Vision Research 165 (2019) 98–108

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

One-shot categorization of novel object classes in humans

Yaniv Morgenstern*, Filipp Schmidt, Roland W. Fleming

Department of Experimental Psychology, Justus-Liebig University Giessen, Giessen 35394, Germany

ARTICLE INFO

Keywords:
Visual perception
Categorization
Classification
Objects
Shape
Computational modelling

ABSTRACT

One aspect of human vision unmatched by machines is the capacity to generalize from few samples. Observers tend to know when novel objects are in the same class despite large differences in shape, material or viewpoint. A major challenge in studying such generalization is that participants can see each novel sample only once. To overcome this, we used crowdsourcing to obtain responses from 500 human observers on 20 novel object classes, with each stimulus compared to 1 or 16 related objects. The results reveal that humans generalize from sparse data in highly systematic ways with the number and variance of the samples. We compared human responses to 'ShapeComp', an image-computable model based on > 100 shape descriptors, and 'AlexNet', a convolution neural network that roughly matches humans at recognizing 1000 categories of real-world objects. With 16 samples, the models were consistent with human responses without free parameters. Thus, when there are a sufficient number of samples, observers rely on shallow but efficient processes based on a fixed set of features. With 1 sample, however, the models required different feature weights for each object. This suggests that one-shot categorization involves more sophisticated processes that actively identify the unique characteristics underlying each object class.

1. Introduction

One remarkable characteristic of human vision is the capacity to classify novel objects from few samples. Given a single novel object, we have clear intuitions about what other class members may look like (Fig. 1A). What makes human novel object classification remarkable is that our inferences are often accurate, even early on in our visual development (Gelman & Markman, 1986; Gelman & Meyer, 2011; Gopnik & Sobel, 2000). Novel object categorization based on few samples contrasts strikingly with machines trained on 'big data', which can paint in the style of artistic masters (Gatys, Ecker, & Bethge, 2015), beat Go grandmasters (Gibney, 2016; Silver et al., 2016) and rival humans at object recognition (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014; He, Zhang, Ren, & Sun, 2016; Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2014; Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, & Fergus, 2013; Szegedy et al., 2015). Despite enormous progress in artificial intelligence, it still remains baffling how humans make class inferences from hardly any data.

Previous work formulated the problem of generalizing from few samples in probabilistic terms (Shepard, 1987; Shepard, 2001; Tenenbaum & Griffiths, 2001). Shepard (1987) introduced a "universal" negatively accelerating generalization function that predicts the probability of perceiving two stimuli as belonging to the same class (Fig. 1A). The key assumption is that novel samples are represented as points in a metric, high-dimensional psychological feature space, where an observer judges nearby samples to be in the same class.

Tenenbaum and Griffiths (2001) developed a Bayesian extension that predicts how novel object samples influence the likelihood of items belonging to a common class. Their model predicts that the decision boundary steepens with increases in sample size due to greater certainty about the range of feature values that define the object class (Fig. 1B). For example, a botanist encountering two similarly shaped new flowers (Fig. 1A, orange and green items), might be inclined to classify them as the same species at first, whereas another novel item (in red) is so different that it appears to be a different class from the first encounter. However, given more samples (Fig. 1B), it becomes clear that the orange flower is an outlier and probably belongs to a different class. Another prediction of the likelihood model is that increasing sample variability expands the decision boundary region (Fig. 1C) to include items that differ more from each other. In this case, given such diverse forms, the botanist would likely assign the orange flower to the same species as the others (Fig. 1C).

The evidence for processes underlying these inferences is inconclusive, with theoretical ideas that span a continuum of complexities. At one end, some have argued for heuristic models (e.g., Shepard, 1987; Tenenbaum & Griffiths, 2001; Hegdé, Bart, & Kersten, 2008; Kromrey,

* Corresponding author at: Department of Experimental Psychology, Justus-Liebig University Giessen, Otto-Behaghel-Str. 10F, Giessen 35394, Germany.
E-mail address: Yaniv.Morgenstern@psychol.uni-giessen.de (Y. Morgenstern).

https://doi.org/10.1016/j.visres.2019.09.005
Received 7 May 2019; Received in revised form 17 September 2019; Accepted 18 September 2019
Available online 07 November 2019

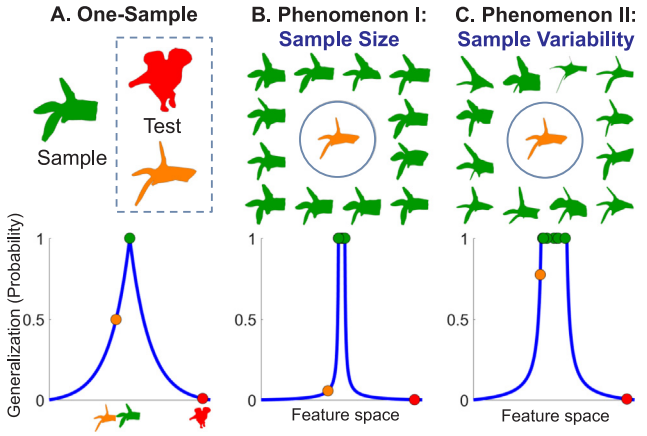

Fig. 1. The influence of sample size and variability on generalization. (A) *Generalization with one sample.* Given a single sample from a novel object class (in green), there is a falloff in the probability that other items are seen as belonging to the same class, as a function of distance in feature space (Shepard, 1987). Thus, the orange test item has a higher probability of being in the same class than an object whose feature values are more different (red). (B) *Generalization with many samples.* With many samples, Tenenbaum and Griffiths (2001) showed that the generalization function steepens, leading to greater certainty about the location of the decision boundary, causing the orange item to appear to belong to a different class. (C) *Generalization with many highly variable samples.* Higher variability across samples shifts the decision boundary outwards, increasing the probability that the orange test item does belong to the same class as the green samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Maestri, Hauffen, Bart, & Hegdé, 2010) based on measuring differences between objects in some pre-established features space. Such models provide a simple and efficient strategy, ubiquitous across human decision making (Gigerenzer & Gaissmaier, 2011). At the other end, others have argued that novel object generalizations in humans is based on more sophisticated models (Fei-Fei, Fergus, & Perona, 2006; Feldman, 1992, 1995, 1997; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Goodman, Tenenbaum, Griffiths, & Feldman, 2008; Lake, Salakhutdinov, & Tenenbaum, 2015; Stuhlmuller, Tenenbaum, & Goodman, 2010) that represent the underlying processes responsible for the object, making it possible to create new samples, and determine which features of an observed sample are most important. Traditionally, in the disciplines of vision science, and, more broadly, psychology, these approaches have been pitted against each other (Gigerenzer & Gaissmaier, 2011; Kingdom, 1997). We suggest that humans likely rely on both approaches. We reason that when there is a sufficient number of novel object class samples, observers compare novel objects using simple heuristics, based on relationships between observed features. In contrast, when samples are sparse, as in generalizations from only one sample, observers invoke more sophisticated inferences, involving a deeper analysis of the shape's characteristics. Here, we sought to (a) characterize generalization of novel object classes from few samples, and, (b) evaluate to what extent human behavior is accounted for by simple strategies based on a fixed weighting of features or more sophisticated strategies involving sample-specific feature weighting.

Testing the predictions of different models is challenging because each participant can only be tested once, yet many judgments are required to map out the decision boundaries around items. To overcome this, we used crowdsourcing to collect a small number of responses from a large number of participants, allowing us to fit psychometric functions to characterize between-subject performance.

Specifically, we generated twenty classes of novel objects defined by complex 2D shapes whose parts varied along multiple spatial dimensions (position, orientation, length, and width). We asked hundreds of participants to judge whether a test shape belonged to the same class as a small set of samples (1 or 16) of related shapes. Then, to tease apart the extent that decisions are better predicted by simple strategies based on a fixed weighting of features or more sophisticated strategies involving sample-specific feature weighting, we compared the responses to two different image-based models: (a) a novel shape model ('ShapeComp') based on over 100 image-based shape descriptors from the computer vision literature (Zhang & Lu, 2004), and (b) AlexNet, a convolutional neural network trained to classify objects in photographs with performance that matches humans (Krizhevsky et al., 2012).

