Modeling Perceptual Learning of Abstract Invariants

Philip J. Kellman <Kellman@cognet.ucla.edu>
Timothy Burke <MizeRai@ucla.edu>
John E. Hummel <Jhummel@psych.ucla.edu>

University of California, Los Angeles
405 Hilgard Avenue, Los Angeles, CA 90095-1563

Abstract

We present the beginnings of a model of the human capacity to learn abstract invariants, such as *square*. The model is founded on four primary assumptions, which we believe to be neurally plausible and generic: Metric space, Topology, Comparison operations (subtraction, greater-than/less-than), and Extraction of vertices. The model successfully learns to discriminate simple planar quadrilaterals, and generalizes that learning across variations in viewpoint and modest variations in shape.

Introduction

A hallmark of human information processing is the ability to detect and respond to abstract invariants. Many of the most important outputs of visual perception and cognition involve shape and spatial relations. As the Gestalt psychologists pointed out early in this century, these are relational notions. A square, or a melody, is not definable in terms of any particular elements, but in terms of invariant relationships.

We not only encode certain characteristic relations, but we discover new ones with experience. This ability makes possible high level, sometimes almost magical, expertise. In chess, the best human player can compete with a computer system that examines 250 million moves per second, although plainly we did not evolve to process the specific relations of shape and arrangement that are important in chess. Human perceptual learning allows the discovery of features and relations that make possible efficient performance in almost every domain of expertise (Gibson, 1969; Goldstone, 1998).

Some key aspects of perceptual learning remain deeply mysterious. A crucial one might be called the *discovery problem*: Expertise in a classification task grows as new stimulus relationships, often quite abstract, become the basis of response. How does the visual system discover abstract invariants, such as squareness, roundness, or parallelism? By abstract invariant we mean a visual property that, while *computable from*, is not *definable in* the vocabulary of primitive features from which it is derived. For example, no logical concatenation—conjunctive, disjunctive or otherwise—of neural responses in primary visual cortex (i.e., local visual properties such as edges, vertices or Gabor components) defines the invariant *squareness*. Squareness is both more and less than any finite set of such features. It is more because some new activation pattern might also form a square, and it is less because many of the attributes of any given activation pattern (e.g., its precise location, size and orientation) have nothing to do with their squareness.

Because an abstract invariant is defined less in terms of the visual primitives that compose it than in terms of the relations among those primitives, it is a mystery how the visual system discovers invariants in the outputs of such primitives. It is possible to build a *local* feature detector by systematically combining the outputs of a finite number of simpler feature detectors (e.g., it is possible to define an edge detector based on a weighted sum of the outputs of a finite number of contrast detectors). Learning such a feature is therefore a relatively simple matter of finding an algorithm to discover the appropriate "wiring diagram" (i.e., weighting terms) from the input (a specific set of contrast detectors) to the output (the local edge detector). Many such learning algorithms exist, including supervised learning algorithms such as back-propagation (Rumelhart, Hinton & Williams, 1986), as well as numerous unsupervised learning algorithms. All these algorithms work precisely because (a) they operate by exploiting statistical regularities in their inputs (or, in the case of supervised rules, regularities in the input-output mappings), and (b) the feature to be learned can be defined in terms of the more basic features of which it is composed. By contrast, because squareness does not correspond to any finite collection of local features, there exists no analogous wiring diagram that can detect all (and only) instances of "squareness". As such, there is no straightforward statistical basis for learning "squareness" based on the outputs of local feature detectors.

We assume that an invariant such as "squareness" is not detected simply by constructing a large (potentially infinite) number of detectors for specific squares and then summing their outputs. As the Gestaltists argued against their structuralist predecessors, we can always devise a new square of a different size, made by arranging some new elements, in a new position. It would still be readily detected as a square. An algorithm geared to learning each and every possible instance of a square would be unwieldy to say the least. Humans, by contrast, can learn many invariants in as little as one exposure and transfer that learning to new instances.

How can we account for this performance? Although perceptual learning of abstract relationships is well documented (e.g., Chase & Simon, 1973; Gibson, 1969; for an excellent recent review, see Goldstone, 1998), little modeling has addressed the learning of abstract invariants. Our aim in the present program of research is to develop models of invariant detection and learning in visual shape

perception. Here we report initial progress toward those goals.

