MetriCat: A Representation for Basic and Subordinate-level Classification

Brian J. Stankiewicz and **John E. Hummel**
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095-1563
bstankie@psych.ucla.edu, jhummel@psych.ucla.edu

Abstract

An important function of human visual perception is to permit object classification at multiple levels of specificity. For example, we can recognize an object as a "car," (the basic level) a "Ford Mustang" (subordinate level), and "Joe's Mustang" (instance level). Although this capacity is fundamental to human object perception, most computational models of object recognition either focus exclusively on basic-level classification (e.g., Biederman, 1987; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996) or exclusively on instance-level classification (e.g., Ullman & Basri, 1991; Edelman & Poggio, 1990). A computational account that naturally integrates both levels of classification remains elusive. We describe a general approach to representing numerical properties (e.g., those that characterize object shape) that simultaneously supports both basic and subordinate/instance-level recognition. The account is based on a general nonlinear coding for numerical quantities describing both featural variables (such as degree of curvature and aspect ratio) and configural variables (such as relative position). Used as the input to a classifier with Gaussian receptive fields, this representation supports recognition at multiple levels of specificity, and suggests an account of the role of attention and time in the classification of objects at different levels of abstraction.

Introduction

One of the most notable properties of human visual perception is our capacity to recognize objects despite variations in the viewing conditions under which the image is presented to the retina (e.g., viewing angle). Numerous models have been proposed in the attempt to account for this property of human object recognition. These models can be divided into two broad classes according to the general strategy they adopt to attack this problem (see Liu, Knill & Kersten, 1995; Tarr, 1995). One class (typically associated with structural description theories of recognition) exploits categorical image properties as the primary basis for object recognition (e.g., Biederman, 1987; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996). On this account, objects are represented in terms of categorical features (including the categorical relations among those features) that remain unchanged as an object's distance or orientation relative to the viewer varies: Because the features remain the same in many views, recognition is unaffected by many changes in viewpoint. The other class of models uses alignment (e.g., Ullman, 1989), view interpolation (e.g., Poggio & Edelman, 1990) or other normalizations (see Hummel & Stankiewicz, 1995) to bring new object views into correspondence with stored views: Here, the normalization serves to correct for variations in the locations of an object's features (in the image) that result from variations in viewpoint.

One notable difference between these approaches is that the former emphasizes the role of categorical image properties (such as categorical features and relations), whereas the latter emphasizes the role of holistic metric properties (specifically, the numerical coordinates of object features). In addition to supporting different algorithms for discounting variations due to viewpoint, these differing approaches to object representation also give rise to different "expertise" at different levels of classification (Bülthoff & Edelman, 1992): Categorical models may provide a better account of recognition at the basic level (e.g., recognizing an object as a "car"; Rosch, Mervis, Johnson, & Boyes-Braem, 1976), while metric models may provide a better account of recognition at the subordinate or instance level (e.g., recognizing an object as a "Mustang" or "Joe's Mustang").

The human is expert at both basic- and subordinate-level classification. It is tempting to speculate that this dual expertise reflects the simultaneous operation of both approaches to recognition: Perhaps categorical features or structural descriptions allow us to classify objects at the basic level while metrically-specific holistic representations allow us to classify objects at the subordinate or instance level (Bülthoff & Edelman, 1992; Farah, 1992). While this account is almost certainly correct for some cases of subordinate-level recognition (e.g., face recognition), it is likely inadequate as a complete account of human multi-level classification. One problem with this account is that it predicts that people will classify objects at the subordinate-level faster than they classify objects at the basic-level (holistic representations can be generated much faster than categorical structural descriptions; see Hummel & Stankiewicz, 1996), whereas people are fastest to classify objects at the basic level (e.g., Rosch, et. al, 1976). A second limitation of this account is that it predicts that subordinate level recognition should be more holistic than basic-level classification. While this is true for face recognition (Tanaka & Farah, 1993), it is not true for all

subordinate-level classification tasks (e.g., Biederman & Schiffrar, 1987). Given these considerations, it seems likely that the human visual system achieves multi-level classification on the basis of something more sophisticated than a simple hybrid holistic-categorical representation of shape.

