AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

2023, Vol. 130, No. 6, 1653–1671
https://doi.org/10.1037/rev0000412

Skeleton-Based Shape Similarity

Nathan Destler, Manish Singh, and Jacob Feldman
Department of Psychology, Center for Cognitive Science, Rutgers University–New Brunswick

Many aspects of visual perception, including the classification of shapes into known categories and the induction of new shape categories from examples, are driven by *shape similarity.* But there is as yet no generally agreed, principled measure of the degree to which two shapes are "similar." Here, we derive a measure of shape similarity based on the Bayesian skeleton estimation framework of Feldman and Singh (2006). The new measure, called *generative similarity*, is based on the idea that shapes should be considered similar in proportion to the posterior probability that they were generated from a common skeletal model rather than from distinct skeletal models. We report a series of experiments in which subjects were shown a small number (1, 2, or 3) of 2D or 3D "nonsense" shapes (generated randomly in a manner designed to avoid known shape categories) and asked to select other members of the "same" shape class from a larger set of (random) alternatives. We then modeled subjects' choices using a variety of shape similarity measures drawn from the literature, including our new measure, skeletal cross-likelihood, a skeleton-based measure recently proposed by Ayzenberg and Lourenco (2019), a nonskeletal part-based similarity model proposed by Erdogan and Jacobs (2017), and a convolutional neural network model (Vedaldi & Lenc, 2015). We found that our new similarity measure generally predicted subjects' selections better than these competing proposals. These results help explain how the human visual system evaluates shape similarity and open the door to a broader view of the induction of shape categories.

Keywords: visual perception, shape, similarity, categorization, Bayesian inference

Shape similarity is the mental impression that two objects are alike in virtue of their shape. It is intuitively obvious that a cat is more similar in shape to another cat, or even to a dog, than it is to a hammer. Indeed, people rely on shape similarity to recognize objects (Biederman, 1987; Hoffman & Richards, 1984), and children use it to induce the meanings of words (Landau et al., 1988). Yet, despite a vast literature, it is still not clear exactly what makes two shapes "similar" and to what degree.

Many measures of shape similarity have been proposed, based on a wide variety of shape features. Some are based on *contour geometry*, that is, on features local to the shape boundary such as curvature extrema (Richards et al., 1988) and other contour features (Belongie et al., 2002; Blake & Isard, 2012; Greene, 2018). Others are based on *axial* or *skeletal* structure, which describe shape structure more globally in terms of component parts and the relations among them. Broadly speaking, skeletal shape representation methods represent shapes as configurations of elongated axes, ideally one to each distinct "part." In their original formulation by Blum (1967, 1973), the axes were computed via a geometric procedure called the *medial axis transform* (MAT), which results in a branching structure in which elongated limbs are represented by their central axes, which are really loci of local symmetry between contours on opposite sides. As Blum argued, the MAT provides a compact and informative representation of shapes, especially biological shapes, which are often composed of distinct articulated limbs. Marr and Nishihara (1978) took up the idea that 3D objects can largely be represented as unions of elongated parts (approximated as generalized cones), an idea that Biederman's (1987) celebrated Recognition-by-Components framework augmented with a finite list of qualitative part types (*geons*). More recently, many more sophisticated procedures for computing axial representations have been proposed (e.g., Bai & Latecki, 2008; Demirci et al., 2006; Feldman & Singh, 2006; Rezanejad & Siddiqi, 2015; Siddiqi et al., 1998; Torsello & Hancock, 2004), most designed to solve problems with Blum's original formulation. Axial representations of shape play an important role in human shape representation (e.g., Chaisilprungraung et al., 2019), and skeletal representations provide a solution to the problem of how to represent the relations among axes in multipart shapes. That is, unlike contour-based representations such as those based on curvature extrema (e.g., Richards et al., 1988), skeletal representations can represent shape "configurally," that is, in terms of the global organization of the entire shape (Kimia, 2003). Empirical support for the importance of skeletal representations has come from psychophysics

This article was published Online First March 6, 2023.

The authors are grateful to Erica Briscoe, Melchi Michel, Ahmed Elgammal, Alfred Yu, and the members of the Jacob Feldman and Manish Singh labs for helpful discussions. They are particularly indebted to Goker Erdogan for assistance in implementing the Erdogan-Jacobs model and to Yaniv Morgenstern for providing data from Morgenstern et al. (2021). An earlier account of these studies was presented to Rutgers University as part of a doctoral dissertation by the first author, entitled *Skeletal Shape Similarity and Shape Classification*, 2019.

Preparation of this article was supported by the U.S. Army Research Lab, the National Institutes of Health (Grant EY021494 to Jacob Feldman and Manish Singh), and the Rutgers Program in Perceptual Science.

Our data and Matlab code for the generative similarity model are available at https://www.dropbox.com/s/9lmzayal9bkqipc/GenerativeSimilarity.zip?dl=0.

The studies reported were not preregistered.

Correspondence concerning this article should be addressed to Jacob Feldman, Department of Psychology, Center for Cognitive Science, Rutgers University–New Brunswick, 152 Frelinghuysen Road, Piscataway, NJ 08854, United States. Email: jacob@ruccs.rutgers.edu

(Ayzenberg et al., 2019; Burbeck & Pizer, 1995; Firestone & Scholl, 2014; Harrison & Feldman, 2009; Kovács et al., 1998; Lowet et al., 2018; Wang & Burbeck, 1998; Wilder et al., 2016), development (Ayzenberg & Lourenco, 2022), neuroscience (Ayzenberg et al., 2022; Hung et al., 2012; Lescroart & Biederman, 2013), and even visual art (Leymarie & Aparajeya, 2017). Several recent studies have found evidence that shape similarity judgments are particularly affected by differences in skeletal structure. Lowet et al. (2018) found that modifications to skeletal structure, such as the introduction or deletion of parts, exert strong influence on subjects' judgments of similarity. However, exactly how skeletal shape representations can be used to construct a shape similarity metric is unclear, and a number of different approaches have been proposed. Ayzenberg and Lourenco (2019), using a skeletal similarity measure based on Feldman and Singh's (2006) Bayesian skeleton estimation framework (described below), also found a strong effect of skeletal differences on similarity judgments but did not compare this similarity measure to other alternative measures from the literature. Destler et al. (2019) used a skeletal cross-likelihood measure (also based on Feldman & Singh, 2006's framework) to model shape discrimination but also did not systematically compare it to other similarity measures. Erdogan and Jacobs (2017) conducted a more systematic comparison of similarity measures and proposed a new part-based measure of their own, which they found fit subjects' judgments better than several competitors, including skeletal cross-likelihood and a metric derived from a convolutional neural network (CNN) model.

All of these shape similarity measures (and several others) will be discussed in more detail below. But to preview, recent literature has enjoyed a flush of new part-based or skeleton-based similarity measures, but not all have been compared to each other, and it remains unclear which measure really most closely corresponds to human judgments. In this article, we propose a new similarity measure called *generative similarity*—also based on the Bayesian skeleton estimation framework introduced by Feldman and Singh (2006)—and present a set of experiments comparing a range of recent shape similarity measures to human judgments in a shape classification task.

Models of Shape Similarity

The first three models we will present require some common theoretical background, which we present first.

Graph Matching Method

Perhaps the oldest method for using shape skeletons to compute shape similarity is to compare the *branching structure* of their respective skeletons. This method arose in the context of *shock graphs* (Siddiqi et al., 1998, 1999; Siddiqi & Kimia, 1996;), which are generalizations of the "grassfire" procedure originally proposed by Blum. Shock graphs yield a hierarchical connectivity relation among symmetry axes representable by a graph. Similarity between shapes can be evaluated by counting the number of "edits" or modifications required to make one shock graph equal to the other, sometimes called the *edit distance*. However, edit distance only reflects "qualitative" changes to part structure, such as the addition or deletion of parts (i.e., branches in the graph), rather than quantitative changes to the shapes of individual parts (such as changes to part length or curvature). Moreover, graph matching methods do not generally include any overt evaluation of the *probability* of shape changes, a central concept in recent human models, as will be explained below. Nevertheless, edit distance has proven effective for shape matching in computer vision (Bai & Latecki, 2008; Demirci et al., 2006; Rezanejad & Siddiqi, 2015; Siddiqi et al., 1998; Torsello & Hancock, 2004), though interest in it has declined in recent years with the advent of image-based comparison techniques such as CNNs. To our knowledge, it has not been compared to human similarity judgments.

The Bayesian Skeleton Estimation Procedure of Feldman and Singh (2006)

Several of the shape similarity measures we consider rely on the Bayesian procedure for estimating the shape skeleton introduced by Feldman and Singh (2006). This approach aimed to address some of the shortcomings of previous medial axis methods by recasting the shape representation problem as a probabilistic inference problem, in which the goal is to estimate the skeletal structure that is most likely (maximum a posteriori; MAP) to have generated the observed shape x, called the MAP skeleton and denoted S_x,

$$S_x = \arg\max_S p(S|x). \tag{1}$$

In the Bayesian framework, each skeletal representation S serves as a *model* of the shape, including both the topological relations among axes (a hierarchical representation of the connections among axes) and the specific parameters of each axis (length, curvature, etc.). The MAP skeleton S_x aims to achieve a balance between simplicity (too few branches—a high prior but a low likelihood) and complexity (too many branches—low prior but high likelihood). The result is a skeleton that is "just right," containing only those axes that correspond to intuitively meaningful parts. Figure 1 gives a visual summary of the framework and several examples of MAP skeletons.

Given a reliable estimate of the shape skeleton, such as the MAP skeleton, how should one compute shape similarity? The first three proposals we discuss answer this question in different ways.

Model 1: Skeletal Deviation

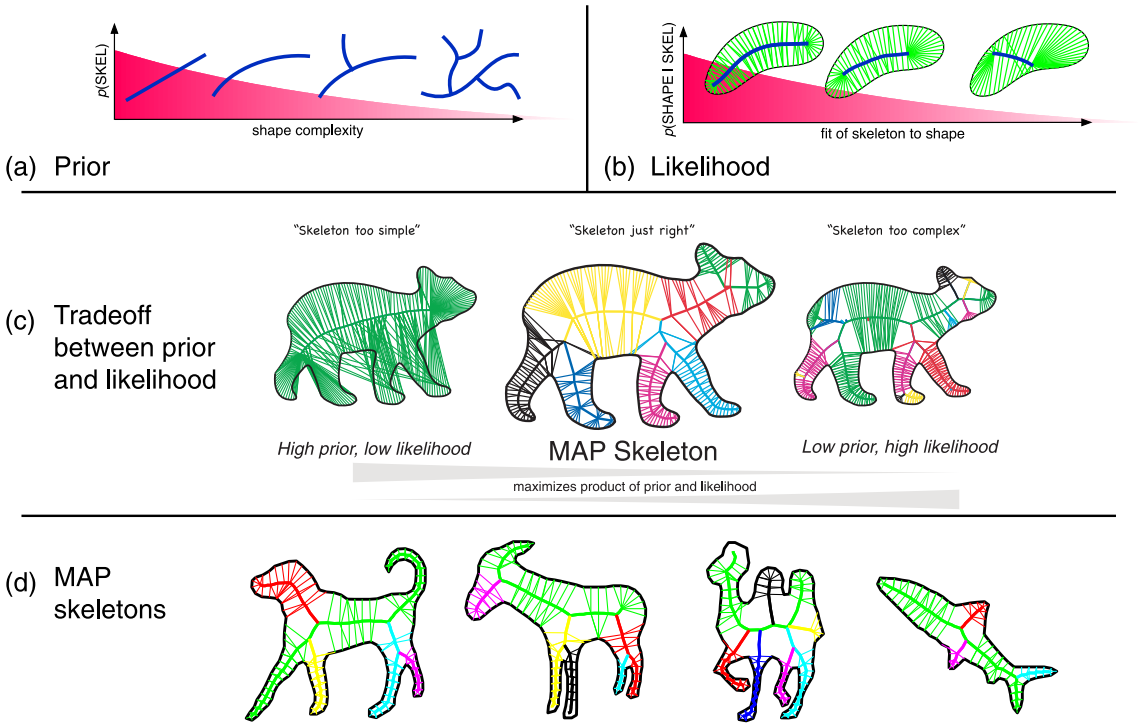
Ayzenberg and Lourenco (2019) evaluated shape similarity by computing MAP skeletons for both shapes and then computing the average Euclidean distance between corresponding points on the two skeletons, which we call *skeletal deviation*:

$$\text{deviation}(a,b) = \frac{1}{n}\sum_{x\in S_a, y\in S_b}^{n} \|x-y\|, \tag{2}$$

where, S_a and S_b denote the MAP skeletons of shapes a and b, respectively, $\|x-y\|$ is the Euclidean distance between points x and y, and n is the number of points on each skeleton. Like edit distance, this method simply directly compares estimated skeletons, albeit pointwise rather than in terms of their connectivity relations. Thus, skeletal deviation does not take into account topological differences

Figure 1

The Bayesian Skeleton Estimation Procedure of Feldman and Singh (2006)



Note. (a) Prior; (b) Likelihood; (c) the trade-off between prior and likelihood, which is optimized in the maximum a posteriori (MAP) skeleton; and (d) some examples of MAP skeletons. See the online article for the color version of this figure.

between skeletons, nor the degree to which the skeletal models actually fit their respective shapes, which can vary enormously from one shape to another. Nevertheless, in Ayzenberg and Lourenco's (2019) study fit human judgments well, though it was not compared to other similarity metrics.