Mixing Architectures in a Principled Fashion

What appears difficult about the problem of detecting invariants may depend on one's starting point. From the standpoint of symbolic descriptions, squareness does not seem too formidable. Given a closed figure, its vertex locations, edge lengths and angle measurements, we can easily write a mathematical test for squareness. But it would be unwieldy to have such a routine for every spatial invariant. Nor is it clear how such static routines could come to learn *new* invariants. On the other hand, traditional connectionist approaches readily support learning, but will not come to classify (correctly) new squares if their initial inputs are limited to concrete elements such as local features.

Toward a Grammar of Form

What initial recodings of concrete inputs could allow both the extraction of abstract relationships and the learning of new ones? At the root of our approach is the idea that the apparent openendedness of the human ability to learn abstract invariants calls for a system that is formally like a grammar with recursive rules. Natural languages can generate an unlimited number of novel sentences based on a fixed set of words and a recursive rule system for combining them. The same idea has been applied to object recognition: Objects may be decomposed into a finite vocabulary of volumetric primitives and spatial relations connecting them (e.g., Biederman, 1987; Marr & Nishihara, 1978). It seems plausible that human perceptual learning of abstract spatial invariants depends on some basic set of relations and some processes for concatenating them.

In the present work, we pursue this approach in building a prototype *shape network* that extracts a small set of important relations early on and uses them as inputs for learning to classify simple shapes. We choose as our domain the names of simple planar quadrilaterals, including squares, rectangles, parallelograms, trapezoids and rhombuses. This domain is useful for several reasons. Most importantly, the labels refer to abstract entities. Changes in retinal position, scale, and planar orientation (for the most part) of the constituent elements do not affect the correct labels. Second, this kind of classification is arguably natural for humans. Young children readily come to distinguish and name squares, rectangles and so on, and generalize naturally across changes in size, position and orientation. Third, planar shape classification—and the subtleties of quadrilaterals in particular—hinge on interesting spatial relations and comparisons. Ultimately, our aim is to encompass richer domains of shape description, including 3-d form, but the planar shape domain is a reasonable choice for confronting the basic challenges of how encoding and learning might cope with abstract spatial relations.

A key challenge is to postulate only those steps for recoding or finding relations that can be justified on independent grounds. That is, success will not consist of finding a special-purpose processor that responds to squareness, but allowing squareness to emerge naturally from basic operations that we would expect, on independent grounds, to be part of visual processing. By implementing only operations that can be justified independently, we can begin to develop the fundamental grammar of invariant processing and ultimately test whether the scope and limits of what is learnable within that grammar approximate those of human visual cognition.

Based on a small set of early recodings that make sense on general grounds, this first shape network can learn abstract notions—specifically various kinds of quadrilateral (4-vertex, planar) figures. It also classifies new instances, such as a square of a size and position not previously seen.

Modeling Assumptions

The vocabulary of a "visual grammar" consists of a finite set of basic operations corresponding to (presumably innate) assumptions on the part of the visual system about the nature of the visual world. For our current purposes, we assume that the visual system comes into the world equipped with (at least) the following knowledge/capacities:

I. Metric Space: We assume that neurons in early visual processing (e.g., retina, LGN, V1) "know" (perhaps implicitly, in the form of their interconnections) about metric space: They know that their receptive fields correspond to finite regions of larger metric space, and they know (approximately) where in that space their receptive fields are located. We assume that this knowledge manifests itself in a neuron's (or hypercolumn's) ability to signal its location independently of any of its other properties, namely by activating other neurons representing location (e.g., in Euclidean coordinates) independently of the nature of whatever visual features happen to reside there.

II. Adjacency (Topology): Implicit in (I), but deserving of mention, we assume that neurons in early visual processing "know about" their adjacency relations (and possibly other topological relations). This kind of knowledge is manifest in the local lateral connectivity among, for example, neurons in visual area V1.

III. Difference Operations: Third, we assume that, given pairs of numerical quantities (e.g., coordinates in a Euclidean space), the nervous system is equipped with routines for performing various kinds of difference operations, including subtraction, and evaluating greater-than and less-than relations.

IV. Vertex Finding: Finally, we assume that early visual routines can find vertices and other local changes in contour curvature based on the outputs of basic local edge computations.