This paper presents our progress toward a model of multi-level classification based on a different kind of hybrid metric-categorical representation of object shape. Following the structural approach of Hummel and Biederman (1992; Hummel & Stankiewicz, 1996), we assume that independent attributes are represented on independent units (i.e., rather than representing attributes and their locations holistically as complete "views"). But in contrast to these models, we assume that shape attributes are not coded in a strictly categorical fashion (e.g., "straight vs. curved", "parallel vs. non-parallel," etc.). Rather, we adopt a representation of numerical quantities (such as degree of curvature and degree of parallelism) that captures both the metric and categorical aspects of those quantities. Like a categorical representation, the proposed representation changes fastest across categorical boundaries (such that the representation of curvature 0.1 [slightly curved] differs more from curvature 0 [straight] than it differs from curvature 0.2 [more curved]). But like a metric representation, it also captures differences between numerical values on the same side of a categorical boundary (e.g., between curvature 0.1 and curvature 0.2). In combination with an architecture for classifying objects on the basis of these metric-categorical ("MetriCat") representations, the result is a model that can classify objects at multiple levels of abstraction simultaneously. The model also suggests a natural account of the role of attention and time in classification at different levels of specificity, and the relationship between view specificity and levels of classification.

The MetriCat Representation of Numerical Values

As described here, the model is addressed only to the representation of properties that can be characterized as real values (or differences of real values) along a single dimension. For example, local curvature can be characterized in terms of a real number ranging from $-\infty$ (infinitely curved in one direction) to 0 (straight) to ∞ (infinitely curved in the opposite direction). Similarly, the expansion in the axis between two straight lines can be described by a real number in the range $-\infty \ldots \infty$, where negative values indicate that the axis narrows from end A to end B, positive values indicate that it expands from end A to end B, and zero indicates that the it remains a constant width along its length (i.e., the lines are parallel). A strictly categorical representation of these values might represent curvature = 0 as "straight" and all curvatures $\neq 0$ as "curved"; likewise, the axes might be represented as simply "parallel" (expansion = 0) or "non-parallel" (non-zero expansion). Such codes change rapidly at a single point (the transition point between adjacent values), and do not change at all in between those values. This property is responsible for the utility of categorical codes for class recognition and for discounting variations in viewpoint (see, e.g., Biederman, 1987), but it is a liability as a basis for instance-level recognition: Two shapes that can only be distinguished by, say, the degree of curvature on a given edge will be identical in a strictly categorical code.

MetriCat represents numerical values in an intermediate fashion, in that it emphasizes differences across categorical boundaries (e.g., straight vs. curved.) without completely discarding differences on the same side of a categorical boundary (e.g., different degrees of curvature). Specifically, we represent numerical variables as a logistic function of their raw numerical value (see also Hummel & Stankiewicz, 1995):

$$C = \frac{1}{1+e^{-R\kappa}}, \qquad (1)$$

where C is the represented value, R is the raw numerical value, and κ is a constant. Like a categorical code, C(R) has the property that it changes fastest across categorical boundaries in R. For example, if R is local curvature, then adding a small degree of curvature, say 1, to R = 0 (a straight line) has a greater impact on the value of C than adding the same amount of curvature to R = 10 (a curved line) (C(0) - C(1) = 0.5 0.731 = -0.231, whereas C(10) - C(11) = 0.99954 - 0.9998 = -0.00044). Thus, like a categorical variable, C changes fastest when the raw value, R, crosses a categorical boundary; but unlike a purely categorical variable, C continues to change even within categorical boundaries of R.

We assume that objects are visually represented in terms of the MetriCat values of each of several numerical quantities (such as the aspect ratio and cross-section curvature of each of their parts; see Biederman, 1987). The value of each variable, C^i, is coded as a vector $\mathbf{c^i}$, where the j th element of $\mathbf{c^i}$ is a unit, c^i_j, with receptive field in C^i that has a specific center, μ^i_j, and a specific width, w^i_j. In the current model, the widths and centers were set to random values in the ranges $0 < w < 0.5$ and $0 \leq \mu \leq 1.0$, respectively.

