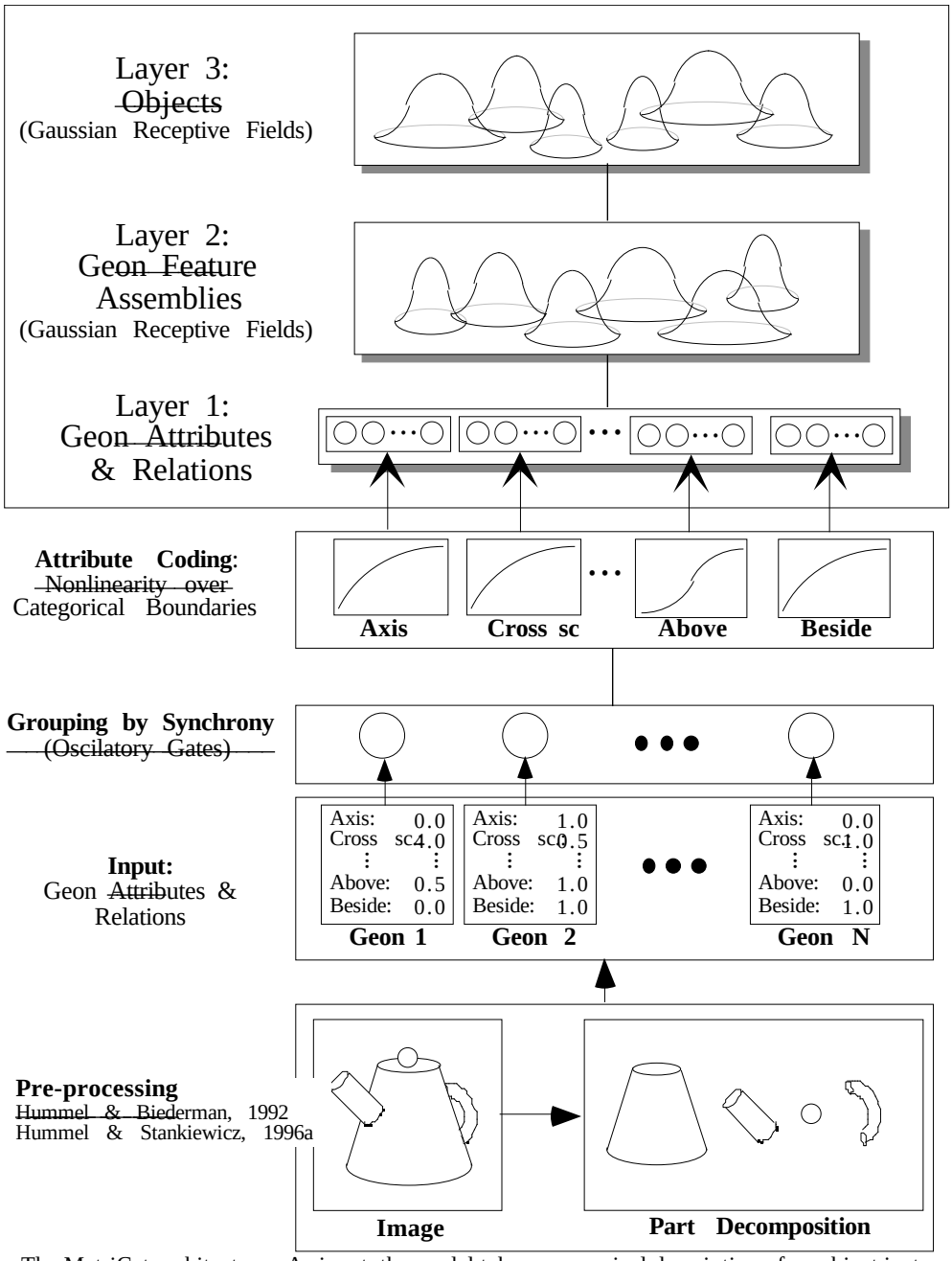

Figure 1. The MetriCat architecture. As input, the model takes a numerical description of an object in terms of its constituent geons and their interrelations (Input). Geon attributes grouped by synchrony induced by the action of oscillatory gates (Grouping by Synchrony). Each geon is characterized by five shape attributes, parallelism of main axis (Ax.Paral.), curvature of main axis (Ax.Cur.), curvature of cross section (XSn.Cur.), aspect ratio (Aspect), and pointedness of main axis (Ax.Point), and two spatial relations, above/below (Above) and centered/beside (Beside). The numerical value of each attribute is passed through a logistic nonlinearity (Attribute Code), the result of which is represented on a bank of 50 units in the model's first layer (Layer 1). Patterns of activation in Layer 1 activate Geon Feature Assembly (GFA) units with Gaussian receptive fields in Layer 2. Patterns of activation on GFA units activate object units with Gaussian receptive fields (Layer 3). GFA and object unit have Gaussian receptive fields with different standard deviations.

means of oscillatory gates associated with the geons: The gates interact so as to "open," sending the properties of the corresponding geon to the model's first layer, one at a time. There is evidence for synchrony-based binding in biological visual systems (see König & Engel, 1995, for a thorough review).

In broad strokes, the model works as follows. When an object is presented for recognition, it's geons begin to fire out of synchrony, generating separate patterns of activation on the attribute and relation units in Layer 1. In the best case, one geon will fire at a time, but as described shortly, the geons' ability to fire cleanly out of synchrony is a function of how much "attention" is directed to the object. The model's second layer (Layer 2) consists of *Geon Feature Assembly* units (*GFA* units; Hummel & Biederman, 1992) that take their inputs from the attribute and relation units, thereby responding to particular geons in particular relations to other geons. The model's third layer (Layer 3) consists of object units that take their inputs from the GFA units. Object units sum their inputs over time, "reassembling" a collection of GFAs (geons in particular relations) into a representation of a whole object.

GFA and object units (*classifier* units) have Gaussian receptive fields (RFs) in their input spaces, where the mean of the Gaussian determines a unit's preferred input pattern, **p**, and the standard deviation, σ, determines the unit's tolerance for deviations from that pattern (see Poggio & Girosi, 1990). GFA units have RFs in the space of geon attributes and relations, and object units have RFs in the space of GFAs. Classifier units with wide RFs (i.e., large σ) respond to a wider range of inputs than units with narrow RFs; that is, wide units are better able to tolerate to deviations from their preferred patterns than are narrow units. This variable-tolerance encoding has several important implications. First, classifier units with wide RFs respond to whole classes of objects, whereas units with narrow RFs tend to respond selectively to individual instances (Stankiewicz & Hummel, 1996). Second, wide (class-level) units are able to respond earlier in processing than narrow (instance-level units) units (as elaborated shortly). As a result, MetriCat predicts that objects will be recognized at the class-level before they are recognized as individual instances. And third, wide units are more robust to noise in their input vectors, and are therefore better able to respond when an object is not fully attended: MetriCat predicts that greater attention (e.g., visual scrutiny) is required to recognize objects at the instance level. Each of these properties is illustrated and elaborated in the *Simulations* section.

