The visual identification of relational categories

Alexander A. Petrov Department of Psychology, Ohio State University, Columbus, OH, USA

Nicholas M. Van Horn Department of Psychology, Ohio State University, Columbus, OH, USA

James T. Todd Department of Psychology, Ohio State University, Columbus, OH, USA

An experiment was performed to investigate the ability of human observers to identify configural relations among three dots. Four stimulus categories were defined on the basis of whether or not the dots were arranged collinearly and whether or not the central dot was equally spaced relative to the two flanking dots. Observers were initially trained with feedback to identify these categories at a single orientation with a fixed uniform background, and then they were tested with variable orientations and backgrounds without feedback. The results revealed almost perfect generalization. We also simulated the same task using a recent feature hierarchy model (J. Mutch & D. G. Lowe, 2008) that is among the most successful for object recognition. This model performed well for fixed orientations and backgrounds, but it could not accurately identify these categories with variable orientations and backgrounds even when given training with those conditions. These findings suggest that feature hierarchy models represent the spatial relations within an image quite differently than human observers.

Keywords: perceptual organization, computational modeling, object recognition, shape and contour

Citation: Petrov, A. A., Van Horn, N. M., & Todd, J. T. (2011). The visual identification of relational categories. *Journal of Vision, 11*(12):11, 1–11, http://www.journalofvision.org/content/11/12/11, doi:10.1167/11.12.11.

Introduction

Human observers have a remarkable ability to perceive and identify the manner in which objects are arranged in space relative to one another. Figure 1 provides an instructive example of this phenomenon for three different objects presented against a random noise background. Note that it is immediately apparent that these objects are distributed in a collinear configuration and that they are equally spaced. The perceptual identities of spatial configurations are surprisingly robust because they remain invariant over a wide range of structural changes. That is to say, they can easily be identified regardless of their sizes, orientations, or positions in space. The constituent elements that make up such a pattern can include virtually anything, such as line segments, small triangles, or images of famous faces. Moreover, within broad limits, these elements can be presented on a wide variety of backgrounds without destroying their configural appearance.

What type of data structure is necessary to adequately represent the global structure of a spatial configuration? One type of representation that has become particularly popular in vision science is to encode the structure of images using histograms of low-level features, such as pixel intensities or the outputs of Gabor filters at different scales and orientations (e.g., Bergen & Adelson, 1988; Dalal & Triggs, 2005; Motoyoshi, Nishida, Sharan, & Adelson, 2007). However, this type of data structure is particularly ill suited for the identification of configural relations because it does not retain the relative spatial positions of those features. Consider an image of three collinear white dots against a black background. If the positions of the black and white pixels were randomly rearranged, the transformed pattern would no longer appear as a collinear configuration, even though its luminance histogram would be identical to that of the original image.

One potential strategy for overcoming this problem might be to employ higher order templates for extracting more complex aspects of local image structure. We know, for example, that the primate visual cortex has a hierarchical structure, in which neurons in the earliest stages behave much like Gabor filters (e.g., De Valois & De Valois, 1988; Hubel & Wiesel, 1968), whereas those farther along the ventral stream are tuned to more complex visual features and exhibit a greater degree of position invariance (e.g., Desimone, Albright, Gross, & Bruce, 1984; Tanaka, Saito, Fukada, & Moriya, 1991; see Ungerlieder & Bell, 2011, for a recent review). Some of the most successful models of object recognition have been designed to mimic this type of organization and are referred to in the literature as *feature hierarchy* models (e.g., Fukushima, 1980; Perret & Oram, 1993; Wallis & Rolls, 1997). These models can be implemented with a

doi: 10.1167/11.12.11 Received August 8, 2011; published October 17, 2011 ISSN 1534-7362

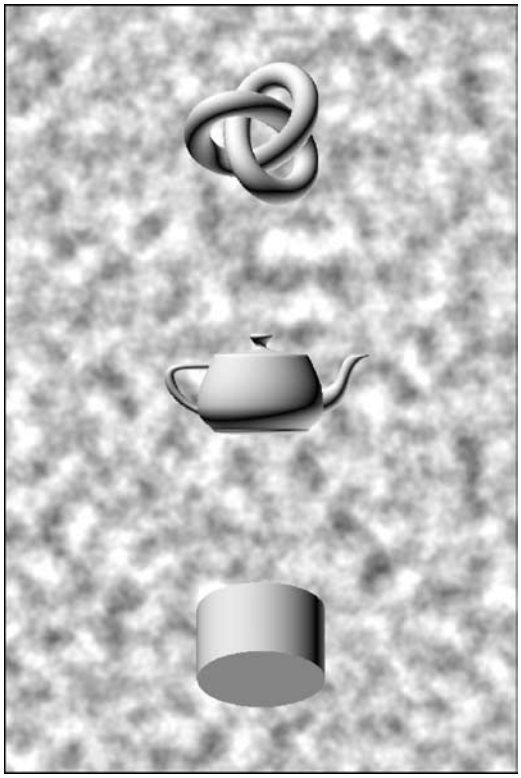

Figure 1. A pattern of three objects against a random noise background. Human observers can easily determine that these objects are distributed in a collinear configuration and that they are equally spaced.

predetermined set of higher order filters (e.g., Riesenhuber & Poggio, 1999), or they can be trained to learn a set of features from a training set (e.g., LeCun, Haffner, & Bottou, 1999; Mutch & Lowe, 2008; Serre, Oliva, & Poggio, 2007). For any given input image, the responses of these filters define a vector in a high-dimensional space, and a standard classifier such as a support vector machine is used to determine the most appropriate response from the set of categories for which the model has been trained (e.g., houses, cars, pianos, etc.). Although these models allow for some distortions of the input, it is not at all obvious how they could successfully cope with the wide range of structural variations for which human observers can identify configural categories.