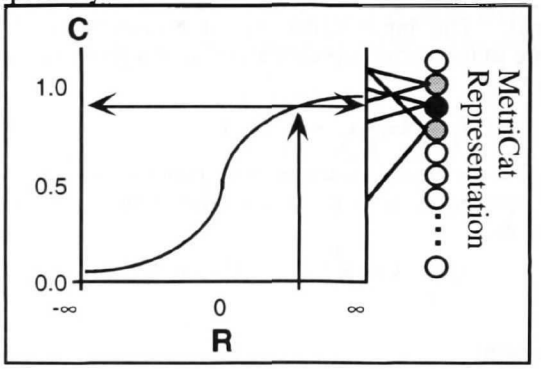

Figure 1: Illustration of the MetriCat representation of a specific value of R. C is a logistic function of R and units (right side) respond to specific values of C.

The bottom-up input to unit c^i_j at time t is:

$$I^i_{j,t} = G(\left|C^i - \mu_j\right|, w^i_j), \qquad (2)$$

where G is the Gaussian, C^i is the input real value input on dimension i, μ_j is the center of receptive field j, and $w^i j$ is the width (standard deviation) of receptive field j. For the purposes of the simulations reported here, we assume that every object is represented by two MetriCat vectors, $\mathbf{c}^1$ and $\mathbf{c}^2$, where each vector encodes the MetriCat representation of one numerical variable. For our current purposes, the precise meanings of these vectors (e.g., "$\mathbf{c}^1$ codes curvature", etc.) is unimportant. Rather, we are interested in the properties of the collection of vectors as a basis for classifying arbitrary objects whose similarity relations are defined to correspond to different basic-level classes (i.e., low similarity, or very different vector representations) and different members of those classes (i.e., high similarity, or similar but non identical vector representations): Will the model treat different members of the same "class" as similar but not identical?

Classification Based on MetriCat Values

To answer this question, it is first necessary to specify an appropriate algorithm to perform classification on the basis of the vectors generated by any given object. For this purpose, the model uses Gaussian radial basis functions (e.g., Poggio & Girosi, 1990; Poggio & Edelman, 1990) in the 50 dimensional space given by the two MetriCat vectors ($\mathbf{c}^1$ and $\mathbf{c}^2$), each with 25 units, $c^i j$ (j = 1..25). Every object is coded in the model's memory as a collection *classifier units* with Gaussian receptive fields in this 50 dimensional space. The center of a given unit's receptive field corresponds to the "preferred" pattern for the corresponding object. Each object, k, is coded by 3 classifier units with the same center but with different standard deviations, σ (σ take values of 0.02, 0.01 and 0.0066). Small σ allow units to tolerate only small deviations from their preferred patterns; such units thus perform instance-level classification. Larger σ permit larger deviations from the preferred pattern, and permit a unit to perform class recognition (responding to multiple, similar patterns). The input value, I_k, of classifier unit k in response to the vector representation, s, of a given stimulus is given by:

$$I_k = \mathrm{G}(\|\mathbf{p}_k - \mathbf{s}\|, \sigma_k), \qquad (3)$$

where G is the Gaussian, and $\mathbf{p}_k$ is k's preferred vector.