Model 2: Skeletal Cross-Likelihood

Another very simple proposal is the *cross-likelihood*, which is the average likelihood of each shape conditioned on the other's skeleton (Briscoe, 2008; Feldman et al., 2013). Given shapes a and b, with MAP skeletons S_a and S_b, the cross-likelihood is given by

$$\mathrm{CL}(a,b) = \frac{1}{2}[p(a|S_b) + p(b|S_a)]. \quad (3)$$

Intuitively, the cross-likelihood is high when each shape's skeletal model also fits the other shape well. But like the Ayzenberg and Lourenco (2019) measure, this measure is somewhat *ad hoc* and tends to break down when the shapes have grossly different part structures. In previous work, we have used the cross-likelihood to model discrimination of shapes along morph spaces (Destler et al., 2019), where it works well. That is, we found that sensitivity to small differences in shape improves inversely with the cross-likelihood between nearby shapes—the more dissimilar shapes are by this measure, the easier it is for subjects to discriminate them. But the shapes compared in the morphing study were always slight variations of each other, where the cross-likelihood is more suitable, and the measure should not be expected to generalize well to shapes with grossly different part structures. It was perhaps for this reason that the cross-likelihood did not perform well in Erdogan and Jacobs's (2017) comparison of similarity measures, which included shapes with grossly different part structures. More broadly, different probabilistic similarity measures (e.g., Erdogan and Jacobs's own model [henceforth Erdogan-Jacobs (EJ)] vs. cross-likelihood) would be expected to perform differently on different shape distributions, reflecting differences in the assumed likelihood model (respectively, the simulation-based likelihood function in EJ vs. the simple generative skeleton formulation underlying MAP skeletons). We will return to this issue below.

To summarize so far: both skeletal deviation and skeletal cross-likelihood are simple and readily computable, but both suffer from the same intuitive shortcoming: they make sense when comparing very similar skeletons, because they capture how much corresponding parts have been modified from one shape to another, but do not have a principled way to handle *configural* changes to the topology of the skeleton. In a sense, this is the opposite problem from that of edit distance, which *only* reflects configural changes but misses metric differences. With all this in mind, the main goal of our new approach is to incorporate both topological and metric shape changes—both of which are ubiquitous in real shapes—in a unified probabilistic framework.

Model 3: Generative Similarity

We begin with an idea first suggested by Kemp et al. (2005): that items should be regarded as similar to the degree that they appear to have a *common generative source*. Here, we take advantage of the fact that the entire Bayesian skeleton framework is based on the idea of generative shape models, meaning models of the skeletal process that produced the shape. The key idea is that shapes should be regarded as "similar" in proportion to the posterior belief that they derive from a common skeletal source—meaning a single latent skeleton that gave rise to both of them—as opposed to two distinct skeletal sources. Most of the mathematical development below aims to define exactly what is meant by a "common" skeleton versus "distinctive" skeletons and to attach probabilities to these as models of the observed shapes.

The posterior belief in a skeletal model A given shape a is given by Bayes' rule, as follows:

$$p(A|a) = \frac{p(a|A)p(A)}{p(a)}. \tag{4}$$

Here, lowercase symbols (e.g., a) refer to shapes, and uppercase symbols (A) refer to skeletal models. Specifically, the posterior belief in a *common* skeletal model C for two shapes a and b is as follows:

$$p(C|a,b) = \frac{p(a,b|C)p(C)}{p(a,b)}. \tag{5}$$

We will generally assume that shapes are independently conditioned on their model, so $p(a,b|C) = p(a|C)p(b|C)$. With this in mind, given two shapes a and b, the support for a common model C relative to two distinct models A and B is given by the Bayes factor (BF)

$$\text{BF} = \frac{p(a|C)p(b|C)}{p(a|A)p(b|B)}. \tag{6}$$

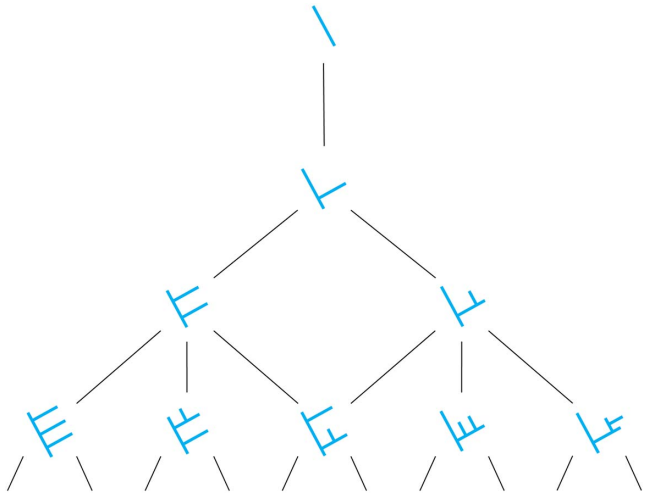
This BF expresses the strength of evidence in favor of a common model relative to distinct models, and we postulate that subjective shape similarity is proportional to it. The key problem for this approach is to define an appropriate common model C and the distinct models A and B and attach probabilities to them. We solve this problem using a formalism called the *shape lattice.*

The Shape Lattice

Recall that each skeletal model is defined by both a skeletal topology—the connection structure among its component axes—and a set of parameters that determine the specific shapes and relations among the axes. These include, for each axis, its length, curvature, and location of origin on its parent axis. Every skeletal model assumes a prior distribution over these parameters and a posterior distribution conditioned on a particular shape or set of shapes. Each potential skeletal topology is thus in effect a *family of models* with its own distinct set of parameters; the number of parameters depends on the topology of the skeleton.

The shape lattice is a hierarchical graph of possible skeleton topologies, ordered by subset and superset relations (Figure 2). At the top is the simplest model, a single axis. Next, down is a two-axis model with a parent and child, that is, a shape with a main part and a subsidiary part. From there on the lattice branches, a third axis can be added in either of two ways: to the parent, yielding a shape with a main part and two subsidiary parts; or to the child, yielding a shape with a main part, a subsidiary part, and a third part protruding from the subsidiary part. Below that the lattice branches further as the possible topologies multiply. In principle, the lattice extends down infinitely as more branches are added, eventually comprising all possible skeletal topologies.

Figure 2
The First Four Levels of the Shape Lattice



Note. Each node represents a class of parent–child skeleton topologies. At each level of the graph, every shape has the same number of axes. The full lattice comprises all possible skeletal topologies, continuing downward infinitely as more axes are added. See the online article for the color version of this figure.

Given any two nodes (topologies) on the lattice A and B, the *join* $A \vee B$ is their least common ancestor, that is, the lowest node on the lattice that connects to both of them from above. This is the skeletal topology that includes all and only the axes they have in common. Similarly, the *meet* $A \wedge B$ is their greatest common descendent, the topology that includes all the branches that either of them has. For two nodes that are in an ancestor–descendant relationship (notated $A \leq B$), the join is the ancestor ($A \vee B = B$) and the meet is the descendant ($A \wedge B = A$). Otherwise, the join is a node above both of them and the meet is a node below both of them. The meet and join exist for all pairs of nodes, and indeed for any set of nodes, in a lattice (Davey & Priestley, 1990). In our context, it is the join that is critical, because it is literally the *common skeletal topology*—the skeletal model that contains all axial components common to both skeletons.

To evaluate probabilities for the common and distinct model probabilities, we need to introduce some notation to indicate how skeletal parameters are fitted. From here on, we use uppercase symbols (A) to refer to skeletal topologies without fitted parameters and hatted uppercase symbols with a subscript ($\hat{A}_a$) to indicate a skeletal model with a parameter distribution fitted to a particular shape or shapes indicated by the subscript. The parameters fitted include the scalar prior on a new axis, the position of each axial

branch along its parent axis (assumed to have a beta distribution, appropriate for proportions), the length of each axial branch (assumed to have a beta prime distribution, appropriate for nonnegative coefficients), and the angular discrepancy between corresponding turning angles along the axis (assumed to have von Mises distribution, appropriate for angles, cf. Feldman & Singh, 2005). More details on parameters are given in the Results section.

With this in mind, we can factor the likelihood of shape a given skeleton A as $p(a|A) = p(a|\hat{A}_a)p(\hat{A}_a|A)$, that is as the product of the probability of a given a fitted skeletal model, times the probability of the fitted parameters under the parametric family A. More specifically, the probability of a under the skeletal distribution $\hat{A}_a$ is the product $p(a|S_a)p(S_a|\hat{A}_a)$. (Recall that S_a is a specific skeleton, the MAP skeleton for shape a, whereas $\hat{A}_a$ is a distribution of skeletons fitted to shape a.) We use this factorization throughout the derivation below.

Given the above, the probabilities of a and b under their common model $C = A \vee B$ can be written respectively as $p(a|C) = p(a|S_a)p(S_a|\hat{A}_a)p(A|C)$ and $p(b|C) = p(b|S_b)p(S_b|\hat{B}_b)p(B|C)$. The last terms $p(A|C)$ and $p(B|C)$ represent the probabilities of the *topologies* A and B of the respective MAP skeletons S_a and S_b given the topology C of the common skeletal model. We define the probability of one skeletal topology conditioned on another, for example, $p(A|C)$, as the probability of A arising by addition and deletion of axes from C. The addition of axes entails a probabilistic penalty in exactly the same manner as in the prior original Bayesian procedure, in which each axis is "born" with a fixed scalar probability and is modified via parametric changes with associated probability distributions. For simplicity, we further assume that the deletion of an axis entails the same probabilistic cost as the addition of one. That is, movement *up* the lattice "costs the same" as movement *down* the lattice. It follows that in general $p(A|C) = p(C|A)$, because all the operations required to transform A into C happen "in reverse" when C is transformed into A.

With all this in mind, the likelihood of shapes a and b under their respective distinctive models is given by the products

$$\begin{aligned} p(a|\hat{A}_a) &= p(a|S_a)p(S_a|\hat{A}_a), \\ p(b|\hat{B}_a) &= p(a|S_b)p(S_b|\hat{B}_b), \end{aligned} \tag{7}$$

and their likelihood under their common model by

$$p(a, b|\hat{C}_{a,b}) = p(a|S_a)p(S_a|\hat{A}_{a,b})p(A|C)p(b|S_b)p(S_b|\hat{B}_{a,b})p(B|C). \tag{8}$$

The ratio of the marginal likelihoods of common versus distinct models is the Bayes factor,

$$\begin{aligned} \mathrm{BF} &= \frac{p(a|\hat{C}_{a,b})p(b|\hat{C}_{a,b})}{p(a|\hat{A}_a)p(b|\hat{B}_b)} \\ &= \frac{p(a|S_a)p(S_a|\hat{A}_{a,b})p(A|C)p(b|S_b)p(S_b|\hat{B}_{a,b})p(B|C)}{p(a|S_a)p(S_a|\hat{A}_a)p(b|S_b)p(S_b|\hat{B}_b)} \\ &= \frac{p(S_a|\hat{A}_{ab})p(A|C)p(S_b|\hat{B}_{ab})p(B|C)}{p(S_a|\hat{A}_a)p(S_b|\hat{B}_b)}, \end{aligned} \tag{9}$$

which expresses the strength of the evidence in favor of shapes a and b having a common skeletal model. Finally, we take the negative log (i.e., description length [DL]) of this BF to provide the final shape dissimilarity measure,

$$\mathrm{dissim}(a, b) \propto -\log \frac{p(a|\hat{C}_{a,b})p(b|\hat{C}_{a,b})}{p(a|\hat{A}_a)p(b|\hat{B}_b)}, \tag{10}$$

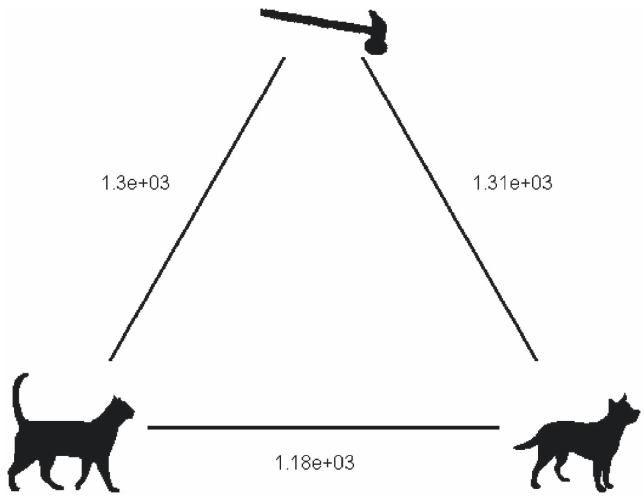
which expresses the weight of evidence in favor of shapes a and b having *distinctive* models. (The direction of the comparison is inverted because the negative log of a ratio is the same as the log of the inverse ratio.)

