Object Knowledge Influences Visual Image Segmentation

Shaun P. Vecera

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890
vecera+@cmu.edu

Abstract

Visual image segmentation is the process by which the visual system groups locations that are part of the same object. Can knowledge of objects influence image segmentation, or is the segmentation process isolated from object information? The use of object knowledge at this stage of vision might seem premature, as the goal of segmentation is to provide input to object recognition. However, purely bottom-up image segmentation has proven a computationally difficult task, suggesting that a "knowledge-based" approach might be required. We addressed this issue using two segmentation tasks: Subjects either determined whether a small 'x' was located inside or outside the region subtended by a block shape, or they determined whether two small x's were on the same shape or different shapes. The familiarity of the shapes was manipulated, and subjects were fastest to segment the visually familiar shapes. These results suggest that image segmentation can be partly guided by information about familiar objects, consistent with knowledge-based image segmentation models.

In everyday vision, observers constantly see objects that overlap and partially occlude one another. In order for the visual system to recognize overlapping objects, each visual object representation must receive input from the regions of space that correspond to that object and only to that object. Visual image segmentation is the process by which the visual system groups locations that belong to individual objects. Is visual image segmentation a bottom-up process, or is object knowledge used to partially guide segmentation?

Each of these alternatives seems equally plausible. Bottom-up image segmentation is consistent with the application of heuristics to the visual field irrespective of the familiarity of objects in the field. This approach would result in unfamiliar objects being segmented as quickly and efficiently as familiar objects. The work of the Gestalt psychologists can be viewed as an attempt to identify bottom-up heuristics for grouping elements of the visual field without the use of object knowledge (Wertheimer, 1923). Some research in computational vision has also focused on developing bottom-up segmentation processes. For example in Marr's (1982) visual processing model, the grouping of features represented in the raw primal sketch corresponds to the full primal sketch. However, the full primal sketch is not influenced by the familiarity of the objects in the visual field. Instead, these visual processes are applied to all objects equally. Other computer vision systems apply processes such as pixel or region classification that make no reference to the object or shape being segmented (Rosenfeld, 1984).

Although some work on image segmentation assumes a bottom-up process, the possibility that object information influences image segmentation is far from a "straw person." Empirical research on higher levels of visual processing suggests that prior knowledge partly guides processing. This suggests that knowledge-based processing may be a general computational principle used at all levels of the visual system. A classic example of this influence of knowledge is the word superiority effect (Reicher, 1969), in which

This research was supported by a Sigma Xi Grants-in-Aid of Research award to Shaun P. Vecera and by ONR grant N00014-91-J1546, NIMH grant R01 MH48274, NIH career development award K04-NS01405, and grant 90-36 from the McDonnell-Pew Program in Cognitive Neuroscience to Martha J. Farah.

the perception of an individual letter is improved when it occurs in the context of a word, as compared to when it appears either in a non-word or alone.

Recent connectionist implementations of image segmentation have also suggested that knowledge can influence segmentation. Mozer and his colleagues (Mozer et al., 1992) have trained a connectionist network to segment images consisting of two overlapping objects. The network discovers grouping principles based on the objects to which it has been exposed, rather than receiving grouping heuristics *a priori*. More traditional approaches to computer vision have also recently emphasized the importance of using knowledge to guide lower level visual processing (Lowe, 1985). These approaches suggest that previously acquired knowledge can influence image segmentation.

Clearly, previous findings suggest that both bottom-up and knowledge-based models of image segmentation are plausible. Surprisingly, there has been no direct attempt to address this issue using the methods of experimental psychology. In the present experiments, we asked which processing strategy the human visual system actually uses in performing image segmentation. In Experiment 1 subjects performed a simple figure/ground segmentation task. Stimuli were simple block shapes, and subjects were asked to determine whether a small probe 'x' fell inside or outside the region defined by the shape. In order to perform this task the subject must determine whether the location of the x is among the locations encompassed by the shape; that is, subjects must group the locations of the region together and determine whether the probed region is among the grouped locations (part of the 'figure') or not among these locations (part of the 'ground').

