Visual Similarity is ObViS

Michel E. Brudzinski (brudzm@rpi.edu)
Chris R. Sims (simsc@rpi.edu)
Wayne D. Gray (grayw@rpi.edu)
Michael J. Schoelles (schoem@rpi.edu)
Cognitive Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180

Abstract

Visual search for a target is affected by visual similarity. Research on visual similarity has primarily focused on the high-level features of objects. Real-world objects are composed of low-level features that can be harder to measure and categorize. We have developed ObViS, an algorithm that measures the visual similarity of objects, based on Rao & Ballard (1995). ObViS calculates a high-dimensional vector that represents the low-level features of a real-world object. The algorithm was applied to a library of real-world object images in order to calculate the similarity of each object to every other object in the library. Two experiments evaluated the ability of our algorithm to predict the effects of visual similarity on visual search behavior.

Keywords: similarity; visual similarity; visual search.

Introduction

The visual similarity of objects in a scene is an important exogenous factor driving visual search behavior. Theories of visual search acknowledge the importance of similarity but do not specify how the visual search system uses these exogenous factors to guide search (Wolfe, 2007).

Most models of visual search, such as Treisman's Feature Integration Theory (FIT) (Treisman & Gelade, 1980), propose separate serial and parallel search mechanisms. A limited set of features such as color and size can be processed in parallel. Parallel visual search for a target that can be distinguished solely on the basis of one of those features is fast and efficient. Serial visual search for a target that is defined by multiple features is slower and requires attention to bind together the features of objects.

Six of the eight phenomena that Wolfe lists as affecting visual search response time entail some form of visual similarity: (1) target-distractor similarity, (2) distractor heterogeneity, (3) flanking/linear separability, (4) search asymmetry, (5) categorical processing, and (6) guidance (Wolfe, 2007). Each of the six types of visual similarity is specific to a low-level feature such as size, color, or orientation.

In addition to low-level features, research on visual similarity has also focused on high-level features. Approaches to the study of similarity include: geometric, feature-based, alignment-based and transformational measures of similarity (Goldstone & Son, 2005). Each of these approaches requires a form of reductionist representation of low-level properties, features, or elements.

Research on visual search has focused on issues surrounding the deployment of the serial or parallel processes. Visual search in the laboratory has used simple stimuli, manipulated set size or a few high-level features, tested search for a single feature, or the conjunction of two features, and used response time as their primary measure. Visual search in the real-world involved conjunctions of many low-level features that can be difficult to measure or categorize.

Statistical models of visual search, such as Itti & Koch (2001), have demonstrated the potential use images of real-world scenes in visual search experiments. Their mathematical model simulates the role of bottom-up saliency in guiding visual attention. It extracts low-level visual features such as color, intensity, and orientation from real-world images. The model calculates the conspicuity at every point in a scene and thus is able to provide bottom-up guidance to direct visual attention. Such models predict where the eye is attracted to in a visual scene and, thereby, have an important but limited role to play in explanations of visual search.

Top-down guidance may dominate bottom-up guidance, such as saliency, when there is a search target (Chen & Zelinsky, 2006). The goal of a visual search for a particular target is to find a location in a scene that matches the visual features of the target. The visual search system must be able to maintain a representation of the features of the target and compare those features with features at locations in the scene.

Rao & Ballard's Active Vision Architecture describes two primary visual routines: one for object identification and one for object location (Rao & Ballard, 1995). Statistical models of visual similarity, such as Rao & Ballard's, share with saliency models their reliance on low-level visual features. They differ from saliency models in that they define vectors of low-level visual features for a known target and for the locations in a scene. They then compare the similarity of the target vector with the vectors computed for locations in the scene. Top-down guidance is based on the statistical similarity of scene locations to the search target.

Cognitive architectures, such as ACT-R (Anderson & Lebierre, 1998), are used to model visual search behavior. Visual attention guidance in ACT-R suffers from the same reliance on high-level visual features that limits most theories of visual search. Statistical models of both bottom-up and top-down guidance would greatly increase the ability

of cognitive models to model real-world visual search. We have implemented a variation of Rao & Ballard's model and applied it to the study of visual search with real-world objects.

