COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2016) 1–18

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12400

Improving Human-Machine Cooperative Classification Via Cognitive Theories of Similarity

Brett D. Roads, Michael C. Mozer

Department of Computer Science and Institute of Cognitive Science, University of Colorado

Received 15 September 2015; received in revised form 1 March 2016; accepted 5 April 2016

Abstract

Acquiring perceptual expertise is slow and effortful. However, untrained novices can accurately make difficult classification decisions (e.g., skin-lesion diagnosis) by reformulating the task as similarity judgment. Given a query image and a set of reference images, individuals are asked to select the best matching reference. When references are suitably chosen, the procedure yields an implicit classification of the query image. To optimize reference selection, we develop and evaluate a predictive model of similarity-based choice. The model builds on existing psychological literature and accommodates stochastic, dynamic shifts of attention among visual feature dimensions. We perform a series of human experiments with two stimulus types (rectangles, faces) and nine classification tasks to validate the model and to demonstrate the model's potential to boost performance. Our system achieves high accuracy for participants who are naive as to the classification task, even when the classification task switches from trial to trial.

Keywords: Perceptual expertise; Categorization; Interactive classification; Similarity; Human-computer interface; Optimization; Cognitive modeling

1. Introduction

Many challenging machine-learning problems involve humans in the loop (e.g., Branson, Van Horn, Wah, Perona, & Belongie, 2014; Deng, Krause, & Fei-Fei, 2013; Jia, Abbott, Austerweil, Griffiths, & Darrell, 2013; Wah, Maji, & Belongie, 2015). Individuals may provide data to machine learning systems in the form of ratings, preferences, judgments, and labels. Machine learning systems can organize and structure information presented to individuals. To build systems that can predict, interpret, and guide human behavior, it is critical to leverage our understanding of human perception and cognition.

Correspondence should be sent to Brett D. Roads, Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder, CO 80309-0430. E-mail: brett.roads@colorado.edu

For example, when designing a human-assisted robot navigation system, human working-memory capacity must be considered (Ahmed et al., 2014); and to interpret human movie ratings on a site such as Netflix, sequential dependencies known to contaminate judgments must be considered (Mozer, Pashler, & Link, 2011).

In this paper, we consider the design of decision support systems that enable individuals to perform difficult image classification problems via image comparison. Our inspiration is a prototype system developed by dermatology researchers for diagnosing skin lesions (Aldridge, Glodzik, Ballerini, Fisher, & Rees, 2011; Brown, Robertson, Bisset, & Rees, 2009). The system requires users to make a visual comparison between a *query* image of an unknown skin lesion to a set of *reference* images, which have ground-truth labels. In a series of screens containing the query and a set of references, users are asked to select the reference most similar to the query. Based on the labels of the selected references, the query can be assigned an implicit class label. This approach of implicit classification via similarity judgments was remarkably effective: Naive participants achieved a diagnostic accuracy of 96% on a five-way classification task, whereas medical students who had completed a 10-day dermatology attachment were only 51% correct when making explicit diagnoses.

Aldridge et al. (2011) did not detail the exact procedure used for selecting references, but it appears to have been heuristic. Machine learning researchers have recently taken a more formal approach to the design of interactive classification systems in which references are chosen to maximize efficiency and accuracy. Ferecatu and Geman (2009) propose an algorithm to search a large, unannotated image database for an instance of a semantic category by asking individuals to make a series of similarity judgments between reference images and a mental image of the target category. Wah et al. (2014) developed a system that allows users to perform fine-grain categorization of a query image via similarity comparisons. Their work focuses on distinguishing among many categories (e.g., 200 bird species) but presumes that instances of a category are interchangeable. Thus, like Ferecatu and Geman (2009), the problem is essentially formulated as an image retrieval task where the image represents a distinct class and the challenge stems from the sheer number of images/classes.

Our goal is to develop an interactive classification system for the classic problem of discriminating between two or a small number of alternatives, for example, the skin lesion task of Aldridge et al. (2011). This problem differs significantly from image retrieval in that the difficulty for humans comes from intra-category variability and determining which features in a potentially high-dimensional space are relevant to the task (Pashler & Mozer, 2013). However, we hypothesize that optimizing the selection of references in interactive classification systems, like that Aldridge et al. (2011), will boost novice performance in the way that it has in interactive image-retrieval systems. We would like to emphasize that our work is largely orthogonal to large literature on training strategies used in visual category learning (e.g., Birnbaum, Kornell, Bjork, & Bjork, 2013; Carvalho & Goldstone, 2014; Kang & Pashler, 2012). Our objective is not to train novices, but to make it easier for naive users to correctly classify difficult unknown images by leveraging psychological models of human similarity judgments.

The systems we have described draw inferences from human similarity judgments. These inferences facilitate the selection of additional references or help determine the query's identity or class label. To glean the most information from human judgments, it is exceedingly helpful to have a cognitive model of how those judgments are produced. This *response model* consists of a theory of representation and a theory of how the representations lead to the selection of responses. Although Ferecatu and Geman (2009) and Wah et al. (2014) noted the importance of modeling human perception and decision making, their formulations of a response model appear to be based on the authors' intuitions. In the present work, we focus on the question of what response model provides the best characterization of human behavior. We leverage the psychological literature to direct the space of explorations. The key premise of this work is that in order to predict and interpret human behavior, one must understand the underlying cognitive mechanisms and incorporate these mechanisms as biases or constraints in machine learning systems.