In order to test between bottom-up and knowledge-based image segmentation models, we manipulated the familiarity of the shapes being segmented. A bottom-up model would predict no effect for the familiarity of the region; the image would be segmented by using properties of the stimulus itself (e.g., good continuation). The familiarity of the shapes should have no effect. However, a knowledge-based model would predict an effect for the region's familiarity. More familiar regions should be segmented faster than less familiar regions. That is, image segmentation would be partially guided by using previous knowledge about the shapes.

Experiment 1

We manipulated the familiarity of stimuli that were to be segmented by presenting upright letters, letters rotated 180°, and non-letter shapes (derived from letters by moving one feature), as shown in Figure 1a. This manipulation allowed us to vary the familiarity of the stimulus, with an upright letter being visually the most familiar. The critical test between bottom-up and knowledge-based segmentation models would be to compare performance between the upright and rotated letter conditions because the bottom-up information (i.e., the information provided to the visual system) is identical in these two conditions. However, the two clearly differ in the degree of visual familiarity. Upright letters are seen much more frequently than rotated letters. If image segmentation can be guided by previous knowledge of shapes, then we would expect upright letters to be segmented faster than rotated letters. In contrast, if segmentation is a bottom-up process, then there should be no effect of familiarity. Finally, it should be noted that comparisons to the non-letter shapes are somewhat ambiguous because the bottom-up information has not been controlled.

Method

Subjects. Subjects were 12 Carnegie Mellon University undergraduates and staff. All had normal or corrected vision and were native English speakers.

Stimuli. Stimuli consisted of 12 block shapes, six letter stimuli and six non-letter stimuli (Figure 1a). The letters used were uppercase A, F, K, L, T, and Y. Subjects viewed the stimuli from a distance of approximately 60 cm. All shapes were 3.8 cm wide and 5.0 cm tall. Non-letter shapes were created by altering the relationships among the features of the letters. Letters were presented in their upright orientation condition and in a 180° rotated orientation condition; the non-letters were presented in only one orientation.

Subjects' task was to determine whether a small probe 'x' fell inside or outside the region bounded by the block shape. The x appeared in a 12 point, bold Helvetica font and was the