Representation of Attributes and Relations

Each attribute (or relation) is represented by a bank of units that coarsely code the numerical value of that attribute (Figure 2). The coded value for any attribute is obtained by passing the attribute's raw numerical value through a logistic function (Hummel & Stankiewicz, 1996b; Stankiewicz & Hummel, 1996). For example, consider the relation *Above(i, j)*, which codes the location of geon *i* relative to geon *j* on the vertical image axis, Y. $P_Y(i, j)$ is the scaled location of *i* relative to *j* on Y:

$$P_Y(i,j) = (Y_i - Y_j)/(l_{Yi} + l_{Yj}), \quad (2)$$

where Y_i and Y_j are the Y-coordinates of the centroids of *i* and *j*, respectively, and l_{Yi} and l_{Yj} are the lengths of *i* and *j* along Y. $P_Y(i, j)$ is unbounded. It will be negative whenever *i* is below *j*, and positive whenever *i* is above *j*. Scaling P by the parts' lengths makes it scale-invariant: P will not change with the absolute size of the object's image. Human object recognition is also scale-invariant in this way (Biederman & Cooper, 1992). For an image of a fixed size, $P_Y(i, j)$ changes linearly with the location of geon *i* relative to geon *j* along Y. Above(*i, j*), the above/below location of *i* relative to *j* that is used for shape classification, is computed by passing $P_Y(i, j)$ through the logistic function:

$$\text{Above}(i,j) = \frac{1}{1 + e^{-\kappa P_Y(i,j)}}, \quad (3)$$

where κ is a scaling constant. Although $P_Y(i, j)$ is unbounded, Above(*i, j*) is bounded between 0 and 1; it will be less than 0.5 whenever *i* is below *j* (i.e., when $P < 0$) and greater than 0.5 whenever *i* is above *j* (when $P > 0$). Like a categorical relation, Above() changes fastest about the categorical

boundary, P = 0 (where $Y_i = Y_j$). However, as noted previously, Above() does not discard all metric differences within categorical boundaries. κ determines the steepness of Above() about the categorical boundary P = 0; when $\kappa = \infty$, Above() is a step function that evaluates to 0 for all P < 0 and to 1 for all P > 0.

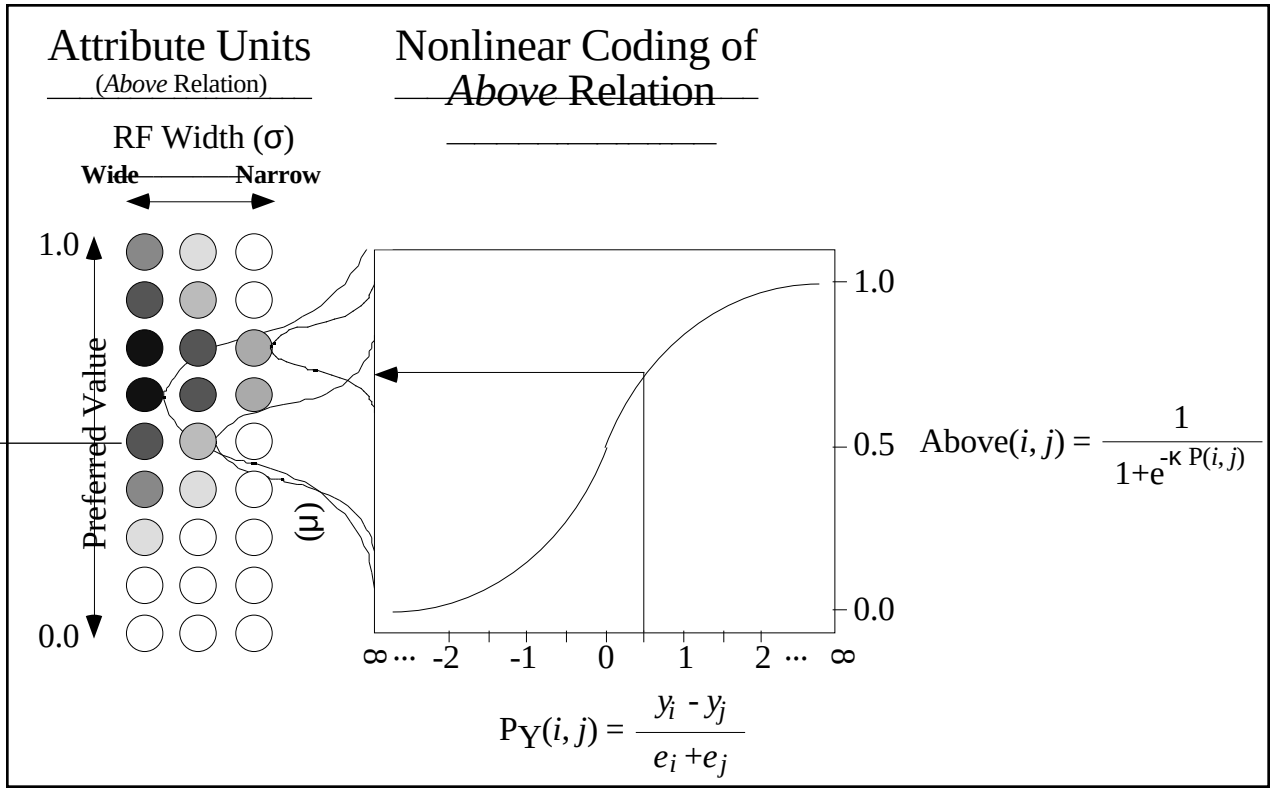

Figure 2. Detail of the representation of object attributes (in Layer 1), illustrated with the relation Above. The raw value of an attribute or relation is passed through a logistic nonlinearity, and the resulting value is coarsely coded on 50 units with Gaussian receptive fields in the space of logistic values (0..1). Units are depicted as circles, receptive fields are depicted as Gaussian curves attached to the circles, and activation as shades of gray, with more active units depicted in darker shades. The relation represented corresponds to $P_Y(i, j) = 0.5$, or "geon *i* is above geon *j* at a distance that is $^1/_2$ the sum of their heights."

The left-right relation, *Beside*, is computed in the same way as the Above relation, except that left/right distinctions are discarded, so the function effectively computes centeredness vs. off-centeredness with respect to the horizontal axis:

$$\text{Beside}(i, j) = 2\left(\frac{1}{1 + e^{-\kappa |P_X(i,j)|}} - 0.5\right) \quad (4)$$

As a result, Beside(*i*, *j*) is invariant with left-right reflection (Hummel & Biederman, 1992). It evaluates to zero when geon *i* is exactly centered on geon *j* (on the X axis), and to values closer to 1.0 as *i* moves farther to the left or right *j*.

Equations (3) and (4) express the functions for computing all attributes in MetriCat. Eq. (3) expresses the form for any attribute with two continuous regions separated by a categorical boundary. These include the Above relation (one geon can be more or less extreme in its location either above or below another geon) and aspect ratio (a geon can be more or less squat [e.g., a coin is a squat cylinder] or more or less elongated [e.g., a pipe is an elongated cylinder]). Eq. (4) expresses the form for an attribute with one continuous region and one singularity. For example, in the case of the Beside relation, *off-center* (or "beside") is a continuous region and *centered* is a singularity (a geon can be more or less distant from the midline of another, but there is only one

point where it is exactly centered). Most of the properties in MetriCat are of this latter variety. These include *parallel* sides (singularity) vs. *non-parallel* sides (continuous), *straight* cross section (singularity) vs. *curved* cross section (continuous), *pointed* major axis (singularity) vs. *truncated* major axis (continuous), and *straight* major axis (singularity) vs. *curved* major axis (continuous).

The model's input layer consists of seven banks of units (one bank for each attribute), each with 50 units. Units within a bank respond to logistic values given by (3) and (4) (depending on the attribute). Units have Gaussian receptive fields over the range of logistic values (0..1) with varying centers (means) and receptive field sizes (standard deviations). As illustrated in Figure 2, the result is a population coding for the logistic value of each attribute (Stankiewicz & Hummel, 1996).

Model Operation

The model's operation is described here only in general terms. The details of its operation (including equations and operating parameters) are given in the Appendix.

