Classifying Objects Based on Their Visual Similarity to Target Categories

Wei Zhang (wzhang@cs.sunysb.edu)
Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA

Dimitris Samaras (samaras@cs.sunysb.edu)
Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA

Gregory J. Zelinsky (Gregory.Zelinsky@stonybrook.edu)
Departments of Psychology and Computer Science, Stony Brook University, Stony Brook, NY 11794 USA

Abstract

Visual similarity relationships underlie a host of human behaviors, and determining these relationships is crucial both to the understanding of these behaviors and the construction of automated systems designed for human use. We conducted a large-scale web-based experiment in which subjects rank ordered random objects according to their visual similarity to object classes. We then constructed a computational model, using Adaboost with color, texture, and shape features, to quantify the visual similarity between these objects and the target classes, and to perform the same similarity ranking task. The model and subjects showed good agreement in the objects judged to be most and least similar to the target categories. Our data also suggest that color, texture, and shape features are all useful for classification, and that the specific weighting of these features depends on the target object class. Moreover, we show that these target-like properties constitute a learnable class, as it was possible to train a classifier on target-like objects (i.e., without positive samples) to recognize actual targets.

Keywords: Cognitive science; Computer science; Concepts and categories; Machine learning; Computer simulation; Human experimentation.

Introduction

Knowing the similarity relationships between objects is key to understanding performance in many tasks. This is true for both human behavior and computer vision. The construct of similarity has been used to describe human behaviors ranging from attentional selection and visual search (Duncan & Humphreys, 1989; Raymond, Shapiro, & Arnell, 1995) to change detection (Zelinsky, 2003), recognition (Ashby & Perrin, 1988; Edelman, 1998) and categorization (Medin, Goldstone, & Gentner, 1993; Oliva & Torralba, 2001).

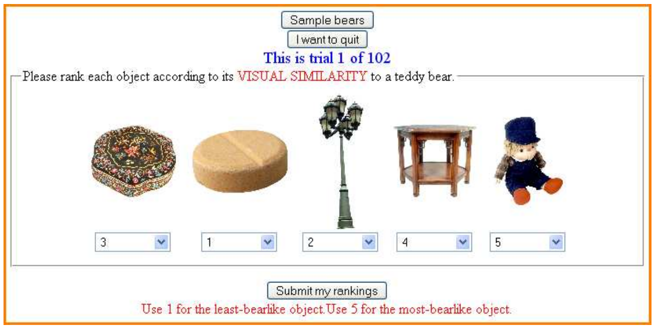
Establishing similarity relationships is also a core operation in many object recognition methods in the computer vision literature. Here, similarity relationships are discussed in the context of specific features, typically color (Swain & Ballard, 1991), texture (Dalal & Triggs, 2005; Lowe, 2004; Serre, Wolf, & Poggio, 2005), and shape (Belongie, Malik, & Puzicha, 2002; A. C. Berg, Berg, & Malik, 2005; Opelt, Pinz, & Zisserman, 2006).

Yet despite its importance to a wide range of fields, relatively few attempts have been made to computationally describe the features underlying human visual similarity judgments. The above mentioned computational work has made great progress in quantifying visual similarity in terms of a variety of features, but these estimates have not been validated in terms of human behavior. Conversely, quantitative theories of human behavior have been developed that describe psychologically meaningful multidimensional spaces (Ashby & Perrin, 1988; Nosofsky, 1992), but these approaches have used relatively simple patterns as stimuli so as to isolate the relevant feature dimensions (for notable exceptions, see (Oliva & Torralba, 2001; Schyns, Bonnar, & Gosselin, 2002; Zelinsky, Zhang, Yu, Chen, & Samaras, 2006)). The question of how visually similar people believe a random object, such as a coffee cup, is to a complex class of objects, such as teddy bears, is largely unknown.

Our goal in this study is to bridge the behavioral and computer vision communities by using methods of similarity estimation from computer vision to describe human visual similarity judgments. Behavioral similarity estimates were obtained from a web-based experiment in which participants rated the visual similarity of random realistic objects to two object classes, teddy bears and butterflies. A web experiment is perfect for this task, as a large number of subjects are required to obtain stable similarity estimates. We then used a machine learning technique with multiple heterogeneous features to similarly classify these objects into those most and least like the target classes. By comparing similarity estimates from the human and model, we can evaluate the usefulness of current state-of-the-art computer vision methods in capturing human visual similarity judgments, and possibly to learn more about the visual features used by humans in arriving at these estimates.

