



Feature Selection for Inductive Generalization

Na-Yung Yu,^a Takashi Yamauchi,^a Huei-Fang Yang,^b Yen-Lin Chen,^b
Ricardo Gutierrez-Osuna^b

^a*Department of Psychology, Texas A&M University*

^b*Department of Computer Science and Engineering, Texas A&M University*

Received 2 February 2009; received in revised form 2 April 2010; accepted 7 April 2010

Abstract

Judging similarities among objects, events, and experiences is one of the most basic cognitive abilities, allowing us to make predictions and generalizations. The main assumption in similarity judgment is that people selectively attend to salient features of stimuli and judge their similarities on the basis of the common and distinct features of the stimuli. However, it is unclear how people select features from stimuli and how they weigh features. Here, we present a computational method that helps address these questions. Our procedure combines image-processing techniques with a machine-learning algorithm and assesses feature weights that can account for both similarity and categorization judgment data. Our analysis suggests that a small number of local features are particularly important to explain our behavioral data.

Keywords: Similarity perception; Machine learning; Feature selection; Inductive generalization

1. Introduction

How do people form similarity in their minds, and how does it affect their generalization? A prevalent assumption in similarity research is that people perceive similarity by selectively weighing matching/mismatching features of objects (Pothos, 2005; Sloutsky & Fisher, 2004; Tversky, 1977; see Hahn, Chater, & Richardson, 2003; Markman & Gentner, 1993, for different approaches). But how do people select features?

Among the most successful techniques to study the psychological process of feature extraction is additive clustering (Shepard & Arabie, 1979). Additive clustering is a method of identifying latent features from clusters of stimuli that are generated from an

Correspondence should be sent to Na-Yung Yu, Department of Psychology, Mail Stop 4235, Texas A&M University, College Station, TX 77843. E-mail: nayungyu@gmail.com

$n \times n$ similarity matrix. This technique was developed to enhance psychological feasibility of the geometric model of similarity (Shepard, 1962a,b). Recently, this method has been improved drastically with machine-learning algorithms, such as Expectation Maximization (Tenenbaum, 1996), regularization functions (Lee, 1998, 1999), Bayesian Information Criterion (Lee, 2001), geometric complexity criterion (Navarro & Lee, 2004), and nonparametric Bayesian statistics (Austerweil & Griffiths, 2009; Navarro & Griffiths, 2008).

Despite these improvements, the methods have a number of drawbacks. First, the additive clustering technique relies on behavioral data to measure pairwise comparisons of similarity (which requires $n(n-1)/2$ pairs of stimuli). Because pairwise comparisons require a large number of trials, it is cumbersome to integrate the method with inductive generalization tasks (see Lee & Navarro, 2002; Zeigenfuse & Lee, 2010 for notable exceptions). Second, except for the recent work by Zeigenfuse and Lee (2008, 2010), many of the studies are based on simple stimuli (e.g., geometric figures, cartoon faces, Arabic numerals, or capital letters). For this reason, it is unclear whether conclusions drawn from these studies fare well when stimuli are more complex and naturalistic (see Carandini et al., 2005; Yuille & Kersten, 2006, for a similar argument regarding biological models of visual perception).

Because of these technical hurdles, several theoretical questions remain unanswered. Is similarity represented by a small number of discrete features (Lee & Navarro, 2002; Shepard & Arabie, 1979; Tversky, 1977), rather than by geometric multidimensional space, particularly when stimuli are natural images and their features vary continuously? What level of feature description, local brightness, spatial frequency, edges, components, or overall shape is important for similarity judgments (Ullman, Vidal-Naquet, & Sali, 2002; Yamauchi et al., 2006)? Because the additive clustering method leaves an observer to interpret the presence of latent features from extracted clusters, the technique makes it difficult to address these questions.

In this article, we present a feature selection method pertaining to a *triad-based similarity* judgment task (e.g., Gelman & Markman, 1986; Sloutsky & Fisher, 2004; Yamauchi & Yu, 2008; Yu, Yamauchi, & Schumacher, 2008) and investigate the following questions: (a) Do a small number of features explain similarities obtained from triad-based similarity judgments? If so, what mechanism mediates this phenomenon? and (b) Which features are selected for similarity judgments involving realistic visual images? In brief, our computational analysis suggests that people are likely to make generalizations by focusing on a few task-relevant features (Lee, 2001; Shepard & Arabie, 1979; Sloutsky, 2003; Tversky, 1977; Ullman et al., 2002) while paying little attention to overall face contours or a constellation of widely distributed local features. Our study further suggests that discrete feature-based similarity is likely to arise from an attention modulation process.

In what follows, we introduce two behavioral tasks—a similarity judgment task and a categorization task—and their results, followed by a description of our computational method and the theoretical implications of our analyses.

2. Behavioral study

We collected two sets of behavioral data from triad-based generalization tasks, a similarity task and a categorization task. The data from the similarity task were used exclusively for feature selection, and the data from the categorization task were used exclusively to test the generalization capacity of our feature selection method. A total of 130 undergraduate students at Texas A&M University participated in this study for course credit ($n = 72$ in the similarity task and $n = 58$ in the categorization task).

2.1. Similarity task

We collected images of 10 original animal faces, scaled them to a size of 307×307 pixels, and converted the images to grayscale with pixel values between 0 and 256. From these images, we created five pairs of animal faces. We paired faces whose appearance was reasonably similar (Fig. 1).

For each animal pair we generated 18 morphed pictures, which changed gradually from the original animal face to the other (Morph Man 4.0). For example, the source image (e.g., the original bear picture in Fig. 1) was tagged with an average of 111.4 markers ($SD = 14.22$) that defined the positions of specific facial features, such as the mouth, nose, and ears. These markers were adjusted for the positions of the corresponding features in the target image (e.g., the original fox picture in Fig. 1). Differences between the positions of corresponding markers in the source and target images were computed by “feature interpolation” morphing (Liu, 2000, p.2). Thus, the morphed pictures of the two original animal faces had different degrees of intermediate positions. We obtained 18 different morphed pictures per pair and designated the source image as the 1st picture and the target image as the 20th picture. Thus, the 18 morphed pictures of each pair (e.g., bear-fox) had 18 different degrees of similarity with the original source (i.e., the bear) (Fig. 1).

Using these morphed images, we examined how people judged the similarity between original and morphed pictures. Participants viewed two original pictures of each pair at the top of a computer monitor and one morphed picture of the pair at the bottom (Fig. 2A). Their task was to decide which original picture, left or right, was more similar to the morphed picture (i.e., a triad task). The dependent measure was the proportion of participants selecting one designated original picture (e.g., bear; the source picture) over the other (e.g., fox; the target picture). The stimuli remained on the computer screen until participants made a response by pressing one of two designated keys.