These four assumptions, along with the way they are embodied and the ways in which they interact, form the theoretical foundation of the current modeling effort. Our goal in this paper is to demonstrate that, embodied in an appropriate architecture, these asumptions are sufficient to "bootstrap" the learning of abstract invariants such as "square."

The Shape Network Model

The assumptions are instantiated in a four-layer "neural"-style network. Units in the first layer represent the retinal coordinates of the vertices in an image of a quadrilateral. We assume that the vertices are detected and their spatial coordinates registered by an early preprocessing stage (Assumptions IV and I, above). (We acknowledge that simply "handing" the model the coordinates of vertices is a strong simplification. We are currently working to relax it and equip the model with a more realistic front-end.) Units in the second layer compute the pair-wise Euclidean distances between the coordinates coded in the first layer (Assumptions I and III). In the current implementation, each unit represents one distance (e.g., there is one unit for the distance between vertex 1 and vertex 2, another for vertices 2 and 3, etc.), and distance is represented as activation. That is, in Layer 2 layer, distance is rate-coded. Coordinates are "read into" the model in a fixed order, starting from some corner on the stimulus, and proceeding around the figure clockwise. This convention is a simplified implementation of our more general assumption that the system knows which vertices are connected to which by virtue of an intervening contour (Assumption II, topological relations). (In the current model, implicit knowledge of sequential order of vertices in a connected figure is important, although the particular starting point in the sequence is not.)

Units in the third layer take their inputs from pairs of distance units (i and j) in the second layer, and compare the distances for their equality (Assumption III). Specifically:

$$e_{ij} = \begin{cases} 1 & \text{if } m\left|d_i - d_j\right| < \delta \\ \dfrac{\delta}{m\left|d_i - d_j\right|} & \text{otherwise,} \end{cases} \quad (1)$$

where e_{ij} is the activation of equality unit ij, d_i and d_j are distances i and j, δ is a difference threshold that determines the rate at which e drops off as the absolute difference in distances increases, and $m = \max(d_i, d_j)$, serves as a scaling factor. δ was set to .02 in the simulations reported here. e_{ij} takes a value of 1 when d_i and d_j are within δ of being equal, and falls off toward zero as $|d_i \quad d_j|$ approaches infinity.

The resulting pairwise distance comparisons serve as the input to the fourth (output) layer of units, which learn to classify their inputs as representing various four-sided geometrical figures (squares, parallelograms, trapezoids, etc.). The net input to output node i is simply the dot product:

$$n_i = \Sigma_j \, a_j \, w_{ij} \quad (2)$$

where a_j is the activation of node j in layer 3, and w_{ij} is the weight on the connection from j to i. The activation of output node i is given by the logistic function:

$$a_i = 1/(1 + e^{-n_i}). \quad (3)$$

At the beginning of training, the connections between the third and fourth (output) layers were initialized to zero. On each training trial, one four-sided figure was presented at a time to the network (as detailed below), and layers of units were updated in sequence (layer 1, then layer 2, etc.) until a pattern of activation was generated on the fourth (output) layer. The connections between the third and fourth layers were modified according to the difference between the actual activation of output node i (a_i) and the desired output (d_i) in response to the training pattern (Rumelhart et al., 1986):

$$\Delta w_{ij} = \eta a_j \, (1 - a_i)(a_i - d_i), \quad (4)$$

where η is a learning rate.

Simulations

The present simulations were to designed to test the model's ability to learn invariants such as square, rectangle, etc., from a small number of examples (typically one) and to explore its ability to generalize to new instances that differ from the training examples in their location, size and orientation. We also explored the model's ability to generalize across small distortions in the coordinates of a figure's vertices.

Training

We trained the model to classify six types of quadrilaterals. They are shown in Figure 1 along with the coordinates used for training. The model was trained on one example of each, except for the trapezoid, of which there were two examples.