Oscillators and Visual Attention: An object's geons are induced to fire out of synchrony by means of a collection of oscillatory gates. There is one gate associated with each geon. By virtue of the interation between an excitatory unit (E_i) and an inhibitory unit (I_i), each gate, *i*, oscillates between a state of activity ($E_i > 0$), in which the gate is "open", and inacitvity ($E_i = 0$), in which the gate is "closed" (see Hummel & Stankiewicz, 1996a; von der Malsburg & Buhman, 1992). When a gate is open (i.e., $E_i, > 0$), the attributes of the corresponding geon are represented on the attribute units. If more than one gate opens at a time, then the attributes of multiple geons will be superimposed on the attribute vectors. The result is a binding error (or "superposition catastrophe"; von der Malsburg, 1981), because it is impossible to determine which attributes belong together as properties of the same geon. The gates act to prevent such errors by inhibiting one another, thereby opening (or "firing") at different times (i.e., out of synchrony with one another; Figure 3c).

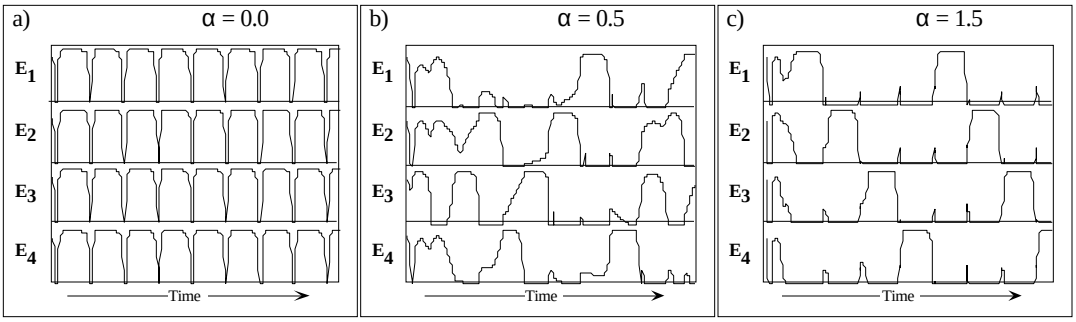

Figure 3. Simulation results illustrating the effect of attention (α) on MetriCat's ability to make separate geons fire out of synchrony. Each frame shows the activation of four gates ($E_1..E_4$) as a function of time and α. Activation ($E_i = 0..1$) is depicted on the ordinate of each curve. (a) With α = 0 (no attention) all four gates (geons) fire in synchrony with one another. (b) With α = 0.5 (limited attention), the gates fire out of synchrony, but there is substantial overlap in their activity. (c) With α = 1.5 (strong attention), the gates fire cleanly out of synchrony.

The model's ability to keep separate geons firing out of synchrony with one another is dependentr on the gates' ability to inhibit one another. In turn, the lateral inhibition between the different gates is modulated by a global parameter, α. Figure 3 illustrates the effect of α on the gates. When α is large (e.g., 1.5), the gates inhibit one another strongly and therefore fire cleanly out of synchrony (Figure 3c); when α is zero, the gates cannot inhibit one another at all, and therefore never fire out of synchrony (Figure 3a); at intermediate values of α, the gates fire somewhat out of synchrony (Figure 3b). We assume that α is proportional to the amount of attention directed to an object. Our claim is that the function of attention (at least in the context of

object recognition) is to enable separate groups of units to fire *out* of synchrony with one another. In so doing, attention serves both to control the dynamic binding of geon attributes and relations (Hummel & Biederman, 1991, 1992; Hummel & Stankiewicz, 1996a), and to control the signal-to-noise ratio of attended objects (e.g., LaBerge & Brown, 1989): Note that the pattern of asynchrony is cleaner for "strongly-attended" objects (Figure 3c) than for "weakly-attended" objects (Figure 3b). The effect of α on model performance is discussed in greater detail in the *Simulations* section.

Coarse Coding of Attribute Values: Each geon attribute is represented as a pattern of activation distributed over a bank of 50 *attribute* units (Figure 2). Attribute units have overlapping receptive fields over the range of logistic values of the corresponding attribute (Eq. (3) or (4), depending on the attribute). A complete geon is represented by a 350-dimensional (50 units X 7 attributes) vector, **a**.

Part Classification: GFA units have Gaussian RFs in the 350-dimensional space of attribute units. Each GFA unit receptive field has a mean, **p**, a vector that corresponds to the unit's preferred stimulus pattern, and a standard deviation, σ, which defines the width of the unit's receptive field (i.e, its tolerance for deviations from **p**). The simulations reported here were run with three sizes of GFA unit RFs, σ = 0.10, 0.25, and 0.35. Units with small σ respond to very specific attribute vectors (i.e., to parts with specific shapes in specific relations); those with larger σ respond to a broader range of vectors (and thus to a broader range of part shapes and relations).

Object Classification: GFA units sum their outputs over time. Object units have Gaussian RFs in the space of GFA unit outputs. As a result, object units respond to collections of GRA units, reassembling an object's parts into a representation of the object as a whole (see Hummel & Biederman, 1992). There were two sizes of object unit RFs in the simulations reported here, σ = 0.10 and 0.5.

Simulations

Simulation Procedure

We trained the model to recognize 16 objects, comprising four instances of each of four classes. As illustrated in Figure 4, we will refer to the classes as *teapots*, *lamps*, *cars*, and *nonsense objects*. Table 1 shows the attribute values of each part of each object. Each instance differed from the next nearest member of its class by 0.5 raw[2] units on one critical attribute dimension (Table 1). The teapots differed in the aspect ratio of the truncated cone (raw values 1.0, 1.5, 2.0 and 2.5 for $teapot_1$..$teapot_4$, respectively), the lamps differed in the degree of non-parallelism of the shade (raw values 1.0..2.5), the cars differed in the aspect ratio of the chassis (-2.5..-1.0), and the nonsense objects differed in the curvature of the curved cylinder (1.0..2.5).

Training took place in two steps. The model was first trained to classify the individual object parts by presenting each part one at a time, and running the model for 100 iterations. At the end of this time, we compared the attribute unit vector, **a**, to the preferred vectors, $\mathbf{p}_i$, of all existing GFA units, *i*. If the Euclidean distance between **a** and $\mathbf{p}_i$ was greater than 0.05 for all *i*, then we created three new GFA units (one at each σ) with **p** = **a**. Next, we trained the object units by presenting each object part for 100 iterations and allowing it to activate all the GFA units. After all the parts of an object had been presented, we created two object units (one at each σ) whose means (**p**) were set to the value of the GFA output vector. All the simulations reported in the next section were run by presenting a single object as input and allowing the model to run for 500 iterations. Except where noted otherwise, α was 1.5 in all simulations. The data reported are means over ten simulation runs. Simulations were run with multiple objects, but because the results were always qualitatively very similar for all objects, each set of results is reported in terms of a set of runs with one object.

---

[2] "Raw" units are units that serve as the input to (rather than the output of) the logistic functions (Eq. (3) and (4)).