2. The psychology of similarity

The rich psychological literature on human and animal generalization (Shepard, 1987; Tenenbaum, 1999) explores the conditions under which responses associated with one stimulus transfer to another, or properties associated with one stimulus are ascribed to another. The more similar stimuli are, the more likely generalization is to occur. Similarity is based not on external properties of the stimuli, but rather on an individual's internal representation. We refer to this internal representation as the *psychological representation*. For now, we take this psychological representation as a given, but we discuss its basis shortly. Various psychological theories (e.g., Jones, Maddox, & Love, 2005; Nosofsky, 1986; Shepard, 1987; Sinha & Russell, 2011) quantify similarity in terms of a weighted ρ-norm distance metric: the distance between two stimuli q and r whose psychological representations are denoted by N-dimensional feature vectors $\mathbf{z}_q$ and $\mathbf{z}_r$, is:

$$d(\mathbf{z}_q, \mathbf{z}_r) = \left(\sum_{i=1}^{N} w_i |z_{qi} - z_{ri}|^{\rho} \right)^{1/\rho} \tag{1}$$

where $\mathbf{w}$ is a relative weighting of the features, constrained by $||w||_1 = 1$ and $w_i \geq 0$. The degree of generalization from one stimulus to the other is then cast as a monotonically decreasing function of distance (Jones, Maddox, & Love, 2006; Jones, Love, & Maddox, 2006; Nosofsky, 1986; Shepard, 1987). Integrating these models into their most general form, we obtain:

$$g(d) = \gamma + \exp(-\beta d^{\tau}) \tag{2}$$

where γ, β, and τ modulate the gradient of generalization.

In the human experiments we will describe, we ask participants to choose one of M reference images most similar to a query image. The literature on psychological models of similarity-based selection is not as well developed as the literature on generalization, but 1-of-M choice is often modeled by a rule in which probability of selecting a candidate is proportional to its strength. In our studies, the strength of a candidate r is simply its strength of generalization from query q, leading to:

$$P(r|q) \propto g(d(\mathbf{z}_q, \mathbf{z}_r)) \tag{3}$$

When individuals judge similarity, the weighting of separable feature dimensions ($\mathbf{w}$ of Eq. 1) is flexible and depends on the focus of attention. Concept learning can be characterized in terms of the adaptation of weights to emphasize discriminative features (e.g., Jones et al., 2005). Other psychological phenomena suggest that weights must vary in a more dynamic, immediate manner. Tversky (1977) demonstrated a set of phenomena that are problematic for similarity measures based on distance metric, but which can be explained if the weights are modulated by contextual factors such as the framing of a question or the set of choice alternatives.

No single psychological theory at present can prescribe the weighting that an individual will adopt on a particular trial given the recent and current trial context. While multiple models exist that specify how an error-signal can be used to update attentional weights during learning (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004), it is less clear how to jointly model the various context-driven shifts in attention that occur outside of learning. Although numerous models exist for handling specific choice scenarios (e.g., Mozer et al., 2011; Trueblood, Brown, & Heathcote, 2014), there is no single theory capable of handling all potential sources of context-driven changes in attention. In modeling human choice based on data over a series of trials, one could simply assume the weights were fixed and estimate their maximum likelihood values. In our work, we improve on this average-best approach by treating the weights as a random vector and marginalizing over their uncertainty. This Bayesian approach has the advantage of steering us away from situations in which the uncertainty in the weights maps to uncertainty in the model's predictions.

To capture the uncertainty in the attention weights, we assume that the weights $\mathbf{w}$ are sampled from a Dirichlet distribution, $\mathbf{w} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where the elements of the N-dimensional concentration vector $\boldsymbol{\alpha}$ are drawn from a Gamma density, $\alpha_i \sim \text{Gamma}(\kappa, \phi)$ (see Fig. 1). The Dirichlet distribution has support over all convex combinations of weights and thus provides a natural distribution for describing the allocation of attention across the N dimensions. The elements of the N-dimensional concentration vector $\boldsymbol{\alpha}$ control the allocation of probability mass in the Dirichlet distribution (e.g., a bias toward uniform weights). To allow for maximum flexibility, we use the Gamma distribution which has support from zero to infinity and permits a wide range of possible distributions for the α values. In this work, the same Gamma distribution is used for all concentration parameters, and the hyperparameters characterizing the Gamma distribution are set to $\kappa = 3$ and $\phi = 1$. There is nothing special about these specific hyperparameter

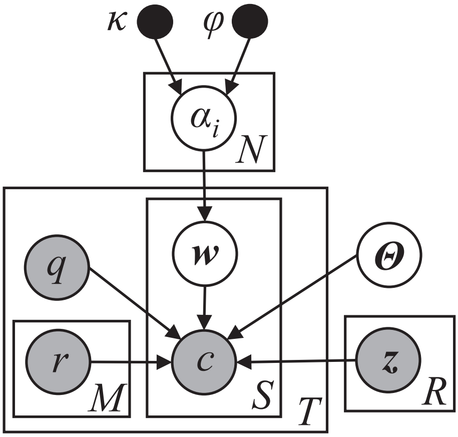

Fig. 1. A graphical model of similarity-based classification. Our model posits that attentional weights w can vary by participant S and by trial T. The variation in weights can be modeled as Dirichlet distribution with N concentration parameters α_i (one for each dimension of the psychological representation). The probability of making a given classification is given by Eq. 4. The parameter vector $\mathbf{\Theta} \equiv \{\gamma,\beta,\tau,\rho\}$ modulates the gradient of generalization in Eq. 2.

values. Rather, what is important is that low values of κ and ϕ produce a Gamma distribution that has low probability mass at low and high values of α_i. In other words, these hyperparameters act as a regularizer when learning the concentration parameters, discouraging values of α_i that are large or near zero. Such regularization is important in order to prevent overfitting to a particular set of observed behavior and enable generalization across different contexts. The results we present are robust to the choice of these hyperparameters.

3. Similarity-based classification

In this article, we explore similarity-based classification using a basic experimental task in which participants are asked to select one of four reference images that is most similar to a query image. Fig. 2a–c depicts sample displays with rectangle stimuli; the query is colored blue and is surrounded by the references, colored black. Unbeknownst to participants, each reference is associated with a class label, and selecting a reference is equivalent to implicit classification. For example, consider a classification task involving discrimination on the basis of rectangle height. In Fig. 2b, the upper and lower pairs of references are tall and short, respectively. If a participant judges the upper-right reference most similar to the query, the query would be classified as belonging to the tall category.