Behavioral Methodology

Human similarity ratings were obtained using a web-based behavioral experiment (interested readers can participate in the actual experiment at: http://www.cs.sunysb.edu/˜rankings/start.html). Subjects were 142 Stony Brook University students. Upon linking to the experiment, subjects were randomly assigned to either a teddy bear or butterfly/moth group; the two target classes used in this study. The experiment consisted of a training phase and a ranking phase. During training subjects were shown 200 examples of either teddy bear or butterfly/moth objects (not both). This was done to familiarize subjects with the types of objects constituting the target class, as well as to expose them to the feature variability among these objects. Except for the objects shown during training, and the instructions indicating the target class, subjects in the bear and butterfly experiments performed the identical task.



This is trial 1 of 102

Please rank each object according to its VISUAL SIMILARITY to a teddy bear.

3 1 2 4 5

Use 1 for the least-bearlike object.Use 5 for the most-bearlike object.

Figure 1: Screenshot of one trial in the visual similarity ranking phase of the web-based experiment. Subjects had to rank order the objects based on their visual similarity to teddy bears. A corresponding trial existed for subjects participating in the butterfly ranking task.

Similarity estimates were obtained during the ranking phase. Figure 1 shows a screen-shot of the ranking phase for one representative teddy bear trial. Five nontarget objects were presented on each trial, and the subject's task was to rank order the objects by assigning each a rank score (1-5) indicating its perceived visual similarity to the target class (either teddy bear or butterfly, depending on the condition). Note that a ranking task is preferable to having subjects assign an independent similarity score to each object, as this tends to produce many low similarity estimates due to subjects not using the full range of the rating scale. A ranking method avoids this problem by requiring, for each trial, an estimate of the least target-like object (rank of 1), the most target-like object (rank of 5), and three intermediately ranked objects.

Each subject performed 100 ranking trials, yielding similarity estimates for 500 objects. These 100 trials were randomly selected from a fixed set of 400 trials. Over subjects, 71,000 separate similarity estimates were obtained for 2000 common nontarget objects spanning a range of categories. All of these objects were selected from the Hemera object database, as were the target objects used in the butterfly/moth class. The teddy bear objects were obtained from (Cockrill, 2001).

Computational Methodology

We used color histogram features, texture features (SIFT), and global shape features in this study.

Color histogram

A histogram of hues was used to describe the global color feature of an object, similar to the approach used by (Swain & Ballard, 1991). Each sample image was first transformed into the HSV color space; background (white) and achromatic pixels were excluded from the histogram by setting a threshold on the saturation channel ($S<0.15$). The hue channel was evenly divided into 11 bins, and each pixel's hue value was assigned to these bins using binary interpolation. The final color histogram was normalized to be a unit vector. The similarity between a given pair of color histogram features, CH_1 and CH_2, was measured using the χ^2 statistic:

$$\chi^2(CH_1,CH_2) = \sum \frac{[CH_1(i) - CH_2(i)]^2}{CH_1(i) + CH_2(i)} \qquad (1)$$

where $CH(i)$ is the value of i^{th} dimension.

Scale Invariant Feature Transform (SIFT)

The texture feature of an object was described by a set of local SIFT descriptors applied at image coordinates indicated by an interest point detector. Following (Lowe, 2004), we localized interest points by finding local extremes on Difference-of-Gaussian (DoG) maps. A SIFT feature for a point encodes gradient information (orientation and magnitude) for all pixels within a 16×16 image patch surrounding the interest point. Each patch is further divided into smaller regions, with each subregion represented by an orientation histogram. The SIFT descriptor has been shown to be robust to rotation, translation and occlusion (Lowe, 2004).

To estimate the similarity between a SIFT feature, P, and a sample object, S, we found $\min D(P,Q_i)$, where $\{Q_i\}$ refers to the set of SIFT features from sample S, and $D(.)$ computes the Euclidean distance between a pair of SIFT features.

Shape context

We represented shape using the global shape context feature descriptor (Belongie et al., 2002). For each image, we sampled a fixed number of edge points evenly distributed along the object's contour. The distribution of these points was described by a coarse histogram feature consisting of uniform bins in log-polar space. The origin of the space was set to the center of the image. By counting the number of edge points grouped by discretized log-distances and orientations, each histogram captures the global shape properties for a given sample. The similarity between shape context features was measured by χ^2 distance, similar to the metric used for the color histogram feature (Eq. 1).