2. Experiment 1: Generalization from few samples in humans

2.1. Materials and methods

2.1.1. Participants

For the main experiment, 500 paid observers were recruited through ClickWorker (a crowdsourcing platform) at a rate of 75 eurocents for a 5-minute task. Participants provided informed consent by clicking "I agree" to an online consent form, approved by the ethics board at Justus-Liebig-University Giessen and in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Participants reported normal or corrected-to-normal vision. An additional 135 observers participated in a second experiment to estimate between-subject consistency.

2.1.2. Stimulus generation

We created 20 base shapes by scaling and stitching parts of two different 2D shapes from the MPEG-7 database (Fig. S1A; Latecki, Lakamper, & Eckhardt, 2000). We generated novel samples by transforming the base shape's skeletal representation (Fig. S1B; Feldman & Singh, 2006) to produce new shapes with limbs that varied in length, width, spatial position, and orientation (Fig. S1C). The *length* of each limb varied by resizing its axial branch (Fig. S1B). The *length* of the root contour (main branch; in green) was kept constant. The *width* of each limb varied by scaling the lengths of all the ribs attached to its axial branch by a common factor. The spatial position of each limb, defined as where the axial branch attached to the root contour, varied along the main branch. The root contour did not vary in its spatial position. The orientation of the limbs also varied relative to the main branch.

There are many other ways to create novel shapes (most of which would also be more-or-less heuristic). We also think that the conclusions are not an artifact of these stimulus generation decisions, but this is an open question.

The experimental stimulus frame for each base shape included 1 test shape presented centrally and either 1 (one) or 16 (many) sample shapes presented in the surround (Figs. 3, S1E, and S1F). The number of surrounding shapes (1 or 16) was chosen to test whether and how novel object category decisions for a very sparse sample (1) contrasts with less sparse samples (16) to test the hypothesis that humans evoke more class-specific than general processes when samples are sparse. Specifically, we reasoned that when 16 samples are presented, observers receive sufficient information about the distribution of features values across samples from the class, to base their judgments on a generic weighting of the various features. In contrast, when only a single sample is presented, there is a high degree of ambiguity about the true distributions of feature values. We hypothesized that to overcome this ambiguity, observers may identify and assign greater weight to certain features from the exemplar shape (i.e, a class-specific weighting of features).

We selected test and sample shapes based on their *shape dissimilarity* from the base shape, measured in terms of the differences in their skeletal parameters (i.e., the width, length and angle of the shapes' limbs). When shapes are similar, their skeletal structure tends to be similar and thus the underlying skeletal parameters are nearer to each other. As the difference in skeletal parameters between two shapes increases, they appear progressively more dissimilar from one another (see Figs. S1D, S2 and 3B). Given that we varied shapes within a class

with the shape skeleton algorithm, we used this intuition to define *shape dissimilarity* as the Euclidean distance between the skeletal parameters of two shapes (see also Destler, Singh, and Feldman (2019) for another definition of shape similarity). The sample shapes were similar to the base shape; sample shape parameters were randomly chosen from a Gaussian centered around the base shape parameters and with small variance (Table S1). The test shapes varied in their similarity relative to the base shape in terms of 25 distance bins (see supplemental methods for more details). There were 10 trials for each distance bin, with different test and sample shapes chosen for each trial. In total there were 250 trials per each base shape (each shown to a different participant), or a total of 10,000 trials across base shapes and sample size conditions.

2.1.3. Stimuli varying in sample variability

The 20 base shapes varied in the number of skeletal limbs from 4 to 7. Base shapes with more limbs tended to produce novel shape samples that showed greater class variability (Figs. 2, S3, and S4). In the many samples condition, more limbs produced greater variation in sample shape within and across trials. In the one sample condition, more limbs produced greater variation in sample shape only across trials, since one sample was present for any given trial. In our data analysis, the 13 base shapes with 4 or 5 limbs were classified as the low variability group, and the 7 base shapes with 6 or 7 limbs were classified as the high variability group.

2.1.4. Procedure

Each observer participated in 22 trials. On each trial, observers judged whether an object presented in a central *test* region was in the same class as 1 or 16 object(s) presented in the surround (Fig. 3A). The specific instruction given to observers was the following: "Does the shape inside the circle belong to the same class/group as the shape(s) in the surround?" Supporting data and code are made available through Data-In-Brief. The first 20 trials showed one of the 250 experimental stimulus frames, selected randomly, for each base shape. In order to prevent observers from using newly learned inferences underlying the variation in the sample shapes, the stimuli with one sample shape in the surround were presented before the sixteen sample shapes. This step ensures human responses in the 1-sample condition are not swayed by the variation across objects in the 16-samples condition. Otherwise, the presentation order was random. The last two 'catch' trials determined if observers were paying attention and understanding the task. One catch trial presented identical test and sample shapes. Another catch trial presented a target shape coming from a different base shape than the sample shapes. Roughly 5% of participants failed at least one of the catch trials and their data was not analyzed. New participants repeated trials belonging to participants who failed the catch trials. No participant was allowed to repeat the experiment.

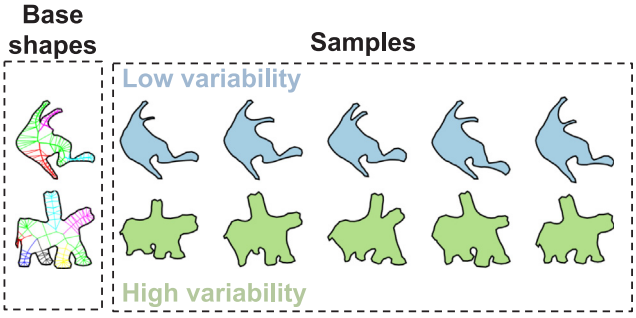

Fig. 2. Base shapes and examples of synthesized samples. The base shape skeleton (left) was used to synthesize new shapes, including the experimental samples (right) surrounding the test stimulus. The skeletal parameters for the samples were chosen from a normal distribution centered on the base shape parameters with small variance leading to samples that tended to appear similar to the base shape. The shape samples produced by the top base shape with 5 skeletal limbs appear to vary less in their appearance across samples than the bottom base shape with 7 skeletal limbs. Samples with less skeletal limbs (< 6) were classified as low variability samples. Samples with more skeletal limbs ($> = 6$) were classified as high variability samples. Example shape samples for the remaining novel object classes are available in the supplementary materials.

2.1.5. Characterizing category boundaries

Here we characterize the location and uncertainty in the category boundary with a maximum likelihood fit of a cumulative Gaussian (with psignifit 4; Schütt, Harmeling, Macke, & Wichmann, 2016) to the human responses as a function of the average shape dissimilarity between the test and the surrounding samples for a given trial.

2.2. Results and discussion

Fig. 3A shows the results for an example shape for the one and many-samples condition, where 'YES' indicates the target belonged to the same class as the samples (data for all 20 classes shown in Fig. S5). Shape dissimilarity was defined as distance in terms of the underlying skeletal parameters used to define the stimuli (see Methods). YES responses fall off with increasing shape dissimilarity, consistent with Shepard (1987) generalization function. Fig. 3B gives an impression of the shapes that observers tended to judge as within the same class.

One potential hypothesis is that observers abstract more in the one-sample condition by extending their decision boundary to include shapes that are more dissimilar. However, the psychometric function parameters show that the novel object decision boundary (DB; where the stimulus was said to be within the category at a rate of 50%) is almost identical for the one- and many-sample conditions (Fig. 3A), meaning that observers tend to switch from YES to NO responses at roughly the same shape dissimilarity index irrespective of context. The population tendency to agree on object class membership (i.e., slope), however, is lower for the one-sample condition, meaning that across observers the boundary is fuzzier (i.e., there is less consistency in the boundary location across observers). These findings are consistent with the Tenenbaum and Griffiths likelihood model (which predicts a steeper generalization function with more samples)—suggesting that with more samples it becomes clearer that a probe stimulus is within the same class as the samples.