The activation of a given classifier unit k at time t changes as:

$$\Delta A_k^t = 0.9\left(1 - A_k^t\right) I_k^t - 0.25 A_k^t. \qquad (4)$$

Algorithm

In addition to its bottom-up input (Eq. 1), each MetriCat unit, $c^i j$, also receives lateral excitation from other units in $\mathbf{c}^i$. Unit j excites unit i to the extent that its center lies within i's receptive field. The input from j to i is:

$$LE_{ij} = A_j \, \mathrm{G}(|\mu_i - \mu_j|, w_i), \qquad (5)$$

where G is the Gaussian, and μ_i and μ_j are the centers of receptive fields i and j and A_j is the activation of unit j. Broad units, which will tend to have many other units in their receptive fields, will tend to receive more lateral excitation than narrow units, which will have fewer other units in their receptive fields. As a result, MetriCat units with broad receptive fields tend to become active faster than units with narrow receptive fields: Coarse information about an object's shape becomes available earlier than information about its fine metric details. The utility of this property is that the coarse information is more robust to noise than is fine information. Noise may originate in both the stimulus and the system. Stimulus-induced noise may result from changes in viewpoint (i.e., producing slight deviations from the expected values of an object's metric properties); system-induced noise may result from random variations in the magnitude of neural impulses (or myriad other sources). At the MetriCat level of representation, lateral excitation makes coarse noise-tolerant information available rapidly; at the classifier level, the initial absence of fine metric information has a greater adverse impact on classifier units with narrow receptive fields than it has on units with wide receptive fields. As a result, class recognition precedes instance recognition. This property is apparent in the simulation results.

Simulations

Simulations were run with four one-part objects, A1, A2, B1, and B2. Each object was defined by real values on two dimensions, C^1 and C^2. Objects were created in pairs (A and B) such that members of the same pair were more similar to one another than to either member of the other pair. Object A1 had values [0.25, 0.1] (on C^1 and C^2, respectively), A2 had [0.25, 0.3], B1 had [0.75, 0.1], and B2 had [0.75, 0.3]. Note that members of the same pair have identical values on C1 and, and each member of one pair shares the same value of C2 with one member of the other pair. But overall, objects are more similar within than between pairs. This arrangement permitted us to observe three properties of the model: (1) Can it distinguish highly similar objects? (2) Will classifier units with broad receptive fields respond to both members of a class? and (3) What is the time course of the model's ability to make within- vs. between-class distinctions?

Simulations were run in two phases: 2000 (unsupervised) learning iterations followed by 1000 test iterations. Objects were presented to the model by means of oscillatory gates (Hummel & Stankiewicz, 1996) that controlled the input to the MetriCat units. Each gate was associated with one object (i.e., part), with the result that (a) the properties of one object "fired" (were passed to MetriCat) out of synchrony with the properties of other objects, and (b) there was random noise in inputs to the MetriCat units, especially during transitions between different objects (see Figure 3, lower frame). During learning, new classifier units were recruited whenever (i) the average ΔA_i over all of MetriCat units, i, was less than 0.03, and (ii) the Euclidean distance between the current MetriCat pattern of activation and that of all preferred classifier patterns was greater than 0.1.

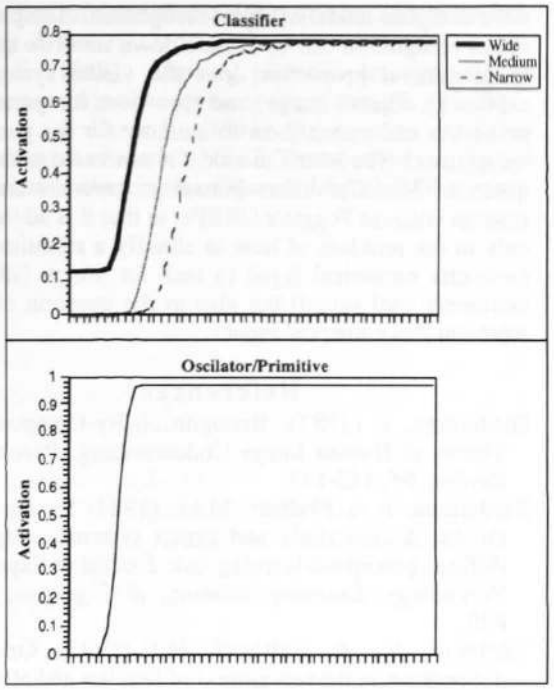

Figure 2: Model responses (over time) on a representative test run. Upper: Activation of three classifier units (one wide, one medium, and one narrow). Lower: Activation of the corresponding oscillator.