Boosting with heterogeneous features

In our method, each color histogram, SIFT, and shape context feature obtained from positive training samples becomes a candidate feature that can be selected and used to classify target from nontarget objects. To select the most discriminative features for classification from this training set, we use a popular machine learning technique, AdaBoost (Freund & Schapire, 1997). The application of AdaBoost, or boosting, refers to the general method of producing a very accurate prediction rule by combining relatively inaccurate rules-of-thumb (Viola & Jones, 2001). In this study we use AdaBoost with heterogeneous features, as described in (Zhang, Yu, Zelinsky, & Samaras, 2005). This method is similar to AdaBoost, except that the different features are processed independently. This means that separate similarity scores are computed between each sample and each feature type, resulting in separate feature-specific classifiers. Two classifiers were learned and used in this study; one discriminating teddy bears from non-bears, and the other discriminating butterflies from non-butterflies. The original sources should be consulted for additional details regarding the AdaBoost method.

Experimental results

Behavioral data

Subjects varied considerably in the objects they ranked as being similar to the target classes. Figure 2 summarizes this variability by showing the number of objects for each ranking, grouped by the level of agreement among subjects. Two patterns are evident from this analysis. First, and as expected, the number of consistently ranked objects decreases as the agreement criterion becomes stricter. For example, there were 150 objects that 60% of the subjects ranked as being most bear-like, but only 33 of these objects were ranked as most bear-like by 80% of the subjects. Second, subjects were most consistent in their rankings of the most target-like objects (rank 5), and second most consistent in their rankings of the least target-like objects (rank 1). There was generally less consistency among the objects ranked as neither most nor least target-like (ranks 2-4). Given these patterns, all subsequent analyses will only include those objects ranked most or least target-like based on a 60% level of subject agreement. We chose this agreement criterion because it afforded a relatively large number of objects while still yielding a statistically significant level of consistency, $p < .005$. Note also that subjects ranking objects for bear similarity were more consistent in their estimates than subjects ranking objects for butterfly similarity, a finding that may reflect greater variability in the butterfly/moth target class compared to teddy bears.

Human and model similarity rankings

We used our multi-feature model to classify the objects ranked as most target-like (rank 5) and least target-like (rank 1) for the bear and butterfly target classes. Again, we limited this effort to only the objects ranked consistently by human subjects, based on a 60% level of agreement. For each category the model was trained on the 200 positive samples of targets shown to subjects during the training phase of the behavioral experiment. For negative samples we used 800 random objects that were not used in the web ranking experiment.

There was good agreement between the objects selected by subjects as most and least target-like and the corresponding objects ranked by the model. Figure 3 illustrates this agreement by plotting the mean rankings for the most and least target-like objects from Figure 2. For the teddy bear category, the mean ranking scores for the least bear-like objects were 1.54 and 1.98 for the human subjects and the model, respectively; the corresponding scores for the most bear-like objects were 4.49 and 4.10. For the butterfly/moth category, the mean ranking scores for the least target-like objects were 1.67 and 1.91 for subjects and the model; the most target-like scores were 4.38 and 3.71. Differences between the least and most target-like objects were highly significant for both subjects and the model in both target classes (all $p < .0001$, by two-tailed t-test). However, when the human and model estimates were compared directly, we found that the average model estimates fell outside of the 95% confidence intervals for the corresponding behavioral means. In general, human rankings tended to be more extreme compared to the model, with slightly higher rankings for most target-like objects and slightly lower rankings for least target-like objects.

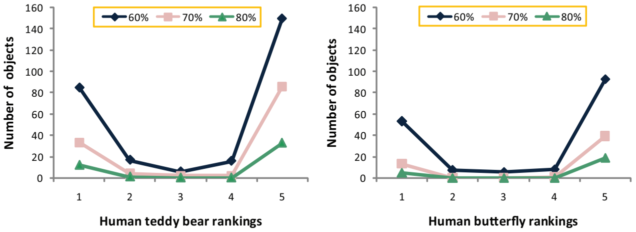


Figure 2: The number of consistently ranked objects for different levels of agreement. Left panel, bear rankings; right panel, butterfly rankings.