The research described in the present article was designed to achieve two goals: First, to demonstrate the ability of human observers to identify configural categories in novel contexts and, second, to compare the performance of human observers with that of a recent feature hierarchy model (Mutch & Lowe, 2008) available in the public domain. We first trained observers to identify four possible configurations among three circular dots that are shown in Figure 2. These categories were defined based on two binary characteristics: (1) whether or not the dots were arranged collinearly (e.g., Westheimer & McKee, 1977) and (2) whether or not the central dot was equally spaced relative to the two flanking ones (e.g., Klein & Levi, 1985). Human observers can make these distinctions with very high acuity (see Morgan, 1991; Westheimer, 1981, for reviews), which provides strong evidence that these particular configural relations are of fundamental importance to human perception. Observers were initially trained with feedback to identify the four categories in Figure 2 at a single orientation with a fixed uniform background. They were then tested without feedback with variable orientations and backgrounds, and they exhibited almost perfect generalization. When the same task was simulated using Mutch and Lowe's (2008) model, it performed quite well for fixed orientations and backgrounds. However, it could not accurately identify these categories with variable orientations and backgrounds even when given training with those conditions.

Experiment

Methods

Participants

Twelve students from Ohio State University, naive for the purposes of the experiment, participated for course

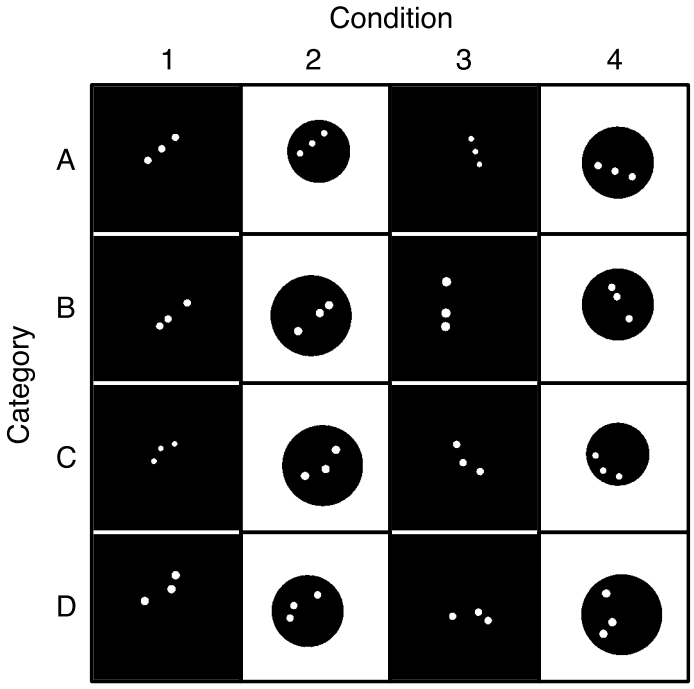

Figure 2. Sample stimuli used in the behavioral experiment and the model tests. The stimuli varied along six independent dimensions: collinearity, bisection, context, orientation, size, and position. The first two dimensions defined the categories (rows) for the 4-way classification task. The context and orientation dimensions defined 4 conditions (columns) that were used for training (1) and subsequent generalization tests (2–4). The size and position varied randomly within each block of trials.

credit. Their visual acuity was assessed by a Snellen chart and was 20/20 or better.

Stimuli and apparatus

The critical part of each stimulus was a constellation of 3 white dots (Figure 2). The stimuli varied along six independent dimensions: collinearity, bisection, context, orientation, size, and position. The first two defined category membership as follows. Category A had collinear, evenly spaced dots. Category B had collinear, unevenly spaced dots. The middle dot was offset either up or down the imaginary *constellation axis* connecting the centers of the flanker dots. Category C had non-collinear (Vernier), evenly spaced dots. The middle dot was offset either to the left or to the right of the axis. In the final category, D, the dots were neither collinear nor evenly spaced. The middle dot occupied one of four positions depending on the signs of two orthogonal offsets.

The context and orientation dimensions defined four "conditions" that were used for generalization tests. On "context" trials, the white dots were embedded in a big black circle on a white screen. On "no-context" trials, the white dots appeared on a black screen. When context and no-context trials were mixed within a block, the background luminance ramped up or down gradually during the intertrial intervals to minimize eyestrain. The orientation of the constellation axis was either fixed or variable in a block of trials. The training condition ("Condition 1") had a fixed orientation and no context. The training orientation was counterbalanced between participants: either 50 deg counterclockwise or 40 deg clockwise from vertical. The behavioral data showed no significant differences between the two groups and they were combined in the analysis. Condition 2 introduced the circular context at the trained orientation. Condition 3 tested the generalization with respect to orientation. It was similar to the training condition in that there was no context around the dots. However, the orientation varied from trial to trial and was sampled at random from a 90-degree-wide sector whose midline was perpendicular to the trained orientation. Finally, Condition 4 tested the generalization with respect to both orientation and context (Figure 2).

The last two stimulus dimensions—size and position—were sources of task-irrelevant variability. The size of each stimulus was defined relative to the length L of the imaginary segment connecting the centers of the flanker dots. It subtended L = 1.54 deg of visual angle for the largest stimuli and 0.77 deg from the smallest ones, with 4 intermediate sizes in between. Each trial was sampled independently and at random from the 6 possible sizes. All distances were measured in relative units: the dot diameter was $0.22L$, the diameter of the circular context (when present) was $2.00L$, and both bisection and Vernier offsets were $0.20L$. For example, the offset at the smallest size was (0.20)(0.77 deg) ≈ 9 arcmin, many times greater than the bisection and Vernier discrimination thresholds (e.g., Klein & Levi, 1985; Westheimer & McKee, 1977). The position of the constellation of dots varied randomly by up to $0.60L$ from the center of the circular context, left or right along the line perpendicular to the constellation axis. The edge of the circle never clipped or touched any dots. In addition, the stimulus as a whole was translated randomly by up to 0.31 deg horizontally and/or vertically relative to the fixed elements of the display.

All stimuli were generated using Matlab (The MathWorks, 2009) and the Psychophysics Toolbox (Brainard, 1997) and were presented at 96 Hz on a gamma-corrected 21″ NEC AccuSync 120 color CRT. They were viewed binocularly with the natural pupil from a chin rest fixed 93 cm away from the screen in a darkened room.