Figure 3 provides an intuition about what the resulting −log BFs (DLs) mean, using the cat/dog/hammer example from the beginning of the article. As can be seen in the figure, the cat–dog comparison has a large DL, meaning that these two shapes probably stem from distinct skeletal models (which of course they do). But the cat–hammer and dog–hammer comparisons have even *larger* DLs, because they are even more likely to have distinct skeletal models, since their skeletal topologies are grossly different. The DLs quantify that cats and dogs are closer to being in the same shape class than cats and hammers or dogs and hammers. Note that DLs are in log space and must be exponentiated to recover BFs, so the difference in DLs shown in the figure translates to a very large difference in generative similarity between cat/dog and cat/hammer.

Edit Distance Interpretation

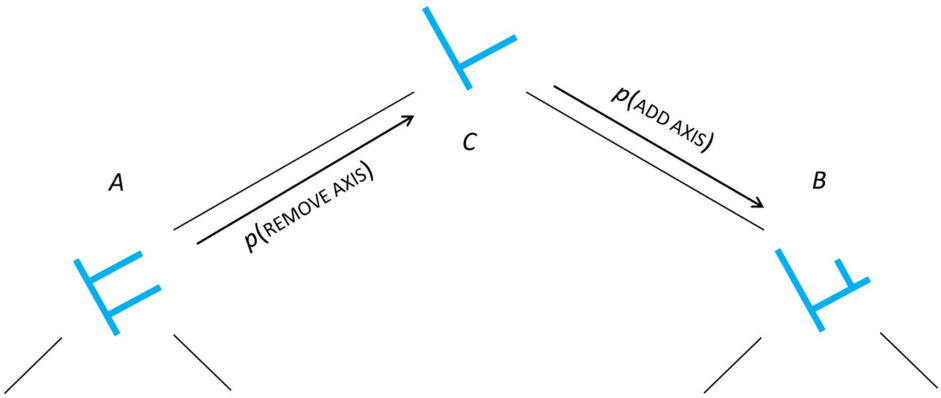
We note that the numerator of the Bayes factor (Equation 6, and in parameterized form in Equation 9), which expresses the evidence in favor of a common model for shapes a and b, has a natural relationship to the edit distance, the "cost" of transforming one shape representation into another (Figure 4). (The denominator,

Figure 3
DLs (−log BFs) According to the Generative Similarity Model Between Cat, Dog, and Hammer, Showing That Cats and Dogs Are More Similar to Each Other (Lower DL) Than Either Is to a Hammer



Note. DL = description length; BF = Bayes factor. DLs are in log space and must be exponentiated to yield BFs. For example, the numbers in the figure mean that the cat is about $e^{(1300 - 1180)} \approx 1.3 \times 10^{52}$ times more similar to the dog than it is to the hammer.

Figure 4
The Edit Path Between Two Distinct Skeletal Topologies A and B on the Shape Lattice



Note. Each addition or deletion of an axis is one step along the path, reducing the probability of the transformation. See the online article for the color version of this figure.

which expresses the evidence in favor of explaining each shape separately, does not depend on the similarity between the shapes, and in effect serves as a normalization; the numerator is the critical term that varies with shape similarity.) Specifically, the product $p(A|C)p(B|C)$, which appears in the numerator of Equation 6, can be interpreted as the probability of the *edit path* between skeletons A and B. Recall that by assumption $p(A|C) = p(C|A)$, so $p(A|C)p(B|C)$ (the numerator) is the same as $p(C|A)p(B|C)$ (the probability of the transformation $A \to C \to B$).

More broadly, the full numerator of Equation 6, which expresses the likelihood of the two shapes a and b under a common model, $p(a|C)p(b|C)$, is closely related to the probability of transforming shape a into shape b via their shared skeletal structure. (Recall that uppercase A, B, … denote skeleton topologies, and lowercase a, b, … denote shapes.) Formally, the edit path probability is $p(C|a)(b|C)$, whereas the numerator of the BF is $p(a|C)p(b|C)$. That is, the edit path is $a \to C \to b$, and the BF numerator is $(C \to a) + (C \to b)$—they differ only in the direction of the leg between C and a, which is forward in the former and inverse in the latter. But by Bayes' rule, $p(C|a) = p(a|C)p(C)/p(a)$, so it follows that

$$\overbrace{p(a|C)p(b|C)}^{\text{BFnumerator}} = \overbrace{p(C|a)p(b|C)}^{\text{Editpathprobability}} \times p(a)/p(C). \tag{11}$$

This is somewhat more intuitive if we take negative logs, so the probabilities turn into *description lengths* (DLs), and the products turn into sums,

$$\overbrace{\mathrm{DL}(a|C) + \mathrm{DL}(b|C)}^{\text{log BF numerator}} = \overbrace{\mathrm{DL}(C|a) + \mathrm{DL}(b|C)}^{\text{Probabilistic edit distance}} + \mathrm{DL}(a) - \mathrm{DL}(C). \tag{12}$$

That is, the numerator in our BF is simply the *probabilistic edit distance* (the DL of the edit path probability) plus an additive factor of $\mathrm{DL}(a) - \mathrm{DL}(C)$. For example, the probabilistic cost of transforming one skeletal axis into another incorporates the total log von Mises penalty on the discrepancies between corresponding angles (see discussion of distributions above). In this sense, $\mathrm{BF}_{\mathrm{skel}}$ includes a probabilistic version of edit-distance-based similarity metrics as a side effect of the Bayesian formulation. This mathematical relationship helps provide an intuition about what the Bayes factor means: It incorporates the relative plausibility of one shape *transforming into the other* rather than each arising independently. This helps connect it to the body of previous work on shape similarity using edit distance (Bai & Latecki, 2008; Demirci et al., 2006; Rezanejad & Siddiqi, 2015; Siddiqi et al., 1998; Torsello & Hancock, 2004). More broadly, it suggests that our similarity metric is resonant with Hahn et al.'s (2003) proposal that similarity (in general) can be understood as proportional to the complexity (in our formulation, the DL) of the transformation from one stimulus to another.

Model 4: The Erdogan–Jacobs Model

The model proposed by Erdogan and Jacobs, henceforth the EJ model, is similar in some respects to ours and shares its Bayesian formulation. The EJ model begins by inferring the 3D structure of a shape from a single rendered image, using a prior over shape models, and a likelihood function that quantifies the probability of a given image being generated from a given shape model. Shapes are represented as sets of segments, somewhat analogous to shape skeleton axes, and segment endpoints, though without the overtly axial statistical process in our model. The prior over shape models assume a uniform distribution over the number of segments and endpoint positions, and the likelihood model assumes Gaussian error conditioned on the 3D shape. Once 3D shapes have been estimated, the EJ model computes shape similarity using a likelihood comparison similar to the cross-likelihood (including the probability of each shape conditioned on the other's model, as well as the average of these two probabilities, which is what is used in the cross-likelihood model).

Erdogan and Jacobs found their model to fit human judgments better than the skeletal cross-likelihood, a CNN, and a number of other models not included here, though it was not compared to skeletal deviation (introduced later, in Ayzenberg & Lourenco, 2019) or to our generative similarity model (introduced above).

Model 5: Active Contour Model

To ensure a comprehensive comparison, we included one similarity model that is entirely contour-based, the *active contour model*. The active contour model (based on the contour-identification model of that name proposed by Blake & Isard, 2012) identifies the minimum-distance order-preserving correspondence between contour vertices in the two shapes—that is, the shape alignment with the smallest total discrepancy between corresponding contour points. Similarity in the model is proportional to the total probability of this correspondence, integrating over all points on the two contours, assuming a Gaussian distribution centered at zero over distance discrepancies. This model is thus similar to the contour component of our generative similarity model, except with discrepancies computed over intervertex distances rather than difference in turning angle. The very simple model serves as a representative contour-based similarity metric.

Model 6: CNN Model

In recent years, convolutional neural network (CNN) models have achieved impressive performance on image classification benchmarks (see Farabet et al., 2010; Jacobs & Bates, 2019), although some studies (e.g., Baker et al., 2017; Heinke et al., 2021) have found that their performance is based largely on local features (e.g., texture) and is substantially insensitive to global shape. Nevertheless, given their success, it is reasonable to ask how our subjects' performance compares to that of a suitably trained network. As explained below, in our experiments, subjects were given only a small number (1, 2, or 3) of training shapes, far fewer than the thousands, millions, or even billions on which CNNs are commonly trained. A CNN evaluated on our subjects' actual stimulus set would be a straw man, since CNNs by design are not capable of learning from such small samples (a point for which they are often criticized, e.g., Bates & Jacobs, 2019; Lake et al., 2015). But given subjects' prior experience with other shapes, it seems reasonable to include a CNN that was fully trained on a similar class of shapes and ask whether the resulting shape representations can explain our subjects' classifications of new shapes.

We used the CNN MNIST (Vedaldi & Lenc, 2015) based on handwritten digits, a close analog to our random axial shapes. The model was pretrained on the Modified National Institute of Standards and Technology (MNIST) handwritten digit database, which consists of 60,000 training images and 10,000 test images. Each image is of a single handwritten digit, 0–9. A handwritten digit database was chosen because it was judged that handwritten characters were similar in structure to the axial 2D shape stimuli used here. Like our stimulus shapes, the character images are of single, isolated objects with some degree of variation, and the 10-digit categories are defined largely by the arrangement of parts relative to one another. To compare similarity models, we used the second-to-last layer of the CNN (which serves as the input to the final classifying layer, whose nodes represent specific character categories). As a dissimilarity measure, we use the cosine distance between the node vectors for the two shapes.

We note that there are many other CNNs that might be applied to our data, and probably other ways to apply their output to explaining human similarity judgments. However, it was not our goal to undertake a comprehensive study of CNN performance in our task. By their nature, CNNs can be expected to fit human judgments when trained on a representative sample of those judgments. However, as discussed above, in the case of shape representation they appear to do so using principles quite different from those used by the human visual system. We include MNIST as a representative example of this class of models simply in order to get a general sense of their fit to human judgments, in order to better understand the relative fit of other more psychologically motivated models.

Experiments

With these five shape similarity models in hand, we conducted a series of shape classification experiments and asked how well each model predicts human responses. In each trial of the experiments, the subject is shown a small number of novel shapes (1, 2, or 3 depending on the condition) with a category label (e.g., "This is a blicket/These are blickets") and is asked for judgments about which *other* shapes are also members of the same category (cf. Landau et al., 1988). We used "nonsense" shapes, meaning shapes that are randomly constructed and thus not necessarily associated with nameable categories, so that subjects' classification judgments would be driven primarily by pure judgments of shape similarity rather than by ex post facto judgments of category membership. We then compared the various similarity models in terms of their ability to predict which shapes the subjects chose.

Our choice of a classification task reflects the importance of classification in the use of similarity judgments. There are of course other kinds of data that are useful for evaluating similarity metrics, including overt similarity ratings (e.g., Cortese & Dyre, 1996; Hahn et al., 2009; Morgenstern et al., 2021) and confusion matrices (Ashby & Lee, 1991). Nevertheless, we would argue that shape classification is the primary function of shape similarity, in that one of the core premises of the categorization literature is that objects within a category are generally more similar to each other than objects in different categories (Mervis & Rosch, 1981; see Panis et al., 2008). Hence, we would argue that a classification task is an especially acute test of a similarity metric.