How should references be selected? A sensible approach would be to choose references so as to maximize the expected information gain about the target class. Ferecatu and Geman (2009) propose a heuristic approximation to this objective, also adopted by Wah

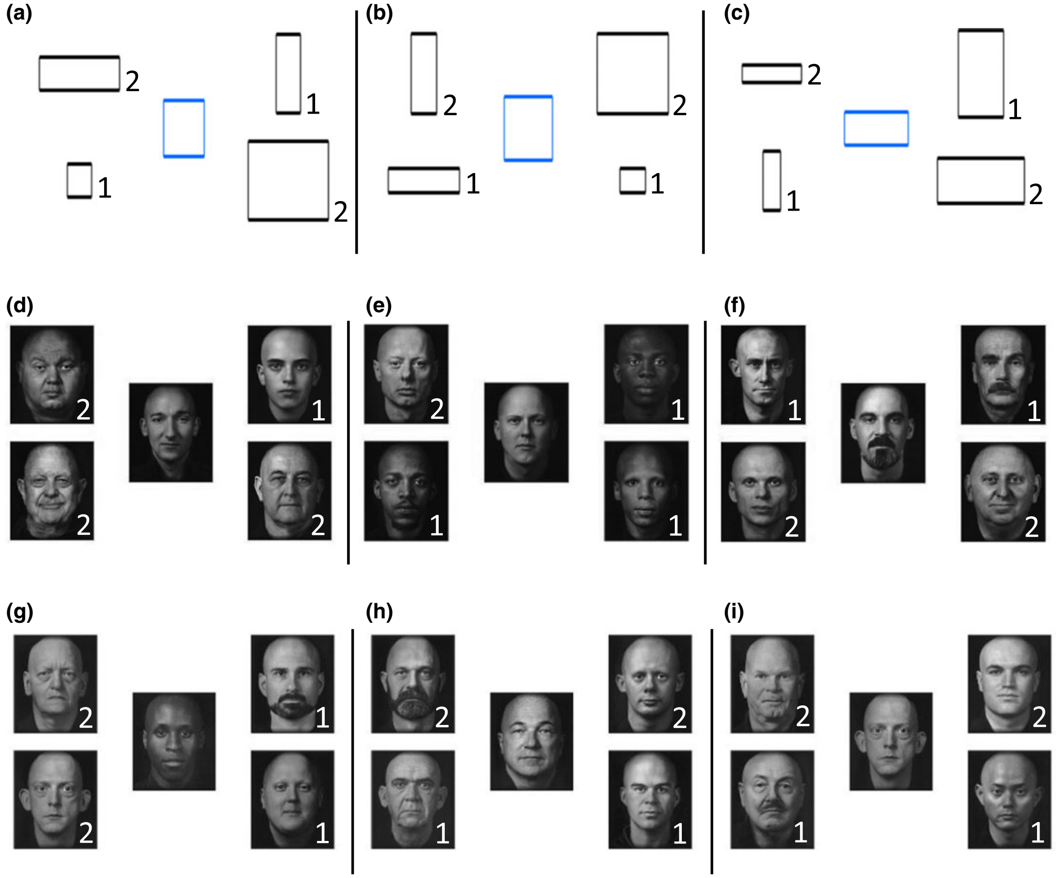

Fig. 2. Sample displays used for implicit categorization. Each display consists of a query image in the center surrounded by four reference images. The numerals indicate category membership and do not appear in the actual displays. Displays (a)–(c) involve rectangles. Displays (d)–(i) involve faces. Each display corresponds to a different implicit classification task, as elaborated in the article.

et al. (2014), that involves searching for a set of references that have roughly equal probability of being selected given the history of user responses on the current trial. Although this heuristic works well for the image retrieval task investigated by Ferecatu and Geman (2009), it is inadequate in our domain due to the small number of classes. Instead, we search for the set of references that maximize classification accuracy.

We start with a candidate set of R reference images. These images have been classified by experts, and thus we can assume that each reference $r \in 1, \ldots, R$ has a known class label, c_r. For the time being, we also assume a known psychological representation, $\mathbf{z}_r$. We assume that the reference instances can be used to define an approximate input density of queries q, that is, $p(q) \approx \sum_{r=1}^{R} v(r)\delta(\mathbf{z}_q - \mathbf{z}_r)$, where δ is the Dirac delta and $v(r)$ is a prior probability on reference r, typically $v(r) = \frac{1}{R}$.

Given a candidate set of M references, $\mathbf{r}$, we can use the response model (Eq. 3) to estimate the probability that a query q will be matched to a reference of class c:

$$P(c|\mathbf{r}, q) = \sum_{i=1}^{M} P(r_i|q) \mathbb{1}_{c_{r_i}=c} \tag{4}$$

Leveraging the query density estimate, we can define the optimal reference set, $\mathbf{r}^*$, as the one that maximizes the expected classification accuracy of some query q with target class c_q:

$$\mathbf{r}^* = \mathrm{argmax}_{\mathbf{r}} \sum_{q=1}^{R} v(q) P(c_q|\mathbf{r}, q) \tag{5}$$

Fig. 2 shows optimal references chosen by this criterion for nine different classification tasks, which we explain shortly. Given the relatively small search space in our tasks ($R < 150$, $M = 4$), we found that a hill climbing search with restarts is adequate to identify the optimal sets. Our technique is readily extended to integrate similarity judgments across multiple screens. That is, an initial set of references is presented, and conditioned on the participant's selection, a new set of references is presented. To support this sequential decision procedure, we later present results from a greedy procedure in which a posterior $v(q|s)$ is determined for a given reference selection s, using the prior $v(q)$ and the likelihood $P(s|q)$ (Eq. 3).

