Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 7354–7358, October 1987
Psychology

Discriminating figure from ground: The role of edge detection and region growing

(vision/computer vision/segmentation/parsing algorithms)

D. MUMFORD†, S. M. KOSSLYN‡, L. A. HILLGER‡, AND R. J. HERRNSTEIN‡

Departments of †Mathematics and ‡Psychology, Harvard University, Cambridge, MA 02138

Contributed by D. Mumford, June 11, 1987

ABSTRACT **Three general classes of algorithms have been proposed for figure/ground segregation. One class attempts to delineate figures by searching for edges, whereas another class attempts to grow homogeneous regions; the third class consists of hybrid algorithms, which combine both procedures in various ways. The experiment reported here demonstrated that humans use a hybrid algorithm that makes use of both kinds of processes simultaneously and interactively. This conclusion follows from the patterns of response times observed when humans tried to recognize degraded polygons. By blurring the edges, the edge-detection process was selectively impaired, and by imposing noise over the figure and background, the region-growing process was selectively impaired. By varying the amounts of both sorts of degradation independently, the interaction between the two processes was observed.**

One of the fundamental purposes of vision is to allow us to recognize objects. Recognition occurs when sensory input accesses the appropriate memory representations, which allows one to know more about the stimulus than is apparent in the immediate input (e.g., its name). Before visual input can be compared to previously stored information, the regions of the image likely to correspond to a figure must be segregated from those comprising the background. The initial input from the eyes is in many ways like a bit-map image in a computer, with only local properties being represented by the activity of individual cells; only after the input is organized into larger groups, which are likely to correspond to objects and parts thereof, can it be encoded into memory and compared to stored representations of shape. Thus, understanding of the processes that segregate figure from ground is of fundamental importance for understanding the nature of perception.

Researchers in computer vision have been faced with the problems of segregating figure from ground, and in this report we explore whether the human brain uses some of the algorithms they have developed. In computer vision, the input is a large intensity array, with a number representing the intensity of light at each point in the display. Two broad classes of algorithms have been devised to organize this welter of input into regions likely to correspond to objects. One class contains *edge-based* algorithms (1–3). These algorithms look first for sharp changes in intensity (i.e., maxima in first derivatives or zero crossings in the second derivative of the function relating intensity to position), which are assumed to correspond to edges. In the Marr–Hildreth theory (3), these changes are observed at multiple scales of resolution and, if present at each, are taken to indicate edges (and not texture or the like). The local points of sharp change are connected, resulting in a depiction of edges that are assembled into the outlines of objects. The other class contains the so-called *region-based* algorithms (4–7). These algorithms construct regions by growing and splitting areas that are maximally homogeneous; they compute not derivatives of intensity but rather homogeneity measures, such as intensity variance. In short, the first algorithm tries to delineate regions by discovering edges, whereas the second delineates edges by discovering regions.

Investigations of the neurophysiology of vision provide strong evidence that mammalian brains use algorithms in the first class. Hubel and Wiesel's (8) "simple cells" in striate cortex seem to be part of an implementation of an edge-based algorithm (compare ref. 9). These cells detect sharp changes in intensity. However, both the linking of local points of sharp change into larger edges and the growing of regions are processes that require a more global organization of the image. Recent work (10) suggests that some such global processes are carried out in area V2, but the findings do not indicate clearly which algorithm is implemented here.

The experiment reported here uses a psychological approach to investigate whether one or both of these algorithms better models the way humans segregate figure from ground. This experiment was designed to discriminate among six alternative hypotheses: the human brain organizes visual input solely by an edge-based algorithm; solely by a region-based algorithm; by whichever algorithm is successful most quickly; by neither algorithm; by both algorithms, with one following the other; or by using both algorithms simultaneously and interactively. In addition, it provides numerical evidence for evaluating various models of simultaneous functioning of the two algorithms.