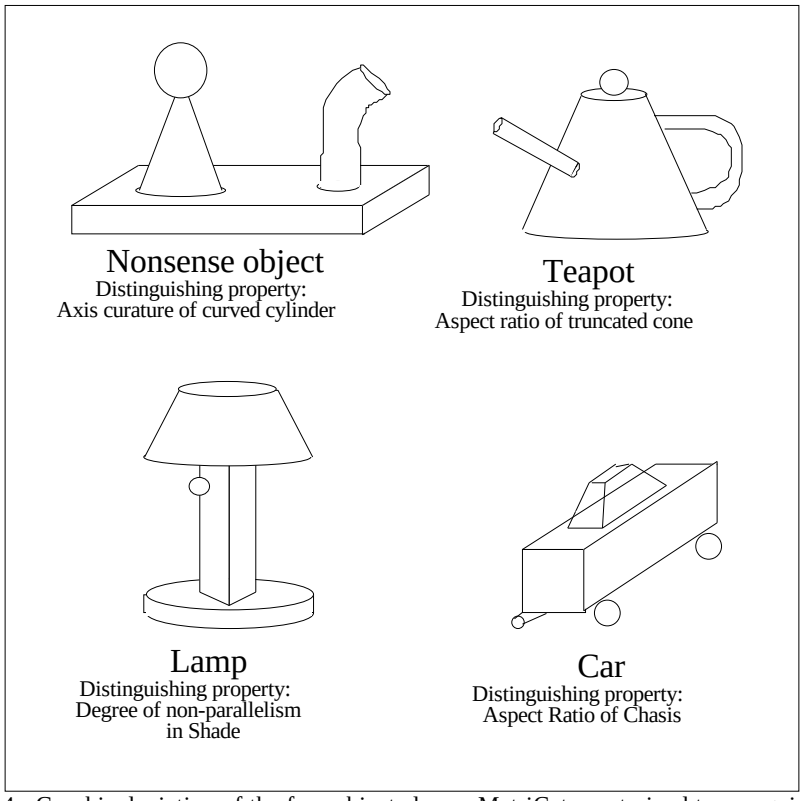

Figure 4. Graphic depiction of the four object classes MetriCat was trained to recognize.

Multi-level Classification

The most basic test of the model's capacity for multi-level classification is its response to a stimulus on which it was trained. If MetriCat can recognize objects at both the instance and class-levels, then the following results should obtain. Given a stimulus such as car_1 (the first member of the category *car*), the model should activate both the instance-level (i.e., small σ) and class-level (i.e., large σ) object units recruited during training with that stimulus. It should also activate the class-level units recruited for different instances in the same class (here, car_2, car_3, and car_4), but it should not activate the instance-level units recruited for those objects. It should activate neither the class- nor instance-level units for any other objects (i.e., $teapot_1$..$nonsense\text{-}object_4$). Figure 5 shows the model's response to the stimulus car_1. Figure 5a shows the mean responses (over ten runs) of the instance-level unit (triangles) and class-level unit (circles) recruited for car_1. Figure 5b shows the mean responses of the instance- and class-level units for car_2, the nearest other instance of the same object class. (Recall that car_1 and car_2 differ by only 0.5 raw units in the aspect ratio of one part.) Figure 5c shows the mean responses of the instance-level and class-level units for the central-most instance in the class *lamp* ($lamp_1$). These results indicate that the model correctly recognized car_1 at both the instance- and class-levels of abstraction: Both the instance- and class-level units recruited for car_1 became active; the class-level unit, but not the instance-level unit for car_2 became active; and none of the units corresponding to $lamp_1$ became active.

Figure 5 also shows the time course of recognition at the class- and instance-levels. Note that class-level units (circles) become active faster than instance-level units (triangles): Consistent with human recognition performance (Jolicoeur, Gluck & Kosslyn, 1984; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), MetriCat recognizes objects at the class-level faster than it recognizes them at the instance-level. This property is a natural consequence of MetriCat's multi-level approach to object recognition. Recall that class-level units have wider receptive fields (larger σ) than instance-level units. (This is true of both object units and GFA units.) As a result, a given deviation (distance) between a stimulus pattern (**a**, in the case of GFA units and **o** in the case of object units) and a unit's preferred stimulus pattern (**p**) has a greater impact on the response of a

Table 1. Attribute values of the 16 objects MetriCat was trained to recognize.[a]

	Object Part	Axis Parallelism	Cross-Section	Aspect Ratio	Axis Curvature	Axis Trucation	Above	Beside
Nonsense Object	Curved Cylinder	4	-4	4	**1, 1.5, 2, 2.5**	0	4	4
	Brick	0	4	-4	0	4	-4	4
	Cone	4	-4	4	0	0	4,-4	0
	Sphere	4	-4	0	0	0	4	0
Teapot	Truncated Cone	4	-4	**1, 1.5, 2, 2.5**	0	4	4,-4	4
	Curved Cylinder	0	-4	-4	4	4	0	4
	Sphere	4	-4	0	0	0	4	0
	Cylinder	0	-4	4	0	4	0	4
Lamp	Truncated Cone	**1, 1.5, 2, 2.5**	-4	4	0	0	4	0
	Vertical Brick	0	4	4	0	4	4,-4	0
	Sphere	4	-4	0	0	0	0	4
	Horizontal Brick	0	4	-4	0	4	-4	0
Car	Truncated Pyramid	4	4	**-1.0,-1.5, -2.0,-2.5**	0	4	4	0
	Brick	0	4	4	0	4	4,-4	4
	Horizontal Cylinder	0	-4	-4	0	4	-4	4
	Sphere	4	-4	0	0	0	-4	0
	Sphere	4	-4	0	0	0	-4	4

[a]The four instances of each class are distinguished by their values on one critical attribute (depicted in bold in table cells).

narrow (instance-level) unit than on the response of a wide (class-level) unit[3]. Stimuli tend to deviate more from the trained patterns (i.e., the units' preferred patterns) earlier in processing than later in processing. At the level of the GFA unit inputs (**a**), this early deviation results from the time it takes for an object's geons to get out of synchrony with one another (see Figure 3b and 3c), causing the properties of different geons to be "mixed" early in processing (as elaborated below). At the level of object unit inputs, this early deviation results both from the initially weak activation of the GFA units and from the time it takes for all an object's geons to have the opportunity to fire. Early in processing, the GFA outputs, O_i, will be zero for all GFA units, *i*, whose preferred geons have not yet fired, causing **o** to differ from the object unit's **p**. Here, too, class-level units, which have wide receptive fields, will respond more strongly to these imperfect inputs than instance-level units, with their narrow receptive fields.

Recognition of Novel Instances of Known Classes

People easily recognize novel instances of known object classes, appreciating both that the novel instance both belongs to the familiar class and that it nonetheless differs from all previous

[3]Consider two Gaussian distributions, G_1 with σ_1, and G_2, with σ_2, where $\sigma_1 > \sigma_2$ (i.e., G_1 is wider than G_2), and let G_1 and G_2 be normalized to have equal heights at their means. For any distance, $d > 0$, from their means, the height of the wide Gaussian, G1, will be greater than the height of the narrow Gaussian, G2. That is, $G_1(d) > G_2(d)$ when $d > 0$ [$G_1(d) = G_2(d)$ when $d = 0$].