2.2. Categorization task

The materials and procedure used in the categorization task were identical to those described in the similarity task except for two points. First, the source and target pictures shown in the similarity task were replaced with the category names of the source and target pictures. For example, in the trials for bear-fox pictures, the source picture (bear) and the

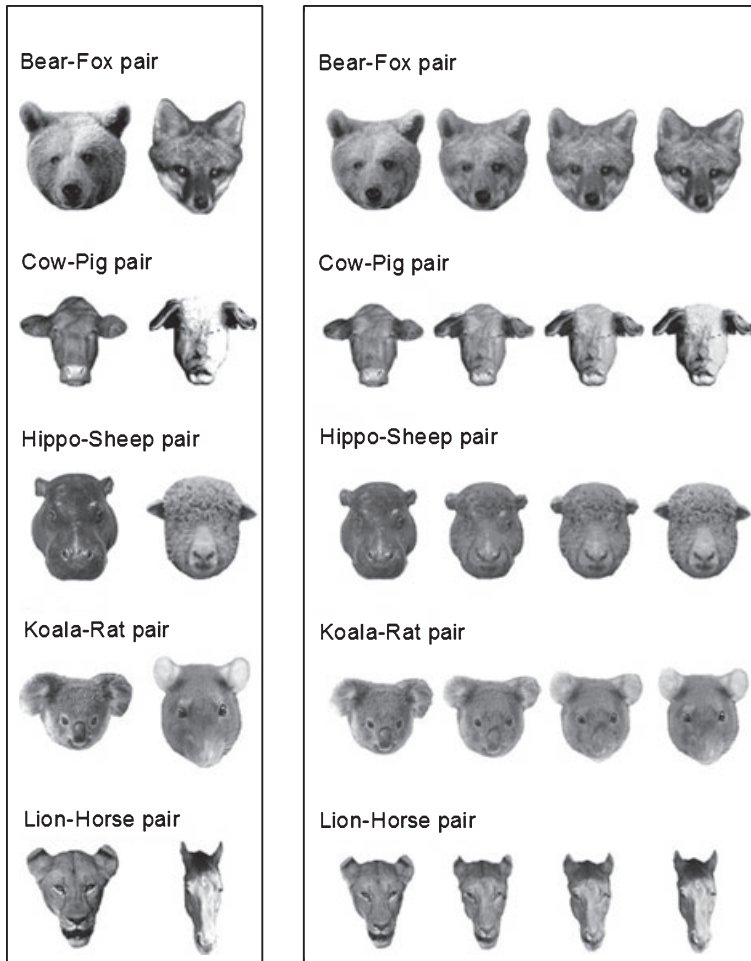


Fig. 1. Sample stimuli used in the behavioral experiment. The images shown in the left box are original animal pictures; those shown in the right box are samples of morphed images.

target picture (fox) were replaced with verbal labels “Bear” and “Fox” (Fig. 2B). Second, participants judged whether each morphed picture belonged to one of the two categories (“Is this Fox or Bear”). Except for these points, the categorization task was identical to the similarity task.

2.3. Results

Fig. 3 shows a summary of the results obtained from the two tasks. As the figures reveal, there was reasonable variability of responses in the animal pairs. These behavioral data were used for our computational analysis.

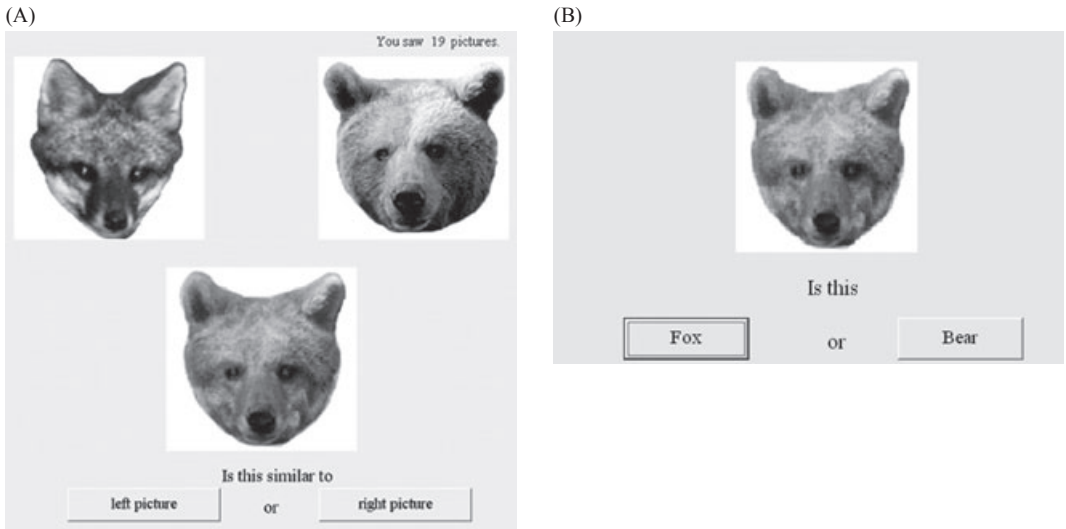


Fig. 2. A sample trial of the similarity task (A) and of the categorization task (B).

3. Computational analysis

3.1. Candidate features

To obtain a computational analog of the behavioral data, we first identified 37 potential facial features that our participants might have used. Our assumption is that the set of features that are used by subjects can be identified by fitting our model to the behavioral data. Our online-appendix (http://people.tamu.edu/~takashi-yamauchi/cog_sci/appendix.doc) describes the procedure of obtaining image features in detail. Below, we briefly summarize the 37 candidate features we preprocessed.

The candidate features included texture, brightness, size, and shape of the animal faces, which were extracted from local face areas and the entire face areas (Table 1 and Fig. 4). To obtain textural information, we computed Gabor-based textures (Manjunath & Ma, 1996) and the rate of co-occurring features (Haralick, Shanmugan, & Dinstein, 1973; Howarth & Ruger, 2004), which have been commonly used in content-based image retrieval. Brightness can also be an important feature in discriminating images. To measure brightness, we averaged the gray values of the image. To capture the relative size of the faces, we computed the ratio of width to height of individual face pictures. Finally, we computed features related to the contour of the faces by measuring the distance from the center of each face to the contour. To extract contour features, we identified the outermost pixel relative to the center of mass of the foreground in angular increments of 1 degree and stored the distance between center mass and the radially spaced outermost pixels in a 360-dimensional feature vector. Finally, we employed principal component analysis (PCA) to identify the directions of maximum variance in the 360-dimensional contour vector, and we kept the top three principal components as contour features.¹ Fig. 5 shows the resulting