Fig. 4A shows the data pooled across 20 object classes with low (in blue) and high sample variability (in green) for the one- and many-sample conditions. These results are consistent with the Tenenbaum and Griffiths likelihood model. The likelihood model predicts that increasing sample variability expands the decision boundary region (Fig. 1C). Our results confirm that highly variable objects have significantly expanded DBs; a two-way ANOVA with the DB as the dependent variable reveals that the psychometric functions for highly variable object classes have higher DBs ($F_{1,39} = 11.31$, $p < 0.01$) with no significant interaction between the number of samples and their variability ($F_{1,39} = 0.28$, $p = 0.60$) and no main effect for number of samples ($F_{1,39} = 0.02$, $p = 0.90$) (Fig. 5A). The likelihood model also predicts how sample variability influences confidence in boundary location across participants, with lowered confidence with smaller sample size, and with higher sample variability. A two-way ANOVA confirms that boundary confidence (indicated by slope of the psychometric function) decreases with less samples ($F_{1,39} = 55.63$, $p < 0.01$) and with more highly variable samples ($F_{1,39} = 13.61$, $p < 0.01$), with no interaction between these factors ($F_{1,39} = 0.31$, $p = 0.58$) (Fig. 5B). These findings support a Bayesian generalization strategy that is strongly influenced by changes in the likelihood term which is based directly on comparing features from one object with that of the other (Tenenbaum & Griffiths, 2001). These findings, and the likelihood model's predictions, are also consistent with strategies that predict increases in confidence with more visual stimuli.

These changes in the psychometric functions reveal how generalization varies across conditions: with fewer samples, observers' uncertainty about DB location increases, and with greater variability

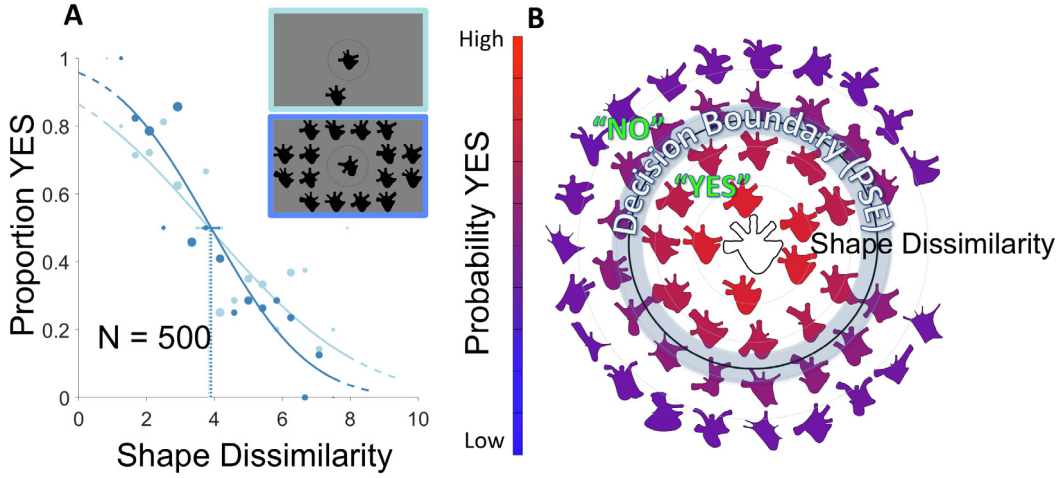

Fig. 3. Crowd-sourced responses for an example object. (A) Probability YES as a function of shape dissimilarity in the one- (light blue) and many-sample (dark blue) conditions (500 observers, one response each). The error bars show the decision boundary location (PSE) ± 1 standard deviation (estimated with 1000 bootstrap simulations). (B) Polar plot showing potential target shapes as a function of shape dissimilarity from the many-samples condition in (A). The radial axis shows Euclidean distance in generative parameters between original base shape and synthesized shapes. Note that the shape boundary falls along a one-dimensional shape similarity axis, as shown in (A). The angular position in (B) is employed for illustration purposes to show more shape samples, and does not depict meaningful dimensions in a high-dimensional space. Shapes closer to the center appear more similar to the sample shapes, yielding a higher proportion of YES responses (red), than more distant objects, which tend to yield NO responses (blue). The decision boundary shows the 50% YES point from the psychometric function fits in A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

across samples, the DBs expand, as does the uncertainty in its location. Thus, consistent with Fig. 1, when presented with only a single sample, participants are less certain whether a novel test stimulus belongs in the same class as the samples; and with increasing sample variability, observers generalize more, such that items that differ more from one another are classed together, although, again, they have greater uncertainty about their decision. It is important to note that for the single-sample condition, the 'high variability' condition refers simply to the greater variability across trials (performed by different participants), and thus the shallowness of the psychometric function reflects the statistics of the stimuli rather than a change in internal representation of the participants.

3. Modelling: Generalization from few samples in ShapeComp and AlexNet

What remains unclear is whether one-shot categorization is based on simple heuristics or more flexible decision strategies. A common assumption is that stimuli are represented in a perceptual feature space acquired during learning of familiar objects. In such a feature space, a given stimulus is represented as having a particular set of feature values, and decisions about whether two stimuli belong to the same or different categories are based simply on their distance in the feature space (Shepard, 1987). Importantly, with the simplest heuristic account, the same strategy (or feature set) operates across both familiar and unfamiliar objects, across object classes, and irrespective of set size. Such approaches are highly efficient, but lack flexibility to adjust decision criteria to the distinctive characteristics of new stimuli (except, perhaps in the very long-term, as extensive experience alters the features; e.g., Adams, Graf, & Ernst, 2004; Green & Bavelier, 2007; de Beeck & Baker, 2010). Alternatively, it could be that observers' decisions are context sensitive, based on a different weighting of features for each object class.

To test the decision strategy underlying novel object categorization from few samples, we implemented two feature spaces for novel-object categorization that were derived from different image-based models. The purpose of using two feature spaces (rather than one) is to evaluate whether the decision strategy is reproducible rather than preconditioned on a specific set of features. In other words, we sought to test whether the conclusions depend on the specific features used to describe the images, or whether multidimensional representations in general can capture the main tendencies. One feature space—ShapeComp—was based on > 100 hand-selected features known to be useful shape descriptors in computer vision. A second feature space was derived from AlexNet, a Convolutional Neural Network (ConvNet) trained to categorize natural objects, and with neural structure and response characteristics reminiscent of visual cortical processing.

With these feature spaces in hand, we then sought to test the way in which observers combined information from different features. Given a single exemplar of a novel object class, how do observers work out whether another object belongs to the same class or not? In the absence of any prior knowledge about the world, this is fundamentally ill-posed, because objects in the same class could be arbitrarily similar or different from one another. Yet despite this, we find that observers have systematic intuitions about whether two objects that they have never seen before belong in the same class or not. The question is: how? Here we test a number of alternative hypotheses.

One possibility is that observers use a constant, fixed set of feature weights, related to their previous experience of the variance of feature values within and between classes. The essential idea is the following. Having learned many classes of objects in the past, observers have internalized how much each shape feature tends to vary within and between classes (i.e., they have estimated the distribution of values for each feature). They then measure the feature values for exemplar(s) and target and compare the distance in feature space with their stored knowledge about how feature values are distributed in general. For example, if test and sample objects differ much more than members of the same class normally do in terms of how *elongated* the shape is, this would provide evidence that the two objects likely belong to different classes. This evidence may be reinforced or overruled by the difference between the objects in terms of other features (difference in overall size, number of limbs, curvature properties, etc.). However, according to this first hypothesis, the thresholds for each feature—for determining whether the objects are in the same or different classes—are fixed and constant across all stimuli that observers are shown, because they are predetermined by prior experience across many classes of objects and are not based on the sample object itself. We test this hypothesis with