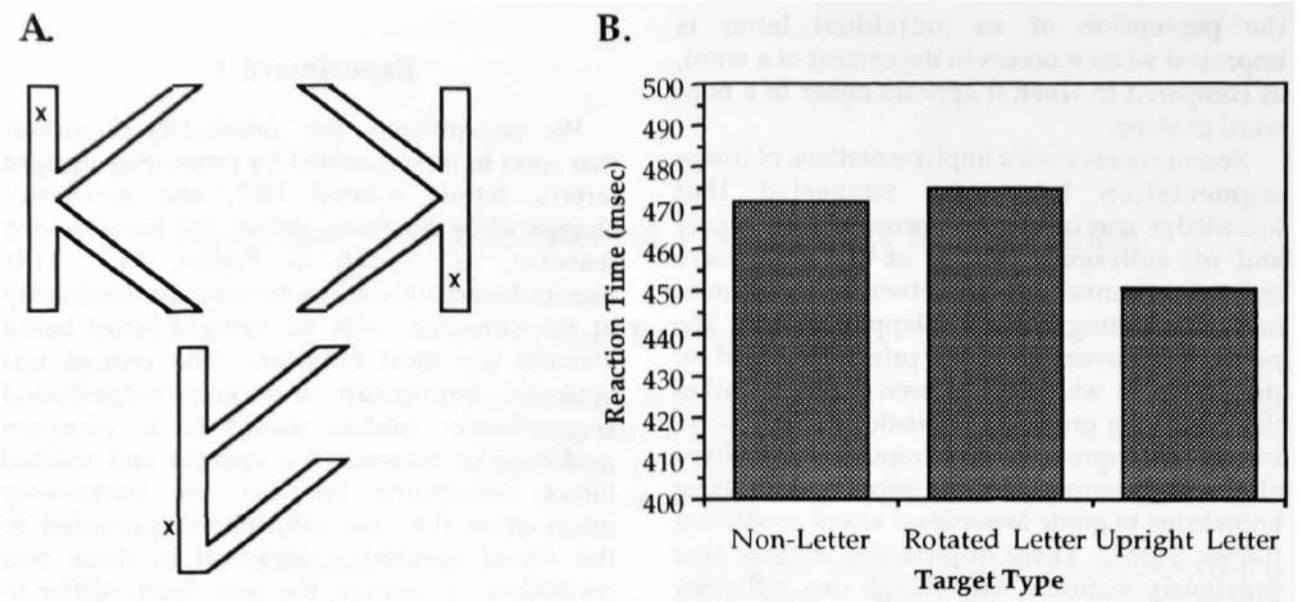

Figure 1. (A) Examples of stimuli used in Experiment 1. (B) Results of Experiment 1. Subjects are fastest to determine the position of the 'x' when it is either inside or outside an upright letter.

same distance from the edge of the shape whether it was inside or outside the shapes. The x's were in approximately the same spatial location whether they were inside or outside.

Procedure. Stimuli were presented via a Macintosh Plus computer. Each subject received six blocked presentations; target type remained constant within a block. Subjects were told the target type of each block in advance. There were 48 individual trials within each block, 24 with the x falling inside the shape and 24 with the x falling outside the shape. An individual trial began with five asterisks appearing on the screen in a plus (+) pattern. Subjects initiated a trial by pressing the space bar; the shape and the small x were then simultaneously flashed for 100 msec. The screen was then blank while the subject responded via a keypress.

Results and Discussion

Only correct reaction times were used in the analyses. Reaction times over 1500 msec and under 100 msec were also excluded. Subjects' median reaction times for each condition were analyzed with a two-factor analysis of variance (target type by 'x' location).

The mean reaction times for upright, rotated, and non-letter shapes appear in the Figure 1b. For reaction times, the main effect for target type was significant, $F(2, 22) = 4.869$, $p < 0.02$, and the main effect for inside versus outside was marginally significant, $F(1, 11) = 4.6$, $p < 0.06$, with "inside" responses being faster than "outside" responses. There was no interaction, $F(2, 22) = 0.037$, $p > 0.50$. Planned pairwise comparisons between the target type factor level means resulted in significant differences between upright letter and non-letter targets, $t(22) = 6.121$, $p < 0.03$, and between upright letter and rotated letter targets, $t(22) = 8.317$, $p < 0.01$. There was no statistically significant difference between the non-letter targets and the rotated letter targets, $t(22) = 0.168$, $p > 0.50$. A correlation between reaction time and percent correct suggested that subjects did not sacrifice accuracy for speed, $r = 0.162$.

These results are consistent with predictions made by a knowledge-based image segmentation model. Namely, the familiarity of a shape in the visual field, in this case a familiar letter, allows the visual field to be segmented into figure and ground more rapidly than when the shape is a less familiar rotated letter or a non-letter.

Although these data argue for knowledge-based segmentation, two points warrant brief discussion. First, one might wonder why there is not a significant response time superiority of rotated letters over non-letters in this task. If image segmentation is knowledge-based, then shouldn't we expect an advantage for the rotated

letters over the non-letters? As mentioned earlier, comparisons between the non-letter stimuli and the letter stimuli are ambiguous because of the fact that visual complexity has not been controlled. The non-letter shapes may be visually less complex than the set of letters used in this experiment. Alternatively, the non-letters may be sufficiently similar to the letters that they do partially activate letter representations.