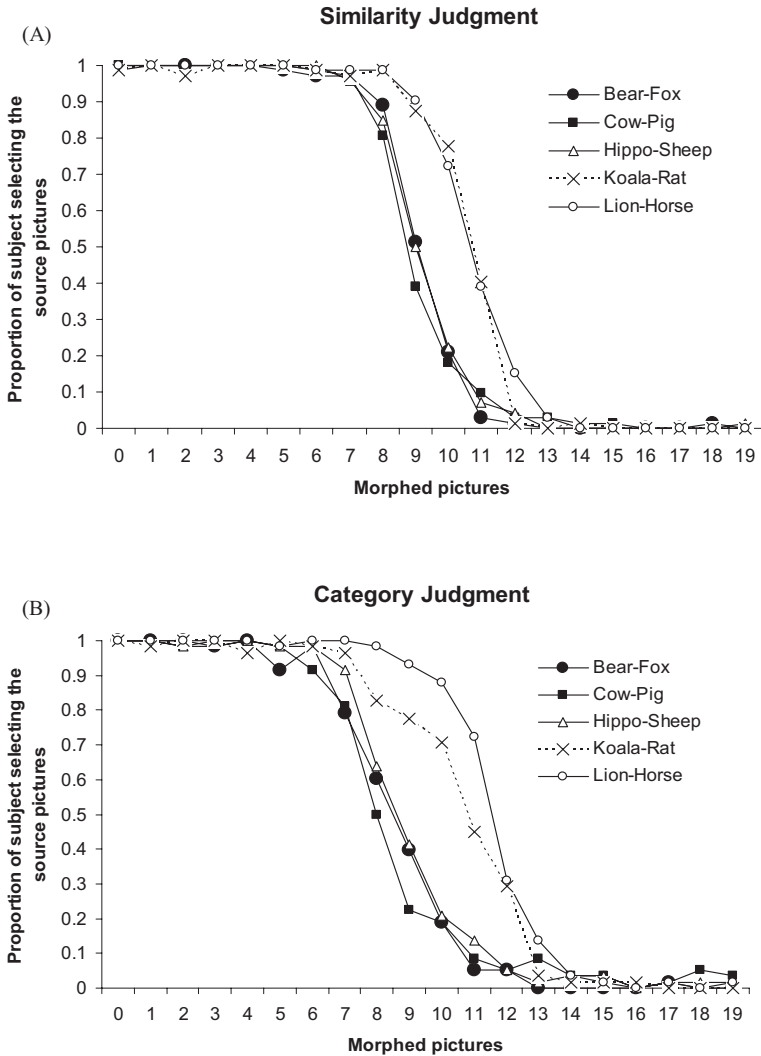


Fig. 3. A summary of the main results from the similarity task (A) and the categorization task (B). The x-axis represents the animal pictures. 0 represents the source picture; 19 represents the target picture; indices from 1 to 18 represent 18 morphed pictures of the source and target pictures. The y-axis represents proportions of participants selecting one original picture (source) as more similar to the morphed picture (input).

average contour (solid) superimposed onto the contour +3 (dotted) and -3 standard deviations (dashed) away from the average along each of the three principal components.

Because people can extract features from the entire face as well as from a part of the face (such as nose, mouth, and eyes), texture, brightness, and size features were extracted from the whole image as well as from nine subregions, which roughly corresponded to different parts of an animal’s face (Fig. 4; Edelman & Intrator, 2000; Heisele, Serre, Pontil, Vetter, & Poggio, 2002; Schyns, Bonnar, & Gosselin, 2002; Ullman et al., 2002). We defined the

Table 1
Thirty-seven candidate features

Characteristic	Description	Feature ID	Description
Holistic features	Brightness	1	Averaged grayscale value computed from the whole image
	Size	2	Ratio of width and height of the whole image
		3	The number of pixels above a threshold counted from the whole image
		4	Gabor-based texture feature extracted from the whole image
	Texture	5–7	Co-occurrence based texture feature extracted from the whole image
		8–10	The distance between the geometric center and the border of the whole face
Local features	Brightness	11–19	Averaged grayscale value computed from each subregion
	Size	20–28	The number of pixels above a threshold counted from each subregion
	Texture	29–37	Gabor-based texture feature extracted from each subregion

Note. The Feature ID was used to represent individual features in Fig. 7. Characteristic denotes whether the feature was extracted from the whole image (i.e., holistic features) or a specific part of the image (i.e., local features). The Gabor-based texture feature was computed using Gabor filters (Manjunath & Ma, 1996). Co-occurrence-based texture features were computed using gray-level co-occurrence matrices (GLCM) (Howarth & Ruger, 2004).

features extracted from the entire image as “holistic features” and the features extracted from one of the subregions as “local features,” and we examined the relative contributions of holistic and local features as well (Table 1).

We preselected these features because variants of these features have been used successfully in biologically inspired object-recognition models (e.g., Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007) and standard face recognition methods (e.g., Brunelli & Poggio, 1993; Kanade, 1973). In addition, these features are likely to be assessed during similarity judgments. For example, the human visual system processes pixel-level brightness at the retina and lateral geniculate nucleus, spatial frequency and edges (V1), implied lines (V2), and geometric components and their combinations (anterior and posterior inferotemporal cortex—TE and TEO) (Kreiman, 2007; Pessoa, Tootell, & Ungerleider, 2008;



Fig. 4. Nine subregions used for feature extraction, which roughly correspond to different parts of animal faces such as ears, eyes, and cheeks.

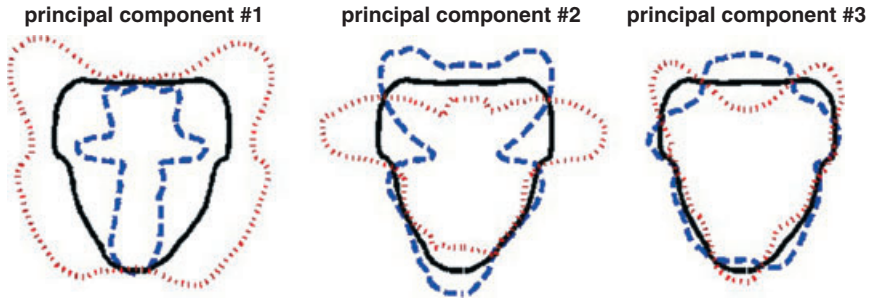


Fig. 5. The top three directions of variance on the contour of an animal's face; the top principal component captures the relative roundness of the face, whereas the second and third principal components capture the horizontal and vertical protrusion of ears, respectively.