4. Stimuli and representation

Our experiments used two sets of stimuli: rectangles and male faces (Fig. 2). For each domain, we require a psychological representation—the human internal representation of the external stimulus properties. For rectangles, we considered four candidate representations: {width, height}, {log width, log-height}, {area, aspect ratio}, {log area, log aspect ratio}. Krantz and Tversky (1975) obtained psychological evidence for {area, aspect ratio}; Borg and Leutner (1983) argued for {log width, log height}. (We defined aspect ratio as height divided by width.) We expanded these two options into the full Cartesian product of possibilities and found that the {log area, log aspect ratio} representation was by far a better predictor of human responses. Results of the representation fitting procedure are shown in Table S1 in the Supplemental Materials. The procedure we used for determining the representation is similar to the procedure we will describe below for fitting model parameters. Our experiments used 169 rectangle stimuli that were uniformly spaced on a 13 × 13 grid in width-height space.

The face stimuli in our experiments consisted of a set of 104 grayscale bald male faces. We used a six-dimensional psychological representation previously derived by Jones and Goldstone (2013) using non-metric Euclidean multidimensional scaling according to the method of Goldstone (1994), based on human similarity judgments. Our inspection of the first three dimensions of the psychological representation suggests that these dimensions *roughly* correspond to (1) age/shape, (2) skin color, and (3) machismo. The next three dimensions are not readily summarized by verbal labels. The dimensions are ordered from explaining the most to least variance in similarity judgments.

5. Experiments

We conducted three sets of human experiments in which participants were asked to select the most similar reference to a query stimulus. Each set is composed of a *Random* experiment and a follow-up *Optimized* experiment. The Random experiment used randomly selected references in order to collect behavioral data for a model-fitting procedure. The fitting procedure allowed us to determine the form of the model that best accounts for human choices. The best fitting model was then used to identify an optimal set of reference examples (Eq. 5), which were then used in the Optimized experiment. We expect to observe better performance with the Optimized references than the Random, but that does not go far in validating the references as truly optimal. We also show that our model accurately predicts human choices on individual trials in the Optimized experiment. To the degree that the model can predict trial-to-trial behavior, it is valid to use the model as a proxy for humans in optimization, and we have some assurance that the Optimized references are indeed that.

For each stimulus domain, the corresponding psychological representation was used to construct multiple implicit categorization tasks. To create a categorization task, all psychological points were first projected onto a single dimension. A category boundary was defined on this projection such that half of the projected points were on one side of the category boundary. Points very close to the boundary were removed. For the rectangle stimuli, three category boundaries were created based on (a) width, (b) height, and (c) aspect ratio. Ignoring color for the moment, Fig. 3a–c shows a visualization of points representing rectangle stimuli in the psychological representation along with the boundary for each categorization task. Note that because the stimuli are sampled uniformly in {width, height} space, they are rotated and non-uniform in {log area, log aspect ratio} space. A visualization showing the correspondence between points represented in {width, height} space and {log area, log aspect ratio} is shown in Fig. S1 in the Supplemental Materials.

For the face stimuli, six tasks were defined by projecting the psychological points onto each of the six axes of the representation. Fig. 3d–i shows the points representing faces in a two-dimensional subspace of the six-dimensional psychological representation. Each scatterplot shows two dimensions of the six-dimensional representation; one of the dimensions is the critical dimension for the classification task.

One set of experiments was based on rectangle stimuli and the three rectangle category boundaries. Two sets of experiments were based on the face stimuli, one with category boundaries along dimensions 1–3 and one with boundaries along dimensions 4–6. To summarize, we had three sets of experiments, each of which tests three category boundaries; each set consisted of a Random and the Optimized condition run between subjects.