Procedure

Each participant completed 16 blocks of 32 trials each. Each block contained 8 examples from each category in a random sequence. The whole session lasted less than 1 h. The participants were instructed only that the stimuli were drawn from four different categories that remained stable throughout the experiment and that it was possible to achieve very high accuracy. The participants learned through trial and error to perform a 4-way classification task. There was feedback during the *training phase* (blocks 1–6) and no feedback during the *test phase* (blocks 7–16). All stimuli during the training phase were in Condition 1—fixed orientation and no context. The test phase assessed all 4 conditions, randomized and counterbalanced within each test block.

Four colored squares (red, green, yellow, and blue, arranged in a 2 × 2 mosaic, Figure 3) served as category labels. The assignment of colors to categories was randomized between participants but remained consistent for each individual. Each trial began with the word "Ready!" printed on the screen. The participant pressed a key to trigger the stimulus presentation. There was no time pressure—the stimulus remained visible until the end of the trial. The participant clicked on one of the colored squares to enter their classification response. During the training phase, feedback was given in three modalities: smiley faces, bonus points (visible onscreen at all times), and beeps. If the response was correct, a smiling face confirmed the chosen response and the bonus increased (Figure 3A). If the response was incorrect, a frowning face marked the chosen response and there was an unpleasant beep. Either way, the square with the correct classification remained in color while the other 3 squares were grayed out. The stimulus remained visible for 2.5 s after the mouse click, allowing the participant to associate the dot configuration with the correct category label. The feedback was discontinued during the test phase: The bonus was displayed as "XXXXX," there were no beeps, and a

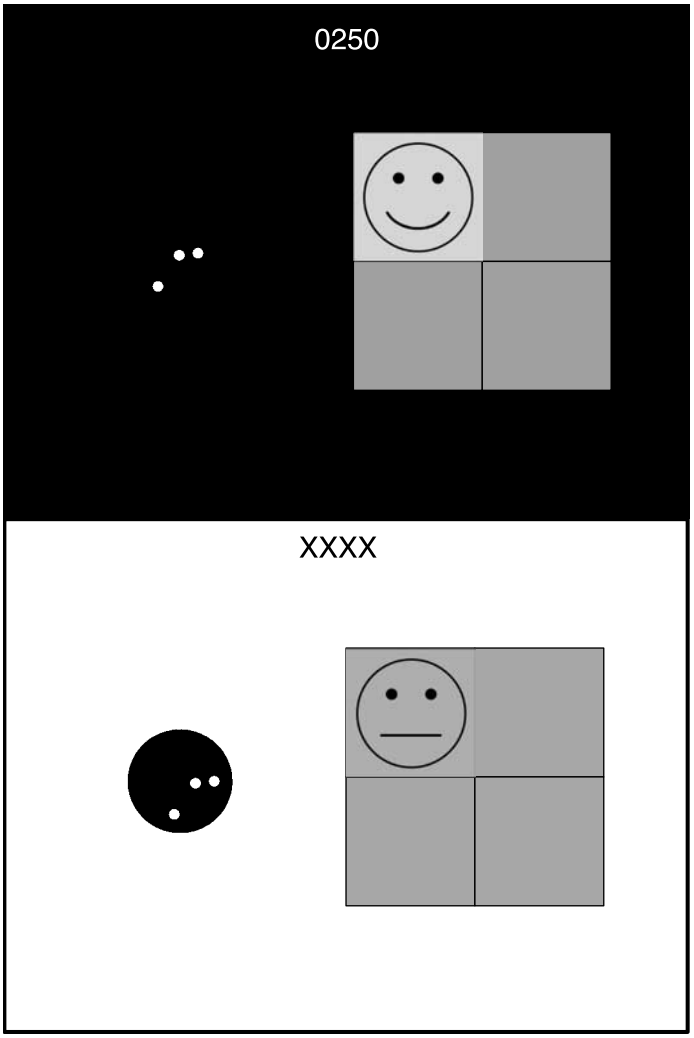

Figure 3. Schematic illustration of the experimental displays at the end of a trial with feedback and no context (top) and with no feedback and circular context (bottom). The smiley face (top) indicates that the chosen response was correct. The neutral face (bottom) confirms that a response was chosen but does not reveal whether it is correct. The four squares were colored in red, green, yellow, and blue.

neutral face confirmed all choices regardless of their correctness (Figure 3B). The stimulus still remained visible for 2.5 s after the mouse click, but all 4 squares retained their colors at all times.

Results and discussion

The participants learned the 4-way classification quickly and easily. All 12 participants showed the same pattern, which is evident in the group average in Figure 4. The mean accuracy rapidly increased during the initial training blocks (open squares) and reached near-perfect levels by the beginning of block 5—that is, after no more than 32 exemplars from each category. Several individuals reached near-perfect accuracy by the end of block 2 (i.e., less than 8 exemplars). This near-perfect performance was maintained after the feedback was discontinued in block 7 (solid squares). Moreover, there was complete transfer to untrained contexts (Condition 2, solid circles) and/or orientations (Conditions 3 and 4, solid diamonds and triangles). The classification accuracy remained robust (M = 99%) across all test conditions and all participants. These results suggest that the human visual system represents our stimuli in a manner that makes the relational information explicit, accessible, and independent of orientation and context.

Could it be that our participants relied on slow, language-mediated reasoning rather than the quick, visual processing characteristic of object recognition? We performed a follow-up experiment to evaluate this possibility. The salient feature of this experiment was that it used backward masking to control the time available for visual processing (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; Serre et al., 2007). Over the course of two sessions on separate days, the stimulus onset asynchrony (SOA) between the target stimulus and a pattern mask was reduced gradually from 3000 ms down to 250 ms. The test phase (the second half of session 2) fixed the SOA at 250 ms, discontinued the feedback, and introduced novel orientations as in the main experiment. Eleven new participants learned the 4-way classification task quickly