That said, similarity and classification are of course deeply intertwined, and the influence between them flows in both directions (Edelman, 1998; Nosofsky, 1986; Shepard, 1987). Similar objects tend to be categorized together, but impressions of category membership can also make objects seem more similar (Schmidt et al., 2020; see also Lupyan, 2008). Our use of nonsense shapes is intended to isolate, to the extent possible, "pure" shape similarity judgments, that is, similarity judgments uncontaminated by semantic classifications. But some degree of resemblance to previously learned categories is unavoidable. For example, some of our nonsense shapes might remind subjects of this or that natural category (an animal, a leaf, etc.), and their similarity judgments might reflect these associations to some degree. Though we have attempted to minimize such influences by using a randomized shape-generating processes and by randomly rotating the resulting shapes, some residual categorical effects probably remain.

General Method

In each of the experiments, subjects were asked to classify shapes as belonging to, or not belonging to, a particular novel category based on a set of example shapes. On each screen, subjects saw one or more example shapes labeled as belonging to a novel named

category (positive examples, labeled, e.g., "This is a blicket"/"These are blickets") or not belonging to it (negative examples, labeled, e.g., "This is not a blicket"). There were 1–3 positive examples depending on the experiment, and 0 or 1 negative (details below). The subject's task was to choose other "blickets" from a grid of other shapes ("Which of these are blickets?"). The candidate shapes were displayed in white in a 6×6 grid at the center of a computer screen, whereas the positive examples were shown on the left in blue, and the negative examples (if any) were shown on the right in white (Figure 5). The alternatives typically included exact matches to the examples (as catch trials, since plausible response patterns should include these choices), but the subjects were encouraged to choose a larger number of examples (the instructions indicated "Most subjects choose about 7") to encourage generalization.

Subjects selected shapes by clicking on them. Once a shape was selected, its color changed from white to blue, the same color as the positive example(s), indicating its judged membership in the novel category. Additional clicks on the same shape would toggle the shape's membership status and color. Once the subject was satisfied with the selected shapes, the subject would confirm their choices and move on to a new screen with new category (new example shapes and new candidate responses). To minimize confusion between categories, each new shape category was associated with a different nonsense word, each starting with a different letter of the alphabet (*aben, blicket, coricle, dax … zem*) in random order. Each subject saw one full round of 26 novel shape categories (one for each letter of the alphabet). Subjects took about 1 min to complete each screen, for a total of about 20–30 min per subject for a full set of 26.

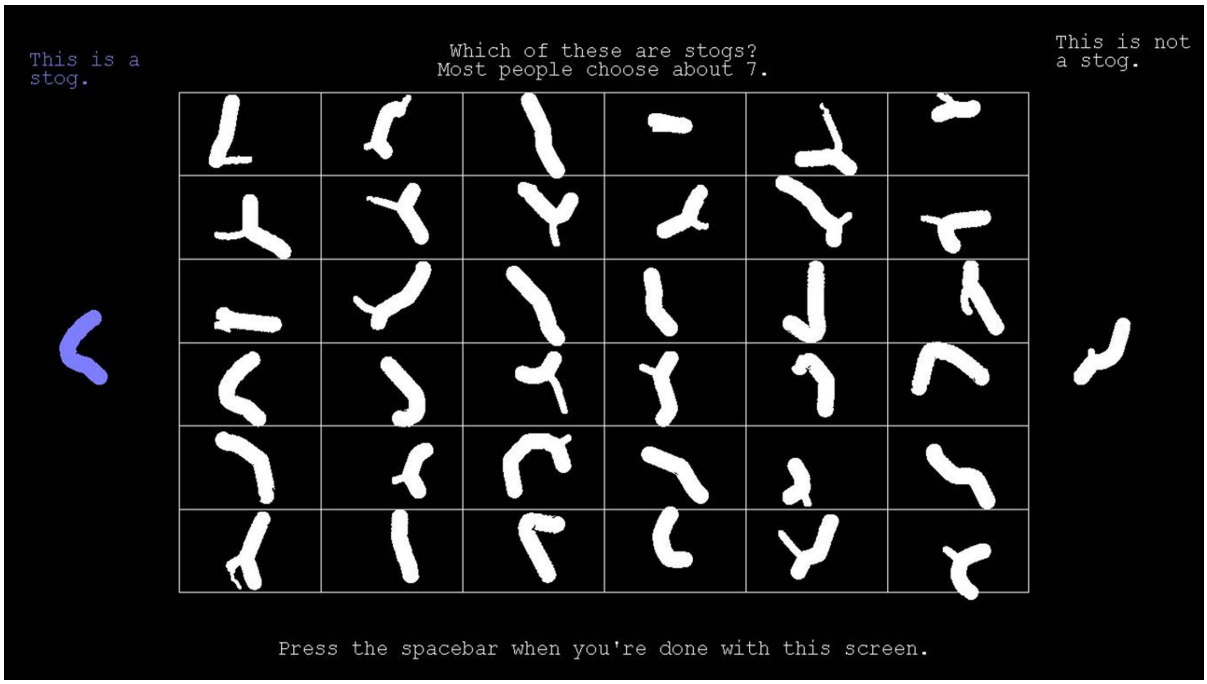
Note that this is a purely subjective task in that there is no correct answer; it simply serves to probe subjects' intuitions about which shapes are similar enough to the example(s) to include in the novel category. The main dependent measure is which shapes the subjects chose to include in the induced category as a function of similarity to the examples(s) as measured by various competing similarity metrics.

Conditions

There were 10 experiments, which varied in the number of positive and negative examples and whether the shapes were 2D (Experiments 1–5) or 3D (Experiments 6–10). We varied the number of examples for several reasons. Although people can induce categories from single positive examples (Feldman, 1997), most contemporary models of categorization involve a comparison between an item's fit to one category and its fit to alternative categories (e.g., via the Luce choice rule; see Jäkel et al., 2008; Luce, 1959; Nosofsky, 1986). Intuitively, without a negative example, subjects might not know how far to extend the class induced from the positives. Hence, we wondered what effect the presence or absence of a negative example might have.

In addition, we wanted to explore the influence of the number of positive examples. Most theories of similarity compute it pairwise, comparing two items at a time. But with multiple positive examples, one has to consider how to integrate the similarity of a candidate to *all* of the examples. Indeed, this is one of the core ways in which categorization models differ, that is, by making different assumptions

Figure 5
An Example Screen Showing One Trial of Our Experiments (1:1 Case)



Note. In this trial, there is one positive example (left margin, blue), and one negative (right margin, white). The subject's task is to click on whichever among the candidate shapes in the center grid they judge to belong to the same category as the positive example(s). See the online article for the color version of this figure.

about how similarities to multiple examples are integrated. As discussed below, our model has a built-in mechanism for handling multiple examples. In order to evaluate this aspect of the model, we wanted to explore cases with multiple positive examples.

With these considerations in mind, we ran the following conditions. Experiments 1–5 used 2D shapes and differed only in the number of positive and negative examples. Experiment 1 used one positive and zero negative (henceforth denoted 1:0), Experiment 2 is 2:0, Experiment 3 is 3:0, Experiment 4 is 1:1, and Experiment 5 is 2:1. Experiments 6–10 used 3D shapes but were otherwise similar and used the same sequence of numbers of examples, respectively, 1:0, 2:0, 3:0, 1:1, and 2:1.

Shape Generation

Shapes for the experiments were randomly generated using a three-step process (Figure 6). First, we randomly selected a series of six 2D points from a uniform distribution over a rectangular region, discarding any shapes found to have self-intersections. Next, we interpolate points to create a polygon of 30 points and smooth the points to yield a smooth random blob. We then compute the MAP skeleton of the blob using the tools introduced by Feldman and Singh (2006). Next, we "inflated" this skeleton into a new random shape by choosing maximum-likelihood random deviates from the implied generative model (called "ribs" in the framework) and joining their endpoints, effectively running the Bayesian generative model forward. (Note that this procedure results in shapes with a variety of skeletal structures.) The 3D inflation model is a simple generalization of the 2D inflation model in which ribs extend outwards in all directions lying in the plane orthogonal to the axis, like spokes on a wheel (Figure 6b; see Feldman et al., 2013), resulting in a random 3D volumetric object with a known generating skeleton. Final rendered objects were displayed at 15° slant from the frontal plane.

Subjects

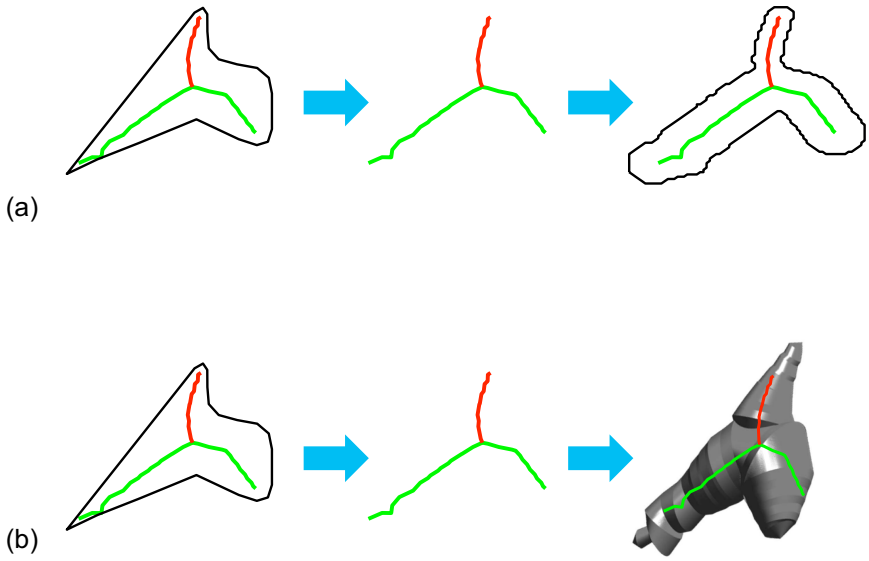
Ten subjects participated in each of the 10 experiments (100 subjects total), a number that piloting suggested was sufficient to reveal consistent trends in subjects' judgments. We used Amazon Mechanical Turk to collect subject responses (Crump et al., 2013). All participants had at least 95% positive ratings on the Mechanical Turk system, were recruited from within the United States, and were compensated $4 each for their participation.

Ethical treatment of human subjects in this study was approved by the Rutgers University Institutional Review Board under a protocol entitled "Human categorization of visual forms." Our data and Matlab code for the generative similarity model are available from https://www.dropbox.com/s/9lmzayal9bkqipc/GenerativeSimilarity.zip?dl=0.

Results

Subjects chose an average of 6.3 (sd 2.1) shapes per concept overall. The number of shapes chosen rose with the number of positive examples (Figure 7a, $BF_{10} = 3.695e8$) but was not affected by the number of negative examples ($BF_{10} = .147$), nor by the dimensionality (2D vs. 3D) of the experiment ($BF_{10} = .06$). The effect of the number of positive examples has several possible

Figure 6
Shape Generation Process, Showing (a) 2D Case and (b) 3D Case



Note. In the 2D case, a random shape is first generated (left), its maximum a posteriori (MAP) skeleton is computed (middle), and then a shape is generated from the skeleton using the forward generative model from Feldman and Singh (2006). In the 3D case, a random (2D) shape is first generated (left), and its (2D) MAP skeleton is estimated. Then the skeleton (reinterpreted as a planar 3D skeleton, i.e., with $z = 0$) is "inflated" using the 3D generalization of the likelihood model described in Feldman et al. (2013) to produce a random 3D shape (see text). See the online article for the color version of this figure.

explanations. It might suggest that subjects took the number of examples as a cue for the expansiveness of the category (cf. Heit, 2000). But because our example shapes were chosen at random, the number of examples was necessarily confounded with the *diversity* of the examples (the internal dissimilarity within the training set), which is also known to influence generalization (Heit, 2000; Tenenbaum & Griffiths, 2001), including in the case of shape (Morgenstern et al., 2019), so we hesitate to draw a strong conclusion. There was also a complex 3-way interaction among these factors ($BF_{10} = 3.682e + 16$, plotted in Figure 7b), in which the rise in the number of shapes chosen was somewhat different in the 3D case compared to the 2D case.