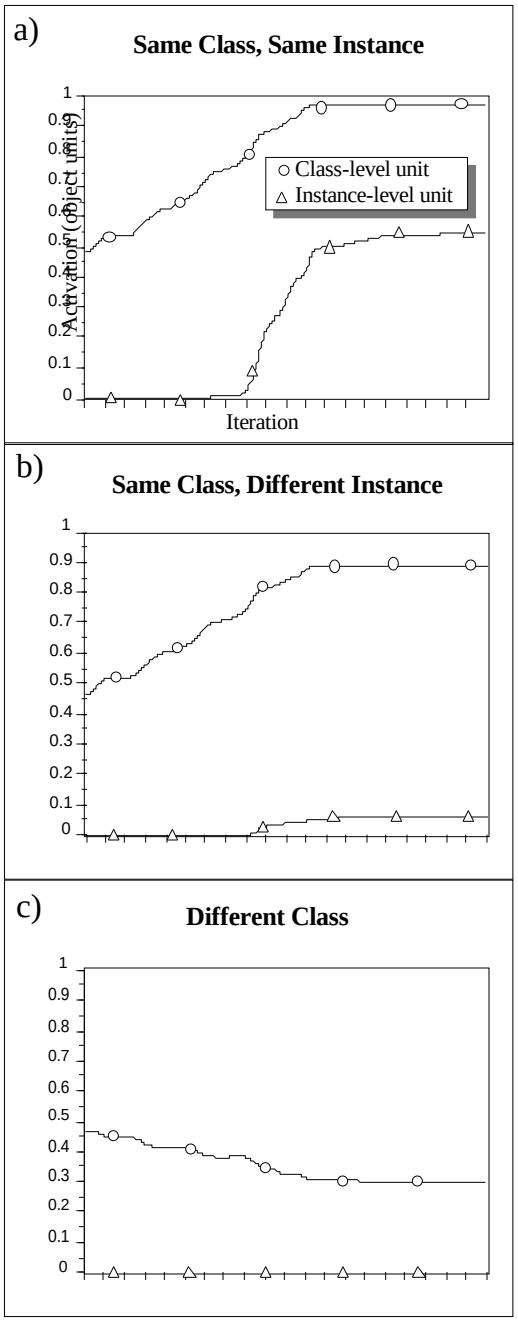

Figure 5. Response of various object units to the stimulus *car*$_1$. (a) Activation of the class (circles) and instance (triangles) object units recruited in response to *car*$_1$ as a function of time (iterations). (b) Activation of the class and instance object units recruited in response to *car*$_2$ as a function of time. (c) Activation of the class and instance object units recruited in response to *lamp*$_1$ as a function of time.

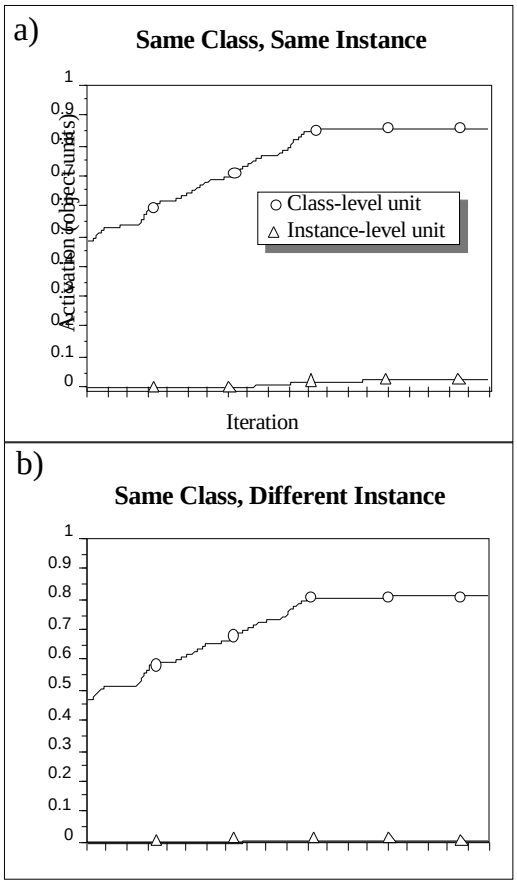

Figure 6. Response to a novel instance of a known class. (a) Activation of the class (circles) and instance (triangles) object units recruited in response to the stimulus object as a function of time (iterations during the simulation). (b) Activation of the class and instance object units recruited in response to a different member of the same class.

instances of that class. We tested MetriCat's ability to do this by presenting it with a new member of the class *nonsense object*. We created this instance by distorting nonsense object$_1$. Distortions were introduced by changing the raw values of each attribute by a random amount between -0.01 and 0.01 (20% of the distance between adjacent members of the class on the critical attribute). Figure 6 shows the model's response to the resulting novel instance. Circles indicate the mean activation of the class-level unit for nonsense-object$_1$, and triangles indicate the response of the instance-level unit for that object (which is the closest instance to the novel one). Note that the model correctly recognized the novel instance as a member of the same class, but it did not mistake it for a familiar instance (as indicated by the low activation of the instance-level unit).

Sensitivity to Noise

MetriCat is subject to intrinsic system-level noise introduced by the operations that establish asynchronous firing of an object's geons (Figure 3). This noise manifests itself as random variations in the patterns of activation generated on the attribute units. MetriCat's ability to recognize trained objects in spite of this noise shows that the model is at least somewhat noise-tolerant. To further observe the model's sensitivity to noise, we tested its ability to recognize stimuli corrupted by variable, random input noise. We generated these stimuli by starting with a trained object and varying each attribute value by a random amount between -0.01 and 0.01. The noise was re-randomized on each iteration. The resulting variable noise can either be thought of as stimulus noise, such as static in the image, or as additional system-level noise. Because the noise was re-randomized on each iteration, the mean value of each attribute (over several iterations) tends to approach the value of that attribute in the original stimulus. If MetriCat can tolerate this kind of noise, then it should recognize the stimulus at both the class and instance levels (although recognition should take longer than in the no-additional-noise case). Figure 7 shows the model's response to a noisy version of nonsense-object$_1$. As expected, the model recognized the stimulus at both the class and instance levels, although classification took longer and did not reach as high a level as in the basic simulations reported above. These results are especially interesting in comparison with the results of the simulations for *Novel Instance* simulations reported above. The only procedural difference between the current simulations and the *Novel Instances* simulations is that the current simulations varied the noise on every iteration, whereas the deviations in the *Novel Instances* simulations did not vary. MetriCat responded appropriately in both cases: It treated the constant noise as a property of the object, classifying the stimulus as a novel exemplar of a known class; and it treated the variable noise as noise, recognizing the stimulus as a familiar (albeit noisy) instance of a familiar class.

Attention in Instance-level Classification

In MetriCat, attention modulates the lateral inhibition that causes an object's geons to fire out of synchrony. If two (or more) geons drift into synchrony with one another, then the properties of one will interfere with the interpretation of the other. As noted previously, GFA and object units are not all equal in their capacity to tolerate this type of processing noise: Units with wide RFs (large σ) are better-suited to tolerate deviations from their trained patterns than are narrow (small σ) units. At the same time, narrow units are better suited to discriminate objects at the instance-level than are wide units. Together, these properties predict that, with reduced attention, class-level recognition will be possible but instance-level recognition will not (especially if the object is depicted in a novel view; see Hummel & Stankiewicz, 1996a). This property is illustrated in Figure 8a, which shows the response of the model to nonsense-object$_1$ when the attention parameter, α, was set to 0.5 (rather than 1.5, the value used in the previous simulations). Note that MetriCat recognized the stimulus as a member of the class *nonsense-object*, but did not recognize it as the instance nonsense object$_1$.

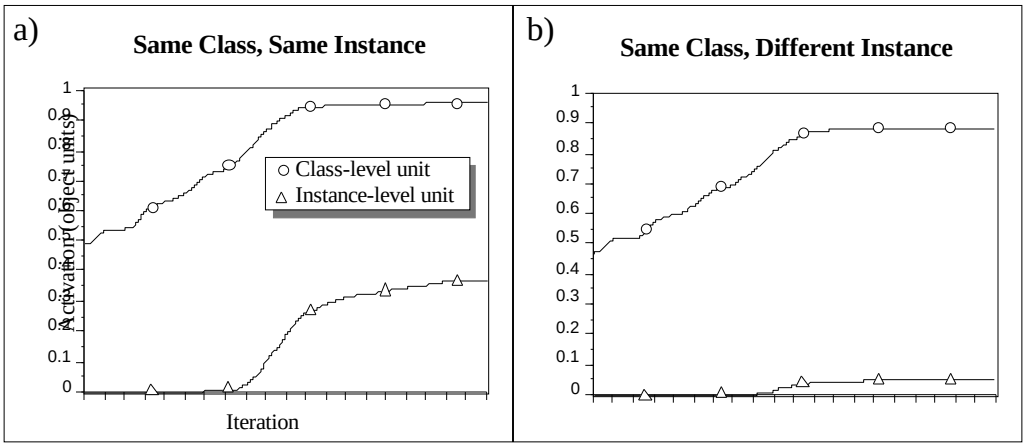

Figure 7. Response to a stimulus created by adding variable noise to a trained instance. (a) Activation of the class (circles) and instance (triangles) object units recruited in response to the stimulus object as a function of time (iterations during the simulation). (b) Activation of the class and instance object units recruited in response to a different member of the same class.