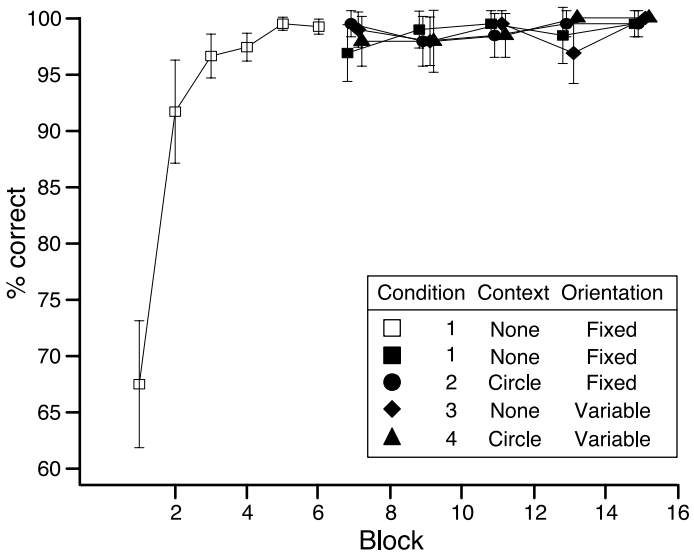

Figure 4. Behavioral data. Mean accuracy on the 4-way classification task, averaged across the 12 participants. (Each individual showed the same pattern.) The training phase (open squares) had feedback and presented stimuli at a fixed orientation and no context. The test phase (solid symbols) discontinued the feedback and introduced new orientations and contexts. Figure 2 illustrates the stimuli in the four conditions. Error bars are 90% confidence intervals (within subjects). All blocks were of equal size (32 trials). Test blocks were merged in pairs for purposes of plotting.

and easily and could perform it with near-perfect accuracy (M = 95%) at the 250-ms SOA. The full transfer to untrained orientations was replicated. Finally, one motivated and experienced observer (the second author, NVH) was tested at a range of fast SOAs. His accuracy was 0.97, 0.93, 0.88, 0.70, and 0.29 at SOAs of 125, 105, 85, 65, and 45 ms, respectively. His median response time at the 65-ms SOA was 986 ms from the stimulus onset, including the time for moving and clicking the computer mouse.

The behavioral data confirmed the phenomenological reports of our participants, who told us that this was an easy task to perform. As we shall see, however, the same task poses a serious challenge to existing feature hierarchy models.

Model tests

Our experimental task embodies one of the central challenges in vision—the need to simultaneously satisfy the contradictory objectives of invariance and specificity. On the one hand, the stimulus categories are invariant with respect to the 4 conditions in Figure 2. Thus, large variations in size, orientation, absolute position, and context must be ignored. On the other hand, displacing the middle dot by only 9 arcmin can change the classification of the stimulus. What kind of computation can extract such small differences in relative position from the image, while ignoring massive differences in absolute position, orientation, and many other dimensions? What kind of representational scheme can support such computation?

This article examines one influential and well-articulated proposal for meeting these objectives. Feature hierarchy models (e.g., Fukushima, 1980; Mutch & Lowe, 2008; Perret & Oram, 1993; Riesenhuber & Poggio, 1999; Serre et al., 2007; Wallis & Rolls, 1997) rely on disjunctive pooling operations to promote invariance and on conjunctive template-building operations to maintain specificity. The two types of operations alternate in a deep hierarchical neural network. The operations are performed in gradual, interleaved steps across multiple layers of units, so that their effects complement each other. The HMAX model (Riesenhuber & Poggio, 1999) illustrates these ideas. The units in the first layer—S1—are Gabor filters modeling the tuning properties of simple cells in the primary visual cortex (V1, Hubel & Wiesel, 1968). The second layer units—C1—pool the responses of the corresponding S1 units within a small neighborhood. This disjunctive (MAX) operation imbues the C1 units with a measure of translation invariance, similar to that of complex cells in V1. The third layer—S2—constructs higher level features (e.g., corners, junctions) as conjunctions of the simpler features (edges) in the C1 layer. They are in turn pooled across space by the C2 units in the fourth layer. Thus, receptive fields grow in size and features become more complex as one ascends the hierarchy. The ventral visual stream exhibits the same gradients (Ungerlieder & Bell, 2011). Models based on these principles achieve state-of-the-art object recognition performance with natural images on a variety of benchmark databases (see, e.g., Riesenhuber, 2009, for a review).

How would feature hierarchy models perform on our experimental task? On the one hand, there are reasons for skepticism because, in the final analysis, the stimulus representation at the top of the hierarchy is just a "bag of features" (also known as feature vector, population code, or distributed representation). No relational or configural information is represented explicitly. Given that our stimulus categories are defined in terms of configural relations, it is far from clear whether such models can handle them (cf. Fodor & Pylyshyn, 1988; Hummel & Stankiewicz, 1996). On the other hand, the models use *redundant*, *overcomplete* feature sets. That is, they are designed to contain a much greater variety of conjunctive features than are minimally necessary. Each element of the image activates multiple features of various shapes, whose receptive fields overlap and interlock. Thus, the configural information could be represented implicitly. The pieces of a jigsaw puzzle provide a useful analogy. Even though they are scrambled in a box, it is possible to reconstruct the global configuration because there is only one way to fit the pieces together with no gaps or overlaps. Analogously, the redundant, composite features at the top of the hierarchy have been proposed to constitute a "generic and universal dictionary … that can support several different recognition tasks and in particular the recognition of many different object categories" (Serre et al., 2007, p. 6428).

Given these conflicting arguments and authoritative opinions on either side of the debate, we set out to test the question experimentally. We performed a series of computer simulations with a representative feature hierarchy model (Mutch & Lowe, 2008). This particular model was chosen for three reasons. First, it is based on the so-called "standard model" (Riesenhuber & Poggio, 1999; Serre, Wolf, & Poggio, 2005) and is representative of the mainstream feature hierarchy framework. Second, it is a recent refinement that carries this framework a step further, performs competitively on various benchmarks, and its Matlab implementation is available under the GNU General Public License. Third, the model restricts the region of the visual field in which a given composite (S2) feature is searched for. Thus, the top-level (C2) features retain some positional information—a property that can be useful for our task. For these reasons, Mutch and Lowe's (2008) model seems as well suited as any model currently in existence to demonstrate what the feature hierarchy framework can do on our 4-way categorization task.