Quiroga, Reddy, Kreiman, Koch, & Fried, 2005; Riesenhuber & Poggio, 1999; Tanaka, 1993; Ullman et al., 2002). Correspondingly, 2D Gabor filters are known to simulate the activity of simple cells in the primary visual cortex of cats (Jones & Palmer, 1987), geometric face features such as the width and height of a face are among commonly processed facial features for automated face recognition systems, and direct gray level comparisons of rectangular fragments are shown to be effective for object classification (Ullman, 2001; Ullman et al., 2002).

3.2. Selecting salient features

Because the behavioral task was to judge whether the input picture was similar to the source picture or to the target picture, we adopted Nosofsky's Generalized Context Model (i.e., GCM: Nosofsky, 1984, 1986; Nosofsky & Zaki, 1998) because of its success in accounting for performance for perceptual categorization, stimulus identification, and recognition tasks. To model the behavioral data, we first used the weighted Minkowski metric and measured the distance between an input image and a source image $d(X^i, S)$, and the distance between an input image and a target image $d(X^i, T)$.

$$d(X^i, S) = c \left[\sum_{j=1}^{37} w_j |x_j^i - s_j|^r \right]^{1/r} \tag{1A}$$

$$d(X^i, T) = c \left[\sum_{j=1}^{37} w_j |x_j^i - t_j|^r \right]^{1/r} \tag{1B}$$

where X^i denotes the 37-dimensional feature vector of the i -th input image ($X^i = [x_1^i \ x_2^i \ \dots \ x_{37}^i]$), S denotes the feature vector from the source image ($S = [s_1 \ s_2 \ \dots \ s_{37}]$), and T is the feature vector from the target image ($T = [t_1 \ t_2 \ \dots \ t_{37}]$). The index j stands for feature ID (Table 1), whereas the index i indicates an image in a given animal pair ($i = 1$ is the source, $i = 20$ is the target image, and $i = 2-19$ are morphed

images). The parameter c is a scale parameter and represents the overall discriminability of stimuli ($0 \leq c < \infty$). Because our focus was to find appropriate weight distributions for the 37 candidate features, we fixed the scale parameter (either $c = 1$ or 2) and measured the distance using the city-block ($r = 1$) or Euclidean metric ($r = 2$) in our actual analyses (the implications of fixing these parameters are discussed later in this section). Thus, to determine the relative salience of the 37 features, we treated the weight vector W ($W = [w_1 \ w_2 \ \dots \ w_{37}]$, $0 \leq w_j \leq 1$, $\sum w_j = 1$) as the sole free parameter and adjusted the vector such that our computational measure of similarity between face images paralleled the similarity judgment data obtained in the behavioral study.

We adopted an exponential decay function to link stimulus similarity to psychological distance (Nosofsky, 1986; Shepard, 1986, 1987).

$$\eta_{X^i S} = e^{-d(X^i, S)} \quad (2A)$$

$$\eta_{X^i T} = e^{-d(X^i, T)} \quad (2B)$$

Our behavioral data represented the probability that one input image (X^i) was judged to be more similar to the source (S) than to the target (T) (Fig. 2A). To simulate participants' probability scores, we transformed the distances in Eqs. 2A and 2B into an estimated probability measure $\overline{\text{Pr}}(S|X^i)$ by applying Luce's choice model:

$$\overline{\text{Pr}}(S|X^i) = \frac{b(\eta_{X^i S})^p}{b(\eta_{X^i S})^p + (1-b)(\eta_{X^i T})^p} \quad (3)$$

where parameter p denotes the power function of the measured distances (Luce, 1963, p. 113; Maddox & Ashby, 1993, p. 54; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994, p. 359). Parameter b captures subjects' bias in selecting a source or target picture, which we defined as 0.5 ($b = [1-b] = 0.5$) with the assumption that there is no a priori bias in favoring either a source or target picture. Our goal was then to identify the pattern of feature weights that minimized the sum of squared errors (SSE) between the estimated similarity (i.e., $\overline{\text{Pr}}(S|X^i)$ in Eq. 3) and behavioral data (i.e., $\text{Pr}(S|X^i)$ obtained from the similarity judgment task:

$$\text{SSE} = \sum_{i=1}^{20} [\overline{\text{Pr}}(S|X^i) - \text{Pr}(S|X^i)]^2 \quad (4)$$

To identify such a weight vector we employed simulated annealing (SA) (Kirkpatrick, Gelatt, & Vecchi, 1983). Our SA algorithm started by assigning 37 random numbers to the 37-dimensional weight vector (W in Eq. 1), the similarity between the input and target vectors was estimated, and the SSE was calculated for each animal pair. From the 37 features, one feature was randomly selected and its weight was randomly increased or decreased by a fixed amount u (i.e., update parameter), resulting in a new 37-dimensional weight vector. Using the new vector, a new SSE was computed as defined in Eq. 4. If the new SSE was larger than the previously estimated SSE, the updated weight vector was accepted (i.e., an uphill move) with the following probability,

$$\text{Pr} = e^{-\Delta/T} \quad (5)$$

where Δ is the difference between the new and previous SSEs, and T denotes the “temperature” parameter. Following Monticelli, Romero, and Asada (2008), the temperature parameter was initially set so that the probability of uphill moves was approximately .85, and the temperature parameter was reduced gradually in order for the probability of an uphill move to decline as the search process continued. The SA algorithm was implemented as outlined by Duda, Hart, and Stork (2001, p. 353).

3.3. Search protocols

In each animal pair, SA was run for 10,000 iterations. This process, which we called a “run,” was repeated 368 times, each with a different setting {46 different update parameter values ($u = [0.001, 0.01]$ in increments of 0.002), two scale parameter values ($c = 1$ or 2 ; Eq. 1), two different power parameter values ($p = 1$ or 2 ; Eq. 3), and two different metrics ($r = 1$ —city-block, or $r = 2$ —Euclidean metric; Eq. 1) = 368 runs for each animal pair}. The city-block metric generally corresponds to feature dimensions that are psychologically separated, while the Euclidean metric corresponds to features with psychologically integrated dimensions (Attneave, 1950; Garner, 1970; Goldstone & Son, 2005). For deterministic responses, a large power parameter p (>1) is usually favored (Nosofsky et al., 1994). The scale parameter c represents the discriminability of stimuli. For stimuli that are easily discriminable, a large c is given. Because the goal of this study is to identify feature weights for similarity judgments, we fixed these parameters as constant and analyzed their effects by repeating the SA algorithm in the 368 different settings.

3.4. Predictions

The presence of a discrete feature-based similarity representation (i.e., Lee & Navarro, 2002; Shepard & Arable, 1979) can be revealed by selected weights concentrated on a small number of features (two or three features), rather than widely distributed weights. If holistic features are preferred over local features, the features representing contours, overall size, and overall spatial frequency features should have larger weights (feature IDs 1–10). If local features are favored, the features obtained from local face areas (IDs 11–37) should collect larger weights.