Second, although these data are consistent with knowledge-based image segmentation, an alternative explanation exists. These results may be due to subjects performing, in effect, two tasks: image segmentation and recognition of the shape being recognized. If recognition is faster for upright letters than it is for rotated letters, and if segmentation speed is influenced by the speed of the recognition process, then even if segmentation is bottom-up, the pattern of results in Experiment 1 would be expected. This pattern of results would be due to a kind of dual task interference from the recognition process. However, there are empirical data that suggest dual task interference is not the case. Corballis and his colleagues have shown that the latency to name a letter is generally independent of the angular rotation of the letter (Corballis et al., 1978). Finally, we conducted an additional experiment (not reported here due to space limitations) that demonstrated naming a letter does not influence segmentation.

Experiment 2

The first experiment established that figure/ground segregation is influenced by the familiarity of the object being segmented. However, figure/ground segregation is only one image segmentation paradigm. In Experiment 2 subjects observed two overlapping transparent shapes, as shown in Figure 2a. The stimuli were the non-letter shapes, rotated letters, and upright letters used in Experiment 1. Two small x's appeared on the stimuli and could either be on the same shape or on different shapes; subjects determined whether the x's were on the same shape or on different shapes.

Again, bottom-up and knowledge-based models of image segmentation make differing predictions as to subjects' performance. A bottom-up model of image segmentation would again predict no effect for the familiarity of the shape that was to be segmented. Segmentation would only operate on the data provided in the image, perhaps according to Gestalt laws or other grouping heuristics. A knowledge-based image segmentation model would predict that subjects could use their knowledge of objects in order to help guide segmentation.

Method

Subjects. Sixteen Carnegie Mellon University staff and students served as subjects. All were native English speakers and had normal or corrected vision.

Stimuli. The same block letters used in Experiment 1 were used. All possible letter pairs were formed and superimposed on each other for a total of 15 stimuli. The overlapping letters were on average 4.43 cm wide and 5.21 cm tall. The overlapping non-letters were on average 4.95 cm wide and 5.85 cm tall. Shapes were initially overlapped randomly, but the stimuli remained the same for all subjects. The overlapping was performed such that (a) two x's could be fit onto the letters and (b) the display was visually not too complex.

Half of the time the x's were on a single shape ("same" condition) and half of the time one x appeared on each shape ("different" condition). The x's appeared in a 12 point bold Helvetica font and were the same distance from each other in the "same" and "different" conditions. The superimposed letters were presented in their upright orientation and in a 180° rotated orientation. These stimuli were identical except for the rotation. Non-letters again only appeared in one orientation.

Procedure. Stimuli were presented via a Macintosh Plus computer. Each subject received nine blocked presentations; the target type was constant within a block. Prior to each block, subjects were told the target type of that block, but they were told that they could ignore the target type and that they should focus on determining whether the x's were on the same shape or on different shapes.

There were 60 individual trials within each block, 30 with the x's falling on the same shape and 30 with the x's falling on different shapes. Individual trials began with a fixation of five asterisks appearing on the screen in a plus (+) pattern. Subjects started a trial by pressing the space bar. The shapes and x's were then simultaneously flashed for 200 msec. The screen