Simulation details

We used the publicly available Matlab implementation of Mutch and Lowe's (2008) model (see Appendix A for details). We deliberately treated the model as a black box and manipulated only three things: the training images, the test images, and the model parameters. The model uses a combination of unsupervised and supervised learning algorithms to extract features from a set of training images and then build a classifier on the basis of these features (Mutch & Lowe, 2008; Serre et al., 2005). We performed two parallel series of simulations that differed in the training images for the initial, unsupervised learning stage. The resulting model variants are referred to as *Model 1* and *Model 2* below. They performed very similarly.

Model 1 was trained and tested exclusively on images generated by the same Matlab routine that generated the stimuli for the behavioral experiment (Figure 2). The images were rendered on a "canvas" of size 256 × 256 pixels, but the model downsized them internally to 140 × 140 pixels. If presented on the experimental apparatus and viewed from the chin rest used by the human observers, the side of each square image would subtend 5.54 deg of visual angle. Recall that the smallest offset that could affect the classification of the experimental stimuli was 9 arcmin, which corresponded to approximately 3.8 pixels in the model. Thus, the input images had sufficient resolution for the classification task.

On a given run, the model was trained on images drawn from one designated experimental condition. Each training set contained 60 samples from each category, 240 images total. (Larger training sets led to negligible improvements in accuracy, whereas the computational cost rose steeply[1] and became prohibitive even on a computer cluster.) Then, the trained model was tested on all conditions. The test set contained 3200 new images (200 exemplars × 4 categories × 4 conditions). As the training protocol involved random sampling, the accuracy varied across runs. We ran batches of 10 runs to estimate the mean and variance of the model performance. One complete simulation included 1 such batch for each training condition—a total of 40 training runs and 160 tests.

We searched the parameter space to optimize the model performance (see Appendix A and Van Horn, 2011 for details). Briefly, very little tuning was required. With one exception, the default parameter values (Mutch & Lowe, 2008; Serre & Riesenhuber, 2004) worked well for our stimuli. The exception was the parameter that controlled the receptive field size of the Gabor filters on the first (S1) layer. Its default value (11 × 11 pixels) produced suboptimal training accuracy and suboptimal generalization to novel orientations. We used receptive fields of 27 × 27 pixels in all simulations.

The training images for Model 1 contained nothing but dots and circles. Admittedly, this is a very impoverished visual environment that does not match the rich perceptual experience of human observers. To alleviate this problem, we repeated the simulation with a model that was pretrained on natural images from the Caltech 101 database (Fei-Fei, Fergus, & Perona, 2004). This database is widely used in computer vision research and contains examples of 100 categories such as airplanes, bicycles, flowers, and butterflies. Model 2 sampled 3060 Caltech images during its unsupervised learning phase. It extracted a reusable dictionary of 4075 fuzzy C1 templates similar to that used by Mutch and Lowe (2008) and Serre et al. (2007). The pretrained Model 2 was then given 240 images generated by our Matlab routine for a designated experimental condition. The supervised learning phase then proceeded as in Model 1. Both models trained an all-pairs support vector machine (Schölkoph & Smola, 2002) that classified the images into 4 categories on the basis of the Caltech-derived features. Finally, the classification accuracy of the fully trained Model 2 was tested on all 4 experimental conditions as described above. Again, we replicated everything in batches of 10—another 40 training runs and 160 tests.

Simulation results and discussion

Table 1 reports the mean classification accuracy of Model 1. Each row represents a batch of runs trained on the same condition. The first row corresponds to the behavioral experiment—training in Condition 1 with feedback followed by testing on all conditions without feedback. The data from the first row are plotted in Figure 5A with solid circles. The open circles plot the analogous results for Model 2. The near-perfect human generalization results are also reproduced for comparison.

First of all, both models were able to learn the 4-way classification task in Condition 1. All stimuli in this condition had the same fixed axis and no context, but the dots varied in size and position. Recall also that the test set was generated independently from the training set. The models thus demonstrated successful transfer to novel

Training condition			Test condition			
	Context	Orientation	1	2	3	4
1	None	Fixed	95 ± 1	42 ± 7	32 ± 6	25 ± 7
2	Circle	Fixed	59 ± 6	75 ± 2	28 ± 4	30 ± 4
3	None	Variable	56 ± 4	25 ± 11	59 ± 2	25 ± 11
4	Circle	Variable	33 ± 8	38 ± 3	33 ± 7	38 ± 2

Table 1. Classification accuracy of Model 1 (mean ± 90% confidence intervals estimated from batches of 10 independent runs). On each run, the model was trained with stimuli from one designated condition (row) and then tested on all 4 conditions (columns). The data from the first row are plotted in Figure 5A, and the data along the diagonal are plotted in Figure 5B.

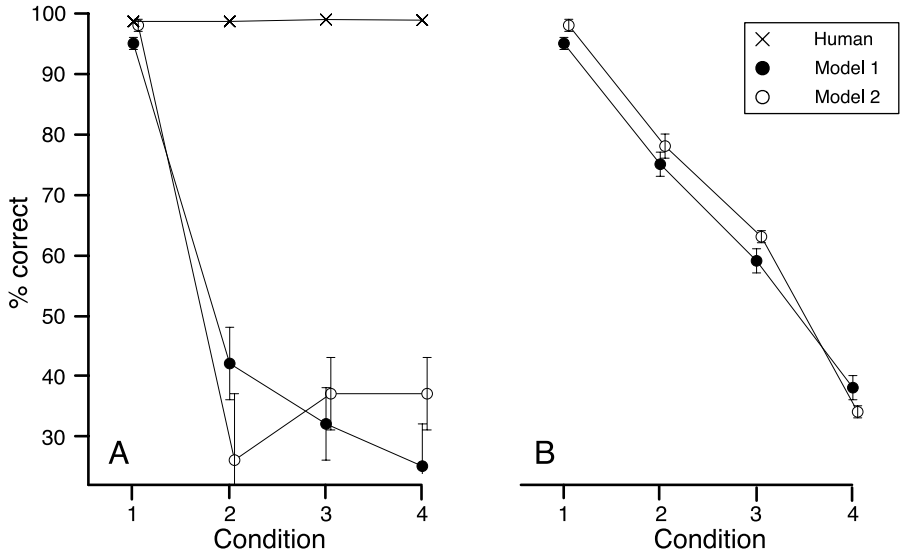

Figure 5. Simulation results. Mean accuracy on the 4-way classification task for two model variants: Model 1 was trained exclusively on the experimental stimuli, whereas Model 2 was pretrained on natural images. Panel (A) mimics the procedure of the behavioral experiment—the models were trained in Condition 1 only and then tested on all 4 conditions. The human generalization data are reproduced from Figure 4 for comparison. In panel (B), each point represents a separate model, trained in its respective condition and then tested in the same condition. Error bars are 90% confidence intervals estimated from batches of 10 independent runs.

images drawn from the same distribution, as well as across a range of sizes and positions. This result also confirmed that the model performance was not limited by the resolution of the input images.