Model Comparisons

The main analysis is the relative fit to the human data of the five models presented above—skeletal deviation, cross-likelihood, generative similarity, Erdogan–Jacobs, active contours, and MNIST. We used the following procedure to compute a (log) likelihood for each model as an explanation of the shape choices made by the subjects (i.e., on each trial, which shapes they selected as other "blickets" based on the given example[s]). Each model provides a shape dissimilarity measure, the details of which depend on the individual model. For each model, we computed D from each of the positive or negative training examples for each of the candidate shapes. We assume that selection probability decays exponentially with distance, for example, the probability that shape x_i will be chosen as an instance of the category exemplified by a positive example p_j decays with the distance D_{ij} between x_i and p_j,

$$p(x_i|p_j) \propto e^{-cD_{ij}}, \tag{13}$$

(Luce, 1959; Shepard, 1987). Here, the parameter c modulates the similarity decay rate, and it was fitted to subjects' choices along with other parameters of each model.

For each of the models other than our own, for experiments with multiple positive examples (2:0, 2:1 and 3:0), we used a standard exemplar-based combination rule in which the probability of choosing shape x_i depends on the average[1] similarity to the positive examples $P = \{p_i\}$,

$$p(x_j|P) \propto \sum_i e^{-cD(p_i, x_j)}/|P|, \tag{14}$$

where $D(p_i, x_j)$ is the distance from the ith positive example to the jth candidate shape, and $|P|$ is the number of positive examples (=1, 2, or 3). Similarly, the probability that candidate shape will be classified with the single negative example n (and thus not selected) decays exponentially with the distance $D(n, x_j)$ between n and x_j,

$$p(x_j|n) \propto e^{-cD(n, x_j)}. \tag{15}$$

Putting these together, the probability of the subject's picking candidate shape x_j depends on its relative similarity to the positive and negative examples,

$$p(x_j|P, n) \propto \frac{p(x_j|P)}{p(x_j|n)}, \tag{16}$$

where the denominator $p(x_j|n)$ is treated as unity in conditions containing no negative examples (1:0, 2:0, and 3:0). Finally, we normalize over all the candidate shapes to yield the actual probability of a given shape being chosen,

$$p(\text{choose } x_j) = \frac{p(x_j|P, n)}{\sum_j p(x_j|P, n)}. \tag{17}$$

Unlike the other models, generative similarity includes a built-in combination rule, based on the lattice join of the example shapes (which applies to $|P| > 2$ shapes just as it does to two shapes, see Davey & Priestley, 1990). So for generative similarity, we used this combination rule in place of average exemplar similarity to yield a measure of similarity between a given candidate shape x_j and the set of positive examples, and between x_j and the single negative example.

In summary, for all models, this procedure assigns a probability to each potential candidate shape in proportion to its similarity to the positive examples (modulo a particular similarity metric), relative to its similarity to the negative example, if there is one. The parameter fitting procedure (details below) sets the parameters of each model to maximize the probability (minimize the negative log likelihood) of the selections the subject actually made. As trials are assumed independently conditioned on the model, negative log likelihoods can be summed over trials to arrive at a cumulative negative log likelihood for each model as an account of subjects' responses.

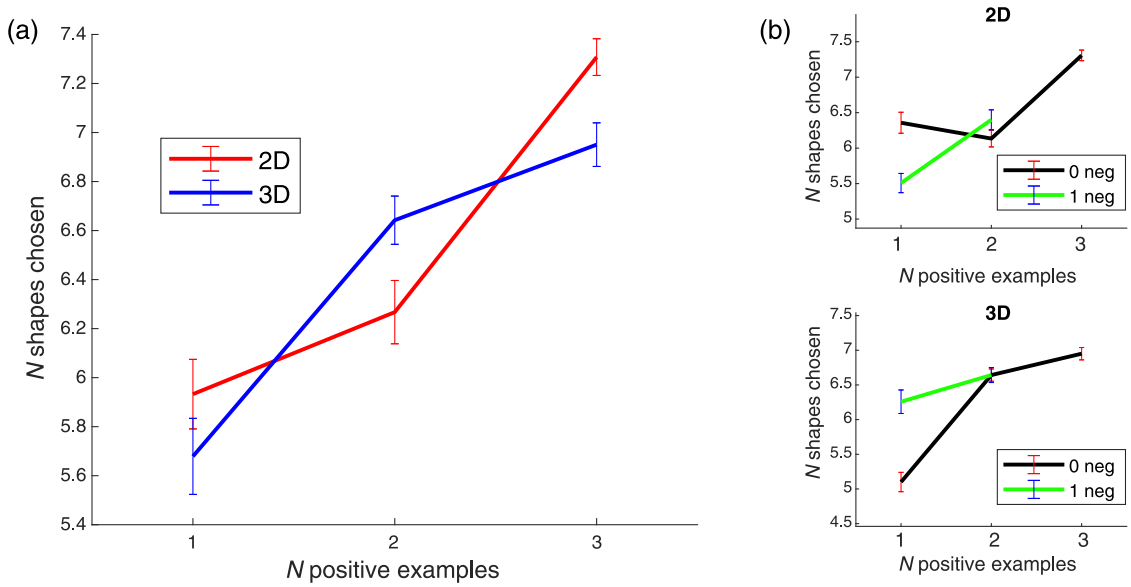
Parameter Fitting

For each model, we fitted parameters by maximizing likelihood relative to the ensemble of subjects' responses, fitting each experiment separately. Parameters shared among all models include the sensitivity parameter c, and (in conditions with a negative example) a parameter modulating the weight of similarity to the negative example relative to the positive(s). The generative similarity model has seven or eight fitted parameters, depending on the experimental condition. Six of the parameters correspond to different components of the similarity calculation, such as branching location, axis curvature, and the penalty for a missing axis on one of the two otherwise corresponding skeletons. Each of these parameters is used to weigh its corresponding component when combining these elements into the final dissimilarity measure. To fit the parameters, we aggregated data across all subjects in each condition and fit the parameters using the Matlab genetic algorithm (Chipperfield & Fleming, 1995).

Fitted values of all the model parameters are given in Table 1. Several conclusions can be drawn from the fitted values. First, the weight on negative examples is generally very low (approximately 0 in Experiments 4, 9, and 10, small in Experiment 5), suggesting that subjects were influenced almost exclusively by the positive example (cf. the "parity effect," which similarly implies that positive examples are more influential in inductions than negative examples; see Feldman, 2000). Second, the parameter called shape2skel, which is the weight of the (log) probability of a shape given a model, is substantially more than zero in about half the experiments. As mentioned above, previous similarity models based on graph matching methods included terms corresponding to skeleton modification,

[1] We also tried a version of the exemplar model using *maximum* instead of *mean* similarity, which has sometimes been found to perform better (Tenenbaum, 2000); but with our data, the mean-similarity version generally fit the data better.

Figure 7
The Effect of the Number of Positive Examples



Note. (a) The number of shapes chosen rises with the number of positive examples (aggregated over all experiments). (b) The same data are broken down by 2D versus 3D, showing the interaction among all three factors. See the online article for the color version of this figure.

but in the absence of a full probabilistic model did not include terms for the probability with which shapes actually correspond to the skeletal models in question. The fitted values of shape2skel suggest that this component of shape comparisons is important at least some of the time.

Once the parameters were fitted, we computed Akaike information criterion scores (AICs), which compensate for the different number of fitted parameters in each model, thus finally arriving at a suitable comparison of model fits among models.[2] Lower values of Akaike information criterion (AIC; lower minus log likelihood after complexity correction) indicate better fit to the data. As is conventional (see Burnham & Anderson, 2002) when displaying the model, we subtract out the AIC of the "winning" (minimum-AIC) model, leaving only the AIC relative to the best fit available, referred to as ΔAIC. (Note that because AIC is a logarithmic measure, only AIC differences, not ratios, are meaningful.) In such a comparison, the winning model "disappears" (ΔAIC = 0). Note ΔAICs can be exponentiated ($e^{\Delta AIC/2}$) to indicate the relative weight of evidence of one model compared to another (again see Burnham & Anderson, 2002). Figure 8 shows the fits of each model to the data from Experiments 1–5 (2D shapes), and Figure 9 from Experiments 6–10 (3D shapes). Note that the scales on the ordinates of these plots are large, so even a visually small difference can indicate a large evidence ratio between models.

Finally, we took advantage of the additivity of AIC to create aggregate plots of AICs summing across experiments, giving cumulative AICs to indicate the aggregate degree of fit to the entire ensemble of data. Figure 10a shows an aggregate plot of the AIC comparisons summing Experiments 1–5 (2D), and likewise Figure 10b for Experiments 6–10 (3D). Finally, Figure 10c gives a single aggregate comparison for all 10 experiments.

Discussion of Model Fits

Overall, the generative similarity measure performed very well, achieving the minimum AIC across all models (ΔAIC = 0) in all 10 experiments. Figure 8 shows model fits for Experiments 1–5 (2D shapes), and Figure 9 for Experiments 6–10 (3D shapes), giving ΔAIC for the generative similarity, skeletal cross-likelihood, active contours, Erdogan–Jacobs, CNN MNIST, and skeletal deviation models. Figure 10 shows fits aggregated across experiments, including the 2D cases (Figure 10a), 3D cases (Figure 10b), and overall aggregate (Figure 10c). The ΔAIC margin between generative similarity and the second-best model (usually the active contour model) was very large (mean ΔAIC over experiments = 2064, minimum = 735), corresponding to a likelihood ratio of at least $e^{735/2}$, or about 4×10^{159}, in favor of generative similarity.

The fit of the models to subjects' data followed a consistent order across experiments. In the 2D experiments, it was always generative similarity > active contours > CNN MNIST > cross-likelihood > skeletal deviation > EJ. In the 3D experiments, it was always generative similarity > active contours > CNN MNIST > EJ > skeletal deviation.

A sense of the "absolute" performance of the generative similarity model is provided by Figure 11, which compares generative similarity to subjects' chosen shapes in Experiment 1 (2D 1:1). In the figure, each candidate shape is colored to indicate the frequency with which subjects chose it (darker red means more often chosen), and the seven shapes with the greatest generative similarity to the

[2] We also analyzed all of our data using BIC, which imposes a somewhat heavier complexity penalty than AIC. This did not change any of the results reported below, for example, the ordering among models, so for clarity of presentation we only present AIC results.

Table 1
Fitted Parameters for the Generative Similarity Model

Experiment	c	Branch pos.	Branch angle	α	Axis length	Axis pnlty.	shape2skel	Neg. example
Experiment 1, 2D 1:0	8.8423	0	0.2645	0.7355	0	79.6986	0	—
Experiment 2, 2D 2:0	6.6704	0.1915	0.1276	0.1074	0.205	67.6631	0.3685	—
Experiment 3, 2D 3:0	6.2213	0.2193	0.0833	0.0801	0.2314	64.4172	0.3859	—
Experiment 4, 2D 1:1	5.945	0.3	0.2384	0.1616	0.3	35.7593	0	0
Experiment 5, 2D 2:1	2.3429	0	0.0653	0.7619	0.0001	58.6194	0	0.1728
Experiment 6, 3D 1:0	0.0837	0.2911	0	0.6014	0	1.143	0.1075	—
Experiment 7, 3D 2:0	$1.0E-05$	0.2079	0.3997	0.3584	0.034	98.7818	0	—
Experiment 8, 3D 3:0	8.4005	0.4718	0.1733	0.1335	0.0045	18.32	0.2169	—
Experiment 9, 3D 1:1	0.0598	0	0.3338	0.3326	0.3336	10.1941	0	0
Experiment 10, 3D 2:1	2.3171	0.2439	0.3046	0.2720	0.0376	96.7593	0.1418	0.0001

Note. Explanations of parameters: c = sensitivity parameter in the exemplar model; branch pos. = weight on location on parent axis at which branching occurs; branch angle = weight on angle relative to parent axis at which branching occurs; α = weight on differences in turning angle at corresponding axis points; axis length = weight on axis length; axis pnlty. = penalty for the absence of a corresponding axis (DL units); shape2skel = weight on probability of the shape given the skeleton; neg. example = weight on the negative example; DL = description length.

positive example (left) marked by green boxes (7 because subjects were instructed that "most people choose about 7"). As can be seen in the figure, shapes that have high generative similarity to the positive example (green boxes) tended to be chosen with high probability (dark red coloring), and vice versa. In this sense, the DLs provided by the generative similarity model seem to give intuitive results. At the same time, as also illustrated in the figure, shapes that perfectly matched a training example were not necessarily chosen. Thus, though selection was influenced by similarity to the training examples, actual choices were still probabilistic.