4. Results

4.1. Results

The upper panel of Table 2 shows a summary of SSE scores obtained from the similarity task. Overall, the city-block and Euclidean metrics produced analogous results. The scale

Table 2
Average sum of squared errors^a

Metrics	<i>c</i>	<i>p</i>	Bear-fox	Cow-pig	Hippo-sheep	Koala-rat	Lion-horse
Similarity data							
City block	1	1	0.571	0.309	0.723	0.808	0.249
City block	1	2	0.129	0.052	0.199	0.181	0.012
City block	2	1	0.128	0.053	0.199	0.181	0.013
City block	2	2	0.007	0.012	0.011	0.014	0.001
Euclidean	1	1	0.617	0.356	0.722	0.813	0.275
Euclidean	1	2	0.149	0.067	0.218	0.193	0.017
Euclidean	2	1	0.149	0.067	0.217	0.194	0.017
Euclidean	2	2	0.008	0.012	0.013	0.017	0.001
Categorization data							
City block	1	1	0.386	0.302	0.509	0.511	0.236
City block	1	2	0.087	0.147	0.093	0.047	0.149
City block	2	1	0.087	0.146	0.092	0.047	0.149
City block	2	2	0.107	0.158	0.033	0.070	0.174
Euclidean	1	1	0.410	0.339	0.503	0.507	0.259
Euclidean	1	2	0.091	0.152	0.114	0.045	0.147
Euclidean	2	1	0.091	0.152	0.114	0.045	0.146
Euclidean	2	2	0.102	0.155	0.031	0.055	0.175

^aCalculated over 46 runs.

parameter *c* and power parameter *p* influenced SSE scores considerably. However, as Fig. 6A shows, analogous sets of features were selected under different settings.

Figs. 7 and 8 illustrate the selected salient features. Note that these features were selected by fitting the model to the behavioral data obtained from the similarity task only. The two most salient dimensions for the bear-box pairs were the brightness of their right cheeks and the Gabor texture of their heads. For the cow-pig pairs, the two most salient dimensions were the relative size of the center (around the eyes and nose ridges) of the pairs and the Gabor texture of the heads, respectively. For the hippo-sheep pairs, the brightness of their heads, the center of the faces, as well as the Gabor texture of the right ears were important. For the koala-rat pairs, the two most salient dimensions were the size of the right ears and the texture of their right cheeks. For the lion-horse pairs, the size of the right ears and the size of their nose ridges were salient.

Our analysis suggests that, given our stimulus set, a small number of local features were more salient than holistic features or constellations of widely distributed features. The dominant features were almost always local features and were confined narrowly to two or three features (Fig. 7). For example, nearly 94% of the weights were given to two local features in both the bear-fox pairs and the koala-rat pairs. A similar trend was present in the hippo-sheep pairs as well as in the cow-pig pairs. In the lion-horse pairs, our algorithm selected some holistic features, but the weights of these holistic features were not dominant. Note that other nonselected features were also highly diagnostic in dividing the animal pairs. For example, some holistic features were highly correlated with the morphing steps (Fig. 7), but these features were not selected. These results suggest that our participants were most

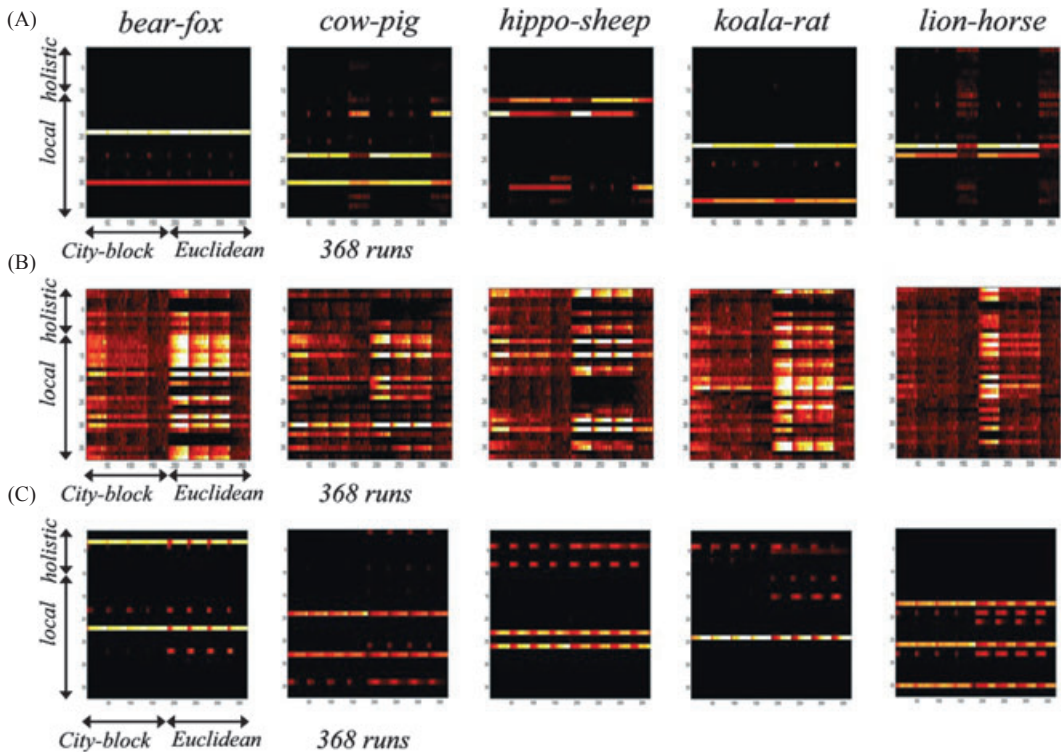


Fig. 6. (A) These heat maps show the best weight distributions identified for each of the 368 runs. The x -axis represents 368 runs made in eight different parameter settings (city block/Euclidean, $c = 1/c = 2$, $p = 1/p = 2$). The y -axis represents 37 feature dimensions. As the illuminations of the maps illustrate, a small number of local features were selected for each animal pair under the 368 different parameter conditions. (B) However, when the normalization function ($w_i = 1/\sum w_i$) was removed, widely distributed features were selected. (C) When the same image features were fitted to the normalized morphing values (1–20 morphing steps were converted to a 0–1 scale) rather than actual behavioral data, our algorithm selected some holistic features in bear-fox, hippo-sheep, and koala-rat pairs. This suggests that the feature selection shown in (A) was not just “artifacts” of our analysis procedure.

likely to make similarity judgments based on narrowly defined local characteristics rather than overall characteristics of morphed images. The next subsection discusses the psychological implications of these results.