However, the accuracy collapsed to near-chance levels on all tests that involved novel contexts and/or orientations (Figure 5A, Conditions 2–4). This contrasted sharply with the effortless generalization of human observers. It seems that the representations used by the model do not encode configural relations in the same flexible format that humans do. Yes, it is possible to encode relational information implicitly in a redundant vector of overlapping composite features. Enough relational information can be preserved in this way to classify the stimuli in Condition 1. However—and this is very important—this representation is brittle and easily disrupted by irrelevant background elements such as the circular contexts in Condition 2. It also fails to generalize to novel orientations.

Could the model learn the 4-way classification task in Conditions 2–4 with the aid of explicit training in these conditions? In other words, granted that it could not transfer as the humans did, perhaps the model could at least scale up. The relevant data are reported along the diagonal of Table 1 and plotted in Figure 5B. Model 2 (open circles) replicated the pattern. Each data point summarizes a separate batch of runs, trained in a given condition, and then tested with novel exemplars drawn from the same condition. A clear overall trend emerged—the accuracy deteriorated rapidly as the stimulus set became more diverse. For fixed-axis constellations of dots on a black background (Condition 1), the accuracy was near ceiling, but it dropped to 75% when a surrounding circle was added to the stimuli (Condition 2). Random rotations of the constellation axis within a 90-degree sector (Condition 3, black background) eroded the performance even more. Finally, when all sources of variability were present simultaneously (Condition 4), the accuracy collapsed to 38%, barely above chance. So, in the end, the model could not scale up.

The interpretation of these results is straightforward. The model's classification strategy is conceptually similar (though not identical) to table lookup. Essentially, the model accumulates a catalog[2] of image fragments and their respective category labels. This is sufficient to get the job done at first, as long as the diversity of the stimulus environment does not exceed the representational capacity of the system. The redundant, overlapping feature vectors preserve enough configural information to differentiate equal from unequal spacing and collinear from non-collinear constellations of dots along a fixed axis. Furthermore, the disjunctive pooling across space builds enough positional invariance to cope with the irrelevant variability in Condition 1. Thus, to its credit, the model successfully learned this condition. The object recognition benchmarks (e.g., Caltech 101) are analogous to Condition 1, except that they contain hundreds of categories rather than just 4. Our results are, therefore, consistent with the good performance of feature hierarchy models on such benchmark tests (e.g., Mutch & Lowe, 2008; Serre et al., 2007). However, the fuzzy templates are not orientation invariant and are affected by the irrelevant circular contours introduced by our context manipulation. This prevents generalization to novel orientations and contexts (Figure 5A).

Explicit training in Conditions 2 and 3 (Figure 5B) improves performance because it provides labeled exemplars in larger swaths of the stimulus space. However, the model fights an uphill battle that, in the end, it cannot win. As more and more irrelevant variability is introduced, the category boundaries become increasingly convoluted and tangled (DiCarlo & Cox, 2007) and it becomes increasingly difficult and inefficient to represent them as weighted feature combinations. Several mechanisms in the model are designed to cope with this problem: disjunctive pooling for positional invariance, coarse coding for greater representational capacity (cf. Cer & O'Reilly, 2006), and a support vector machine for flexible classification. These mechanisms, however, simply delay the inevitable. In the end, the model succumbs to the combinatorial explosion.

General discussion

The research described in the present article provides compelling evidence that configural categories based on collinearity and equal spacing can easily be identified over a wide range of conditions. Observers were initially trained with feedback to identify these categories at a single orientation with a fixed uniform background. When they were then tested with variable orientations and backgrounds without feedback, the results revealed almost perfect generalization. We also simulated the same task using Mutch and Lowe's (2008) feature hierarchy model. Although it performed very well for fixed orientations and backgrounds, it could not accurately identify these categories with variable orientations and backgrounds even when given training with those conditions.

These results suggest that the model uses a classification strategy that is fundamentally different and inferior to the strategy used by humans. The behavioral data and the phenomenological reports of our participants indicate that they employed an abstract classification rule expressed in terms of collinearity and equal spacing. These configural relations apparently are represented explicitly by the human visual system. This allowed the observers to induce the correct rule from just a few training examples and then generalize fluently to arbitrary contexts, orientations, sizes, and positions. By contrast, Mutch and Lowe's (2008) model can represent relations only implicitly as jigsaw puzzles of overlapping features. Although sufficient for certain constrained environments, this neither transfers nor scales up to more complex environments. The generalization failure is an inevitable consequence of the model's inability to represent relations independently of the other properties of the image (Hummel, 2003). The configural information in feature hierarchy models is inextricably mixed with irrelevant attributes such as orientation and nearby contours.

A similar pattern of results has been reported by Hayworth, Yue, and Biederman (2007). They created line drawings of objects with deleted contours in complementary pairs that had no contours in common. In a match-to-sample task, observers had no difficulty recognizing that these complementary images depicted the same object, but they could not accurately match a given image with a transformed version with randomly repositioned contours. They also simulated this same task using another feature hierarchy model by Serre et al. (2007, 2005). The results were the opposite of those obtained for humans—that is, the model was unable to match complementary images of the same object that had no contours in common, but it was able to accurately match scrambled and intact images of an object. Although Mutch and Lowe's (2008) model would have similar difficulties with complementary images, it is likely that it would be more impaired by image scrambling than the model developed by Serre et al. (2005). The reason for this is that Serre et al.'s model discards all position information at the highest level of the hierarchy, whereas Mutch and Lowe's model does not. Thus, it is able to provide a coarse representation of the relative positions of higher order features.