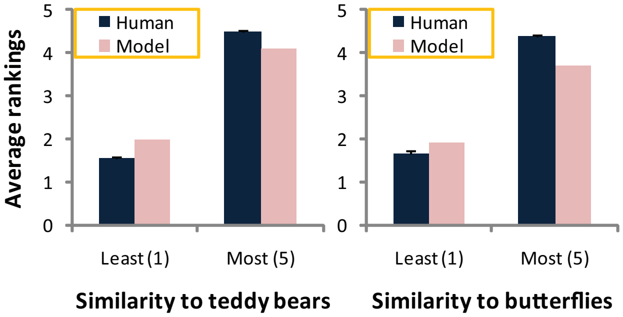


Figure 3: Average human and model rankings for the least target-like and most target-like objects, based on a 60% level of agreement. Left panel, bear rankings; right panel, butterfly rankings. Error bars indicate one standard error of the mean (SEM).

Figure 4 shows a more detailed breakdown of the model rankings, by target category. As clearly indicated by the crossover interactions, our model was able to correctly distinguish between the most target-like and least target-like objects. For the objects ranked least target-like by 60% of our subjects, the model correctly assigned a rank of 1 to approximately 50% of these objects. The chance probability of a 1 ranking is .2, given the 5 objects per trial in the web experiment. The figure also shows a monotonic decrease in the probability of the model assigning a target dissimilar object a higher rank. For example, the probability of the model misclassifying a least target-like object as target-similar (rank 4 or 5) was only about .1. A similar pattern was found for the most target-like objects. The model ranked these objects as 5 with high probability (.50 and .37 for the bear and butterfly categories, respectively), and ranked these objects as 1 with low probability (.1 or less). Considering the fact that up to 40 of the subjects failed to agree on these rankings, the classification rates generated by the model are highly representative of human behavior.

The above analyses demonstrated good agreement between our multi-feature model and human behavior with respect to visual similarity ranking, but are some features better than others in describing human behavior, and do these features depend on the target category? To address these questions we conducted additional computational experiments in which

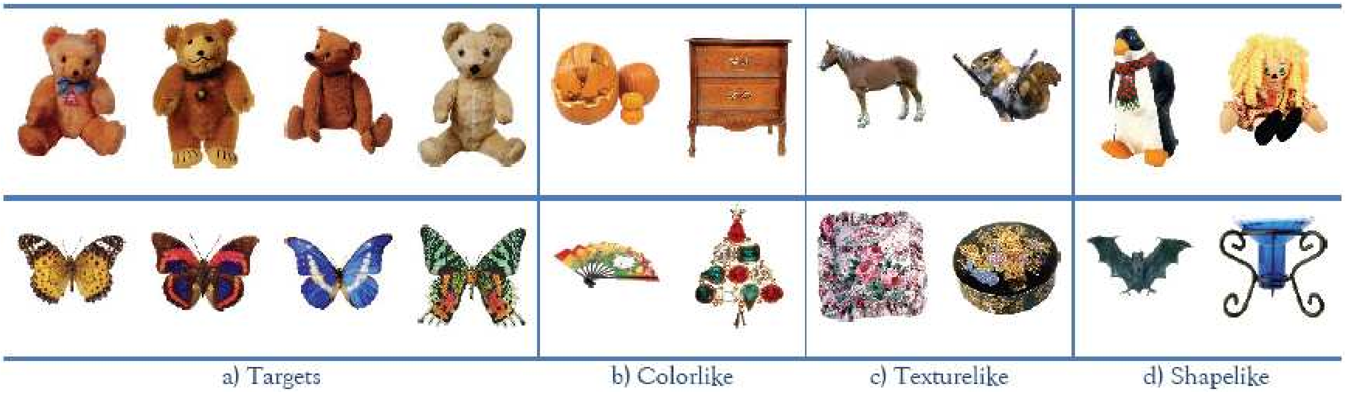


Figure 6: a) Representative bear and butterfly targets. b) Representative objects ranked as most target-like by human subjects and a version of our model using only a color histogram feature. c) Objects ranked as most target-like by a texture-only model. d) Objects ranked as most target-like by a shape-only model.

we attempted to classify least target-like and most target-like objects using either color histogram, SIFT, or shape context features, rather than a combination of the three. These results are shown in Figure 5, along with data from the combined feature model for comparison. In general, models using any single feature alone do not describe human behavior as well as the full multi-feature model. There is also evidence for features contributing differently to the bear and butterfly tasks. Shape features are most discriminative for the bear category, and produce the closest agreement to the bear-like objects ranked by subjects. For butterflies the SIFT feature was most discriminative, which suggests that human subjects relied most on texture when ranking objects as either least or most butterfly-like. Interestingly, the contribution of color was relatively minor in the model's butterfly rankings. This may be due to the fact that the color histogram feature estimates similarity by computing distances between distributions of hues, which fails to capture the color variability in an object that may be diagnostic of the butterfly object class.