We trained the model to perform two types of classification. *Inclusive* classifications required the model to respond to each stimulus with every label that applied to it (e.g., a square is also a rectangle, a rhombus, etc., so in this condition, the model was trained to respond with all these labels given a square as a stimulus). *Exclusive* classifications required the model to respond only with the most specific label corresponding to a stimulus (e.g., a square would activate only the square unit). The exclusive classification is arguably the more humanly natural, and serves as the basis of the majority of the tests reported here. During training, stimuli were presented one at a time, and the weights at the output layer were corrected in response to the model's output, as described above. The weights were updated "in batch", after all stimuli had been presented. Training proceeded until the mean square error of the model's response at the output layer fell below .01. The classification tasks were trained separately and stored as separate matrices of connection weights. The model's responses to the trained stimuli are shown in the second column of Table 1 (*Training*) for the purposes of comparison with the results of the other simulations. Not surprisingly, the model learned to classify the patterns on which it was trained.

Generalization Across Viewpoint

We next tested the model with translated and scaled version of the stimuli on which it had been trained. Table 1 shows the model's responses to scaled, oriented and distorted versions of the training stimuli using the exclusive response criteria.

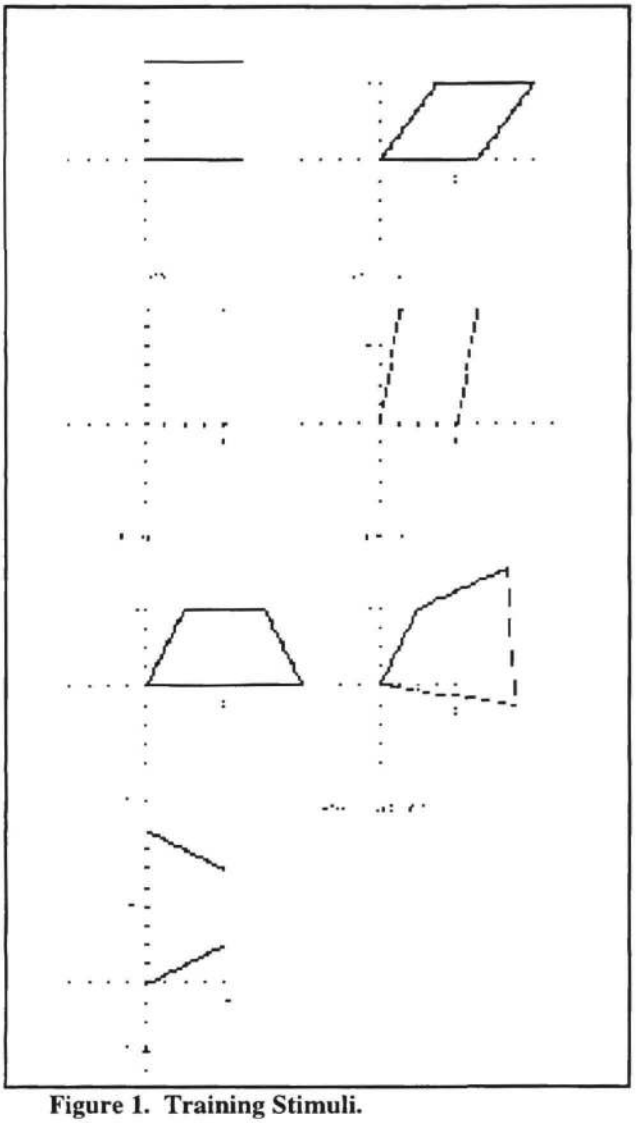

Figure 1. Training Stimuli.

Results were similarly successful in the inclusive labelling condition: generally, the model correctly assigned all relevant labels to a given test display.

These results demonstrate that the model treats the trained figures as invariants, responding equivalently to them regardless of their location in the visual field, size, or orientation. This kind of strong invariance with orientation illustrates a limitation of the current simplified model. To the human visual system, a square rotated 45 in the picture plane may be a diamond, not a square (Mach, 1897).

Tolerance for Distortion

Human shape classification includes some tolerance for distortion. A square with one side slightly too long may still look squarelike. As asymmetry increases, squareness decreases and other classifications become more probable. We do not have quantitative data about human observers' tolerance for distortion, or whether it varies by task, etc. Qualitatively speaking, however, we would expect a plausible model to accept some distortion but not too much in maintaining a shape response.

In the model, this tolerance is controlled by the parameter δ, which modulates the magnitude of response for departures from equality of given length pairs.