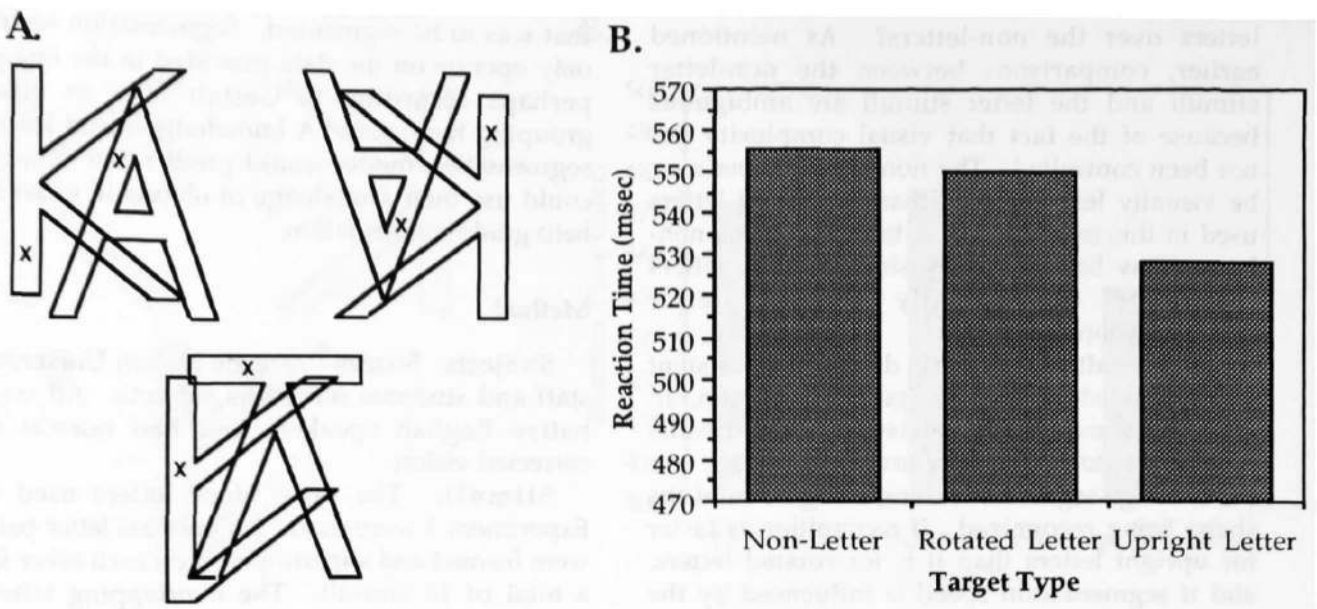

Figure 2. (A) Examples of stimuli used in Experiment 2. (B) Results of Experiment 2. Subjects are fastest to determine location of the two x's when the shapes are upright letters.

was then blank while the subject responded via a keypress.

Results and Discussion

Only correct reaction times were used in the analyses. Responses faster than 100 msecs or slower than 2000 msecs were excluded. Subjects' median reaction times for each condition were analyzed with a repeated measures two-factor analysis of variance (target type by 'x' location). Planned pairwise comparisons were conducted between the target type factor means.

The mean reaction times for upright, rotated, and non-letter shapes appear in Figure 2b. As before, knowledge-based image segmentation predicts a main effect for target type, which was significant, $F(2, 30) = 5.18$, $p < 0.02$, as was the main effect for x location, $F(1, 15) = 6.32$, $p < 0.03$, with "same" responses being faster than "different" responses. The interaction was not significant, $F(2, 30) = 2.81$, $p > 0.08$. Planned pairwise comparisons were performed between the target type factor means. The reaction times to upright letters were significantly faster than reaction times to non-letters, $t(30) = 3.03$, $p < 0.005$. Reaction times to upright letters were also significantly faster than reaction times to rotated letters, $t(30) = 2.44$, $p < 0.02$. There was no significant difference between reaction times to non-letter targets and rotated letter targets, $t(30) = 0.59$, $p > 0.40$.

The error data were also analyzed with a two factor repeated measure ANOVA. The pattern of results was similar to that observed in the reaction time data. The main effect for target type was significant, $F(2, 30) = 3.99$, $p < 0.03$, as was the main effect for x location, $F(1, 15) = 4.22$, $p < 0.001$, with "different" responses being more accurate than "same" responses. The interaction was not significant, $F(2, 30) = 0.39$, $p > 0.50$. Planned pairwise comparisons on the target type factor level means revealed significant differences between the non-letters and upright letters, $t(30) = 2.78$, $p < 0.01$, and the rotated letters and upright letters, $t(30) = 1.92$, $p < 0.06$. There was no significant difference between the non-letter shapes and rotated letters, $t(30) = 0.81$, $p > 0.40$.