As an additional evaluation of the predictive power of the generative similarity model, we next applied the model to data from a recent study of shape similarity by Morgenstern et al. (2021). This study collected similarity ratings for four groups of animal shapes. We ran the generative similarity model on these shapes and found that resulting DLs correlated strongly with the reported human similarity judgments (respectively $r = 0.2153$, 0.1630, 0.3255, and 0.4596, all BFs $> 9{,}000$), though admittedly not as strongly as Morgenstern et al.'s (2021) own similarity measure. Thus, while our primary data involve classification, this secondary analysis confirms that our measure is predictive of overt similarity ratings as well.

Next in overall model fit after generative similarity was CNN MNIST, the deep neural network. As discussed above, visual classifiers based on CNNs have been found to perform well on a variety of benchmark databases (Bates & Jacobs, 2019), and such models have become extremely prevalent in computer vision. However, our results suggest that in a direct comparison, such models do not account for human similarity judgments as well as more "structural" models such as generative similarity. As mentioned above, CNN classification performance is generally insensitive to structural aspects of images such as shape and part structure (Baker et al., 2017; Heinke et al., 2021). But our results suggest that human similarity judgments are best accounted for by a model that *is* sensitive to these factors, such as a skeleton-based measure.

Less easy to explain is the poor performance of the EJ model, which performed very well in a similar comparison to Erdogan and Jacobs (2017). Like our model, the EJ model is probabilistic and part-based, albeit with a very different likelihood function. The EJ likelihood function computes the probability of rendered images given 3D models based on solid volumes, whereas ours computes the probability of shape boundaries (approximated as 2D or 3D polygons) conditioned on skeletons. Some of the difference in performance presumably reflects the difference in likelihood functions and the differences in the actual shapes tested in the respective studies. Note that the idea that similarity depends on the statistical characteristics of the shapes under consideration is itself an argument in favor of probabilistic models generally—including both ours and EJ—because only in a probabilistic framework is such a dependence to be expected. That said, our stimuli are relatively naturalistic shapes (i.e., shapes with smooth boundaries and highly articulated part structure), which we would argue provide a reasonable test case.

Another important difference between the EJ model and ours is the mathematics of the similarity measure, which in the EJ model is essentially cross-likelihood (the probability of each shape conditioned on the other model), whereas ours is a quantification of the evidence in favor of a *common* model. The idea that shapes are similar to the extent that they appear to have common causal origins, which is the essential novelty in our approach, is not present in the EJ model. It is not unreasonable to suppose that some of the difference in performance reflects this difference.

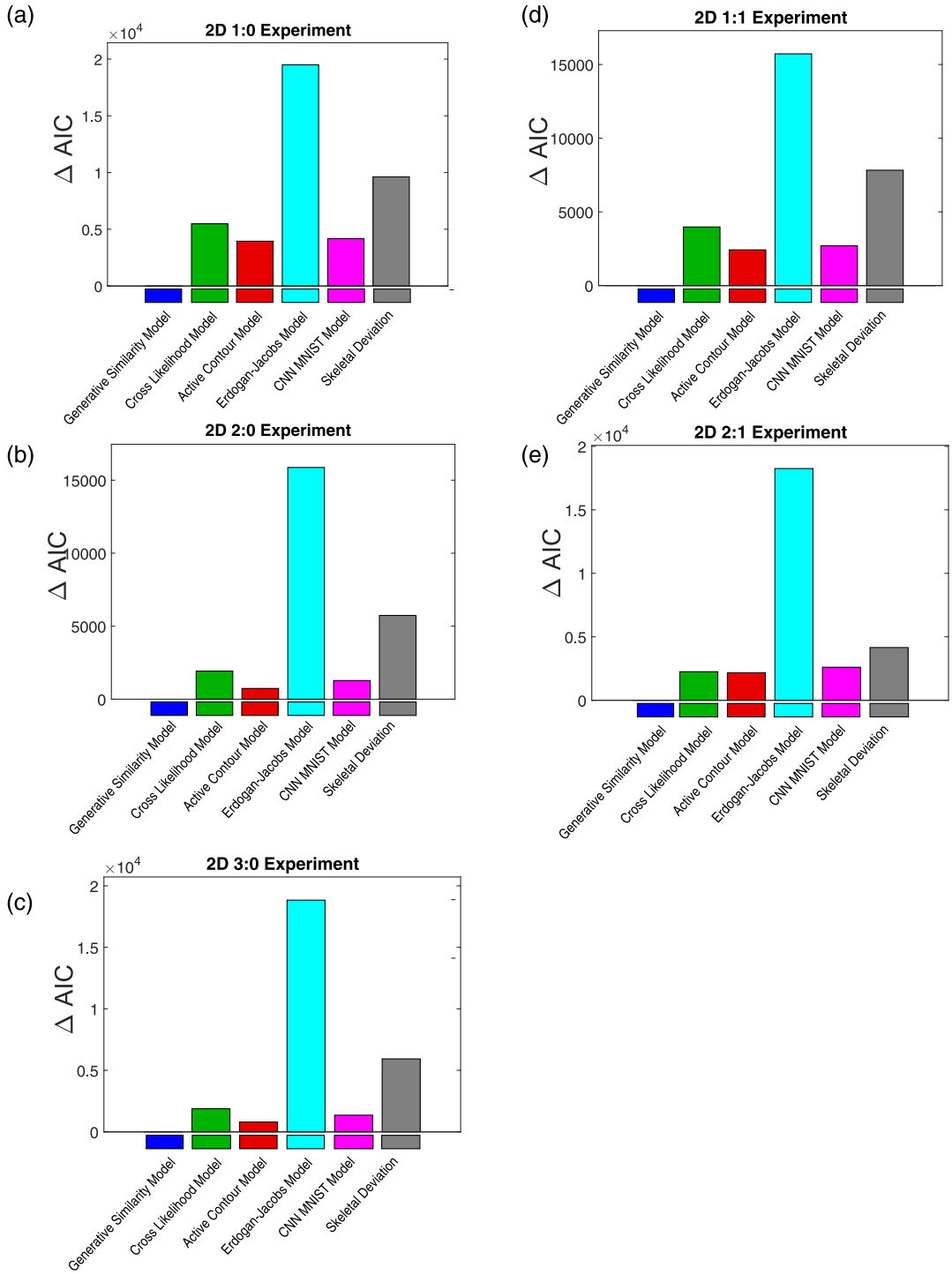
Contour-Based Similarity

One aspect of these results that deserves further comment is the consistently good performance of the active contour model, which came in second in most experiments. This model simply measures the squared deviation between corresponding points on the two contours and has no regional or "shape" component whatsoever. This model would be presumed by most contemporary shape researchers to be too simplistic to model human shape judgments. But our results suggest that—at least for simple 2D shapes—it actually fits human data rather well, better than many more nominally sophisticated models.

Indeed, a number of previous results (Basri et al., 1998; Fleming & Schmidt, 2019), including some of our own findings (Wilder et al., 2015, 2016), suggest that contour structure makes a substantial contribution to shape representation independent from that of region structure. That is, two shapes can seem similar because their *boundaries* have similar characteristics—for example, they are

Figure 8

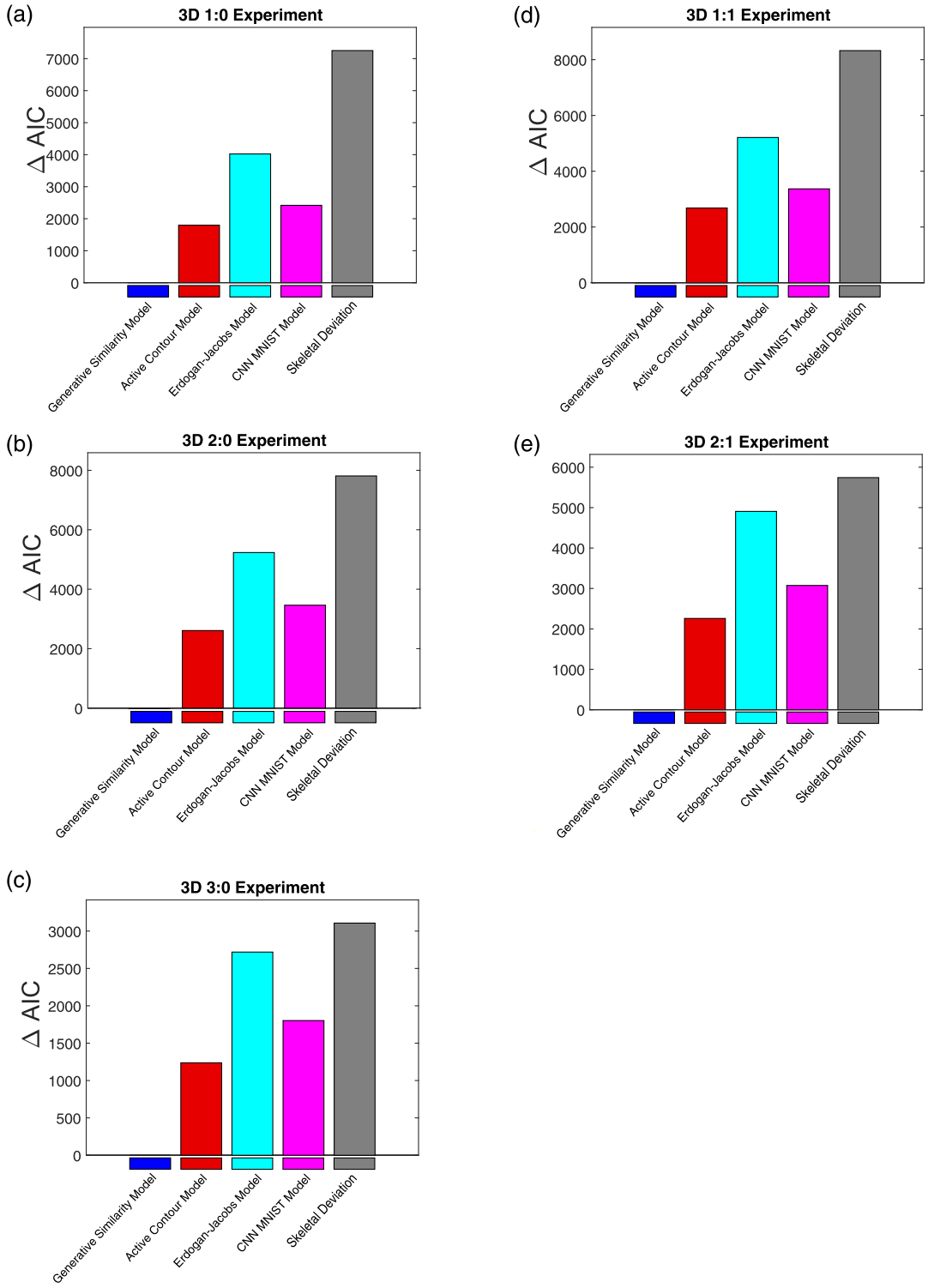
Results From Experiment 1–5 (2D Experiments)



Note. (a) Experiment 1 (1:0). (b) Experiment 2 (2:0). (c) Experiment 3 (3:0). (d) Experiment 4 (1:1). (e) Experiment 5 (2:1). Note differences among scales on the ordinates. See the online article for the color version of this figure.