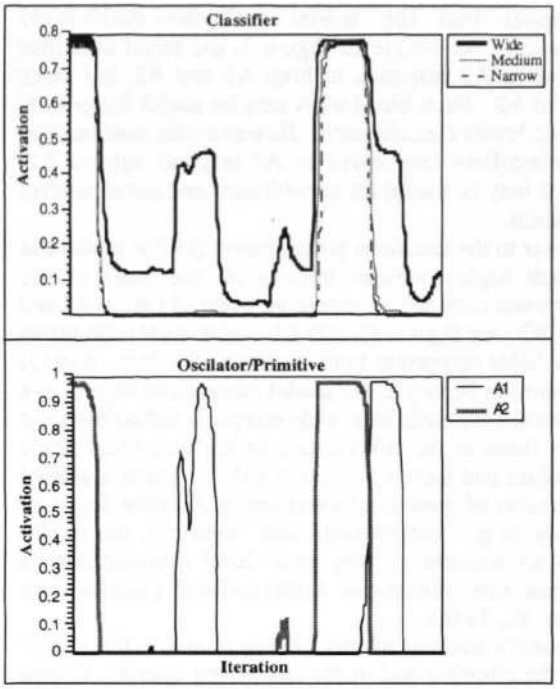

Figure 3: Model responses (over time) on a representative test run. Upper: Activation of three classifier units that respond preferentially to A2. Lower: Activation of the A1 and A2 oscillators.

Three different classifier units were recruited in response to each learned MetriCat pattern. The three units for a pattern have the same **p** (preferred pattern) but different σ (receptive field width). Test iterations were run in the same manner as the training iterations except that no learning took place. Figures 2 and 3 show activation as a function of time for three classifier units with the same preferred pattern but different σ (top row) and the corresponding oscillator activations (bottom row). The temporal ordering of classifier responses is apparent in Figure 2. Note that the classifier with the widest receptive field reaches asymptote first, followed by the medium unit and finally the narrow units.

The coarse classification behavior of the wide units is apparent in Figure 3. The wide classifier shown in the figure was recruited to respond primarily to A2. Note that this classifier responds most strongly to A2 but also responds to A1. Although it does not appear in the figure, the wide unit recruited for A1 showed the complementary response pattern. Neither unit responded to B1 or B2. The medium and narrow units responded very little to non-preferred inputs (e.g., the narrow units for A1 did not respond at all to A2). As apparent in Figures 2 and 3, the model classifies its inputs at a coarsest level first and later at finer level: Those units that become active rapidly (Figure 2) are the same as those that respond to patterns that deviate from their preferred patterns (Figure 3).

Discussion

The MetriCat model represents numerical values at multiple levels of specificity and combines the properties of both categorical and metric representations of numerical variables. Coupled with an appropriate classifying routine (e.g., Gaussian basis functions), this approach to the representation of numerical values has a number of desirable properties as a basis for multi-level classification. The preliminary simulations reported here are consistent with this claim. First, information about the general properties of a stimulus are made available faster than specific information. The utility of this property is that coarse information, which becomes available first, is also more robust to noise (e.g., resulting from changes in viewpoint; Biederman, 1987) than is metrically precise information. This property permits rapid recognition that is robust to noise in the input (e.g., as a result of variations in viewpoint) and noise in the system (e.g., as a result of the oscillators). The rapid availability of coarse information also suggests an account of our ability to categorize an object at the basic-level (e.g., "car") faster than we can classify it at the subordinate-level (e.g., "Mustang").

A second important property of this architecture is its ability to capture the hierarchical similarity relations among different stimuli. In some ways this capacity is property of any vector coding of a population of stimuli. However, the architecture here takes this capacity one step further and (by the activity of the classifier at different scales) explicitly tags the level at which to stimuli are similar or different. It is in

this respect that the model performs multi-level classification. As visible in Figure 3, the broad classifier responsive to A2 responds to both A1 and A2, but more strongly to A2. Such broad units may be useful for general (e.g., basic level) classification. However, the medium and narrow classifiers responsive to A2 respond only to A2. Such units may be useful for subordinate- and instance-level classification.