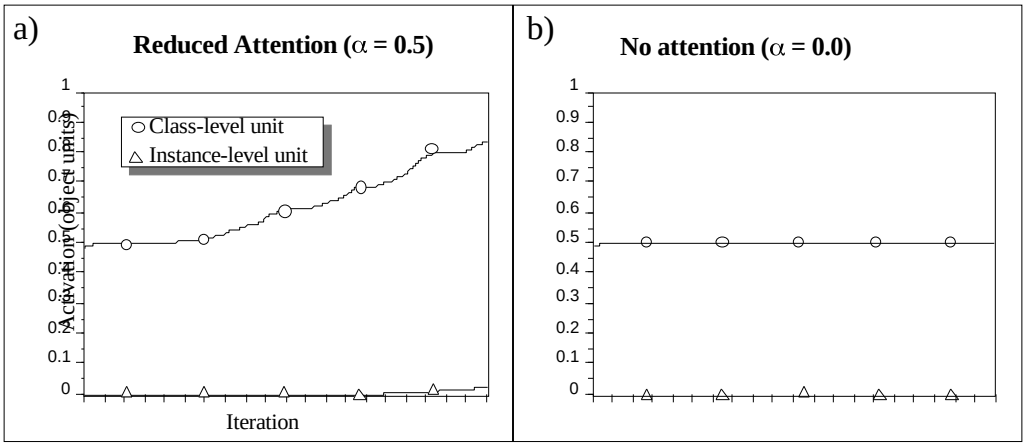

Figure 8. Response to a trained stimulus under reduced attention. Circles depict the activation of the class-level unit recruited in response stimulus. Triangles depict the response of the instance-level unit recruited in response to the stimulus. (a) Model response with reduced attention ($\alpha = 0.5$). (b) Model response without attention ($\alpha = 0$).

Figure 8b shows the model's performance with α set to zero—that is, when the stimulus is not attended at all. Here, the object is recognized neither at the instance nor class levels. This result seems to predict that object recognition requires attention. However, recall that MetriCat is intended, not as a stand-alone model, but as a component of the upper layers of Hummel and Stankiewicz's (1996a) JIM.2. Specifically, Layer 1 of MetriCat corresponds to the *Independent Geon Array* (*IGA*) of JIM.2, and the classification layers of MetriCat correspond to the classification layers of JIM.2. Like Layer 1 of MetriCat, the IGA represents shape attributes independently, is dependent on dynamic binding, and is responsible for recognition by structural description. However, JIM.2 has an additional component, the *Substructure Matrix* (*SSM*), that does not require dynamic binding. For familiar object views, the SSM is sufficient for rapid recognition without attention. The IGA and SSM represent object shape in qualitativelydifferent ways. JIM.2 predicts, not that unattended objects will not be recognized, but that recognition of unattended objects will differ *qualitatively* from the recognition of attended objects (see Hummel & Stankiewicz, 1996a). Stankiewicz et al. (1996; see also Stankiewicz & Hummel, 1997) report experimental findings that are strikingly consistent with the specific predictions of JIM.2. For simplicity, we have left the SSM out of MetriCat, but in principle, the SSM could be part of

MetriCat, just as it is part of JIM.2. Thus, we take the simulation results in Figure 8b to predict that structural description (and recognition on the basis of a structural description) will require attention, not that recognition *at all* will require visual attention.

Discussion

MetriCat is a structural description model of multi-level shape classification. Like its predecessors, JIM and JIM.2, MetriCat represents object shape as a collection of parts in particular spatial relations. Part attributes and relations are represented independently, and are dynamically bound into structural descriptions by synchrony of firing. However, MetriCat departs from its predecessors in that its representation of part attributes and relations is not strictly categorical. MetriCat's representations are categorical in the sense that they are nonlinear over categorical boundaries in attribute dimensions, but they are metric in that they preserve metric differences within categorical boundaries. With the appropriate routines for matching shape representations to object memory (here, Gaussian RFs with varying standard deviations), the resulting metric/categorical representation supports classification at multiple levels of abstraction as a natural consequence. Simulation results show that, counter to popular wisdom, structural descriptions need not be limited to recognition at the level of object classes.

The MetriCat approach to shape classification also suggests a unified account of the role (and mechanisms) of visual attention in shape perception. Due to the logistic representation of metric properties, metric differences within categorical boundaries are reduced in magnitude relative to metric differences across boundaries. As a result, the former are more sensitive to noise than are the latter. MetriCat uses attention to recover such purely metric differences and classify objects as individual instances. In this role, attention serves to increase the signal-to-noise ratio in the representation of a geon, allowing small metric differences to express themselves. The same mechanism serves to maintain the dynamic binding of shape attributes and relations into sets. In both cases, the mechanism is a simple parameter, α, that determines the strength with which geons inhibit one another in the competition for the opportunity to fire: The greater the inhibition, the cleaner the asynchrony, and the higher the signal-to-noise ratio.

Given the benefits of attention to MetriCat, it is reasonable to ask why α should even be a variable: Why not leave it set to a maximum value all the time? The reason is that the privilege of firing cleanly out of synchrony is a finite resource (Hummel & Biederman, 1991). To the extent that one geon is firing all by itself (and thereby maximizing its binding and its signal-to-noise ratio), then by definition, no other geons can be firing at all. All the simulations reported here were run with stimuli consisting of single objects. In such situations, it is sensible to leave α at a maximum value, devoting as much processing as possible to the one object in the visual field. But in more realistic situations there are many objects in the visual field, so it may be sensible as a default to devote only a moderate amount of attention to any one object. Such a default would permit rapid classification at the class-level without necessarily excluding the processing of all other objects in the visual field (cf. Stankiewicz et al., 1997). When it becomes necessary to recognize an object as an instance, it may be worth "turning attention up," temporarily processing that object to the exclusion of others. Clearly, at this point these considerations are merely speculation. But MetriCat suggests a specific mechanism for thinking about the implementation and consequences of visual attention in shape perception.

Relation to Other Models of Object Recognition

MetriCat bears important similarities to current view-based models of object recognition. First, like the models of Poggio and his colleagues (e.g., Edelman et al., 1996; Poggio & Edelman, 1990; Vetter et al., 1994), MetriCat uses units with Gaussian receptive fields as the basis for object classification. However, it is important to recognize that MetriCat differs markedly from any view-based model in the *representational space* over which the Gaussians are defined. In all extant view-based models, shape classification is performed by basis functions defined in the space of the 2-D coordinates of object features: The representational assumption underlying these models is that objects are represented and matched to memory in terms of the coordinates of their features (hence

the name "view-based"). The psychological plausibility of this assumption is questionable (Hummel & Stankiewicz, 1996b). By contrast, in MetriCat, the Gaussian RFs are defined in the space of categorical attributes and relations. As a consequence, the models have very different properties. In particular, it is not clear how view-based models could be adapted to account for the role of categorical properties in shape perception (as summarized in the Introduction; see also Biederman & Gerhardstein, 1995). More importantly, the representations in MetriCat are *structured* (i.e., attributes are represented independently of their relations) whereas current view-based models are strictly holistic. This difference has important implications for the models' ability to generalize from known to novel instances (cf. Fodor & Pylyshyn, 1988; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996b).