Table 2 shows the model's responses with the value of δ used in these simulations (.02). It can be seen that, as would be expected of human observers, small (2%) distortions that technically violate squareness are ignored. Larger deviations do change the response of the model, however. For example, a 12% lengthening of parallel sides leads the network to abandon the "square" response and classify the shape as a rectangle.

Shape	*Training*	*Scale*	*Orientation*	*Proportion*
Square	.94	.96	.94	--
Rhomb.	.93	.97	.93	.93
Rect.	.93	.93	.91	.92, .83
Par.	.91	.92	.92	.93
Trap.	.95, .95	.98	.95	.91, .45[a]
Quad.	.97	.97	.97	.34[b]

Table 1. Classification Results for Exclusive Categorization. Classification scores are shown for transfer tests involving changes in scale, orientation and proportion. Classification scores were calculated from shape network outputs as: c/(c+w), where c is the network's output for the correct response and w is the network's highest response for any incorrect response. Scale changes consisted of a doubling of all lengths. Orientation was changed by 90 deg from the training set except for the square, which was rotated 45 deg. Proportion change displays varied and also included position and orientation changes.

a The network gave a higher score to parallelogram (.51) for this display (vertex coordinates: (2,3), (5,6), (9, 8), (4, 3).)

b The network gave a higher score to parallelogram (.66) for this display (vertex coordinates: (0,0), (5,2), (7, 0), (3, 1).)

Vertex Coord.	Technically Correct Shape		Shape Net's Highest Resp.	
(0.0,0.0)(0.0,5.0) (5.1,5.1)(5.0,0.0)	QUAD	.004	SQU.	.94
(0.0,0.0)(0.0,5.1) (5.0,5.1)(5.0,0.0)	RECT.	.20	SQU.	.80
(0.0,0.0)(0.0,5.0) (5.6,5.6)(5.0,0.0)	QUAD.	.043	RHMB.	.97
(0.0,0.0)(0.0,5.4) (5.0,5.4)(5.0,0.0)	RECT.	.88	RECT.	.88

Table 2. Distortion Tolerance. (See text.)

Discussion

The shape network succeeded in learning to distinguish squares, rectangles, parallelograms and other quadrilaterals based on training with one example of each (two examples of trapezoids). It did so in both inclusive and exclusive response conditions, generally activating strongly all correct shape labels in the former condition and limiting its response to the single most specific (and correct) choice in the latter condition. The model correctly classified new instances that had no overlap in coordinates or edge lengths with the training instances. The basic comparisons built into the early layers of processing made possible learning of abstract invariants involving spatial relationships.

Although the model is simplified in a number of respects and performs a restricted shape classification, it demonstrates how combining early registration of certain relations with learning in a neural network might account for discovery of abstract invariants. The interest of the approach depends on whether the basic operations in the model are *ad hoc* manipulations to perform the task under study or are likely to be part of a basic inventory of relations in a visual "grammar." How plausible are the key ingredients here?

Finding Vertices. The model assumes the ability to locate vertices—points of contour slope discontinuity—in the visual array. This makes sense if the human visual system readily encodes such points and if they are required for a variety of tasks in visual processing. For vertices, this appears to be the case. The mere fact that these points have many names in the literature (e.g., vertices, tangent discontinuities, key points, etc.) is suggestive. Location of vertices appears to be required for many middle and high-level visual processes, such as contour interpolation (Shipley & Kellman, 1990) and object recognition (Hummel & Biederman, 1992). Even so, we do not yet have a completely satisfactory model of vertex finding. One goal of the present research is to develop a suitable method to extract vertices from luminance maps of natural images.

Distance Computations. The model computes distances between vertices. In the present work, retinal extents (rather than real lengths in the physical world) are all that are required. Evidence from a variety of perceptual tasks, including studies of size perception and size constancy, implies that visual space has a metric, and retinal as well as real extents are routinely measured.

Distance Comparisons. Comparing extents would seem to be of the essence of form classification, e.g., it defines the difference between a square and a rectangle. There is also reason to believe that humans are highly sensitive to aspect ratios in shape perception.

In short, each of the three kinds of information extracted by the early stages of the model seems to be not only a type of information potentially available to visual processing, but information that is routinely used for a variety of visual tasks. These are the kinds of information that are plausibly members of a set of basic inputs from which higher-level relationships can be constructed. The model's shape classification abilities emerged more from "off-the-shelf" components than from tools specially engineered for a limited task.