Figure 9

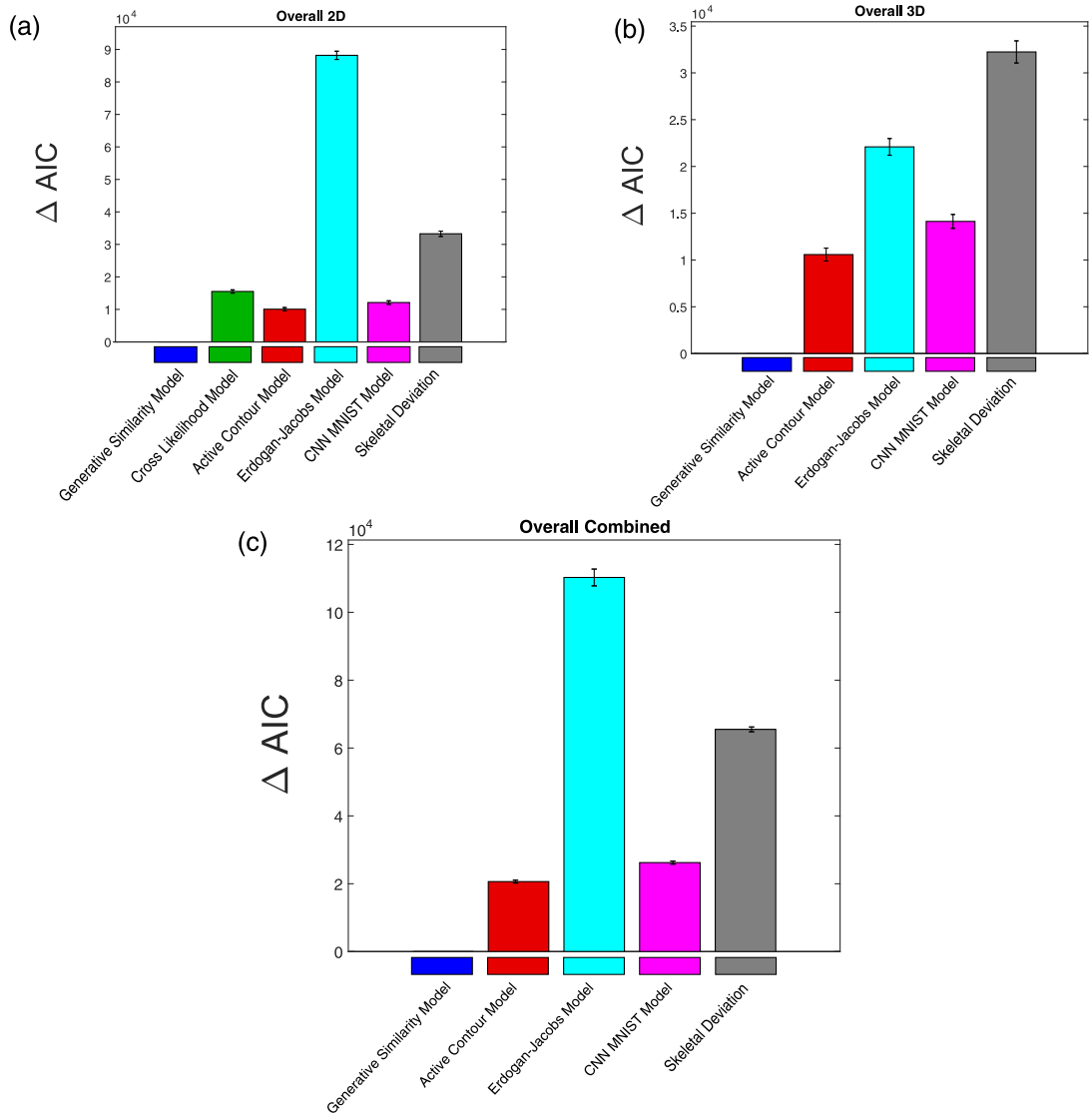
Results From Experiment 6–10 (3D Experiments)



Note. (a) Experiment 6 (1:0). (b) Experiment 7 (2:0). (c) Experiment 8 (3:0). (d) Experiment 9 (1:1). (e) Experiment 10 (2:1). Note differences among scales on the ordinates. See the online article for the color version of this figure.

Figure 10

Aggregate Fits (Summed AICs) From (a) Experiments 1–5 (2D Experiments), (b) Experiments 6–10 (3D Experiments), and (c) All Experiments



Note. *SE* = standard error. Differences among scales on the ordinates. Error bars are ±1 *SE* around the mean aggregating over experiments. See the online article for the color version of this figure.

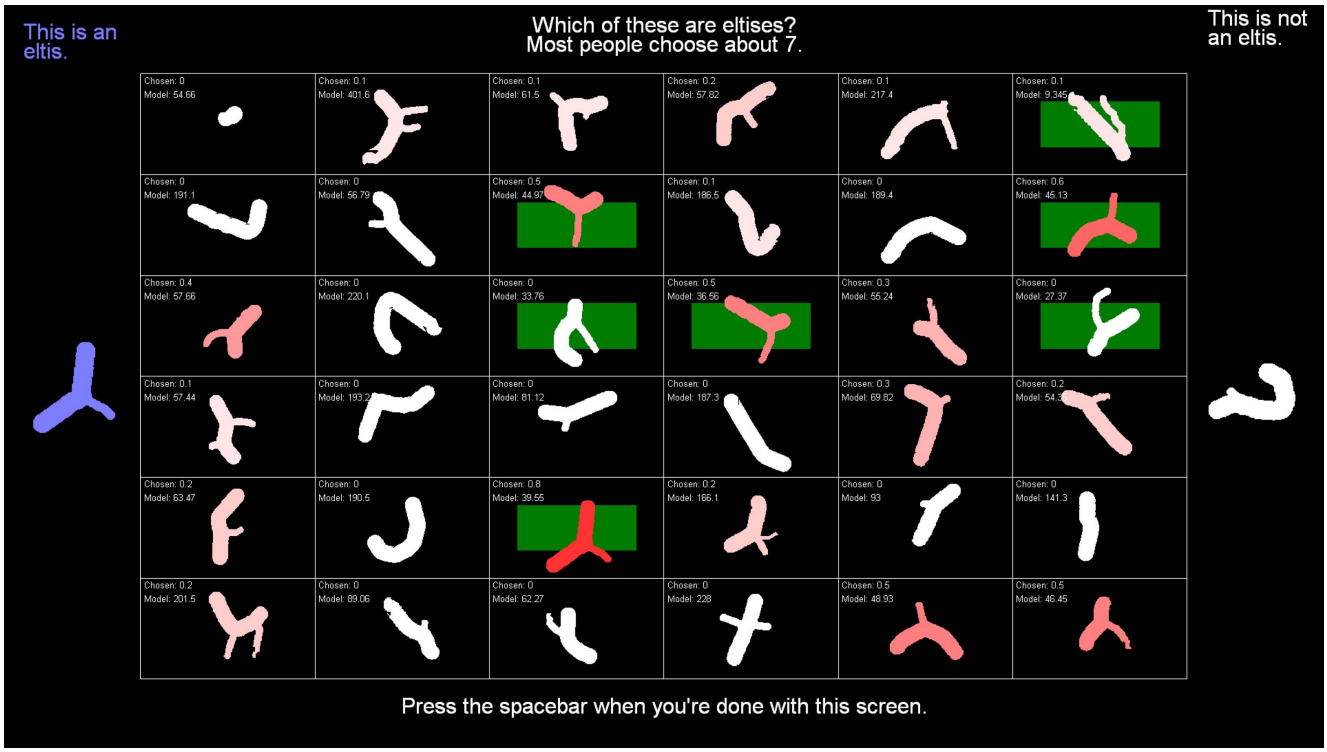
similarly "fuzzy" or "bumpy"—separate from the configural qualities of the enclosed region.

The generative similarity model as presented above is primarily regional, meaning that it represents shapes as regions generated by the shape skeleton, without any specifically contour-based component. (Note, though, that it is *not* based solely on similarity of the skeletons themselves, as are classical graph matching methods; it also includes a measure of the probabilistic degree of fit between each shape and its skeleton.) However, as emphasized in our previous articles (e.g., Feldman et al., 2013), there is a very close mathematical affinity between the *axis*-generating probabilistic process in the skeletal generative model and a *contour*-generating process (e.g., as presented in Feldman & Singh, 2005). This means that it is fairly straightforward to define a similarity measure over contours, analogous to the model above, except with a probabilistic contour-generating boundary process substituting for a region-generating skeletal process. For example, the probabilistic cost of transforming one contour into another would depend on the integrated log probability of the von Mises distribution over differences in corresponding turning angles—just as for axes in the above model. The resulting contour similarity model is almost identical to the skeletal model, except there is no "shape lattice" because all contours have the same topology.

We ran a version of the generative similarity model that incorporated this contour similarity component, with an additional free parameter representing the relative weight of the contour component

Figure 11

Illustration of Absolute Performance of the Generative Similarity Model (Data From Experiment 1, 2D 1:1)



Note. DL = description length. Each shape is colored according to the frequency with which subjects selected it (darker red = higher frequency), and the seven shapes that the generative similarity model judged most similar to the positive example (left) are indicated by green boxes. Also indicated for each shape is the proportion of subjects who chose it and its DL relative to the positive example under the generative similarity model. See the online article for the color version of this figure.

relative to the skeletal component. We found that it did indeed substantially improve the fit of the model, typically by about 200–300 AIC points across the board. This difference does not change the overall conclusion from the model comparisons, namely that generative similarity fits human judgments better than other models whether region-based (skeletal deviation, cross-likelihood, and EJ), contour-based (active contours), or image-based (MNIST). (Recall that ΔAIC between our model and the next best was always more than 700 points; including the contour component would have simply increased this margin to 900–1,000 points.) However, the difference is very large in absolute terms and is thus enough to conclude that—in keeping with the previous literature mentioned above—human shape similarity judgments do indeed include an element of pure contour similarity. We have not included the contour component in our basic model description above, in part to avoid overburdening the article and in part to keep the conceptual focus on skeleton-based similarity models. But the downloadable version of our code (see link above) includes an option to incorporate the contour component in similarity calculations.

General Discussion

Shape similarity is a fundamental and ubiquitous problem, underlying not only shape category formation (Landau et al., 1988), but also shape recognition (Ayzenberg & Lourenco, 2019; Biederman, 1987), superordinate categorization (Tiedemann et al., 2022; Wilder et al., 2011), and shape discrimination (Destler et al., 2019). In many other contexts, similarity is assumed to take the form of exponential decay in some simple metric space (Jäkel et al., 2008; Luce, 1959; Minda & Smith, 2011). However, shape representation is notable for lacking any simple metric structure and is thus not easily associated with a simple probability distribution. That is, there is no single parameter or set of parameters of shape over which to measure distances. Instead, shape similarity, like shape representation itself, depends on numerous subtle and complex configural qualities that are difficult to express in any simple way. This is why shape similarity has proven to be such a difficult problem.

Indeed, Ashby and Perrin (1988) have argued that even when a simple similarity space is available, human judgments of perceptual similarity are not actually well accounted for by simple metric distances. Instead, they argued that similarities reflect the structure of the probability distributions associated with the stimulus space. In our proposal, there is no comprehensive metric space and thus no simple way to define a distribution over one. Instead, we assume that shapes are the product of complex probabilistic generative processes (Ons & Wagemans, 2012; Sprote & Fleming, 2016), modeled by the skeletal prior and likelihood function. Our skeletal probability framework yields shapes with a parametric structure that depends on the topology of the shape skeleton, growing in complexity as the shape skeleton itself grows and branches. In this framework, we can

use Bayes' rule to estimate the likely generative model for a given shape or set of shapes.

The core idea underlying the resulting similarity measure is that two items should be regarded as similar in proportion to the evidence that they share a common generative origin—that is, the same skeletal model. We quantify that evidence via the Bayes factor in favor of a *common* model relative to the alternative hypothesis that the shapes have *distinct* generative models. The similarity measure is high when two (or more) shapes appear to be the result of common generative processes and is low when they appear to be of distinct kinds.

More specifically, the shape lattice illustrated in Figure 2 is a way of conceptualizing the *set of potential common models*. In the lattice, the most likely common model of skeletons A and B lies within the lattice join $A \vee B$, which is the skeletal model containing all the axes common to both shapes. The pathway on the lattice from A to B necessarily passes through the common model: the path from A to $A \vee B$ represents the *removal* of axis branches present in A but not present in $A \vee B$, whereas the path from $A \vee B$ to B represents the *addition* of axes present in B but not present in $A \vee B$. Because the generative model specifies the probabilities of axis additions and removals, the "length" (really, DL) of the path from skeleton A to skeleton B represents the probability of transforming A into B (or vice versa). Once these skeletons are coupled with shapes via the likelihood function $(p(a|A),p(b|B))$, the resulting overall DL represents the probability cost of transforming one shape into another (via their respective MAP skeletons connected by the least common model). This is the link between the "common model" conceptions of generative similarity and the more traditional "edit distance" conception of similarity: The BF in favor of the common model turns out to incorporate the probabilistic penalty associated with transforming one shape into the other.

Our experimental data suggest that this similarity measure predicts human category judgments well, providing the best fit for human judgments in the 2D experiments, the 3D experiments, and overall. Given this strong empirical support, combined with its principled derivation, we would argue that the generative similarity model represents the best available account of human shape similarity. More broadly, these results argue strongly for a skeleton-based account of human similarity judgments, and in particular, one in which shapes are regarded as similar to the extent that they share common generative models.

Conclusion

Why does a cat look more like a dog than it looks like a hammer? Our answer (see again Figure 3) is, first, because cats and dogs have relatively similar skeletal structures: They both have body plans in which a head, forelegs, hind legs, and a tail are attached to a central torso—while a hammer has an elongated handle attached to a small perpendicular head. In this sense, cats and dogs have a relatively specific common model, whereas cats and hammers only share the coarsest of common models, a central elongated body. On the other hand, cats and dogs still look somewhat different because the parameters of their limbs and other parts have somewhat different probability distributions, for example, dogs have longer legs relative to their bodies, more horizontal tails, and so forth.