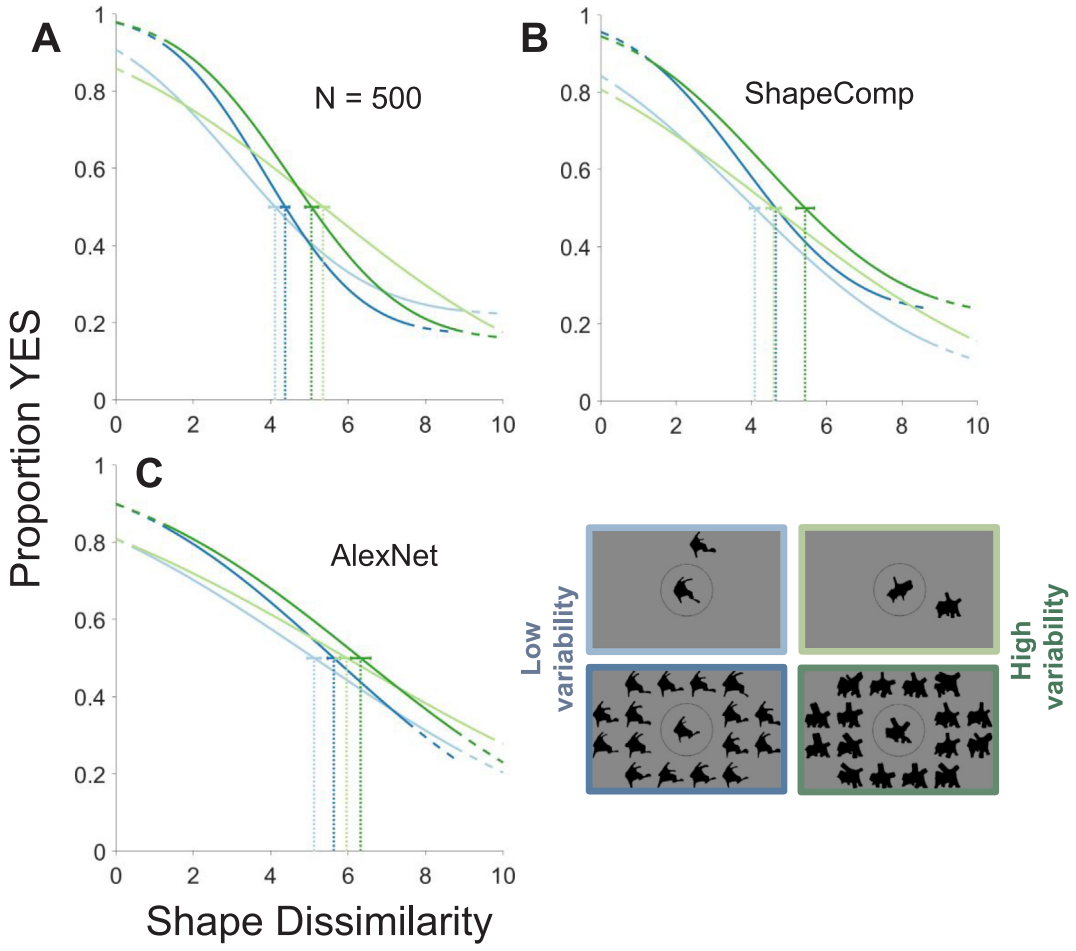

Fig. 4. Crowd-sourced human responses and model responses pooled across objects. (A) Proportion YES as a function of shape dissimilarity in the one (light) and many-samples (dark) under low (blue) and high (green) sample variability. (A) 500 crowd-sourced human observers that responded to each shape only once, (B) responses from ShapeComp, an image-computable model (based on ca. 100 shape descriptors), and (C) responses from AlexNet (Krizhevsky et al., 2012). In both human and model, the decision boundary (50% YES point) tends to be higher in more variable samples, and confidence (i.e., slope) is lower for the one- than many-samples condition. These results are qualitatively consistent with the likelihood model of Tenenbaum and Griffiths (2001). Error bars show the decision boundary location (PSE) ± 1 standard deviation (estimated with 1000 bootstrap simulations). Fig. S6 shows the psychometric function fits above embedded with the data points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the *parameter-free* and *global weighting* models described below, in which the weights of features are held fixed across different classes.

An alternative is that on a case-by-case basis, for each exemplar the observers are shown, observers select which features to base their decision on. For example, suppose that an exemplar has a particularly pronounced (i.e., atypical) feature, like a distinctive curly limb (Fig. 6). The observer may choose to base their decisions about class membership more on that particular feature than on other features of the object. In other words, they modify which features they use to evaluate class membership depending on the characteristics of the exemplar shape. We test this hypothesis with the *class specific* weighting model, described below, in which the weights of features in the model vary from

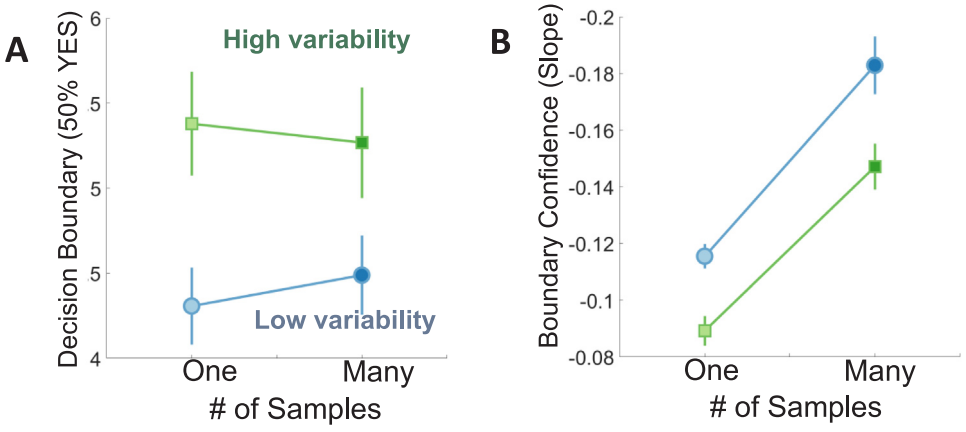

Fig. 5. Changes in the human decision boundary with sample properties. The decision boundary location (A) and confidence (B) estimated from fits to human observer responses for the one- and many-samples condition under high (green) and low (blue) sample variability. Error bars show standard error across objects. Mean decision boundary across objects tends to be higher in more variable samples. Confidence (i.e., slope) is lower for the one- than many-samples condition. These results are qualitatively consistent with the likelihood model of Tenenbaum and Griffiths (2001). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