4.2. Psychological implications

4.2.1. Generalization capacity: Similarity and categorization

Previous research has shown that there is a close relationship between similarity and categorization judgments (Nosofsky, 1984, 1986, 1989; Yamauchi & Markman, 1998; Yamauchi & Markman, 2000; Yamauchi, Love, & Markman, 2002). In fact, Nosofsky’s GCM was originally developed to account for perceptual categorization performance. Thus,

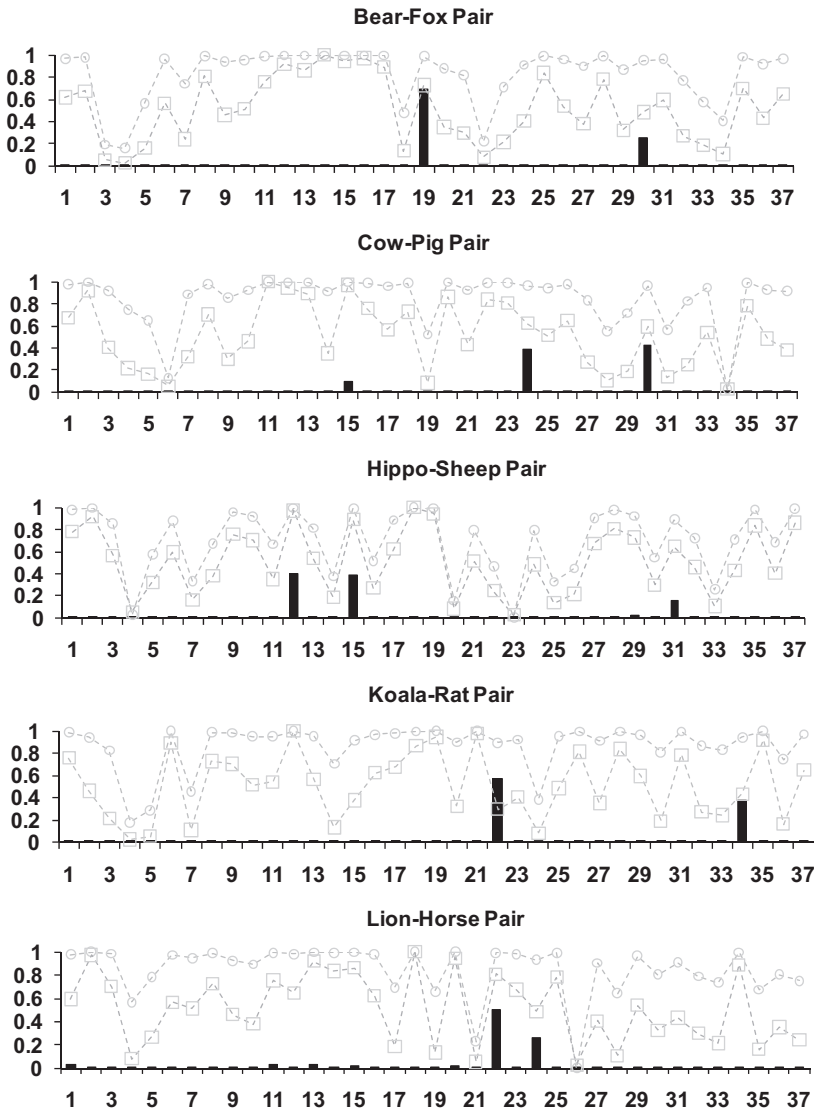


Fig. 7. Graphical representation of weight vectors. The x-axis represents the indices of the individual features listed in Table 1; features 1 through 10 are holistic, whereas features 11 through 37 are local. The black bars indicate the selected features with the y-axis representing their normalized weights. To examine the diagnosticity of “nonselected features,” we calculated the absolute values of the correlation coefficients between individual feature values and the indices of morphing steps (1–20) (shown with circle markers). For example, to create bear-fox stimuli, we generated 18 morphed pictures, which changed gradually from the original bear face (index 1) to the original fox face (index 20). Thus, the morphing steps are highly indicative of two classes of animal faces (e.g., bear and fox), and the diagnosticity of each image feature can be estimated by the absolute correlation between the morphing values (1–20) and the feature values extracted from 20 stimuli of each animal pair. Note that many of the 37 features (including holistic features) are highly correlated with the morphing values, suggesting that these features are diagnostic in separating animal pairs. Square markers show the rank order of the correlation scores of the 37 features in increasing order (from low to high correlation).

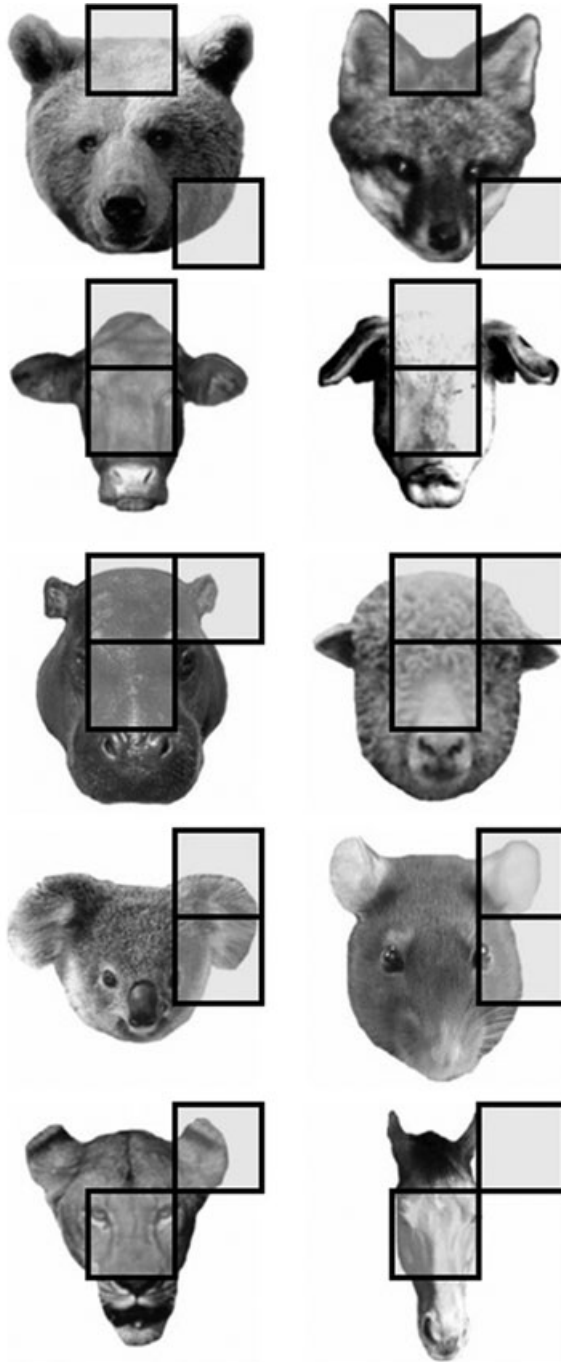


Fig. 8. Locations of salient features selected for each animal pair (see also Fig. 4).

if our feature selection method is valid, the feature weights that we identified from the similarity data should be able to explain categorization data reasonably well.