A second similarity between MetriCat and view-based models is the shared emphasis on metric properties in the representation of object shape. But here, too, the differences are more important than the similarities. The most important difference is that in view-based models, the representational space (feature coordinates) is perfectly linear in (indeed, identical to) the perceptual domain from which it derives (feature coordinates as they appear in the image). By contrast, in MetriCat, the representational space (the logistic representation attribute values) is nonlinear in the perceptual domain from which it derives (the numerical values of the attributes). This difference is important because the nonlinearity plays a critical role in the behavior of MetriCat, and the linearity plays an equally critical role in the view-based models. It is doubtful that the principles underlying MetriCat could be adapted to operate with a linear representation of attributes, and it is equally doubtful that the principles underlying the view-based approach could be adapted to work with a nonlinear representation of coordinates (Hummel & Stankiewicz, 1996b).

Extensions

The simulations reported here suggest that MetriCat has promise as an account of our ability to recognize objects at multiple levels of abstraction. However, numerous additional tests are required to evaluate MetriCat as a general model of object recognition. Doing so will require integrating MetriCat with a front-end (such as the early layers of JIM) that can automatically recover geon attributes and relations from line drawings or gray-level images. At this point, it is nonetheless possible to speculate about some other implications of the MetriCat approach to object recognition. One particularly interesting implication concerns the relationship between levels of classification and sensitivity to variations in viewpoint. In general, categorical attributes and relations are more robust to variations in viewpoint than otherwise equivalent metric attributes (Biederman, 1987). For example, a representation that describes the axis of a curved cone simply as "curved" will remain the same in many views of the cone, whereas a representation that specifies its numerical degree of curvature (e.g., as given by the curvature of the cone's bounding edges) will change as the cone is rotated in depth. MetriCat relies more heavily on precise metric properties for instance-level recognition than for class-level recognition. (This is true because of the way it codes numerical attributes and the way it matches those attributes to memory for recognition.) It therefore predicts that instance-level recognition will tend to be more view-sensitive than class-level recognition. There is some indirect support for this prediction (see Biederman & Gerhardstein, 1995; Bülthoff et al., 1995; Tarr & Bülthoff, 1995).

MetriCat as Model of Visual Attention

As a model of attention, MetriCat is still in an early stage of development. The mechanism it uses to implement attentional selection—enhanced inhibition for asynchrony—is motivated strictly by the problem of generating structural descriptions in a neural network for shape recognition. Conspicuously absent from the discussion so far is any basis for deciding *when* attention should be directed to an object and *where* to direct it. One particularly interesting property of MetriCat is that an answer to the first question (How do we know when to direct attention?) may serve as a partial answer to the second (How do we know where to direct it?). Note that α is a global parameter. It does not operate on any particular location in the visual field, or even on any particular object. Rather, it serves only to enhance the ability of whatever is currently firing to inhibit anything else that might attempt to fire in the near future: α is directed in time rather than

space. Any given object (or part) exists at a particular location in the visual field, and fires at a particular time. Thus, knowing *when* to direct attention is tantamount to knowing *where* to direct it: Directing attention to whatever is firing at a given time implicitly directs attention to a particular object occupying a particular location in the visual field.

How can MetriCat figure out when to direct its attention? Part of the answer to this question must depend on the system's processing goals, and this part of the question is clearly beyond the scope of our current research. However, given that the system knows that it cares about a particular object, there is a straightforward basis for dynamically directing attention to various parts of that object. This basis relates to the relationship between the activation of object units and the utility of object parts for discriminating one object from another. Ambiguity about an object's identity consists of simultaneous activation in multiple, inconsistent object units. For example, if the units for both "teapot" and "lamp" are equally active at some instant, then at that instant, MetriCat is undecided about whether the object is a teapot or a lamp. Imagine that the system is processing some (as yet unrecognized) stimulus, and that, on the basis of the geon that is currently firing, there is no way to decide whether the object is a teapot or a lamp. As long as the current (uninformative) geon continues to fire, the activations of the "lamp" and "teapot" units will remain unchanged. Then the current geon stops firing, and some new geon starts to fire. If this new geon is informative, then the activations of the object units will begin to change—say, "teapot" begins to decline and "lamp" begins to grow. This change in activation can serve as a signal that the geon under consideration is informative, and that attention should be directed to it (so it can inhibit its competitors and remain active longer without interference). Once the informative geon has done all it can in the way of indicating whether the object is a teapot or a lamp, the object units' activations will once again stop changing. Attention should now be moved away from this geon and onto a different one. Thus, one way to dynamically direct attention may be to set α to a value proportional to the change in the activation of object units. We have implemented an early version of this mechanism and the results so far have been promising. Augmented with a learning algorithm for keeping track of which parts are diagnostic for the recognition of which objects, this temporal search mechanism might serve as one way to search for diagnostic features for object recognition (e.g., as discussed by Biederman, 1987).

Of course, it is unlikely that this simple mechanism could constitute a complete solution to the problem of allocating visual attention. Among other things, it provides no clear way to explain phenomena such as visual search or visual neglect, which seem to have a decidedly spatial component. However, it is tempting to speculate that this type of mechanism might constitute at least one basis for attentional selection in human shape perception.

Selective Processing of Visual "Features"

One other extension of MetriCat is worth mentioning here. MetriCat uses Gaussian receptive fields with variable standard deviations (σ) to classify objects at multiple levels of abstraction. In the model's current form, σ for any given unit is a scalar. Implicitly, this convention sets σ to the same value for all attribute dimensions. An extension (and generalization) of this approach would be to make σ a vector that can take different values on different dimensions (as $\mathbf{p}$ is currently). For example, a given GFA might select for cones with a very particular aspect ratio (i.e., by setting σ to a small value on the dimension *aspect ratio*), without being as selective for particular values on other dimensions (σ would be larger on these dimensions). This kind of selective narrowing of the units' receptive fields would allow MetriCat to choose diagnostic stimulus dimensions for instance-level classification, and constitutes another way the architecture might account for the use of diagnostic features for visual scrutiny.

Summary and Conclusions

MetriCat is in an early stage of development, but the approach is promising. The fundamental principles underlying the model are that: (1) Objects are visually represented as collections of part attributes and relations, which are dynamically bound into structural descriptions (Hummel & Biederman, 1992); (2) part attributes and relations are represented in a nonlinear

fashion that emphasizes categorical boundaries without discarding all metric information (Stankiewicz & Hummel, 1996); and (3) attention enables parts to inhibit one another so that they can fire cleanly out of synchrony. Together, these principles provide a preliminary account of the human capacity to classify objects at multiple levels of abstraction, and suggest a theory of the mechanism of attention in shape perception.

References

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*, 595-593.

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.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for 3-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 1162-1182.

Biederman, I. & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: A critical analysis. *Journal of Experimental Psychology: Human Perception and Performance., 21*, 1506-1514.

Biederman, I. & Schiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. Journal of Experimental Psychology: Learning, Memory and Cognition, 13, 640-645.

Bülthoff, H. H. & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science, 89*, 60-64.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 3*, 247-260.

Clowes, M. B. (1967). Perception, picture processing and computers. In N.L. Collins & D. Michie (Eds.), *Machine Intelligence*, (Vol 1, pp. 181-197). Edinburgh, Scotland: 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., Cutzu, F., & Duvdevani-Bar, S. (1996). Similarity to reference shapes as a basis for shape representation. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pp. 260-265.

Farah, M. (1992). Is an object an object an object: Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. *Current Directions in Psychological Science, 1*, 164-169.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. In Pinker, S., & Mehler, J. (eds.) *Connections and Symbols*. MIT Press, Cambridge, MA.

Foster, D. H., & Ferraro, M. (1989). Visual gap and offset discrimination and its relation to categorical identification in brief line-element displays. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 771-784.

Hoffman & Richards (1985). Parts of recognition. *Cognition, 18*, 65-96.