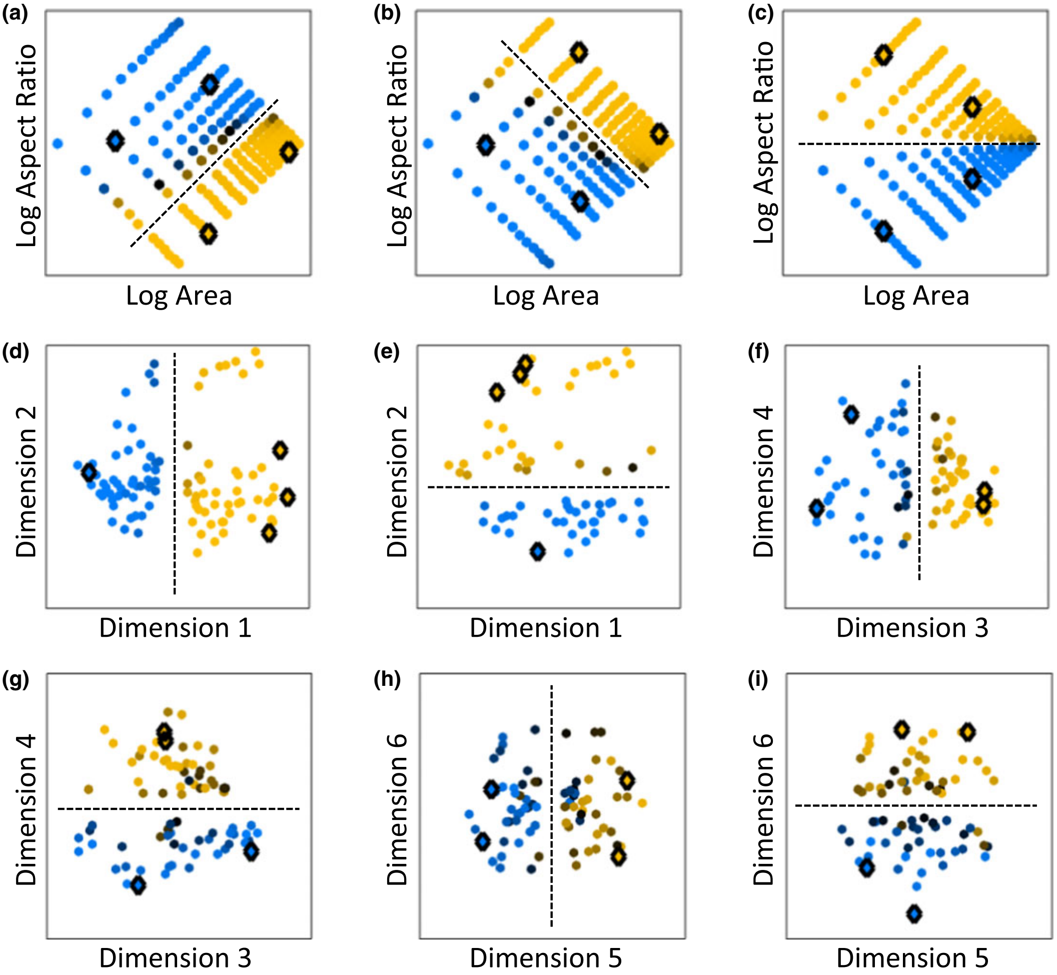

Fig. 3. A depiction of rectangle stimuli (a–c) and face stimuli (d–i) in a psychological representation. Each colored dot in a scatterplot corresponds to a single stimulus. Each scatterplot depicts a different implicit categorization task, with the category boundary shown as a dashed line. For example, the boundaries shown in a, b, and c correspond to a category boundary based on width, height, and aspect ratio, respectively. For each categorization task, the $R = 4$ optimal references have been determined and are depicted by black diamonds. The color of a dot—blue or yellow—indicates the predicted category label of each stimulus given the references. The darkness of a dot indicates the extent to which an item will be classified into either category with equal probability. Black dots indicate items that have exactly equal probability of being classified into either of the two categories.

6. Experiments with randomly chosen references

Experiments were conducted using participants from Amazon Mechanical Turk. Participants in the rectangle experiments were compensated $1.00 for approximately 10 minutes of work. Participants in the face experiments were compensated $1.50 for approximately 15 minutes of work. Although the number of trials was held constant for each participant, judgments of face similarity required more time than judgments of rectangle similarity.

In each of the three Random experiments, 20 participants made similarity judgments between a query image and a reference set on 150 experimental trials (60 different subjects total). Instructions explained that the participant's goal was to select the reference object that was most similar to the query object. Following instructions, participants completed three practice trials.

The 150 experimental trials were organized into 5 blocks, each with 30 trials. Each block showed 10 trials for each implicit categorization task. The order of the 30 trials in a block was randomized such that the implicit categorization tasks were highly interleaved, resulting in an implicit categorization task that varied on a trial-by-trial basis. However, the frequently changing categorization task is irrelevant for the unaware subject because their task remains the same—select the reference object that is most similar to the query object. For each trial, references were drawn at random, with the constraint that there were two references of each category for that trial's implicit categorization task. The query was also chosen randomly with the constraint that queries were drawn with equal frequency from each category. Subjects were given no feedback on the selections throughout the entirety of the experiment.

Each block included three *catch* trials in which the query was identical to one of the references. On the catch trials, participants making an earnest effort should choose the identical reference. Subjects that failed to correctly select the identical reference on at least 50% of catch trials were dropped from further analysis and replaced with a new subject that met the criterion. Across the three Random experiments, one subject was replaced in the rectangles experiment and one subject was replaced in the faces dimension 4–6 experiment. Average accuracy was very high on catch trials for the rectangles (M = 0.92, SD = 0.02), face dimensions 1–3 (M = 0.94, SD = 0.02), and face dimensions 4–6 (M = 0.96, SD = 0.02).

With this procedure for generating displays, two of the choices are "correct" with respect to the implicit categorization task and two are "incorrect." All participants saw the same set of trials in the same block order. Within each block, the order of trials was randomized for each participant. Throughout each block, a block-specific progress bar was displayed at the top of the screen. Between blocks, participants were explicitly notified of their progress in the experiment (e.g., "you have completed 1/5 of the experiment").

To begin each trial, participants clicked on a screen-centered cross. Immediately following the click, the query object and reference set were displayed. Participants clicked on the reference object they considered most similar to the query object. After the selection of a reference object, there was a 500 ms delay period followed by the cross for the next trial. Participants were given no feedback. From the participant's perspective, they were performing a similarity judgment task and category-based feedback would have been odd.

For each implicit categorization task, we computed an average accuracy score across participants. The blue bars in Fig. 4 indicate average accuracy in the Random

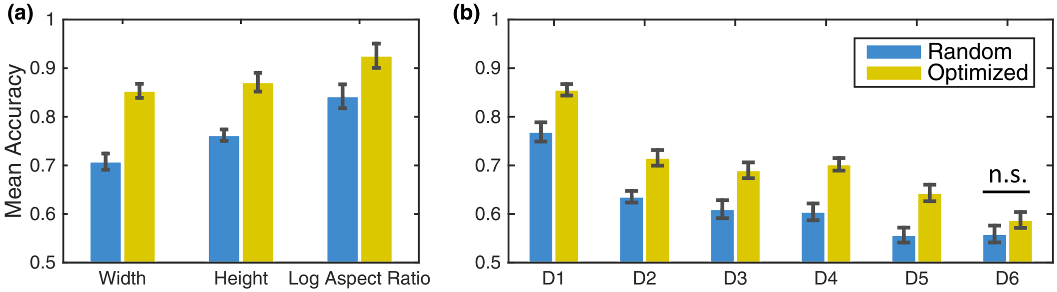

Fig. 4. Empirical classification accuracy on (a) the rectangle tasks and (b) the face tasks. The blue and yellow bars indicate average classification accuracy in the Random and Optimized experiments, respectively. Error bars indicate standard error of the mean.

experiments. For all nine tasks, performance was reliably above the chance rate of 0.5. (The chance rate is the implicit classification accuracy that would be obtained if participants chose references at random from a uniform distribution, given that the queries are balanced between the two categories.)