The lower panel of Table 2 summarizes the results of this analysis. Here, all weights were obtained from the similarity data, and SSEs between the estimated categorization performance (Eq. 3) and the actual categorization performance was calculated directly for each animal pair without adjusting any parameters. Overall, our selected feature weights were able to account for categorization data and similarity data equally well. Given the cow-pig and lion-horse pairs, the average SSEs from the categorization data were significantly larger than the average SSEs from the similarity data; cow-pig pairs, $t(734) = 10.14$, $p < .001$; lion-horse pairs, $t(734) = 19.14$, $p < .001$. However, for all the other animal pairs, the average SSEs from the categorization data were significantly smaller than the average SSEs from the similarity data; bear-fox pairs, $t(734) = 3.68$, $p < .001$; hippo-sheep pairs, $t(734) = 6.03$, $p < .001$; koala-rat pairs, $t(734) = 7.10$, $p < .001$. This result suggests that the salient features identified by our method were general enough to explain both similarity and categorization data.

4.2.2. Local features vs. holistic features

People tend to perceive human faces in a holistic manner (Bukach, Gauthier, & Tarr, 2006; McKone, Kanwisher, & Duchaine, 2007), while local image fragments are important for object classification (Edelman & Intrator, 2000; Ullman, 2006). To investigate the relative significance of local and holistic features, we compared the performance of the local feature weights (ID 11–37 in Table 1) to that of holistic feature weights (ID 1–10 in Table 1). We first divided the 37 feature into two subsets—holistic features (ID 1–10) and local features (ID 11–37), and then applied the SA algorithm in the same manner described above and compared how well the two sets of feature weights could be generalized to explain the categorization data. Again, the feature weights were identified first from the similarity data, and they were applied directly to the categorization data without modifying any parameters.

The local feature weights explained the categorization data better than the global feature weights in bear-fox [$t(734) = 31.76$, $p < .001$], cow-pig [$t(734) = 58.10$, $p < .001$], hippo-sheep [$t(734) = 16.81$, $p < .001$], and koala-rat [$t(734) = 8.16$, $p < .001$] pairs. The difference between local and holistic features was not significant in the lion-horse pairs; [$t(734) = 0.73$, $p < .1$] (Table 3). These results indicate that local features selected by our method were, relative to our holistic features, general enough to explain both similarity and categorization data.

4.2.3. Attention modulation and concentrated feature weights

Sloutsky (2003) argues that the basic mechanism of inductive generalization is attention allocation (see also Goldstone, 1998; Schyns, Goldstone, & Thibaut, 1997). One interesting aspect of attention is modulation: When a particular feature is attended, attention enhances the awareness of that feature, while reducing the awareness of other features (Boyton, 2005; Treue, 2003). In our model, the attention modulation process is implemented by the normalization factor ($w_i = 1/\sum w_i$) (Heeger, 1993; Heeger, Simoncelli, & Movshon, 1996). This

Table 3

Average sum of squared errors obtained from the categorization data^a

Metrics	<i>c</i>	<i>p</i>	Bear-fox	Cow-pig	Hippo-sheep	Koala-rat	Lion-horse
27 local features							
City block	1	1	0.377	0.298	0.498	0.502	0.235
City block	1	2	0.085	0.147	0.091	0.046	0.150
City block	2	1	0.085	0.147	0.091	0.045	0.150
City block	2	2	0.112	0.156	0.034	0.073	0.175
Euclidean	1	1	0.402	0.335	0.494	0.497	0.257
Euclidean	1	2	0.089	0.151	0.112	0.043	0.147
Euclidean	2	1	0.089	0.151	0.112	0.043	0.148
Euclidean	2	2	0.104	0.153	0.032	0.058	0.175
10 holistic features							
City block	1	1	2.242	1.819	1.333	0.689	0.291
City block	1	2	1.226	1.336	0.475	0.202	0.130
City block	2	1	1.226	1.336	0.475	0.201	0.130
City block	2	2	0.372	0.761	0.087	0.019	0.174
Euclidean	1	1	2.258	1.825	1.334	0.690	0.293
Euclidean	1	2	1.257	1.358	0.476	0.270	0.129
Euclidean	2	1	1.258	1.359	0.476	0.270	0.129
Euclidean	2	2	0.406	0.814	0.096	0.026	0.172

^aCalculated over 46 runs.

process ensures that an enhanced weight of a particular feature necessarily reduces the weights of the other features. If the attention modulation process was responsible for selective weight allocations, removing the normalization factor should simultaneously remove selective weight allocations. As Fig. 6B reveals, without the normalization factor ($w_i = 1/\sum w_i$), selected feature weights were distributed widely, implying that the attention modulation mechanism is likely to play a key role in the generation of discrete feature-based similarity representation.

5. Discussion

Inductive judgments involving visual images go through hierarchical processes. In the ventral visual pathway, neurons show progressively longer latencies and larger and more complex receptive fields (Kreiman, 2007; Pessoa et al., 2008; Riesenhuber & Poggio, 1999). In this process, people use information that is most diagnostic for the task at hand (Schyns et al., 2002). When they have limited prior knowledge, their best strategy is to follow the principle of least commitment (Goldstone & Son, 2005; Kersten & Yuille, 2003; Marr, 1982) by holding judgments until enough evidence is accumulated for decision making (Ratcliff, Van Zandt, & McKoon, 1999). Attention enhances the awareness of the diagnostic features, and increased attention decreases the awareness of other nondiagnostic features. It is likely that the feature-based similarity representation emerges as a natural out-

come of the attention process, and our SA method offers a viable tool to capture the salient features used for similarity judgment.

Our SA method can be viewed as a complement to the additive clustering methods developed by Lee and colleagues. One advantage of our method is simplicity. To implement the additive clustering methods, at least $n \times (n-1)/2$ trials are needed. Thus, the additive clustering methods are cumbersome when the number of the stimuli in an experiment exceeds 50 (in this case at least 2,526 trials are needed). Implementing model-specific constraints such as normalization of weight parameters ($w_i = 1/\sum w_i$) is relatively straightforward in our method. However, implementing model-specific constraints is not always straightforward for the search algorithms that employ elaborate procedures such as Bayesian analyses based on Markov Chain Monte Carlo methods (Lee, 2008).