Figure 6 shows representative samples illustrating feature-specific contributions to similarity estimates. Teddy bears and bear-like objects are shown in the top row; butterflies and butterfly-like objects are shown in the bottom row. All of the nontargets were ranked as most target-like by both the human subjects and a version of the model using color histogram (panel b), SIFT (panel c), and shape context features (panel d), individually. Clearly, all three feature types capture dimensions of "bearness" and "butterflyness", and might therefore be useful in deriving similarity estimates to these target classes.

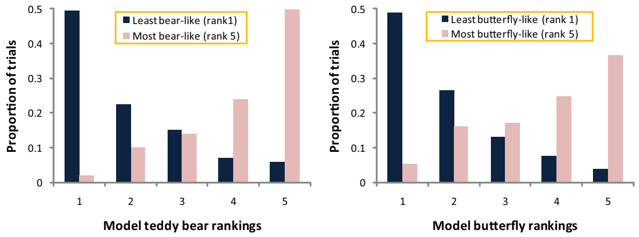


Figure 4: Distribution of model rankings for objects ranked least target-like and most target-like by human subjects. Left panel, bear rankings; right panel, butterfly rankings.

Despite the demonstrated contribution of color, texture, and shape features to the task of similarity ranking, it is undoubtedly the case that subjects used features in addition to these three when making their similarity estimates. One such feature is likely a semantic descriptor. Although subjects were instructed to base their judgments only on an object's visual similarity to the target class, completely excluding semantic similarity from such estimates is difficult. Figure 7 shows some cases where semantic similarity may have influenced the behavioral ranking. These objects were rejected by the model but ranked as most target-like by human subjects, despite having different shape, color, and texture features. If semantic factors affected these similarity estimates, this might explain why the model failed to correctly classify these objects as target-like. However, a strength of our approach is that, as new features are discovered and new computational feature descriptors become available, they can be easily integrated into our multi-feature model.

Learning from target-like objects

To further validate our computational method of estimating visual similarity, we analyzed the false positive errors made by our multi-feature model. If our classifiers truly learned the features of "bearness" and "butterflyness", we would expect higher false positive rates to the most target-like objects

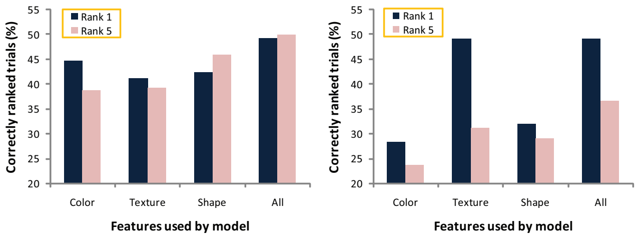
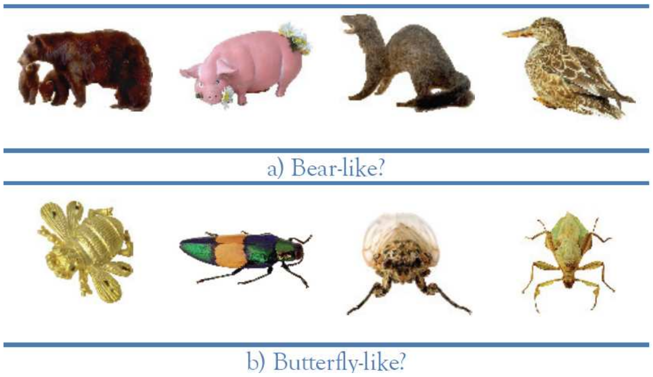


Figure 5: The percentage of trials with correctly ranked least and most target-like objects, as estimated by single feature (color, texture, shape) and multi-feature (all) versions of the model. Left panel, teddy-bear model rankings; right panel, butterfly model rankings.



a) Bear-like?

b) Butterfly-like?