7. Model fitting

Given a parameterization, our similarity model specifies the likelihood of each reference being chosen on a given trial of the experiment (Eq. 3). Assuming independence of choices, we can compute the likelihood of the entire collection of participant responses. (Ignoring the catch trials, there are 20 × 135 = 2,700 total trials.) To fit model parameters, we use this likelihood in a slice sampler (Neal, 2003) to obtain the maximum a posteriori parameters for $\boldsymbol{\alpha}$, β, and γ. For $\rho \in \{1, 2\}$ and $\tau \in \{1, 2\}$ we simply performed an exhaustive search. The search over ρ and τ was constrained in order to reflect the traditional focus of response models (e.g., Jones et al., 2005; Nosofsky, 1986; Shepard, 1987; Sinha & Russell, 2011). A ρ value of 1 and 2 corresponds to 1-norm (city-block distance) and 2-norm (euclidean distance), respectively. A τ value of 1 and 2 corresponds to an exponential and Gaussian generalization function, respectively.

For the rectangle experiment, $\rho = 2$ and $\tau = 1$ obtained the best fit. For each of the face experiments, $\rho = 2$ and $\tau = 2$ obtained the best fit. For the face experiments, the concentration parameters reveal a general trend of decreasing importance going from the first face dimension to the sixth face dimension. The values for the all fitted parameters are shown in Table S2 of the Supplemental Materials.

In addition to fitting the data to our similarity model, we also fit to the related model proposed by Wah et al. (2014). Their model differs from ours in three respects. First, they fix $\tau = 2$. Second, they do not marginalize over the uncertainty in the dimension weighting; instead, they assume equal weighting across dimensions. Third, in contrast to our generalization function (Eq. 2), which is based on psychological theory, they propose a distinct function:

$$g(d) = \max(\epsilon, (1 - \epsilon) \exp(-\beta d^2)) \qquad (6)$$

In our generalization function (Eq. 2), γ modulates the degree of guessing when a probe is far from all references. For a non-zero γ, the contribution of γ starts to dominate over the exponential term when distances are large, homogenizing the choice distribution over references. The parameter ϵ in Wah et al.'s (2014) model has a similar effect, although it kicks in only when distances reach a threshold.

For all three experiments, our model obtained the best log-likelihood scores. However, this result is expected given that our model has more flexibility. A comparison of the models will be more meaningful when we test the models—with parameters fixed—on new data, as we will report shortly. A test on unseen out-of-sample data represents the most informative test of a model.

8. Selection of optimized references

Given the best fitting model for each Random experiment, we determined the optimal references according to Eq. 5. Fig. 3 shows the optimized references for the three rectangle tasks and six face tasks; the references are indicated as black hollow diamonds. Each scatterplot depicts the stimuli, references, and category boundary in the representation space (or subspace, in the case of the face stimuli). The optimized reference points in Fig. 3 correspond to the references shown in the sample displays of Fig. 2. Note that the optimization procedure does not always select an equal number of references from each implicit category. Due to the symmetry of the rectangle stimulus space, the rectangle references are always balanced. However, two of the six face reference sets are unbalanced (Fig. 2d,e).

The color of a dot—blue or yellow—indicates the predicted categorization of each stimulus given the references. The darkness of a dot indicates the predicted probability of categorization, where darker blue and darker yellow indicate items that have a less probability of being classified in a reliable manner. Black dots indicate items that have an equal probability of being classified into either category. Note that for many of the tasks (e.g., a and b), miscategorizations are predicted by the fitted response model.

9. Experiments with optimized references

The Optimized experiments were identical to the Random experiments with the substitution of optimized references for the random references. The Optimized experiments were conducted with new groups of 20 participants for each of the three experiments (60 different subjects total). As in the three Random experiments, we replaced subjects that did not correctly answer at least 50% of the catch trials. Only one subject was replaced for the face dimensions 4–6 experiment. After replacing subjects below criterion, average accuracy was very high on catch trials for the rectangles ($M = 0.96$, $SD = 0.02$), face dimensions 1–3 ($M = 0.96$, $SD = 0.01$), and face dimensions 4–6 ($M = 0.96$, $SD = 0.02$).

For each implicit categorization task, we computed an average accuracy score across the participants. The yellow bars in Fig. 4 indicate average accuracy in the Optimized experiments. To compare accuracy in the Random and Optimized experiments, we performed a two-sample *t* test (two-tailed, unequal variance) for all category boundaries. All *t* tests revealed a significant ($p < .05$) advantage for the Optimized experiment with the exception of the implicit categorization task on face dimension 6, where the difference was in the right direction but not reliable. Table S3 in the Supplemental Materials shows the details of the two-sample *t* test for all implicit categorization tasks.

We tested the models trained on the Random experiments with the unseen behavioral data from the Optimized experiments. On the three Optimized experiments, the predictions of our model obtained a higher log-likelihood score than those of the model of Wah et al. (2014). In further testing where we replaced assumptions of Wah et al.'s (2014) model with those of our model, we found that all three differences between the models—τ, marginalization, and the generalization function—contribute to the improved likelihood. It should be noted that, although the log-likelihood scores favor our model, accuracy predictions (to be described next) using either model are very similar. While the log-likelihood score is technically the correct measure of fit, log-likelihood scores accentuate differences in fit that may not translate to practical differences in behavioral accuracy.