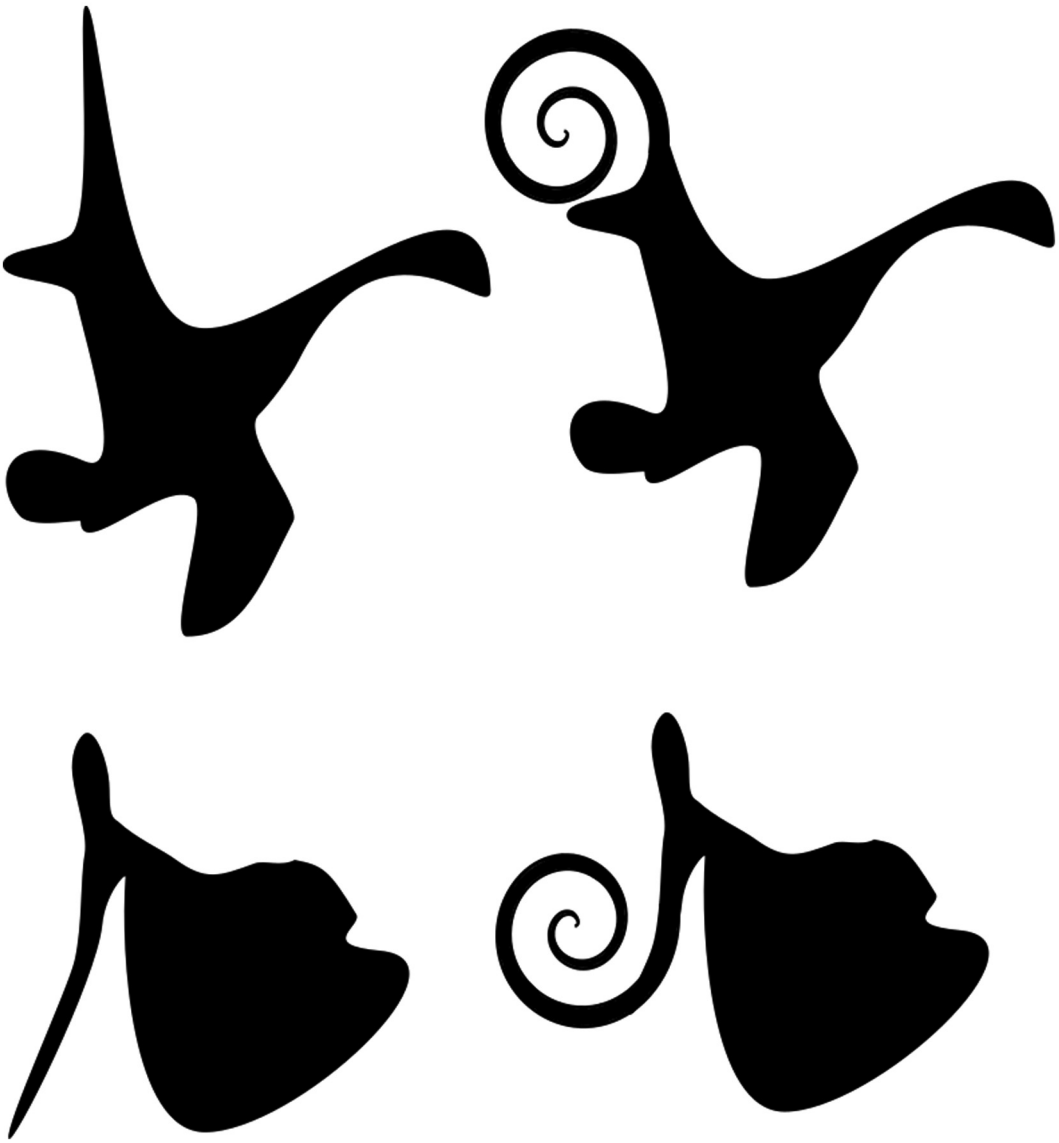

Fig. 6. Feature weighting models. The shape pairs along the rows share many more properties (e.g., curvature, elongation, area) than the shape pair in the right column. Despite these feature similarities, one atypical feature can sway an observer's decision about class membership. In this example, observers may decide to group the shape pair in the right column as belonging to the same class leading to decisions where some features (e.g., the curly limb) are weighted more heavily than others (e.g., curvature, elongation, area).

class to class.

3.1. Materials and methods

3.1.1. Between-subject consistency

To estimate the noise in human observer responses, we ran a second crowd-sourcing experiment with new observers, to create a benchmark for comparing human responses to generic image computable models, which can only account for the variation shared across observers. The human correlation with other humans provides a criterion on good model performance, as we would not expect an image computable model observer to outperform another human as a model for average human behavior. Thus, in this experiment, stimuli were repeated across different observers. We had 300 stimuli repeated 9 times, and computed the human correlation with the mean of the other humans on exactly the same trials. We computed confidence intervals with 10,000 bootstrapped simulations (randomly chosen stimuli with replacement). We did not measure noise within observers, using methods like the double-pass technique (Burgess & Colborne, 1988), because such methods would require gathering human responses to the same stimuli more than once, and this would no longer be a one-shot categorization task.

3.1.2. ShapeComp – Image computable model

We posited a set of 106 features based on descriptions of planar shape known to be important for recognition, synthesis, and perception (Zhang & Lu, 2004). The features spanned a range of complexity from simple (e.g., area) to complex (i.e., shape skeleton).

a) *Simple Shape Descriptors:* Area, Perimeter, Circularity (perimeter^2/Area), Eccentricity (length of major axis/length of minor axis), Major axis orientation, Convexity (ratio of perimeters of the convex hull over that of the original contour), Circularity ratio (area of shape to area of circle, where the circle has the same perimeter), Extent, Solidity, Orientation, Compactness, Convex area, the × and y coordinates of the Centroid, Curviness, Area/Perimeter, and the skewness, kurtosis, and standard deviation of the vertical and horizontal distribution of pixels (as in Paulun, Kawabe, Nishida, & Fleming, 2015).

b) *Correspondence based matching:* A joint histogram of distance (bins = 15) and angle (bins = 15) known as the shape context (Belongie, Malik, & Puzicha, 2002), as well as a histogram of chord lengths (bins = 100) and chord angles (bins = 100).
c) *Shape signature:* Centroid/radius distance, curvature, triangular area from centroid, complex coefficients, tangent angle, and cumulative angle. We also took the Fourier transform of these signatures to be additional feature.
d) *Frequency decomposition:* We used 11 features based on Fourier Descriptors (Zahn & Roskies, 1972; Granlund, 1972; Pavlidis, 1980) that a more commonly used in behavioral (Cortese & Dyre, 1996; Wilder, Fruend, & Elder, 2018) and neural studies (Albright & Gross, 1990; Op de Beeck, Wagemans, & Vogels, 2001). The first 10 low frequency components were included as single features. The last feature compared all the components. We found similar ShapeComp model responses with an Elliptical Fourier Descriptor (FD) representation (Kuhl & Giardina, 1982).
e) *Surprisal:* Information, in Shannon's sense, along contours (Feldman & Singh, 2005), both signed and unsigned.
f) *Boundary moments:* We took the mean, standard deviation, skew, and kurtosis of the shape descriptors in (c) and the histogram of chord lengths and angles in (b).
g) *Shape skeleton* (Feldman & Singh, 2005). We summarized the shape skeleton with the 8 parameters described in (Wilder, Feldman, & Singh, 2011), and also included variance of children branch angles relative to parent branch, variance of branch length relative to root, and the mean, standard deviation, skewness, and kurtosis of rib lengths.

The features were evaluated on each trial by taking the sum RMS error for the test on that feature with the surrounding shape samples. Since different features come from different spaces whose values have vastly different meaning, some feature distances will lead to larger values merely because their values tend to be larger. We put these features on a common scale using a histogram equalization method. There are many other ways to scale these responses; we chose this efficient coding principle in part because it also accounts for the responses of fly neurons to contrast (e.g., Laughlin, 1981), and converts luminance contrast into human lightness and color responses via a host of neural receptive field properties (e.g., Morgenstern, Rostami, & Purves, 2014; Morgenstern, Rukmini, Monson, & Purves, 2014; Purves, Morgenstern, & Wojtach, 2015). Specifically, we normalized each distance to range from [0, 1] by taking the cumulative probability of the feature distance relative to distances of that feature typically found within natural animal categories. This nonlinearly transformed the feature distances in accordance with visual experience, such that the range of values for more common distances across naturally occurring shapes were expanded (given more sensitivity), and, more importantly, outlying distances were compressed (given less sensitivity). To do this, first, we computed the same 106 shape features on a natural animal database consisting of 2000 animal shapes with 100 animals belonging to 20 categories (Bai, Liu, & Tu, 2009). For each feature, we computed the empirical cumulative distribution function (CDF) to all within animal category distances. On any given trial, the feature distance between test and sample was transformed using the empirical CDF. As a consequence, the CDF-transformed feature distances ranged from [0, 1] by compressing outlying larger distances much like Shepard's negatively accelerating generalization function compresses the range of probabilities for larger feature distances. In the *parameter-free* model, the 106 features distances, now transformed into a common space, were converted into responses r for each trial t, as follows:

$$r_t = \frac{\sum_{f=1}^{106}\sum_{s=1}^{n} d_{sf}}{n},$$

where d is the Euclidean distance between test and sample s of n (1 or 16) for feature f (1 out of 106). Thus, in the *parameter-free* model, there is no fitting to the human responses, and the feature weights are constant across all conditions.

Since there are 106 features, the expectation is that these features tap into different shape qualities such that combinations of these features will, between them, represent complementary aspects of the shapes. Thus, some features lead to smaller features distances for a pair of shapes than other features. Given the complementary nature of these features, the r_t values do not occupy the full range of values from 0 to 1; in our experiment the values ranged from 0.34 to 0.87 across all objects. To ensure that the model responses took up the entire range, for each base shape and sample size condition, the responses r were scaled from [0,1] across the 250 trials for each object, as follows:

$$\vec{s} = 1 - \left(\frac{\vec{r} - \min(\vec{r})}{\max(\vec{r}) - \min(\vec{r})}\right),$$

such that the scaled model responses close to 1 indicate lower shape dissimilarity.

In *global weighting* and *class specific* models, a lasso regression (see below) was used to estimate a sparse set of weights w for the 106 CDF transformed features, as follows:

$$r_t = \Phi\left(\sum_{f=1}^{106}\sum_{s=1}^{n} d_{sf} + c\right)$$

where c denotes a constant (intercept) term and Φ is the normal (Gaussian) cumulative distribution function.

The difference between the *global weighting* and *class specific* models is in how the weights were determined. In the *global weighting* model, for each novel object, ShapeComp is based on a weighted combination of features, where the weights were determined by a regression with the human response and features across the remaining 19 novel objects (used in the current experiment). The *global weighting* model tests the idea that there is some constant weighting of feature responses that operates across all human responses to unseen novel objects. In the *class specific* model, the weights were determined by a regression with the human responses and features across the novel object. The *class specific* model tests the idea that observers use different feature weightings for different novel objects when they make their class decisions.