Figure 7: Objects ranked as target-like by human subjects that might have been influenced by semantic similarity.

ranked by human subjects. Using the same training and testing sets described under the computational methods section, we tested this hypothesis by adjusting the classifier thresholds so that they accepted 20% of random objects as positive targets. Such an adjustment was necessary because our classifiers recognized both categories with equal error rates greater than 95%, thereby providing too few false positives to analyze. We then reanalyzed the most and least target-like objects using this more liberal classification threshold and observed the false positive rates (FPRs). These results are shown in Figure 8 for the bear and butterfly classes, as a function of agreement level. The FPR for objects ranked most target-like by human subjects was above 40%, over twice the 20% FPR for random nontargets. Conversely, the FPR for objects ranked least target-like by human subjects was approximately 5%, well below the FPR for random objects. Together, these patterns suggest another point of agreement between our model and human behavior; the objects ranked as most similar to the target categories are the same objects that are most likely to be misclassified by the model.

The previous analyses suggested that "bearness" and "butterflyness" may be learnable classes, which raises the intriguing possibility that actual teddy bear and butterfly targets might be recognized by classifiers trained entirely on bear-like and butterfly-like nontargets. Note that this is very different from a standard category learning problem, in which a classifier is trained from positive samples of the actual target class. Is it possible to learn a classifier for an object category without any positive training samples?

We addressed this question by training classifiers on the most target-like objects from the behavioral rankings. As before, the training set was composed of the most bear-like and butterfly-like objects based on a 60% level of agreement among subjects, as well as random objects from the web experiment which were used as negative samples. The testing set consisted of 100 actual teddy bear and butterfly targets, and random objects that were not used in the web experiment. The results from this experiment are shown in Figure 9. Classifiers trained on target-like objects achieved equal error rates (EERs) of 92% and 70% for the teddy bear and butterfly target classes, respectively (red data). Clearly, the classification rates for both target classes were better than the random classification baseline (green dashed line). For comparison, we also show data from classifiers trained on positive samples and tested on the same dataset. Predictably, these EERs were even higher, especially for the teddy bear target class (EER = 99% for teddy bears, EER = 95% for butterflies). These results provide the first implementation proof that categories of visually complex targets can be recognized by classifiers trained entirely without positive samples of the target classes. They also provide proof positive that "bearness" and "butterflyness" are learnable visual categories.

Conclusions and future work

The estimation of visual similarity is a core operation in many human and computer systems. In this study we collected a large number of behavioral similarity estimates between random objects and two target classes. Through computational experiments using the Adaboost machine learning method with heterogeneous features, we showed that simple color, texture, and shape features can describe the objects ranked as most and least target-like by human subjects. We also provided evidence for a category-specific weighting of these features, and demonstrated that they can be used to define learnable classes of target-like objects. Although we have no reason to believe that AdaBoost describes how human's actually learn discriminative features, it is certainly true that discriminative features are used when making similarity judgments, and that these features are largely unknown for real-world objects. Our study is a first step towards discovering and quantifying these behaviorally-relevant features.

These findings have important implications for both human and computer vision systems. Determining the visual features used to code complex object classes is fundamental to understanding a host of human behaviors in the real world. For example, most day-to-day search tasks are categorical; it is rare for us to have a preview specifying a target's exact appearance. Nevertheless, search in these situations is guided to the categorically-defined target (Yang & Zelinsky, 2006; Zhang, Yang, Samaras, & Zelinsky, 2006). What are the features of the target class used by this guidance operation? The current study suggests potential answers to this question, at least with respect to the teddy bear and butterfly target categories. Future work will determine whether the category-specific fea-

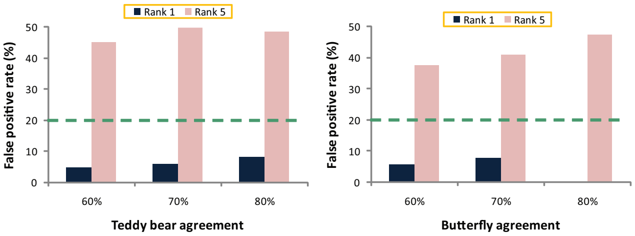


Figure 8: Model false positive rates for the least and most target-like objects, as ranked by subjects at different levels of agreement. The overall FPR for the classifiers was set to 20%, indicated by the dashed green lines.