To compare our model predictions to observed behavior, we examined predicted versus observed implicit categorization accuracy by aggregating data both by task and by individual trial. We plot observed task accuracy against model-predicted accuracy to get a sense of the correspondence of our model to behavior. Fig. 5 shows a scatterplot of predicted versus observed accuracy for each of the nine implicit categorization tasks in (a) the Random and (b) Optimized experiments. The theoretical predictions explain nearly all of the variance in observed accuracy.

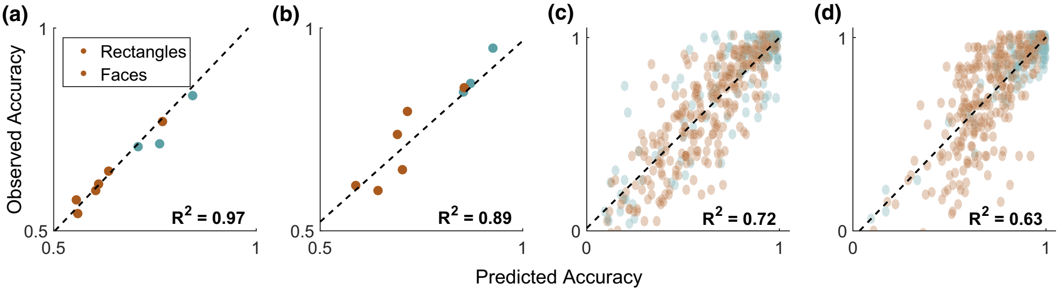

Fig. 5. Predicted versus observed accuracy for (a) the nine implicit categorization tasks in the Random experiments, (b) the nine tasks in the Optimized experiments, (c) individual trials in the Random experiments, and (d) individual trials in the Optimized experiments. Data from the rectangle experiments are displayed in blue and data from the face experiments are shown in red. A slight jitter is introduced along the ordinate to distinguish the points.

Beyond examining model predictions on each implicit categorization task as a whole, we examined model predictions of implicit categorization accuracy on individual trials. Because each participant in each experiment saw the same set of trials, we can average across participants to obtain a trial-based accuracy. Fig. 5 shows predicted versus observed trial accuracy for (c) Random and (d) Optimized experiments. Table S4 in the Supplemental Materials shows that the theoretical predictions explain a significant proportion of the variance in observed accuracy for all experiments.

Although the Random experiments were used to fit model parameters, note that model fitting aimed at predicting the selected reference, not on optimizing implicit categorization accuracy. Thus, it is not entirely trivial that the model does as well as it does on the Random experiment. Regardless, it is impressive that the model is able to predict which specific queries will be classified correctly, in both Random and Optimized experiments.

10. Conclusions

We have described an approach to human-machine cooperative classification that leverages the human's ability to extract high-level visual features and judge similarity, and the machine's ability to predict, steer, and optimize human performance. By utilizing a hierarchical Bayesian model of attention, we have developed a general approach that is robust to arbitrary sources of attention variation, such as sequential effects. The approach allows novices to competently categorize images even when they are unaware of the categorization task, or even more remarkably, when the task switches from moment to moment.

The obvious obstacle to putting this approach into practice is scaling, progress on which has been made for image-retrieval tasks (Ferecatu & Geman, 2009; Wah et al., 2014). To tackle intricate discrimination tasks in high-dimensional feature spaces, a multi-stage approach will be required in which a series of similarity judgments are jointly used to obtain an implicit classification. Once a user makes an initial similarity judgment, a multi-stage approach can show a new set of references that allows the system to hone in on the implicit categorization. We have taken a small step in this direction by considering two-stage judgments. In Fig. 6a, we compare the model's prediction of implicit classification accuracy for a single-stage judgment with 4, 6, 8, or 16 references, as well as a two-stage judgment with four references at each stage. The model makes two interesting predictions. First, the model predicts a diminishing value of additional references, due to an effect of decision noise parameters (γ and β) that grows with the number of alternatives. (Interestingly, decision failure with a large number of alternatives is usually attributed to limits on attention, yet the model provides a natural explanation without requiring an additional mechanism.) These results suggest that even though our approach is computationally capable of optimizing a large number of references, it may be psychologically suboptimal to present the user with a screen packed with different references. Second, the model predicts that a two-stage judgment can yield a significant boost in accuracy. Fig. 6b and c show the selected references at the first and second stages, respectively. The

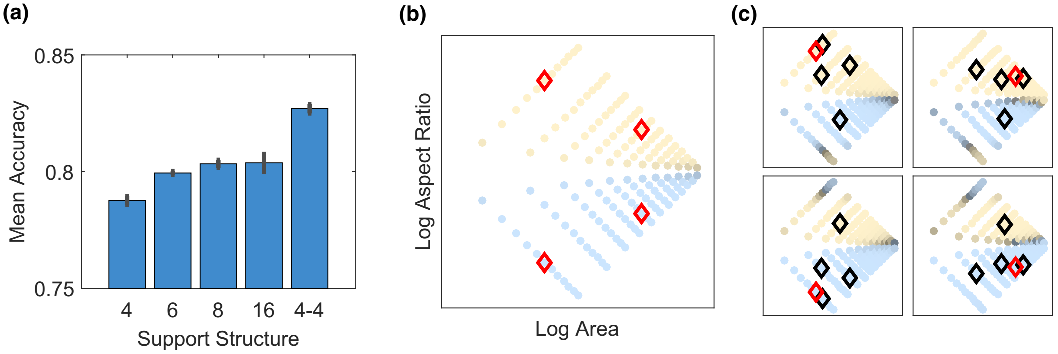

Fig. 6. (a) Predicted performance using single (M = 4, 6, 8, 16) and two-stage (M = 4) judgments. Error bars are repeated measures corrected ±1 standard error (Masson & Loftus, 2003). (b & c) The selected references for two-stage (with M = 4 at each stage) decision support for the aspect ratio task. The references shown at (b) the first stage and (c) second stage given the selection of a first stage reference (shown in red).

reference chosen by the participant at the first stage determines the set of references used for the second stage. These references were obtained with the greedy stage-wise optimization described earlier. Future work will determine if our model correctly predicts human responses for a larger number of references and multiple stages of selection.