Hummel, J. E., & Biederman, I. (1991). Binding by phase locked neural activity: Implications for a theory of visual attention. Paper presented at the Annual Meeting of The Association for Research in Vision and Ophthalmology, Sarasota, Fl. May.

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 analog access and mapping. *Psychological Review*, in press.

Hummel, J. E., & Stankiewicz, B. J. (1996a). 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.

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

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Picture and names: Making the connection. *Cognitive Psychology, 16*, 243-275.

König, P., & Engel, A. K. (1995). Correlated firing in sensory-motor systems. *Current Opinion in Neurobiology*, *5*, 511-519.

LaBerge, D. & Brown, V. (1989). Theory of attentional operations in shape identification. *Psychological Review*, *96*, 101-124.

Pinker, S. (1984) Visual cognition: An introduction. In S. Pinker (Ed.), *Visual Cognition*. (pp. 1-62). Amsterdam: Elsevier.

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

Poggio, T. & Girosi, F. (1990). Regularization algorithms that are equivalent to miltilayer networks. *Science, 277*, 978-982.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Schnieder, W. X. (1995). VAM: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. *Visual Cognition*, *2*, 331-375.

Saiki, J. & Hummel, J. E. (1996). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, accepted pending revision.

Stankiewicz, B. J., & Hummel, J. E. (1996). MetriCat: A representation for basic and subordinate-level classification. *Proceedings of the 18th Annual Conference of the Cognitive Science Society* , pp. 254-259.

Stankiewicz, B. J., & Hummel, J. E. (1997). Automatic priming for translation- and scale-invariant representations of object shape. Manuscript in preparation.

Stankiewicz, B. J., Hummel J. E., & Cooper, E. E. (1997). The role of attention in priming for left-right reflections of object images: Evidence for a dual representation of object shape. *Journal of Experimental Psychoilogy*: *Human Perception and Performance.* in press.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *146A*, 225-245.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, *2*, (1), 55-82.

Tarr, M. J., & Bülthoff, H. (1995). Conditions for viewpoint-dependence and viewpoint-invariance: What mechanisms are used to recognize an object? *Journal of Experimental Psychology*: *Human Perception and Performance*. in press.

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

Tipper, S. P., & Driver, J. (1988). Negative priming between pictures and words in a selective attention task: Evidence for semantic processing of ignored stimuli. *Memory and Cognition*, *16*, 64-70.

Treisman, A. (1982). Perceptual grouping and attention in visual search for objects. *Journal of Experimental Psychology*: *Human Perception and Performance, 8*, 194-214.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

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

Vetter, T., Poggio, T., & Bülthoff, H. H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology, 4*, 18-23.

von der Malsburg, C. (1981). The correlation theory of brain function. Internal Report 81-2. Department of Neurobiology, Max-Plank-Institute for Biophysical Chemistry. Gottingen, Germany.

von der Malsburg, C., & Buhmann, J. (1992). Sensory segmentation with coupled neural oscillators. *Biological Cybernetics, 67*, 233-242.

Appendix

Oscillatory Gates: Each gate, *i*, has two components, an excitor, E_i, and an inhibitor, I_i, which interact to cause E_i to oscillate. E_i changes according to:

$$\Delta E_i = \left[0.9\left(1 - E_i\right)\left(1 - 2I_i\right)\right] - 0.5\left(1 - E_i\right) - \sum_{j \neq i} L_j \,, \tag{5}$$

where I_i is the activation of the corresponding inhibitor, and L_j is the inhibitory strength of gate *j*. I_i changes according to:

$$\Delta I_i = \begin{cases} 0.005\left(E_i\right) \text{ iff } I_i < 0.2 \\ \text{if } r_i \\ 0.5\left(E_i\right) \text{ iff } I_i \geq 0.2 \end{cases} \quad \text{otherwise} \begin{cases} -0.01 \text{ iff } I_i > 0.9 \\ -0.5 \text{ iff } I_i \leq 0.9 \end{cases} . \tag{6}$$

where r_i is a Boolean variable that is set to *true* whenever $I_i < 0.1$ and to false whenever $I_i > 0.995$; r_i remains at its last set value whenever $0.1 < I_i < 0.995$. Together, (5) and (6) cause E_i to oscillate, as illustrated in Figure 3 (Hummel & Stankiewicz, 1996a; Hummel & Holyoak, 1997). In combination with the inhibition between competing gates (L), the result is that separate E_i tend to oscillate out of synchrony with one another. L_i, the inhibitory strength of gate *i* is:

$$L_i = E_i P_i \alpha \,, \tag{7}$$

where P_i —the priority of gate *i*—modulates the ability of *i* to inhibit other gates, *j*, and α modulates the global ability of all gates to inhibit one another. A gate's priority is:

$$P_i = \begin{cases} 0 \text{ if } I_i > 0.2 \text{ and } \Delta I_i = (-0.5), \\ P_i = P_i + \left(0.001 + \left[0.01\left(1 - P_i\right)\right]\right) \text{ otherwise.} \end{cases} \tag{8}$$

By (6) and (8), P_i will go to zero immediately after *i* fires and then gradually grow toward 1.0. P_i helps the gates to "time share" by ensuring that a gate will have little ability to inhibit other gates (and therefore little opportunity to fire) if it has fired in the recent past. α corresponds to the amount of attention directed to an object. When α is large (e.g., 1.5), the various gates inhibit one another strongly (Eq. 7) and therefore fire cleanly out of synchrony; when α is zero, the gates cannot inhibit one another at all, and therefore never fire out of synchrony; at intermediate values of α, the gates fire somewhat out of synchrony.

Attribute Units: Each attribute unit, *i*, has a Gaussian receptive field (RF) with a center, μ_i, in the range of logistic values (0.0..1.0) and a standard deviation, σ_i, in the range 0.0..0.25. The input to attribute unit *i*, from all geons, *j*, is:

$$I_i = \sum_j E_j G\left(\left\| \mu_i - l_j \right\|, \sigma_i\right), \tag{9}$$

where E_j is the value of the excitatory gate on geon *j*, l_j is the logistic value of geon *j* on the corresponding attribute, and G() is the Gaussian. The activation of an attribute unit grows as:

$$\Delta a_i = 0.5\, I_i - 0.05 a_i \,. \tag{10}$$

GFA Units: GFA units have Gaussian RFs in the 350-dimensional (50 units X 7 attributes) space of attribute units. The activation of GFA unit *i* in response to attribute vector **a** is:

$$A_i = G\left(\frac{\|\mathbf{p}_i - \mathbf{a}\|}{\sqrt{d}}, \sigma_i\right), \tag{11}$$

where $\mathbf{p}_i$ is the mean (center) of i's RF, σ_i is its standard deviation, and d is the dimensionality of the attribute vector (here, 350). The output of GFA unit i at time t is:

$$O_i^t = \begin{cases} A_i^t \text{ if } A_i^t > O_i^{t-1} \\ O_i^{t-1} \text{ otherwise} \end{cases}. \tag{12}$$

Object Units: Object units have Gaussian RFs in the space of GFA unit outputs. The activation of object unit i is:

$$B_i = G\left(\frac{\|\mathbf{p}_i - \mathbf{o}\|}{\|\mathbf{p}_i\|}, \sigma_i\right), \tag{13}$$

where $\mathbf{p}_i$ is the mean (center) of i's RF, $\mathbf{o}$ is the vector of GFA unit outputs, and σ_i is the standard deviation of i's receptive field. The normalization by $\|\mathbf{p}_i\|$ (the length of the object unit's receptive field vector) serves to correct for the dimensionality of the GFA vector. (Note that, as more objects are learned, the number of object parts stored in memory grows, changing the dimensionality of both the GFA output vector and an object unit's receptive field.)