Recent feature hierarchy models have been designed explicitly to account for the tuning and invariance properties of neurons in the anterior inferotemporal cortex (e.g., Logothetis, Pauls, & Poggio, 1995), which are believed to be involved in object recognition, but the behavioral data suggest that these models cannot account for the encoding of configural relations. One possible reason for this is that configural relations are processed somewhere else, and there is some evidence from fMRI to support this hypothesis. For example, imaging studies of Vernier and bisection tasks have revealed that these particular relations produce significant activations in the dorsal stream along the intraparietal sulcus (Fink, Marshall, Weiss, & Zilles, 2001; Sheth et al., 2007). Other studies (e.g., Hayworth & Biederman, 2006; Hayworth, Lescroart, & Biederman, 2011) suggest that the lateral occipital complex may also be involved in the encoding of spatial relationships.

Several models of Vernier and bisection performance have been proposed in the literature (e.g., Klein & Levi, 1985; Wilson, 1986), but they would be of little use for the task in the present experiment. These models operate by comparing the overall pattern of responses for aligned and unaligned targets within a bank of filters with varying orientations and scales, and they are able to predict observers' thresholds over a wide variety of conditions. However, different patterns of filter responses are also obtained by the addition of irrelevant background contours (Morgan & Ward, 1985), changing the aligned elements from dots to lines, or altering their spacing. In order to recognize collinearity or equal spacing as general categories, it would be necessary to distinguish all of the possible patterns of filter responses that can be produced by these categories from all the patterns that cannot. This requirement is especially difficult because the number of

different stimulus configurations contained within these categories is extraordinarily large (e.g., Figure 1), and observers can identify them with a high degree of precision.

Pomerantz and Portillo (2011) have argued that some types of configural relations are special in the sense that they form emergent features that are more discriminable than their component parts. They used an odd-man-out paradigm with patterns composed of one to four dots. Their results revealed that response times were fastest when discriminating patterns that differed with respect to collinearity or symmetry. Response times were much slower, in contrast, when comparing patterns that did not differ with respect to these attributes.

Chen (1983, 2005) has proposed an interesting theoretical hypothesis that the relative perceptual salience of emergent features may be systematically related to their structural stability under change, in a manner that is similar to Klein's hierarchy of geometries (see also Todd, Chen, & Norman, 1998). According to this hypothesis, observers should be most sensitive to those aspects of an object's structure that remain invariant over the largest number of possible transformations, such that changes in topological structure (e.g., the pattern of connectivity) should be more salient than changes in projective structure (e.g., straight versus curved), which in turn should be more salient than changes in affine structure (e.g., parallel versus non-parallel). The most unstable properties of all are purely metric changes in length, orientation, or curvature. It is likely that the mechanisms for detecting the most stable emergent features have evolved over eons because they are so important for object recognition. For example, Chen, Zhang, and Srinivasan (2003) have shown that honeybees can learn the abstract concept of topological invariance in visual forms but cannot distinguish between a square and a circle.

There are many other types of emergent features that are important in natural vision. These include parallelism, symmetry, and various types of vertex structures, such as arrows, Ys, and Ts. It is interesting to note that these are the constituent elements used by Biederman (1987) for defining geons in his theory of recognition by components, but he has said very little about how they are detected in actual images. For example, in the neural implementations of this theory (Hummel, 2001; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996), they are hand-coded. Discovering the neural mechanisms for identifying these elementary configural relations in natural vision remains as a fundamental problem for future research.

Appendix A

Parameter search

All simulations with Mutch and Lowe's (2008) model used the FHLib Multiscale Feature Hierarchy Library (version v8, downloaded from http://www.mit.edu/~jmutch/fhlib/). We deliberately treated the model as a black box and manipulated only the input images and the parameters. No algorithms or equations were modified. We did not activate the sliding window mechanism used in Mutch and Lowe's localization experiments because our stimuli were relatively small. This mechanism was not activated in Mutch and Lowe's multiclass experiments (Caltech 101) either.

Given that the main goal of our simulations was to evaluate the best performance that the model can achieve with our stimuli, it was important to use appropriate parameter values. We imposed the constraint that a given simulation should utilize the same parameters in all experimental conditions. Without this constraint, it would have been very difficult to compare the results across conditions. A thorough search of the parameter space was not feasible because the simulations were time consuming. Consequently, we (Van Horn, 2011) explored one parameter at a time.

First, we doubled the number of features from 4075 to 8150. As this changed the accuracy by less than 1%, we adopted the original number of features (4075) for all subsequent simulations. This number is commonly used in the literature (Mutch & Lowe, 2008; Serre et al., 2005). Next, we evaluated a range of values for the receptive field size ("xySize") of the S1 layer. The accuracy increased as the RF sizes increased and then leveled off. For example, the default size (11 pixels) yielded 50% accuracy in one representative condition, whereas sizes greater than 21 yielded 75%. We chose a value (27) in the middle of the plateau and used it for all subsequent simulations. This was the only parameter that was changed in the end. We explored three other parameters in a similar manner, in order: within-layer inhibition ("inhibit," default = 0.5), XY tolerance of the C2 layer ("xyTol," default = 0.05), and scale tolerance of the C2 layer ("scaleTol," default = 1). Their default values could not be improved upon, and so we adopted them for our simulations as well. All in all, Mutch and Lowe (2008) seem to have accomplished their goal "to find parameters that could be used for any data set" (p. 51). Very little tuning was required to apply their model to our stimuli.

On one of the early simulations, we also explored how accuracy scaled up with the number of images in the training set. Our longest runs trained with 300 images per category, 1200 in total. The accuracy did improve with training, as expected, but the slope of the learning curve was extremely shallow—on the order of 1 percentage point per 50 exemplars per category.

Acknowledgments

We thank Jim Mutch and David Lowe for making the implementation of their 2008 model freely available.

We also thank James Pomerantz for his comments on the manuscript. This work was supported in part by NSF Grant BCS-0962119 and by an allocation of computing time from the Ohio Supercomputer Center.

Commercial relationships: none.
Corresponding author: Alexander A. Petrov.
Email: apetrov@alexpetrov.com.
Address: Department of Psychology, Ohio State University, 1827 Neil Ave, Columbus, OH 43210, USA.