3.1.3. AlexNet

AlexNet features were based on the final fully connected layer with 1000 outputs and computed with MATLAB using the neural network toolbox. Like ShapeComp, the *parameter-free* AlexNet model combined features on each trial by taking the sum RMS error for the test and surrounding shape samples for the 1000 outputs or features as follows:

$$r_t = \frac{\sum_{f=1}^{1000}\sum_{s=1}^{n} d_{sf}}{n},$$

where d is the Euclidean distance between test and sample s of n (1 or 16) for feature f (1 out of 1000). Thus, smaller differences between the target shape and sample shapes indicate lower shape dissimilarity. These responses r were scaled from [0,1] across the 250 trials for each object, as follows:

$$\vec{s} = 1 - \left(\frac{\vec{r} - \min(\vec{r})}{\max(\vec{r}) - \min(\vec{r})}\right),$$

such that the scaled model responses close to 1 indicate lower shape dissimilarity.

The *global weighting* and *class specific* AlexNet models were also fit with a lasso regression (see below) to estimate a sparse set of weights w for the 1000 scaled feature responses, as follows:

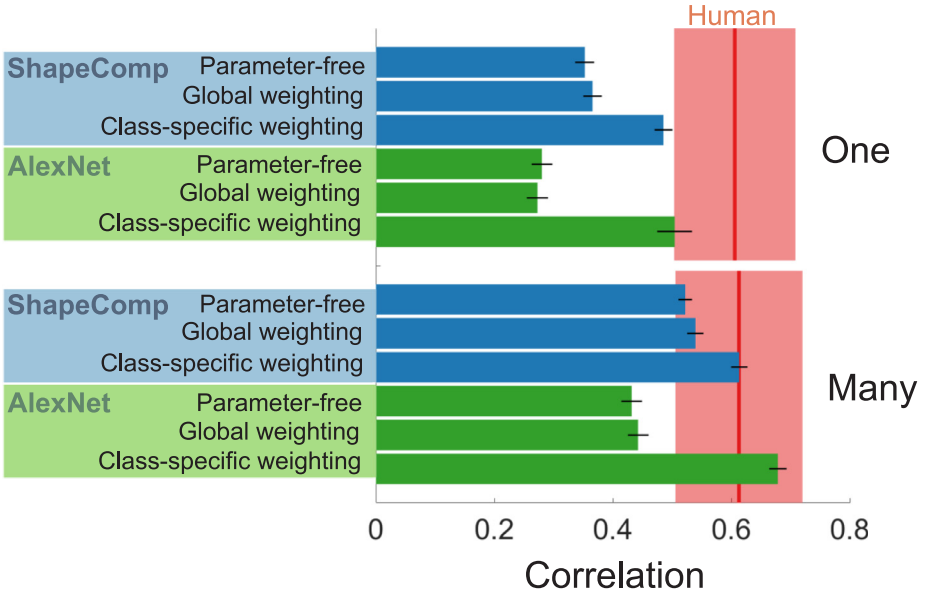

Fig. 7. Correlation of human responses to models. Red vertical bars indicate the human-to-human correlation for the same trials (95% confidence intervals estimated via 10,000 bootstrapped simulations), which provides a criterion for good model performance. The parameter-free and global weighting models better account for the human responses in the many condition than they do in the one condition. In the one-sample condition, a separate weighted model (based on a *lasso* regression) for each object (*class-specific weighting*) better accounts for responses suggesting that the underlying decision strategies are more sophisticated for one-shot categorization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$r_t = \Phi\left(\sum_{f=1}^{1000}\sum_{s=1}^{n} d_{sf} + c\right)$$

where c denotes a constant (intercept) term and Φ is the normal (Gaussian) cumulative distribution function.

3.1.4. Psychometric function fits

Psychometric function parameters were estimated with a maximum likelihood fit of a cumulative Gaussian (with psignifit 4; Schütt et al., 2016) to the scaled model responses as a function of the average shape dissimilarity between the test and the surrounding samples for a given trial (for implementation see supplemental MATLAB code Fig. 4b.m and Fig. 4c.m available with the supporting data through Data-In-Brief).

3.1.5. Regression

For the *class-specific* and *global weighting models,* we used a lasso regularization for a generalized linear model (using lassoglm in MATLAB) with a probit link function. To guard against overfitting, we used 5-fold cross-validation to regress the features from ShapeComp and AlexNet on the human data and we selected the regression coefficients that minimized the deviance across cross-validated data. In the few cases that minimum deviance led to only zero model coefficients, to ensure some feature selection, we selected the lambda regularization term that provided the fewest regression coefficients. The *parameter-free* model uses the same strategy across novel objects and so it is not regressed on the novel object class.

3.2. Results and discussion

3.2.1. Generalization from few samples with ShapeComp

ShapeComp uses 106 'hand-engineered' features to capture a wide variety of geometrical properties that are known to be effective at describing shape. The shape descriptors range from simple scalar features (e.g., area, perimeter, compactness; see also Peura & Iivarinen, 1997) to more complex quantities based on Fourier descriptors, shape signatures, and the shape skeleton (Kuhl & Giardina, 1982; Belongie et al., 2002; Feldman & Singh, 2006; see Methods 3.1.2). We sought to test whether human generalization judgments could be predicted by distances in the space spanned by such image-computable features.

For each trial, we computed root mean square error between target and sample shapes features. We normalized distances across features to the range [0, 1] by taking the cumulative probability of within-category feature distances from a dataset of natural animal shapes (see Methods 3.1.2). As a byproduct of this normalization, within a feature, larger distances were compressed much like they would be in Shepard's negatively accelerating generalization function. The model responses were then the sum of the normalized distances across features.

Comparing human and model responses via fit psychometric functions, we find that, like the human data, ShapeComp shows the two qualitative phenomena predicted by Tenenbaum and Griffith's theory (Fig. 4B). Specifically, both humans and the model tend to see more dissimilar samples as part of the same class with high variability, and are more uncertain about their decisions in the one-sample condition.

To evaluate whether human responses were better accounted for by a fixed or flexible weighting of the features, we then combined ShapeComp's features in different ways. We defined a simple heuristics approach with a model that determines the differences between objects without learning or relying on a different strategy for each object class. Accordingly, we derived two types of simple heuristic models: (a) a *parameter-free* model that compares distances between sample objects within and across features (i.e., model responses from Fig. 4B), and (b) a *global weighting* model where ShapeComp is based on a weighted combination of features, where the weights were determined by a regression with the human response and features across the remaining 19 novel objects. The *parameter-free* ShapeComp model in (a) is a very rigid way of combining features. There could exist, for example, some other weighted combination of features that is preferred across categories. The *global weighting* model in (b) is a stronger test of the simple heuristic model since the model is based on fits to many other objects (not including the actual novel object), which means that the weights do not specify feature preference that is unique to the object in question, but work as a decision strategy across objects. The simple heuristics approaches determine differences between objects without relying on a strategy that is uniquely optimised for each class. We contrasted the parameter-free and global weighting models against a model that allowed *class-specific weighting* of features, to represent a flexible, context-sensitive inference process. That is, we developed models with free parameters that adjusted the relative weights of each feature in the model, separately for each class.

To compare the models with the range of human responses, in another crowd-sourcing experiment, separate participants responded to a subset of the stimuli. However, unlike the main experiment, all participants saw the same stimuli, allowing us to evaluate between-subject consistency. We then measured how well models correlated with

human data relative to the average correlation between observers. Fig. 7 shows the 95% confidence intervals around the mean inter-subject correlation.

The correlation of the *parameter-free* and *global-weighting* real-valued simple heuristic model responses with the human YES-NO data reveal that, at least for the many-samples condition, the distributions of stimuli in the high-dimensional image-computable feature space are sufficient to predict human performance. In the many samples condition, no class specific 'weighting' of features is necessary to work out how to generalize. Importantly, the same model could not predict the same amount of variance in the one-sample condition. This suggests that humans use more sophisticated and computationally demanding processes to generalize from a single sample than a fixed strategy across samples. One obvious possibility is that in the absence of sufficient data about the distribution, participants rely on a generative model consisting of some prior, derived from previous experiences with other object classes (see, for example, Stansbury, Naselaris, & Gallant, 2013), where the prior relies on specific 'weighted' features. The *class-specific weighting* model falls within the confidence intervals of human performance (Fig. 7), suggesting that human responses in the one-sample condition could be based on a stimulus-by-stimulus reweighting of features.