Our similarity measure incorporates all these aspects—differences in skeletal topology, differences in axial parameters, and so forth—in a unified probabilistic framework. Of course, the framework is very simple; more finely tuned similarity comparisons would require more elaborate and carefully chosen distributional assumptions, perhaps tailored to particular classes of shapes. However, given its impressive performance in fitting our subjects' data, especially in comparison to a number of sophisticated alternative proposals, we conclude that our framework provides a way forward in solving the difficult problem of shape similarity.

References

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, *120*(2), 150–172. https://doi.org/10.1037/0096-3445.120.2.150

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*(1), 124–150. https://doi.org/10.1037/0033-295X.95.1.124

Ayzenberg, V., Chen, Y., Yousif, S. R., & Lourenco, S. F. (2019). Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of Vision*, *19*(6), Article 6. https://doi.org/10.1167/19.6.6

Ayzenberg, V., Kamps, F. S., Dilks, D. D., & Lourenco, S. F. (2022). Skeletal representations of shape in the human visual cortex. *Neuropsychologia*, *164*, Article 108092. https://doi.org/10.1016/j.neuropsychologia.2021.108092

Ayzenberg, V., & Lourenco, S. (2022). Perception of an object's global shape is best described by a model of skeletal structure in human infants. *eLife*, *11*, Article e74943. https://doi.org/10.7554/eLife.74943

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), Article 9359. https://doi.org/10.1038/s41598-019-45268-y

Bai, X., & Latecki, L. J. (2008). Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(7), 1282–1292. https://doi.org/10.1109/TPAMI.2007.70769

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2017). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, *14*(2), Article e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Basri, R., Costa, L., Geiger, D., & Jacobs, D. (1998). Determining the similarity of deformable shapes. *Vision Research*, *38*, 2365–2385. https://doi.org/10.1016/s0042-6989(98)00043-1

Bates, C. J., & Jacobs, R. A. (2019). Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science*, *28*(1), 34–39. https://doi.org/10.1177/0963721418801342

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 509–522. https://doi.org/10.1109/34.993558

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, *94*, 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Blake, A., & Isard, M. (2012). *Active contours: The application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*. Springer Science & Business Media.

Blum, H. (1967). A transformation for extracting new descriptors of shape. *Models for the Perception of Speech and Visual Form*, *19*(5), 362–380.

Blum, H. (1973). Biological shape and visual science (Part I). *Journal of Theoretical Biology*, *38*, 205–287. https://doi.org/10.1016/0022-5193(73)90175-6

Briscoe, E. (2008). *Shape skeletons and shape similarity* [Unpublished doctoral dissertation]. Rutgers University.

Burbeck, C., & Pizer, S. (1995). Object representation by cores: Identifying and representing primitive spatial regions. *Vision Research*, *35*, 1917–1930. https://doi.org/10.1016/0042-6989(94)00286-U

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. Springer.

Chaisilprungraung, T., German, J., & McCloskey, M. (2019). How are object shape axes defined? Evidence from mirror-image confusions. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(1), 111–124. https://doi.org/10.1037/xhp0000592

Chipperfield, A., & Fleming, P. (1995). *The matlab genetic algorithm toolbox*. IET.

Cortese, J. M., & Dyre, B. P. (1996). Perceptual similarity of shapes generated from fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 133–143. https://doi.org/10.1037/0096-1523.22.1.133

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), Article e57410. https://doi.org/10.1371/journal.pone.0057410

Davey, B., & Priestley, H. (1990). *Introduction to lattices and order*. Cambridge University Press.

Demirci, M. F., Shokoufandeh, A., Keselman, Y., Bretzner, L., & Dickinson, S. (2006). Object recognition as many-to-many feature matching. *International Journal of Computer Vision*, *69*(2), 203–222. https://doi.org/10.1007/s11263-006-6993-y

Destler, N., Singh, M., & Feldman, J. (2019). Shape discrimination along morph-spaces. *Vision Research*, *158*, 189–199. https://doi.org/10.1016/j.visres.2019.03.002

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Science*, *21*(4), 449–467. https://doi.org/10.1017/s0140525x98001253

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychological Review*, *124*(6), 740–761. https://doi.org/10.1037/rev0000086

Farabet, C., Martini, B., Akselrod, P., Talay, S., LeCun, Y., & Culurciello, E. (2010). Hardware accelerated convolutional neural networks for synthetic vision systems. *Proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 257–260). IEEE. https://doi.org/10.1109/ISCAS.2010.5537908

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*, 145–170. https://doi.org/10.1006/jmps.1997.1154

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633. https://doi.org/10.1038/35036586

Feldman, J., & Singh, M. (2005). Information along contours and object boundaries. *Psychological Review*, *112*(1), 243–252. https://doi.org/10.1037/0033-295X.112.1.243

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014–18019. https://doi.org/10.1073/pnas.0608811103

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. D. (2013). An integrated Bayesian approach to shape representation and perceptual organization. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision: An interdisciplinary perspective* (pp. 55–70). Springer.

Firestone, C., & Scholl, B. J. (2014). "Please tap the shape, anywhere you like": Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, *25*(2), 377–386. https://doi.org/10.1177/0956797613507584

Fleming, R. W., & Schmidt, F. (2019). Getting "fumpered": Classifying objects by what has been done to them. *Journal of Vision*, *19*(4), Article 15. https://doi.org/10.1167/19.4.15

Greene, E. (2018). New encoding concepts for shape recognition are needed. *AIMS Neuroscience*, *5*(3), 162–178. https://doi.org/10.3934/Neuroscience.2018.3.162

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*, 1–32. https://doi.org/10.1016/S0010-0277(02)00184-1

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape-similarity judgments. *Psychological Science*, *20*(4), 447–454. https://doi.org/10.1111/j.1467-9280.2009.02310.x

Harrison, S., & Feldman, J. (2009). Influence of shape and medial axis structure on texture perception. *Journal of Vision*, *9*(6), Article 13. https://doi.org/10.1167/9.6.13

Heinke, D., Wachman, P., van Zoest, W., & Leek, E. C. (2021). A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision. *Vision Research*, *189*, 81–92. https://doi.org/10.1016/j.visres.2021.09.004

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review* , *7*(4), 569–592. https://doi.org/10.3758/bf03212996

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*, 65–96. https://doi.org/10.1016/0010-0277(84)90022-2

Hung, C.-C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*(6), 1099–1113. https://doi.org/10.1016/j.neuron.2012.04.029

Jacobs, R. A., & Bates, C. J. (2019). Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science*, *28*(1), 34–39. https://doi.org/10.1177/0963721418801341

Jäkel, F., Schölkopf, B., & Wichmann, F. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin and Review*, *15*(2), 256–271. https://doi.org/10.3758/PBR.15.2.256

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1785–1790). Lawrence Erlbaum Associates.

Kimia, B. B. (2003). One the role of medial geometry in human vision. *Journal of Physiology-Paris*, *97*, 155–190. https://doi.org/10.1016/j.jphysparis.2003.09.003

Kovács, I., Fehér, A., & Julesz, B. (1998). Medial-point description of shape: A representation for action coding and its psychophysical correlates. *Vision Research*, *38*, 2323–2333. https://doi.org/10.1016/S0042-6989(97)00321-0

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

Lescroart, M. D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex*, *23*(3), 629–637. https://doi.org/10.1093/cercor/bhs046

Leymarie, F. F., & Aparajeya, P. (2017). Medialness and the perception of visual art. *Art & Perception*, *5*, 169–232. https://doi.org/10.1163/22134913-00002064

Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception & Psychophysics*, *80*(5), 1278–1289. https://doi.org/10.3758/s13414-017-1457-8

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*(2), 566–577. https://doi.org/10.1016/j.cognition.2008.03.009

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B*, *200*, 269–294. https://doi.org/10.1098/rspb.1978.0020

Mervis, C., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115. https://doi.org/10.1146/annurev.ps.32.020181.000513

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation. predictions. and limitations. In E. M. Pothos & A. J. Wills

(Eds.), *Formal approaches to categorization* (pp. 40–64). Cambridge University Press.

Morgenstern, Y., Hartmann, F., Schmidt, F., Tiedemann, H., Prokott, E., Maiello, G., & Fleming, R. W. (2021). An image-computable model of human visual shape similarity. *PLOS Computational Biology*, *17*(6), Article e1008981. https://doi.org/10.1371/journal.pcbi.1008981

Morgenstern, Y., Schmidt, F., & Fleming, R. W. (2019). One-shot categorization of novel object classes in humans. *Vision Research*, *165*, 98–108. https://doi.org/10.1016/j.visres.2019.09.005

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Ons, B., & Wagemans, J. (2012). Generalization of visual shapes by flexible and simple rules. *Seeing & Perceiving*, *25*(3–4), 237–261. https://doi.org/10.1163/187847511X571519

Panis, S., Vangeneugden, J., & Wagemans, J. (2008). Similarity, typicality, and category-level matching of morphed outlines of everyday objects. *Perception*, *37*(12), 1822–1849. https://doi.org/10.1068/p5934

Rezanejad, M., & Siddiqi, K. (2015). Flux graphs for 2d shape analysis. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision* (p. 41–54). Springer.

Richards, W., Dawson, B., & Whittington, D. (1988). Encoding contour shape by curvature extrema. In W. Richards (Ed.), *Natural computation* (pp. 83–98). MIT Press.

Schmidt, F., Kleis, J., Morgenstern, Y., & Fleming, R. W. (2020). The role of semantics in the perceptual organization of shape. *Scientific Reports*, *10*(1), Article 22141. https://doi.org/10.1038/s41598-020-79072-w

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Siddiqi, K., & Kimia, B. B. (1996). A shock grammar for recognition. In B. Bhanu, C. Dyer, & K. Ikeuchi (Chairs), *Proceedings CVPR'96, 1996 IEEE computer society conference on computer vision and pattern recognition, 1996* (pp. 507–513). IEEE Computer Society.

Siddiqi, K., Kimia, B. B., Tannenbaum, A., & Zucker, S. W. (1999). Shapes, shocks and wiggles. *Image and Vision Computing*, *17*(5), 365–373. https://doi.org/10.1016/S0262-8856(98)00130-9

Siddiqi, K., Shokoufandeh, A., Dickinson, S. J., & Zucker, S. W. (1998). Shock graphs and shape matching. In D. Goldgof, A. Jain, D. Terzopoulos, & Y.-F. Wang (Chairs), *Proceedings of the sixth international conference on computer vision* (p. 222). IEEE Computer Society.

Sprote, P., & Fleming, R. W. (2016). Bent out of shape: The visual inference of non-rigid shape transformations applied to objects. *Vision Research*, *126*, 330–346. https://doi.org/10.1016/j.visres.2015.08.009

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in neural information processing systems* (Vol. 12, pp. 59–65). MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/s0140525x01000061

Tiedemann, H., Schmidt, F., & Fleming, R. W. (2022). Superordinate categorization based on the perceptual organization of parts. *Brain Sciences*, *12*(5), Article 667, https://doi.org/10.3390/brainsci12050667

Torsello, A., & Hancock, E. R. (2004). A skeletal measure of 2d shape similarity. *Computer Vision and Image Understanding*, *95*(1), 1–29. https://doi.org/10.1016/j.cviu.2004.03.006

Vedaldi, A., & Lenc, K. (2015). Matconvnet: Convolutional neural networks for MATLAB. In X. Zhou (Ed.), *Proceedings of the 23rd annual ACM conference on multimedia conference, MM '15, Brisbane, Australia, October 26–30, 2015* (pp. 689–692). ACM. https://doi.org/10.1145/2733373.2807412

Wang, X., & Burbeck, C. A. (1998). Scaled medial axis representation: Evidence from position discrimination task. *Vision Research*, *38*(13), 1947–1959. https://doi.org/10.1016/S0042-6989(97)00299-X

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, *119*, 325–340. https://doi.org/10.1016/j.cognition.2011.01.009

Wilder, J., Feldman, J., & Singh, M. (2015). Contour complexity and contour detection. *Journal of Vision*, *15*(6), Article 6. https://doi.org/10.1167/15.6.6

Wilder, J., Feldman, J., & Singh, M. (2016). The role of shape complexity in the detection of closed contours. *Vision Research*, *126*, 220–231. https://doi.org/10.1016/j.visres.2015.10.011

Received April 27, 2022
Revision received October 29, 2022
Accepted November 20, 2022 ■