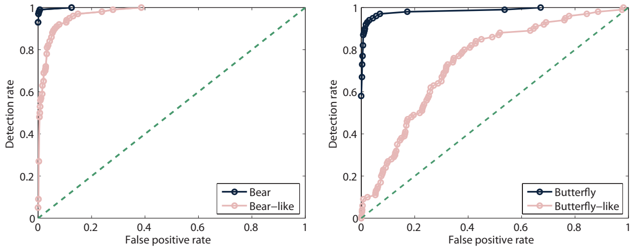


Figure 9: ROC curves from classifiers trained on only target-like objects and tested on actual targets (in red), and from classifiers trained and tested on actual targets (in blue). The dashed green line indicates the baseline classification rate.

tures suggested by this study are the same features used to guide gaze during a categorical search task.

Our findings also have implications for the construction of computer systems designed to interact with human users. For example, our methods of obtaining behavioral similarity estimates and learning from these estimates the features of object classes might be incorporated into common web-search applications. Tagging images with text-labels (e.g., through Flickr.com) has become easy and widespread, and these text labels have proven very useful in the automated understanding of image semantics (T. L. Berg & Forsyth, 2006; Quattoni, Collins, & Darrell, 2007). As demonstrated in this study, using a web-based task to rank images for visual similarity is also relatively easy and reliable, perhaps as easy as adding text labels to images. This raises the possibility that a web-search application, if combined with methods for similarity estimation, could treat as a class the group of objects that humans consider most similar. Rather than simply searching for images that have been labeled as “teddy bears”, it might therefore be possible to search for images based on their visual similarity to the teddy bear object class.

Acknowledgements

This work was supported by grants from NSF (IIS-0527585) and NIMH (R01-MH63748) to G.J.Z. We would also like to thank Hyejin Yang for her help in obtaining the images used in this study.

References

Ashby, F., & Perrin, N. (1988). Toward a unified theory of similarity of recognition. *Psychological Review*(95), 124-150.

Belongie, S., Malik, J., & Puzicha, J. (2002, April). Shape matching and object recognition using shape contexts. *PAMI*, *24*(4), 509-522.

Berg, A. C., Berg, T. L., & Malik, J. (2005). Shape matching and object recognition using low distortion correspondences. In *CVPR05* (pp. I: 26–33).

Berg, T. L., & Forsyth, D. A. (2006). Animals on the web. In *CVPR06* (pp. II: 1463–1470).

Cockrill, P. (2001). *The teddy bear encyclopedia*. New York: DK Publishing, Inc.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR05* (p. I: 886-893).

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*(96), 433-458.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*(21), 449-498.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139.

Lowe, D. (2004, November). Distinctive image features from scale-invariant keypoints. *IJCV*, *60*(2), 91-110.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*(100), 254-278.

Nosofsky, R. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*(43), 25-53.

Oliva, A., & Torralba, A. (2001, May). Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, *42*(3), 145–175.

Opelt, A., Pinz, A., & Zisserman, A. (2006). Incremental learning of object detectors using a visual shape alphabet. In *CVPR06* (pp. I: 3–10).

Quattoni, A., Collins, M., & Darrell, T. (2007). Learning visual representations using images with captions. In *CVPR07*.

Raymond, J., Shapiro, K., & Arnell, K. (1995). Similarity determines the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*(21), 653-662.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! understanding recognition from the use of visual information. *Psychological Science*(13), 402-409.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *CVPR05* (p. II: 994-1000).

Swain, M., & Ballard, D. (1991, November). Color indexing. *IJCV*, *7*(1), 11-32.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR01* (p. I:511-518).

Yang, H., & Zelinsky, G. J. (2006). Evidence for guidance in categorical visual search. *Journal of Vision*, *6*(6), 449a.

Zelinsky, G. J. (2003). Detecting changes between real-world objects using spatio-chromatic filters. *Psychonomic Bulletin and Review*(10), 533-555.

Zelinsky, G. J., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1569–1576). Cambridge, MA: MIT Press.

Zhang, W., Yang, H., Samaras, D., & Zelinsky, G. J. (2006). A computational model of eye movements during object class detection. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1609–1616). Cambridge, MA: MIT Press.

Zhang, W., Yu, B., Zelinsky, G. J., & Samaras, D. (2005). Object class recognition using multiple layer boosting with heterogeneous features. In *CVPR05* (p. II: 323-330).