We are optimistic about the practical application of similarity-based classification in domains such as medicine and biology. Although months or years of training are required to learn difficult classification tasks, similarity-based classification requires no training. Further, we conjecture that similarity-based classification can be improved with a small amount of instruction. Although verbal instruction is ordinarily difficult for novices to operationalize, instruction may be effective in highlighting relevant feature dimensions (e.g., “focus on the irregularity and patchiness of the rash”).

Acknowledgments

This research was supported by NSF grants SES-1461535, SBE-0542013, and SMA-1041755. We thank Matt Jones and Priya Kalra for helpful guidance and feedback. We also thank our anonymous reviewers for their helpful suggestions and feedback.

References

Ahmed, N., de Visser, E., Shaw, T., Mohamed-Ameen, A., Campbell, M., & Parasuraman, R. (2014). Statistical modeling of networked human-automation performance using working memory capacity. *Ergonomics*, *57*, 295–318.

Aldridge, R. B., Glodzik, D., Ballerini, L., Fisher, R. B., & Rees, J. L. (2011). Utility of non-rule-based visual matching as a strategy to allow novices to achieve skin lesion diagnosis. *Acta Dermato Venereologica*, *91*(3), 279–283.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392–402. doi:10.3758/s13421-012-0272-7

Borg, I., & Leutner, D. (1983). Dimensional models for the perception of rectangles. *Perception & Psychophysics*, *34*(3), 257–267. doi:10.3758/BF03202954

Branson, S., Van Horn, G., Wah, C., Perona, P., & Belongie, S. (2014). The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization. *International Journal of Computer Vision*, *108*(1–2), 3–29. doi:10.1007/s11263-014-0698-4

Brown, N. H., Robertson, K. M., Bisset, Y. C., & Rees, J. L. (2009). Using a structured image database, how well can novices assign skin lesion images to the correct diagnostic grouping? *Journal of Investigative Dermatology, 129*(10), 2509–2512.

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481-495. doi:10.3758/s13421-013-0371-0

Deng, J., Krause, J., & Fei-Fei, L. (2013). Fine-grained crowdsourcing for fine-grained recognition. In P. Kellenberger (Ed.), *2013 IEEE conference on* (p. 580–587). Los Alamitos, CA: IEEE Computer Society. doi: 10.1109/CVPR.2013.81

Ferecatu, M., & Geman, D. (2009). A statistical framework for image category search from a mental picture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(6), 1087–1101.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386. doi:10.3758/BF03204653

Jia, Y., Abbott, J. T., Austerweil, J., Griffiths, T., & Darrell, T. (2013). Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *26* (pp. 1842–1850). Curran Associates, Inc. Available at http://papers.nips.cc/paper/5205-visual-concept-learning-combining-machine-vision-and-bayesian-generalization-on-concept-hierarchies.pdf

Jones, M., & Goldstone, R. L. (2013). The structure of integral dimensions: Contrasting topological and cartesian representations. *Journal of Experimental Psychology: Human Perception and Performance*, *39* (1), 111–132.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 1066–1071). Mahwah, NJ: Lawrence Erlbaum.

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In R. Sun (Ed.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 405–410). Mahwah, NJ: Lawrence Erlbaum.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 316–332.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. doi:10.1002/acp.1801

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, *12*(1), 4–34.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309–332.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expèrimentale*, *57*(3), 203–220.

Mozer, M. C., Pashler, H., & Link, B. V. (2011). *An unsupervised decontamination procedure for improving the reliability of human judgments*. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24* (pp. 1791–1799). La Jolla, CA: NIPS Foundation.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*(3), 705–741.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1162–1173.

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

Sinha, P., & Russell, R. (2011). A perceptually based comparison of image similarity metrics. *Perception*, *40* (11), 1269–1281.

Tenenbaum, J. B. (1999). *Bayesian modeling of human concept learning*. In M. J. Kearns, S. A. Colla & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11* (pp. 59–65). Cambridge, MA: MIT Press.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multi-alternative choice. *Psychological Review*, *121*(2), 179–205.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Wah, C., Horn, G. V., Branson, S., Maji, S., Perona, P., & Belongie, S. (2014). Similarity comparisons for interactive fine-grained categorization. In L. O'Connor (Ed.), *(CVPR)* (pp. 859–866). Los Alamitos, CA: IEEE Computer Society.

Wah, C., Maji, S., & Belongie, S. (2015). Learning localized perceptual similarity metrics for interactive categorization. In L. O'Connor (Ed.), *2015 IEEE winter conference on* (pp. 502–509). Los Alamitos, CA: IEEE Computer Society.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Fig. S1. A depiction of the various representations tested for the rectangle stimuli. Throughout the rectangle experiments, we used stimuli that were sampled uniformly in width and height space. The four panels demonstrate the result of representing the same set of stimuli (used in the width category boundary task) in four different representational spaces: (a) {width, height}, (b) {area, aspect ratio}, (c) {log-width, log-height}, and (d) {log-area, log-aspect ratio}.

Table S1. Evaluation of different rectangle representations across various models. These comparisons were made without a hierarchical framework. The Theory-based model is a general form of various similarity models motivated by psychological theory. The Wah et al. model corresponds to the general form of a similarity model used in Wah et al. (2014).

Table S2. Optimal model parameters for each experiment

Table S3. Two-sample *t* test for Random versus Optimized Experiments, using empirical behavioral data.

Table S4. Predicted versus Observed Accuracy. The *F*-statistics and *p*-values show the results of testing if the fitted linear model is significantly different than a constant model. Results are shown when data are aggregated by implicit categorization tasks and by individual trials.