Algorithm

Our algorithm, ObViS, is a variant of the one developed by Rao & Ballard (1995) and Rao et al (2002). The algorithm represents image patches using high dimensional feature vectors, where the computed features consist of the image response to oriented spatial frequency filters. Such filters approximate the receptive fields of simple cells in primary visual cortex, and are also similar to features obtained from the statistics of natural images (Hyvärinen, Huri, & Hoyer, 2009). In our implementation, we used 10 filters defined by the directional derivatives of a 2D Gaussian, using derivatives of up to 3rd order. The 10 combinations of Gaussian derivative order and orientation were as follows:

Table 1: Steerable basis set.

Order of derivatives	Filter orientations used (degrees)
0	0
1	0,90
2	0,60,120
3	0,45,90,135

This set of filters was chosen as it forms a steerable basis set—that is, the filter response at any orientation can be computed by a linear combination of the filter responses in the basis set (Freeman & Adelson, 1991). This property endows the feature representation with some rotation invariance, though we have not explored this property in our current work. In our implementation, the filter kernels were 9x9 pixel discretized versions of the Gaussian derivatives defined above. These 10 filters were applied to 3 color channels extracted from the original image: luminance, red–green, and blue–yellow color opponency channels. In addition, the filters were applied to these color channels at 5 spatial scales, by resampling the image to 25, 50, 100, 200, and 400% of its original size. Thus in total, each image was represented by the ObViS algorithm using a set of 150 measurements (10 filters x 3 color channels x 5 spatial scales). The use of color opponency channels was an extension to the algorithm presented by Rao et al (2002), as their implementation used only grayscale images whereas we are interested in capturing the visual similarity of color images. Finally, to determine the visual similarity between any two images, we determine the feature vector representation for two images, and then calculate the root mean square (RMS) of the difference between the images' respective feature vectors. Images with low RMS difference are highly similar according to the ObViS measure of visual similarity.

Experiments

We conducted two visual search experiments that examined ObViS' accuracy in predicting human visual search behavior. Subjects were asked to find target objects located in a circular array of object images from the Amsterdam Library of Object Images (ALOI) (Geusebroek, Burghouts, & Smeulders, 2005). We compared the timing, and accuracy of responses, and the number of fixations with predictions based on ObViS' calculations of visual similarity. The two experiments differed in how similar the distractors displayed on each trial were to the target object. In experiment 1, for each trial, the distractors in the search array were all either similar, or dissimilar to the target. In experiment 2, for each trial, the distractors in the search array were approximately half similar and half dissimilar to the target.

Methods

Subjects. Thirty-four RPI undergraduates participated in the experiment 1 and twenty-seven RPI undergraduates participated in experiment 2. All subjects received course credit for their participation and signed informed consent forms. Subjects were screened for color blindness using a 10-plated Ishihara test (Ishihara, 1987).

Materials. The experiment was run on a Mac OSX computer and displayed on a 17" flat-panel LCD monitor with a screen resolution set to 1280 x 960 pixels. The software used for the experiment was written in LispWorks 5.0. The object images displayed during the experiment were 192 x 192 pixel images of real-world object from the Amsterdam Library of Object Images. A Cedrus RB-834 response pad was used to collect responses. White noise was played over headphones, using the freeware program Noise, to reduce auditory distractions. All subjects in these experiments were eye-tracked using an LC Technologies eye-tracker that recorded at a rate of 120 Hz. Subjects were asked to rest their chin on a chinrest throughout the experiment.

Design. The same 100 target objects were used as the target objects in experiments 1 and 2. The target object was present in the search array in only half of the trials. Subjects had to respond whether the target objected was present, and if so, identify its location in the circular search array. Experiments 1 and 2 differed in how similar the distractors displayed on each trial were to the target object.