In answer to the questions posed above: (1) The model can distinguish highly similar objects on the basis of the classifier units with narrow receptive fields (A1 from A2 and B1 from B2; see Figure 3). (2) Classifier units with broad receptive fields respond to both members of a class. And (3) as illustrated in Figure 2, the model categorizes objects at a general level (via units with wide receptive fields) before it classifies them at the subordinate or instance level (units with medium and narrow receptive fields). Using a unified representation of numerical variables at different levels of specificity (e.g., "categorical" and "metric"), the model suggests an account of why basic-level classification is faster than subordinate- or instance-level classification (Rosch et. al., 1976).

The model's account of this finding relates to the role of noise in the stimulus and in the classifying system. Coarse MetriCat units are both faster to respond and more robust to deviations from their preferred inputs (e.g., as resulting from noise) than are fine units. As a result, coarse (roughly categorical) information becomes available earlier than fine (more metric) information. The classifier units exploit this difference: Because broad classifiers are more robust to deviations from their preferred patterns than narrow classifiers, they are less sensitive to the initial absence of activity in the fine MetriCat units. Broad classifiers therefore respond earlier in processing than narrow classifiers. As processing proceeds, the fine MetriCat units begin to respond, so the resulting pattern better fits any narrow classifier units that are tuned to respond to it. Noise is inevitable under realistic assumptions about the world and neural information processing. The current approach provides a basis for rapid general classification and subsequent detailed classification even in the presence of such noise.

The model also suggests an account of the role of attention in subordinate-level classification. More time is required to activate fine MetriCat units than coarse MetriCat units. Attention may serve in part to devote the necessary processing time to diagnostic elements of an object's shape. Thus rather than generating a holistic representation of an object for the purposes of subordinate-level classification, the current model suggests that attention may instead direct processing to diagnostic elements. Although this idea is intuitive, the current model provides the first computational account of the representations that may serve as the basis for this selective processing, and the classification routines that may exploit it.

The simplified simulations reported here were run with MetriCat as a stand-alone system. However, the utility of the MetriCat approach lies in its properties as a component of a more general object recognition system. In particular, MetriCat can easily be incorporated as a component of a more complete model of object recognition. The problem of object recognition can be broken down into two questions: What general properties does the visual system make explicit an object's image? and How does it represent those properties and match them to memory for the purposes of recognition? The MetriCat model is addressed to the second question. MetriCat differs from other general classification systems (such as Poggio's GRBFs) in that it is addressed not only to the problem of how to classify a stimulus given a particular numerical input (a task for which GRBFs are extremely well suited) but also to the question of how to represent that numerical input.

References

Biederman, I. (1987). Recognition-By-Components: A Theory of Human Image Understanding. *Psychological Review, 94*, 115-147.

Biederman, I. & Shiffrar, 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, & Cognition, 13*, 640-645.

Edelman, S., & Bulthoff, H.H.(1992) Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*, 2385-2400.

Farah, M. J. (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.

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

Hummel, J. E., & Stankiewicz, B. J. (1995). Coordinates and spatial relations in object memory. *Technical report 95-01, Shape Perception and Memory Laboratory*, University of California, Los Angeles. Los Angeles, California.

Hummel, J. E., & Stankiewicz, B. J. (1996). An architecture for rapid, hierarchical structural description. T. Inui and J. McClelland (Eds.) In *Attention and Performance XVI*. In Press.

Liu, Z., Knill, D.C. & Kersten, D.(1995) Object classification for human and ideal observers. *Vision Research, 35*, 549-568.

Poggio T., & Edelman S. (1990). A network that learns to recognize 3-Dimensional objects. *Nature, 343*, 263-266.

Poggio, T. & Girosi, F. (1990) Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247*, 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.

Tanaka, J.W. & Farah, M.J. (1993)Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 46A*, 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*, 55-82.

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

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