It is important to note that we do not intend ShapeComp to be a simulation of processes in the human visual system; that is, we do not believe that the human visual system computes these specific features. Rather, we see ShapeComp as a proof of principle that generalization from sparse data can be achieved by representing shape in terms of a large set of complementary geometrical properties. To test whether these findings are specific to ShapeComp's feature space or reproducible across feature spaces, we also compared human novel object decisions in a feature space that arises from previous experience with real world objects.

3.2.2. Comparison to performance of object recognition neural network

It could be argued that the shape features that we selected for the image computable model are somewhat arbitrary, and that better or worse performance might be achieved through the use of an alternative feature space. Moreover, the 'hand-engineered' image-computable model does not take into account the massive visual diet of familiar objects that we encounter during learning and development, which potentially structures our visual representations of unfamiliar stimuli. It seems plausible that extensive visual experience leads to visual features that are particularly good for distinguishing between objects, and thus our judgments of unfamiliar objects might reflect characteristics of these features.

To acquire an alternative feature space based on learning from natural stimuli, we took advantage of recent advances in artificial neural network models, which match or even exceed human levels of performance at recognizing objects in photographs (Chatfield et al., 2014; He et al., 2016; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; Szegedy et al, 2013; Szegedy et al., 2015), and also predict a number of human perceptual phenomena (e.g., Lake, Zaremba, Fergus, & Gureckis, 2015; Peterson, Abbott, & Griffiths, 2018; Sanders & Nosofsky, 2018). Specifically, we took an instance of AlexNet (Krizhevsky et al., 2012) —a feedforward convolution neural network, trained on the ImageNet dataset (Deng et al., 2009; Russakovsky et al., 2015) consisting of roughly 1000 images in each of 1000 real-world categories —and measured its responses to the unfamiliar shapes from our experiments. We simulated the trials in our experiment and derived psychometric functions that could be compared to the human performance.

Again, we find the two qualitative phenomena predicted by Tenenbaum and Griffith's theory (Fig. 4C). Specifically, both humans and the model tend to see more dissimilar samples with high variability as part of the same class, and are more uncertain about their decisions in the one-sample condition. Fig. 7 shows that the *parameter-free* and *global weighting* model based on AlexNet's features is much closer to the human observer confidence intervals for the many than the one-sample condition. As with ShapeComp, only a more sophisticated *class-specific weighting* model was within reach of human performance for one-shot categorization. We find that the simple heuristic model based on AlexNet performs worse than ShapeComp, perhaps because ShapeComp is designed to work specifically well with 2D shapes. What both AlexNet and ShapeComp simple heuristic models have in common is worse predictions of the one-sample than the many-sample condition. Moreover, both AlexNet and ShapeComp *class-specific weighting* models exhibit better performance in the one-sample condition compared to their *parameter-free* or *global weighting* counterparts. Thus, the pattern of successes and failures of ShapeComp does not reside in the specific features. Together, these results provide convergent evidence that human performance in one-shot categorization is more sophisticated than measuring distances in some fixed feature space that operates the same way across objects.

3.2.3. ShapeComp versus AlexNet

Our goal was to contrast simple but efficient heuristic models (e.g., the parameter-free model) versus deeper and more computationally costly models (e.g., the class-specific model). We do this with two quite distinct feature sets (ShapeComp and AlexNet), to confirm that the conclusions do not depend on the specific features in the model, but rather depend on *the way those features are combined* (i.e., with either fixed weights or with weights adjusted on a stimulus-by-stimulus basis). The consistent pattern of findings across ShapeComp and AlexNet features suggests similar findings would occur with any feature set that extracts useful information in distinguishing shapes or objects.

Certainly, the two feature spaces can be compared and they have different strengths and weaknesses. ShapeComp features are hand-engineered, and particularly useful at discriminating between shapes. On the other hand, AlexNet features are learned by training a neural network on object classification with massive labeled datasets. Thus, the AlexNet feature set would suggest how extensive prior experiences with natural photographs predict human performance with novel objects. Neither feature sets, however, can predict human performance in one-shot categorization without 'additional' processing.

Another major difference between the two feature spaces is in the number of features; ShapeComp is based on 106 shape descriptors, while AlexNet features were based on the final fully connected layer with 1000 outputs. With fewer features, ShapeComp is better at generalizing, explaining why the parameter-free and global-weighting ShapeComp model is more consistent with human responses than their AlexNet model counterpart (Fig. 7). With more features, a regression is more capable of finding a weighted combination of features that better match the human data, explaining why the class-specific (i.e., weighted) version of the AlexNet model better accounts for the human responses than the ShapeComp counterpart. In line with this intuition, across the 3 models, 20 objects, and 2 sampling conditions, the Akaike information criterion is significantly lower for the ShapeComp ($M = -848.96$, $SD = 55.32$) models than the AlexNet ($M = -763.13$; $SD = 157.81$) models ($t(119) = -8.66$, $p < 0.0001$), suggesting that the lower number of parameters makes ShapeComp a simpler and higher quality model.

4. General discussion

Measuring one-shot categorization in humans is challenging because each participant can perform only a single trial for each novel object class, yet mapping out psychometric functions requires dozens or hundreds of trials. Therefore, we used crowdsourcing to obtain responses from 500 participants, each performing only a single judgment on one of 20 different shape classes. This allowed us to test for the first time some of the central predictions of previous theoretical work on generalization from small numbers of exemplars.

4.1. Systematic changes in generalization with sample properties

Our findings confirm that with just 16 samples, when variability between samples is small, participants use tight criteria to judge whether a new object belongs to a class. When a shape deviates by even a small amount from the range spanned by the samples, participants judge it as belonging to a different class. When variability between samples is larger, participants extend their decision boundary accordingly, being more willing to accept deviations from the sample data as plausible variants. Importantly, even when presented with just a single exemplar, participants exhibit clear generalization gradients. We find that the decision boundary is essentially the same for just a single exemplar as for 16 exemplars: only the certainty about the decision boundary reduces. This is consistent with the predictions of Bayesian models of generalization (Tenenbaum & Griffiths, 2001), and similar strategies where increases in the decision boundary confidence are associated with the presence of more visual stimuli.

4.2. Heuristics versus more-sophisticated decision strategies

The generalization function, and likelihood model approach, assume that generalization decisions are based on estimating distances between items in some kind of metric psychological space. We created ShapeComp, an image-computable model based on 106 shape descriptors—spanning a wide range of different shape properties—to test whether human generalization could be predicted directly from images of objects. We also evaluated features from AlexNet, a ConvNet that achieves human level object categorization. We found that when class ambiguity is low (i.e., many samples condition), the parameter-free models better accounted for human responses with a fixed feature space than when class ambiguity was high. We found similar results for ShapeComp and AlexNet with *global weighting* of features determined by a regression of the human responses and features across the remaining novel objects. These results suggest that when image information is sufficient, observers use a simple heuristic approach that is the same for all objects and is based entirely on the information in the image at hand, as represented by pre-established features. The simple heuristic approach, however, could not account for the human responses when class ambiguity was high (one sample condition), suggesting that the visual system does not use a fixed feature weighting in the 1-examplar condition.

Instead, we find that human decisions were more consistent with a model that could flexibly re-weight its features on a stimulus-by-stimulus basis. This suggests that in one-shot categorization, human observers treated novel objects as having unique structures and characteristics that other novel or familiar objects do not share. While there could be networks more reminiscent of human behaviour than AlexNet (e.g., *ResNet* (He et al., 2016), *Inception-v3* (Szegedy et al., 2013); *See also* Rajalingham et al., 2018), these results for specific-class weightings for novel objects are broadly in accordance with state-of-the-art object recognition models based on deep learning networks that also learn a unique combination of weighted features for each object class when trained on large labeled datasets. Unlike training on large datasets, however, we suggest that to generalize from just a single exemplar, the visual system actively generates hypotheses about what 'variations' of the object might look like. This involves identifying those features that are most likely to vary within the class. Which features are likely to vary and in which ways is specific to the unique exemplar seen so far, which may explain why a flexible weighting of features is required to account for the pattern of human responses. It would be very interesting in future studies to evaluate evidence for explicit generative processes through which the visual system creates and evaluates plausible 'variants' of the exemplar (e.g., via mental imagery).