In experiment 1, subjects performed a visual search for a target in a search array that contained distractor objects that were either similar or dissimilar to the target. All subjects in experiment 1 saw the same 400 trials. In half of the trials, the distractor objects were all similar to the target; in the other half of the trials the distractor objects were all dissimilar to the target (Table 2).

In experiment 2, subjects performed a visual search for a target in a search array that contained approximately half similar and half dissimilar distractor objects (Table 3). The similarity of all distractor objects was based on calculations

from the ObViS algorithm. All subjects in experiment 2 saw the same 400 trials. The locations of all objects, in all trials, for all subjects, were randomized.

Table 2: Experiment 1 trials.

Trial count	Targets	Similar Distractors	Dissimilar Distractors
100	0	8	0
100	0	0	8
100	1	7	0
100	1	0	7

Table 3: Experiment 2 trials.

Trial count	Targets	Similar Distractors	Dissimilar Distractors
100	1	4	3
200	0	4	4
100	1	3	4

Procedure. Experiments 1 and 2 used the same procedures. All task instructions were presented using a Keynote presentation prior to the experiment. Subjects pressed a button on the response pad labeled "next" to begin each trial. A fixation cross, consisting of a white "+" was displayed in the center of the screen (Figure 3a). The trial did not begin until the participant had fixated on the fixation cross for 500 milliseconds. The target image was then displayed in the center of the screen for 300 milliseconds (Figure 3b). A random dot image was displayed for 300 milliseconds (Figure 3c). A circular search array was then displayed until the subject responded by pressing the "Present" or "Absent" button on the response pad (Figure 3d).

Following a response, the random dot image was displayed for another 300 milliseconds (Figure 3e). If the response indicated that the target was present, the subjects were asked to indicate the location of the target. Buttons were arranged on the screen in locations that matched the locations of objects in the search array (Figure 3f). Subjects responded by moving a mouse to and clicking on one of the buttons. Once the participant responded, a progress screen displayed the number of trials completed out of the total number of trials. Subjects pressed the response pad button labeled "next" to begin the next trial (Figure 3).

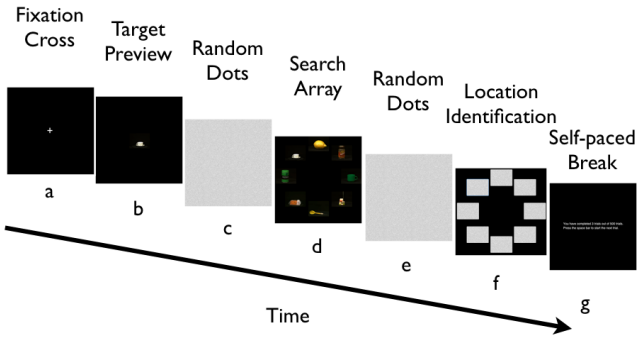

Figure 3: Experimental procedure for experiments 1 & 2.

Measures. The search array was displayed until the participant responded that the target object was either present or absent. The duration of time from the initial display of the search array until the participant responded was measured as the response time. The participant's response was recorded and measured as target presence accuracy.

Eye-tracking data was recorded throughout the experiment. The number of fixations on distractor objects was counted for each trial.

Results

Response time was compared for trials in which the target object was present in the search array and trials in which it was absent.

In experiment 1, subjects took longer to respond on trials when the target was absent (M = 1190.13, SE = 61.76), than on trials in which the target was present (M = 1066.22, SE = 34.05), $t(129)_{\text{two-tail}}$ = 1.716, p = 0.089, marginally significant.

Experiment 2 showed the same trend in response times: subjects took longer to respond on trials in which the target object was absent (M = 1234.83, SE = 54.82), as compared to trials when the target was present (M = 1064.59, SE = 31.76), $t(79)_{\text{two-tail}}$ = 2.872, p = 0.005 (Figure 4).