At the same time, there are limitations of the present model, and some aspects of the results reflect arbitrary simplifications. Many limitations involve the domain of shape classification. Additions to the model will be needed to encompass planar shapes of various numbers of vertices, and even more elaboration may required to perform meaningful classification of smooth forms, beginning with the simple circle, which has no vertices. These challenges may help reveal more about the grammar of shape. Attempts to organize the shape domain have a long and continuing history (e.g., Attneave, 1954; Hochberg & Brooks, 1960; Leyton, 1993), yet no system of general utility has emerged. It is possible that further development of the modeling efforts begun here, combined with research on the abstract relationships learnable in human shape classification can help clarify and constrain the grammar of shape.

Another kind of limitation of the present results involves learning. Although we believe the relations extracted in the early layers of the network are generic and not contrived, our current model has the good fortune to include only these several sorts of information. In natural circumstances, human perceptual learners must discover which among routinely computed or potentially computable basic functions will be relevant to a particular classification task. Suppose our stimulus inputs had included many more concrete and relational features, and feedback in our task was supposed to allow the system to converge on the notion of "square." At a minimum, many more examples would be needed by the network to separate the useful invariants from irrelevant variation. Even more may be needed, however. A system that registers lengths and colors, for example, and has comparisons such as equality/difference may have to learn to compare lengths rather than colors, by sampling possible comparisons or by applying previously learned strategies. Even if some comparisons are automatically computed, as in our simple model, it seems unlikely that all learnable comparisons are carried out all the time. If, as we suspect, the most advanced varieties of perceptual learning involve sensitivity to higher-order relations that are synthesized from new combinations of basic relations, then a fundamental problem will be how the search for useful new combinations is guided.

The model's current architecture is unrealistic in the sense that we postulate a separate unit for every vertex-to-vertex distance (layer 2), and every distance comparison (layer 3). (The model's input is similarly unrealistic in the sense that units are dedicated to particular vertices on the quadrilateral, rather than vertices at locations in the visual image.) These representations are spatially multiplexed, in the sense that identical properties of different entities (e.g., different lengths, length comparisons), are represented by completely separate units in the network. This architectural convention cannot be expected to scale to represent figures with arbitrary numbers of vertices. In future incarnations of the system, we intend to replace this spatial multiplexing with temporal multiplexing, allowing separate vertex coordinates, distances, and distance comparisons to be represented by the same units firing at different times (for similar ideas, and for a summary of neurophysiological support for temporal multiplexing in the visual system, see Hummel & Biederman, 1992, and the references therein).

Our prototype shape network can discover, from plausible building blocks, the abstract invariants that determine a simple shape classification. Building on this foundation, we hope to discover the visual grammar and computational processes that make possible and constrain human perceptual learning.

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*, 181-193.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94* (2): 115-117.

Chase, W. & Simon, W. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Prentice-Hall.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585-612.

Heitger, F. & von der Heydt, R. (1993). A computational model of neural contour processing: figure-ground segregation and illusory contours. *Fourth International Conference on Computer Vision.* Los Alamitos, CA, USA: IEEE Comput. Soc. Press.

Hochberg, J. & Brooks, V. (1961). The psychophysics of forms: Reversible perspective drawings of spatial objects. *American Journal of Psychology, 73*, 337-354.

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

Leyton, M. (1993). Symmetry, causality, mind. Cambridge, MA: MIT Press.

Mach, E. (1897). The analysis of sensations. English translation. New York: Dover, 1959.

Marr, D. & Nishihara, K. (1978).

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323 (6088): 533-536.

Shipley, T.F. & Kellman, P. J. (1990). The role of discontinuities in the perception of subjective contours. *Perception & Psychophysics*, **48,** (3), 259-270.

Author Note

We thank Randy Gallistel for helpful discussions of perceptual learning and shape representation. This research was supported by NSF SBR 9720410 (Learning and Intelligent Systems Program).

Correspondence should be addressed to: Philip J. Kellman, UCLA Department of Psychology, 405 Hilgard Avenue, Los Angeles, CA 90059-1563 or by email to <Kellman@cognet.ucla.edu>.