The simplicity of our method is also liability. Our SA method is a feature-selection method, not a feature-extraction method, and before applying our method, candidate features should be preselected. This means that the result from our procedure depends on how candidate features are determined. Given these pros and cons, the two types of feature identification methods can be used together along with other behavioral methods such as eye tracking.

Note

1. These three principal components explain 92% of the total variance in the contour features.

Acknowledgments

Part of this research formed Na-Yung Yu's Texas A&M University doctoral dissertation. The author would like to thank Wookyoung Jung and Daniel Navarro for helpful comments. The data described in this paper were presented at the 6th International Conference of the Cognitive Science Society in the Asian-Pacific region.

References

- Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, *63*, 516–556.
- Austerweil, J., & Griffiths, T. L. (2008). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, Y. Bengio, D. Schuurmans, & L. Bottou (Eds.), *Advances in neural information processing systems*, Vol. 21. (pp. 97–104). Cambridge, MA: MIT Press.
- Boynton, G. M. (2005). Attention and visual perception. *Current Opinion in Neurobiology*, *15*, 465–469.
- Brunelli, R., & Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(10), 1042–1052.
- Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: The power of an expertise framework. *TRENDS in Cognitive Sciences*, *10*, 159–166.

- Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant J. L., & Rust N. C. (2005). Do we know what the early visual system does? *The Journal of Neuroscience*, 25(46), 10577–10597.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.
- Edelman, S., & Intrator, N. (2000). (Coarse coding of shape fragments) + (Retinotopy) = Representation of structure. *Spatial Vision*, 13, 255–264.
- Garner, W. R. (1970). The stimulus in information processing. *American Psychologists*, 25, 350–358.
- Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23, 183–209.
- Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49, 585–612.
- Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 13–36). New York: Cambridge University Press.
- Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, 87(1), 1–32.
- Haralick, R. M., Shanmugan, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3, 610–621.
- Heeger, D. J. (1993). Modeling simple-cell direction selectivity with normalized half-squared, linear operators. *Journal of Neurophysiology*, 70(5), 1885–1898.
- Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of National Academy of Sciences USA*, 93, 623–627.
- Heisele, B., Serre, T., Pontil, M., Vetter, T., & Poggio, T. (2002). Categorization by learning and combining object parts [electronic version]. *Advans in Neural Inoformation Processing Systems*, 2, 1239–1245.
- Howarth, P., & Ruger, S. (2004). Evaluation of texture features for content-based image retrieval. Paper presented at the International Conference on Image and Video Retrieval.
- Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58, 1233–1258.
- Kanade, K. (1973). *Picture processing system by computer complex and recognition of human faces*. Unpublished PhD thesis, Kyoto University.
- Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13, 150–158.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.
- Kreiman, G. (2007). Single unit approached to human vision and memory. *Current Opinion in Neurobiology*, 17, 471–475.
- Lee, M. D. (1998). Neural feature abstraction from judgments of similarity. *Neural Computation*, 10, 1815–1830.
- Lee, M. D. (1999). An extraction and regularization approach to additive clustering. *Journal of Classification*, 16, 255–281.
- Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, 45, 131–148.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychological Bulletin & Review*, 15, 1–15.
- Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9, 43–58.
- Liu, M. (2000). Morphing. *Instructional Technology Program at the University of Texas at Austin*. Available at <http://www.edb.utexas.edu/multimedia/PDFfolder/Morphing.pdf> Accessed October 13, 2007.
- Luce, R. D. (1963). Detection and recognition. In R. D. Luce (Ed.), *Handbook of mathematical psychology*, Vol. 1. (pp. 103–190). New York: Wiley.
- Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, 53, 49–70.
- Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18, 837–842.

- Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25, 431–467.
- Marr, D. (1982). The philosophy and the approach. In D. Marr (Ed.), *Vision* (pp. 3–38). New York: W.H. Freeman and Company.
- McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *TRENDS in Cognitive Sciences*, 11, 8–15.
- Monticelli, A. J., Romero, R., & Asada, E. N. (2008). Fundamentals of simulated annealing. In K. Y. Lee & M. A. El-Sharkawi (Eds.), *Modern heuristic optimization techniques* (pp. 123–146). Hoboken, NJ: John Wiley & Sons.
- Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, 20, 2597–2628.
- Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*, 11, 961–974.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104–114.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39–57.
- Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, 45, 279–290.
- Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22, 352–369.
- Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, 9, 247–255.
- Pessoa, L., Tootell, R. B. H., & Ungerleider, L. G. (2008). Visual perception of objects. In L. R. Squire, D. Berg, F. Bloom, S. Du Lac, A. Ghosh, N., & C. Spitzer. (Eds.), *Fundamental neuroscience*, 3rd ed. (pp. 1067–1228). San Diego, CA: Academic Press.
- Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, 28, 1–49.
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature Neuroscience*, 435, 1102–1107.
- Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, 106, 261–300.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2, 1019–1025.
- Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, 35, 402–409.
- Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1997). The development of features in object concepts. *Behavioral and Brain Sciences*, 21, 1–54.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 411–426.
- Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27(2), 219–246.
- Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27, 125–140.
- Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General*, 115, 58–61.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820), 1317–1323.
- Shepard, R. N., & Arable, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2), 87–123.

- Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, 7, 246–558.
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, 133(2), 166–188.
- Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, 262, 685–688.
- Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* 8, Vol. 8. (pp. 3–9). Cambridge, MA: MIT Press.
- Treue, S. (2003). Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology*, 13, 428–432.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews of Neuroscience*, 2, 717–726.
- Ullman, S. (2006). Object recognition and segmentation by a fragment-based hierarchy. *TRENDS in Cognitive Sciences*, 11, 58–64.
- Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7), 682–687.
- Yamauchi, T., Cooper, L. A., Hilton, H. J., Szerlip, N. J., Chen, H. C., & Barnhardt, T. M. (2006). Priming for symmetry detection of three-dimensional figures: Central axes can prime symmetry detection separately from local components. *Visual Cognition*, 13, 363–397.
- Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 28(3), 585–593.
- Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124–148.
- Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26(3), 776–795.
- Yamauchi, T., & Yu, N. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition*, 36(3), 544–553.
- Yu, N. Y., Yamauchi, T., & Schumacher, J. (2008). *Category labels highlight feature interrelatedness in similarity judgment*. Paper presented at the 30th Annual Meetings of the Cognitive Science Society, Mahwah, NJ: Erlbaum.
- Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, 10(7), 301–308.
- Zeigenfuse, M. D., & Lee, M. D. (2008). Finding feature representations of stimuli: Combining feature generation and similarity judgment tasks. Paper presented at the Annual Meeting of the Cognitive Science Society.
- Zeigenfuse, M. D., & Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, 133, 283–295.