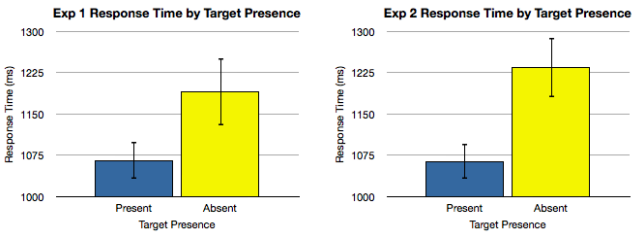

Figure 4. Response time (ms), by target presence, in experiments 1 & 2.

In experiment 1, each trial contained distractors that were all similar or all dissimilar to the target object. Response time was significantly greater for trials with distractors that were similar to the target (M = 1244.35; SE = 56.04) than for trials with dissimilar distractors (M = 1012.00; SE = 39.18), $t(130)_{\text{two-tail}}$ = 1.66, p = 0.0009 (Figure 5). In

experiment 2, each trials contained both similar and dissimilar distractors, so the analogous comparison was not possible.

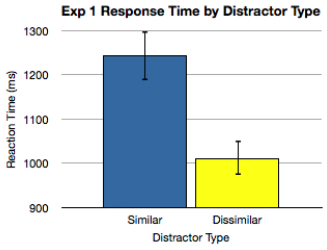

Figure 5. Response time (ms), by distractor type, in experiment 1.

The number of fixations on distractor objects was counted for each trial. In experiment 1, when the target was absent, subjects averaged more fixations on similar distractors (M = 2.99; SE = 0.22), than on dissimilar distractors (M = 1.76; SE =0.17). When the target was present, subjects averaged fewer fixations on distractors, but still had more fixations on similar distractors (M = 0.88; SE = 0.07), than on dissimilar distractors (M = 0.41; SE = 0.04) (figure 6). An ANOVA showed a significant main effect of target presence, (F(1, 124) = 23.15, p < 0.001). There was also a significant main effect of distractor type, (F(1, 124) = 6.46, p < 0.05). There was no significant interaction.

In experiment 2, when the target was absent, subjects averaged more fixations on similar distractors (M = 1.96; SE = 0.11), than on dissimilar distractors (M = 0.94; SE = 0.07). An ANOVA showed a significant main effect of target presence, (F(1, 100) = 70.20, p < 0.001). There was also a significant main effect of distractor type, (F(1, 100) = 25.73, p < 0.001). There was a significant interaction between the two factors (F(1, 100) = 25.73, p < 0.001); there was a larger difference in mean fixation count for trials with similar distractors when the target was absent. When the target was present, subjects averaged fewer fixations on distractors, but still had more fixations on similar distractors (M = 0.54; SE = 0.03), than on dissimilar distractors (M = 0.22; SE = 0.02) (Figure 7).

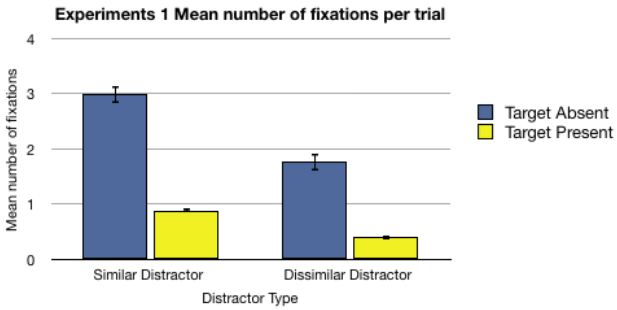

Figure 6. Mean number of fixations on distractor object per trial, by target present and distractor type, in experiment 1.

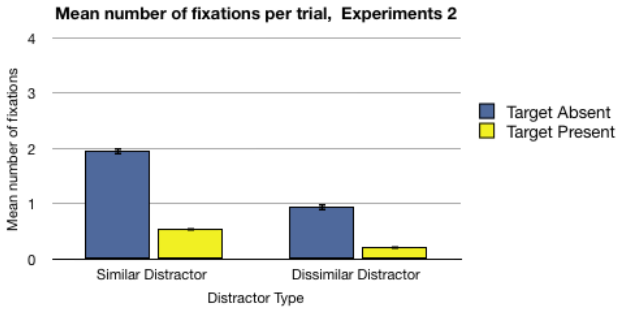

Figure 7. Mean number of fixations on distractor object per trial, by target present and distractor type, in experiment 2.