Footnotes

[1]The model uses an SVM classifier. The time to train a support vector machine scales as the third power of the number of training examples in the worst case (Bottou & Lin, 2007).

[2]Technically, the SVM classifier identifies and leverages the "support vectors" near the decision boundary and disregards the exemplars in the interior. For this and other reasons, the model's classification strategy is not identical to table lookup. These technical details do not invalidate the conceptual argument in the text.

References

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorization. *Vision Research, 45,* 1459–1469.

Bergen, J. R., & Adelson, E. H. (1988). Early vision and texture perception. *Nature, 333,* 363–364.

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

Bottou, L., & Lin, C. (2007). Support vector machine solvers. In L. Bottou, O. Chapelle, D. DeCosta, & J. Weston (Eds.), *Large-scale kernel machines* (pp. 1–27). Cambridge, MA: MIT Press.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Cer, D. M., & O'Reilly, R. C. (2006). Neural mechanisms of binding in the hippocampus and neocortex: Insights from computational models. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Handbook of binding and memory: Perspectives from cognitive neuroscience* (pp. 193–220). New York: Oxford University Press.

Chen, L. (1983). Topological structure in visual perception. *Science, 218,* 699–700.

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition, 12,* 553–637.

Chen, L., Zhang, S., & Srinivasan, M. V. (2003). Global perception in small brains: Topological pattern recognition in honey bees. *Proceedings of the National Academy of Sciences of the United States of America, 100,* 6884–6889.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 1,* 886–893.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience, 4,* 2051–2062.

De Valois, R. L., & De Valois, K. K. (1988). *Spatial Vision*. New York: Oxford University Press.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11,* 333–341.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). *Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories*. Paper presented at the CVPR Workshop on Generative-Model Based Vision.

Fink, G. R., Marshall, J. C., Weiss, P. H., & Zilles, K. (2001). The neural basis of vertical and horizontal line bisection judgments: An fMRI study of normal volunteers. *Neuroimage, 14,* S59–S67.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3–71.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202.

Hayworth, K., & Biederman, I. (2006). Neural evidence for intermediate representations in object recognition. *Vision Research, 46,* 4024–4031.

Hayworth, K., Yue, X., & Biederman, I. (2007). Some tests of the standard model [Abstract]. *Journal of Vision, 7*(9):924, 924a, http://www.journalofvision.org/content/7/9/924, doi:10.1167/7.9.924.

Hayworth, K. J., Lescroart, M. D., & Biederman, I. (2011). Visual relation encoding in anterior LOC. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1032–1050.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195,* 215–243.

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

Hummel, J. E. (2003). The complementary properties of holistic and analytic representations of shape. In M. A. Peterson & G. Rhodes (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 212–234). New York: Oxford University Press.

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

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

Klein, S. A., & Levi, D. M. (1985). Hyperacuity thresholds of 1.0 second: Theoretical predictions and empirical validation. *Journal of the Optical Society of America A, 2,* 1170–1190.

LeCun, Y., Haffner, P., & Bottou, L. (1999). Object recognition with gradient-based learning. In D. A. Forsyth, J. L. Mundy, V. di Gesu, & R. Cipolla (Eds.), *Shape, contour, and grouping in computer vision. Lecture notes in computer science* (vol. 1681, pp. 319–345). Berlin, Germany: Springer.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5,* 552–563.

Morgan, M. J. (1991). Hyperacuity. In M. Regan (Ed.), *Spatial vision* (pp. 87–113). London: Macmillan.

Morgan, M. J., & Ward, R. M. (1985). Spatial and spatial-frequency primitives in spatial-interval discrimination. *Journal of the Optical Society of America, 2,* 1205–1210.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447,* 206–209.

Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision, 80,* 45–57.

Perret, D. I., & Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing, 11,* 317–333.

Pomerantz, J. R., & Portillo, M. C. (2011). Grouping and emergent features in vision: Toward a theory of basic Gestalts. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1331–1349.

Riesenhuber, M. (2009). Object categorization in man, monkey, and machine: Some answers and some open questions. In S. J. Dickinson, A. Leonardis, B. Schiele, & M. J. Tarr (Eds.), *Object categorization: Computer and human vision perspectives* (pp. 216–240). Cambridge University Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.

Schölkoph, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 6424–6429.

Serre, T., & Riesenhuber, M. (2004, November). *Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex (CBCL Paper 239/AI Memo 2004-107)*. Cambridge, MA: Massachusetts Institute of Technology.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In C. Schmid, S. Soatto, & C. Tomasi (Eds.), *Proceedings of the IEEE Computer Science Society Conference on Computer Vision and Pattern Recognition (CVPR)* (vol. 2, pp. 994–1000). San Diego, CA: ICS Press.

Sheth, K. N., Walker, B. M., Modestino, E. J., Miki, A., Terhune, K. P., Francis, E. L., et al. (2007). Neural correlate of vernier acuity tasks assessed by functional MRI (fMRI). *Current Eye Research, 32,* 717–728.

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology, 66,* 170–189.

The MathWorks (2009). *MATLAB user's guide*. Natick, MA: The MathWorks.

Todd, J. T., Chen, L., & Norman, J. F. (1998). On the relative salience of Euclidean, affine and topological structure for 3D form discrimination. *Perception, 27,* 273–282.

Ungerlieder, L. G., & Bell, A. H. (2011). Uncovering the visual "alphabet": Advances in our understanding of object perception. *Vision Research, 51,* 782–799.

Van Horn, N. M. (2011). *Limitations of using bags of complex features: Hierarchical higher-order filters fail to capture spatial configurations*. Unpublished master's thesis, Department of Psychology, Ohio State University, Columbus, OH.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194.

Westheimer, G. (1981). Visual hyperacuity. *Progress in Sensory Physiology, 1,* 1–37.

Westheimer, G., & McKee, S. P. (1977). Spatial configurations for visual hyperacuity. *Vision Research, 17,* 941–947.

Wilson, H. R. (1986). Responses of spatial mechanisms can explain hyperacuities. *Vision Research, 26,* 453–469.