These results are consistent with those from Experiment 1, suggesting that visual image segmentation can be guided in part by knowledge of the shapes being segmented. In the second experiment, subjects are fastest to respond to whether the two locations are on the same shape or on different shapes when the shapes are familiar letters, as opposed to rotated letters or non-letters.

Conclusions

The results of these two experiments, taken together, suggest that image segmentation is a knowledge-based process: Knowledge about the

shapes being segmented can partially guide the segmentation process. However, an unresolved issue concerns the *locus* of these familiarity effects. That is, where is this knowledge of objects coming from? Specifically, are these effects due to top-down influences from internally stored visual memories, or are they due to a processing advantage within the segmentation stage itself?

While we do not have an answer to this question at present, we have previously argued for an interaction between image segmentation and internally stored visual object representations (Vecera & Farah, 1992). This model, shown in Figure 3, suggests that the knowledge effects observed above are the result of cascaded processing (McClelland, 1979): Preliminary results of partial processing at an earlier stage are available to the next stage, and feedback from a later stage in turn guides processing in this earlier stage. Specifically, as the image is segmented, activation is sent to object-level representations stored in visual memory. As matches are made with these object representations (as presumably happens with upright letters), these representations send activation back to the image segmentation stage (the "grouped array" in Figure 3). This top-down activation reinforces groupings that correspond to familiar objects, allowing segmentation to finish faster than it would if there were no such top-down activation.

However, the alternative model in which knowledge is implemented in the segmentation process itself is also plausible. This model has been suggested by Mozer and his colleagues (Mozer et al., 1992). Mozer's model implements knowledge in a connectionist network that is trained to segment images. In this scheme, familiarity effects might be expected on the basis of low-level image statistics, such as familiar combinations of features, not partial matches to objects in visual memory.

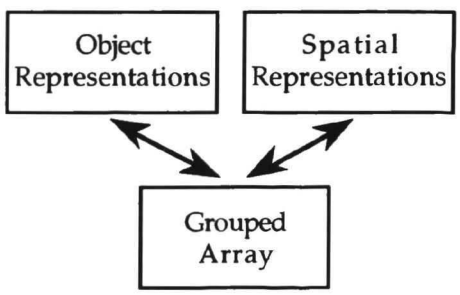

Figure 3. Proposed model of image segmentation. Adapted from Farah (1990).

Although the locus of knowledge remains unresolved, the results of experiments in progress will allow us to distinguish between the two alternatives and to better understand the computational architecture of the visual system.

Acknowledgments

The author wishes to thank Martha Farah for help during all phases of this research. Thanks to Jay McClelland, Mike Mozer, and David Plaut for comments on these experiments. Thanks also to Karen Klein for her technical assistance.

References

Corballis, M. C., Zbrodoff, N. J., Shetzer, L. I., & Butler, P. B. (1978). Decisions about identity and orientation of rotated letters and digits. *Memory and Cognition, 6*, 98-107.

Farah, M. J. (1990). *Visual agnosia.* Cambridge, MA: MIT.

Lowe, D. G. (1985). *Perceptual organization and visual recognition.* Boston: Kluwer.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86*, 287-330.

Mozer, M. C., Zemel, R. S., Behrmann, M., & Williams, C. K. I. (1992). Learning to segment images using dynamic feature binding. *Neural Computation, 4*, 650-665.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 274-280.

Rosenfeld, A. (1984). Image analysis: Problems, progress and prospects. *Pattern Recognition, 17*, 3-12.

Vecera, S. P., & Farah, M. J. (1992). Is visual image segmentation a bottom-up or an interactive process? Manuscript submitted for publication.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. *Psychologische Forschung, 4*, 310-350. [Reprinted in part as Principles of perceptual organization. In D.C. Beardslee & M. Wertheimer (Eds.), *Readings in Perception* (pp. 115-135). Princeton, NJ: Van Nostrand Reinhold, 1958.]