Target presence accuracy was a measure of the accuracy of subjects' response that the target was present or absent. Overall, subjects made few mistakes in both experiment 1 (M = 96.32%; SE = 0.36), and experiment 2 (M = 96.30%; SE = 0.002). In experiment 1, accuracy was significantly higher for trials in which the target was absent (M = 97.06%; SE = 0.36), than when the target was present (M = 95.58%; SE = 0.60), $t(130)_{\text{two-tail}}$ = 2.12, p = 0.018. In experiment 2 there was no main effect of target presence on accuracy; accuracy for trials with the target absent (M = 95.83%; SE = 0.002), was not significantly different than trials with the target present (M = 96.54%; SE = 0.003) , $t(79)_{\text{two-tail}}$ = 1.36, p = 0.17. Each trial in experiment 1 had distractors that were either all similar to the target or all dissimilar to the target. Target presence accuracy was higher for trials with dissimilar distractors (M = 97.76%; SE = 0.60), compared to trials with similar distractors (M = 94.88%; SE = 0.36), $t(130)_{\text{two-tail}}$ = 4.31, p = 0.0001.

General Discussion

We developed the ObViS algorithm in order to measure visual similarity for the top-down guidance of visual search. One of the goals of this work was to extend the study of visual search and visual similarity to real-world objects. We replicated the basic visual search phenomena. The algorithm's calculations were used to manipulate the similarity of visual search distractors. The response time data replicated phenomena typically found in laboratory search tasks using very simple stimuli; subjects took longer to find target when the distractors were similar to the target. They also made more mistakes in their responses when the distractors objects were similar to the target.

Fixation data added an additional source of information that has not typically been used in the study of similarity. Our results demonstrated that the longer response times for visual searches with similar distractors were the result of a greater number of fixations.

There are other visual search phenomena that we did not test. We did not manipulate set size, randomize locations, occlude objects, or place objects in natural scenes. All of

these phenomena could be studied in future work using our measures of visual similarity.

Conclusions

We developed a measure of the visual similarity of real-world objects based on the representation of their low-level visual features. We applied the algorithm to a library of object images. The resulting similarity calculations were used to manipulate the similarity of distractors in visual search tasks. We replicated basic findings on the effects of target presence and distractor similarity, using real-world objects. Further refinement of the ObViS algorithm could improve its ability to predict the effects of visual similarity on visual search. The algorithm could be used to create iconic representations to guide top-down visual search in computational models. ObViS could extend the study of visual search and visual similarity to real-world objects and even provide visual representations for cognitive architectures.

Acknowlegments

This work was supported, in part, by grant N000140710033 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer.

References

Anderson, J.R., & Lebiere, C. (Eds.). (1998) *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Chen, X., & Zelinsky, G.J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research, 46*, 4118-4133.

Freeman, W.T., & Adelson, E.H. (1991). The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(9), 891-906.

Geusebroek, J.M., Burghouts, G.J., & Smeulders, A.W.M. (2005) The Amsterdam library of object images. *International Journal of Computer Vision, 61(1)*, 103-112.

Goldstone, R.L., & Son, J.Y. (2005). Similarity. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning*. New York: Cambridge.

Hyvärinen, A., Hurri, J., & Hoyer, P.O. (2009). Natural Image Statistics: A Probabilistic Approach to Early Computational Vision: Springer.

Ishihara, S. (1987). Ishihara's tests for colour-blindness (concise edition). Tokyo: Kanchara.

Itti, L., & Koch, C. (2001) Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*, 194-203.

Treisman, A., & Gelade, C. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Rao, R.P.N., & Ballard, D.H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78, 461-505.

Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., & Ballard, D.H. (2002). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78, 461-505.

Wolfe, J.M. (2007). Guided Search 4.0: Current progress with a model of visual search. In W. Gray (ed.), Integrated Models of Cognitive Systems (pp. 99-119). New York: Oxford.