4.3. One-shot categorization happens over a large time-scale

How exactly the visual system 'actively' infers or learns which features of a novel object are most important for the class is an intriguing open question. While one-shot categorization is often associated with the process of acquiring knowledge from the recent past, there is not much one can learn about an object category from a single exemplar in only one trial. Since we find class-specific weights are required to account for human category responses for novel objects (i.e., different weightings of feature for each novel object class), decisions in our task must be based on learning over a much longer time scale than a single trial - i.e., learning must somehow occur through evolution or our lifetime experiences and our action decisions with other familiar objects.

4.4. Conclusions

These results imply a broader framework that synthesizes the inconsistent theoretical ideas put forth in previous studies. When learning to classify novel objects, the visual system uses a continuum of strategies depending on the ambiguity of the available information. When there is sufficient data to define category boundaries based on the observed samples, the visual system simply compares feature values to determine whether objects belong in a common class. This is a computationally efficient strategy that exploits experience with familiar objects to enable decisions for arbitrary novel shapes and classes. In contrast, when information is very sparse (i.e., during inferences from just a single example), such an approach is not sufficient to account for human judgments. Instead, the visual system seems to exploit a more sophisticated approach, consistent with a generative model that involves some degree of 'shape understanding' to determine which features are likely to be most important, and perhaps even to predict the range over which those features are likely to vary when new samples are encountered. While this is computationally more demanding, we suggest it enables us to make more reliable inferences when generalizing from only a single exemplar than otherwise. Thus, in contrast to previous work that often pits these several approaches against one another, we suggest that each approach is a valid model of human behaviour. However, they appear to trade-off as a function of the amount of data. Large amounts of data lead to heuristical decisions, while small amounts invoke more sophiscated strategies based on deeper understanding of novel shapes and their characteristics. Thus, rather than being distinct approaches, these findings show how models of human decisions that vary in complexity map onto different ends of a continuum with the location determined by the ambiguity of the data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 222641018–SFB/TRR 135 TP C1 and the ERC Consolidator award "SHAPE" (ERC-CoG-2015-682859). We thank Brendan Nicholls and Hendrik Will for their help setting up the experiments and running initial pilot studies.

Author contributions

All authors conceived and designed the study. YM collected the data and implemented the analyses. All authors wrote the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2019.09.005.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience, 7*(10), 1057.

Albright, T. D., & Gross, C. G. (1990). Do inferior temporal cortex neurons encode shape by acting as Fourier descriptor filters. *Proceedings of the international conference on fuzzy logic & neural networks* (pp. 375–378). .

Bai, X., Liu, W., & Tu, Z. (2009). Integrating contour and skeleton for shape classification. *International conference on computer vision workshops (ICCV Workshops)* (pp. 360–367). IEEE.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 4*, 509–522.

Burgess, A. E., & Colborne, B. (1988). Visual signal detection IV. Observer inconsistency. *JOSA A, 5*(4), 617–627.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531.

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.

de Beeck, H. P. O., & Baker, C. I. (2010). The neural basis of visual object learning. *Trends in Cognitive Sciences, 14*(1), 22–30.

de Beeck, H. O., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience, 4*(12), 1244.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition,* 248–255.

Destler, N., Singh, M., & Feldman, J. (2019). Shape discrimination along morph-spaces. *Vision Research, 158*, 189–199.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4), 594–611.

Feldman, J. (1992). Constructing perceptual categories. *Proceedings of the 1992 I.E.E.E. conference on computer vision and pattern recognition* (pp. 244–250). Los Alamitos, CA: I.E.E.E. Computer Society Press.

Feldman, J. (1995). *Formal constraints on cognitive interpretations of causal structure. Proceedings of the I.E.E.E. workshop on architectures for semiotic modeling and situation analysis, Monterey, CA*.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology, 41*, 145–170.

Feldman, J., & Singh, M. (2005). Information along contours and object boundaries. *Psychological Review, 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.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Gelman, S., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183–209.

Gelman, S. A., & Meyer, M. (2011). Child categorization. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(1), 95–105.

Gibney, E. (2016). Google AI algorithm masters ancient game of Go. *Nature News, 529*(7587), 445.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482.

Goodman, N. D., Tenenbaum, J. B., Griffiths, T. L., and Feldman, J. (2008a) Compositionality in rational analysis: grammar-based induction for concept learning. In M. Oaksford and N. Chater (Eds.). The probabilistic mind: Prospects for Bayesian cognitive science.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 75*, 1205–1222.

Granlund, G. H. (1972). Fourier preprocessing for hand print character recognition. *IEEE Transactions on Computers, 100*(2), 195–201.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88–94.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). .

Hegdé, J., Bart, E., & Kersten, D. (2008). Fragment-based learning of visual object categories. *Current Biology, 18*(8), 597–601.

Kingdom, F. (1997). Simultaneous contrast: The legacies of Hering and Helmholtz. *Perception, 26*, 673–677.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097–1105). .

Kromrey, S., Maestri, M., Hauffen, K., Bart, E., & Hegdé, J. (2010). Fragment-based learning of visual categories in non-human primates. *PLoS One, 5*, e15444.

Kuhl, F. P., & Giardina, D. R. (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing, 18*, 236–258.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). *Deep neural networks predict category typicality ratings for images. Proceedings of the 37th annual meeting of the cognitive science society, Pasadena, CA, July 22–25, 2015*. Cognitive Science Society ISBN:978-0-9911967-2-2.

Latecki, L. J., Lakamper, R., & Eckhardt, T. (2000). Shape descriptors for non-rigid shapes with a single closed contour. *IEEE Conference on Computer Vision and Pattern Recognition'00* (pp. 424–429). .

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c, 36*(9–10), 910–912.

Morgenstern, Y., Rostami, M., & Purves, D. (2014). Properties of artificial networks evolved to contend with natural spectra. *Proceedings of the National Academy of Sciences, 111*(Suppl. 3), 10868–10872. https://doi.org/10.1073/pnas.1402669111.

Morgenstern, Y., Rukmini, D. V., Monson, B. B., & Purves, D. (2014). Properties of artificial neurons that report lightness based on accumulated experience with luminance. *Frontiers in Computational Neuroscience, 8*, 134. https://doi.org/10.3389/fncom.2014.00134.

Paulun, V. C., Kawabe, T., Nishida, S. Y., & Fleming, R. W. (2015). Seeing liquids from static snapshots. *Vision Research, 115*, 163–174.

Pavlidis, T. (1980). Algorithms for shape analysis of contours and waveforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 4*, 301–312.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42*(8), 2648–2669.

Peura, M., & Iivarinen, J. (1997). Efficiency of simple shape descriptors. *Proceedings of the third international workshop on visual Form, Capri, Italy* (pp. 443–451). .

Purves, D., Morgenstern, Y., & Wojtach, W. (2015). Perception and reality: Why a wholly empirical paradigm is needed to understand vision. *Frontiers in Systems Neuroscience, 9*, 156.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 38*(33), 7255–7269.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Sanders, C. A., & Nosofsky, R. M. (2018). *Using deep learning representations of complex natural stimuli as input to psychological models of classification. Proceedings of the 2018 Conference of the Cognitive Science Society, Madison*.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research, 122*, 105–123. https://doi.org/10.1016/j.visres.2016.02.002.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behaviour of Brain Science, 24*, 581–601.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron, 79*(5), 1025–1034.

Stuhlmuller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). *Learning Structured Generative Concepts. Proceedings of the thirty-second annual conference of the cognitive science society*.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) 1030 Intriguing properties of neural networks. arXiv preprint arXiv:13126199.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). .

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behaviour of Brain Science, 24*, 629–640.

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition, 119*, 325–340.

Wilder, J., Fruend, I., & Elder, J. H. (2018). Frequency tuning of shape perception revealed by classification image analysis. *Journal of vision, 18*(8) 9-9.

Zahn, C. T., & Roskies, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers, 100*(3), 269–281.

Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition, 37*, 1–19.