In this experiment, subjects were asked to judge whether light polygons on a dark background were the same or different from a target shape. Holding constant the average intensities inside and outside the figures, the edges of the test stimuli were blurred to a greater or lesser degree, and the amount of variability in the intensity of the points composing the figure and ground was varied by superimposing noise to a greater or lesser degree. If the brain parses using edge detection, then the sharpness of the gradient from ground to figure should be critical, with greater blur resulting in more time and errors. Similarly, if the brain uses region growing, then the overlap in intensity variability between figure and ground should be critical, with greater overlap resulting in more time and errors. Finally, different forms of interactions between the two variables will indicate whether the two algorithms are used independently or interactively.

In the design of this experiment, we were aware that very large amounts of superimposed variability begin to introduce spurious irregular edges all over the stimulus, and very large amounts of blur wipe out the shape of the region. However, these are second-order effects: provided that the noise and blur is not too extreme, properly aligned simple-cell-type edge detectors will respond equally strongly to a sharp edge with or without superimposed noise and weakly to the noise alone. Similarly, with a blurred edge of limited width w,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

region-growing algorithms will immediately group the parts of the figure and ground away from the edge by distance w.

The stimuli consisted of nine simple geometric shapes, such as a triangle and a diamond. The stimuli were initially computed as 512 × 512 images on a VAX computer and displayed on a AED color graphics monitor. The polygons were normalized to a perimeter of 950 pixels (hence varying in area) and centered on the screen. Letting 0 represent black, and 1 represent the brightest output of the monitor, the mean intensity of the interior of the figures was always 0.7, whereas the mean intensity of the ground was always 0.3. The edges were blurred by convolving the image with four Gaussian filters, $g(i)$, with spatial standard deviations of 0, 4, 8, and 12 pixels. A noise signal n was computed by using a Fourier series with independent normally distributed random coefficients $a(i, j)$ such that

$$E(a(i, j)) = 0$$

and

$$E(|a(i, j)|^2) = (i^2 + j^2)^{-1} \quad \text{if} \quad (i, j) \neq (0, 0),$$

where E is the expectation.

The stimuli were made by adding a multiple $l(j)\cdot n$ of the noise to the blurred polygon signal $g(i)*p$ (where p stands for polygon) and passing this through a sigmoidal function to keep the blacks and whites within the range (0 to 1) of the screen.§ The multiples were chosen so that

$$\frac{\|l(j)\cdot n\|}{\|p - \bar{p}\|} = 0, 0.5, 1.0, 1.5.$$

Here $\bar{p}$ is the mean (0.5) of the signal p, and $\|f\|$ represents the strength of the signal f:

$$\|f\| = (\Sigma f(i, j)^2)^{1/2}.$$

Because the mean intensity of figure and ground was kept constant in all conditions, increasing the range of variability had the effect of increasing the amount of overlap in intensity values for figure and ground. All combinations of blur and variability were used, resulting in 16 versions of each of the 9 stimuli. Thus, a total of 144 stimuli were generated. Examples of the stimuli are presented in Fig. 1.

The stimuli were photographed to produce slides. Ektachrome 100 slide film was used at a distance 0.84 m from the computer screen in a darkened room using an exposure and aperture of 0.5 sec and f11.0, respectively. The AED screen was black and white, but color film was used to capture the gray shades produced by the noise and blurriness. Pilot work was done to approximate a linear increase in the subjective impression of the increase in successive levels of blur and variability.

Eighteen adults volunteered to participate as subjects. Written instructions describing the experimental procedure were given to the subjects and then were reviewed orally by the experimenter. The subjects sat about 1 m from a translucent screen, with the slides of the figures being back-projected onto the screen so that the polygons subtended approximately 5° from the vantage point of the subject. (Extending the stimuli into the near periphery was necessary if subjects were to be able to see the figures clearly enough to recognize them even in the highly degraded conditions. Given the question being asked here, it is not necessary that the stimuli be confined to the fovea.) The room was darkened to facilitate slide viewing. The experiment was divided into nine blocks of trials, with a different polygon being used as the target in each block. At the beginning of each block the subjects were asked to remember the shape of a target polygon, which was presented with no blur and no variability. The subjects were told to examine the target polygon until they could make and maintain an accurate mental image of it; this procedure was employed to aid memory. They then were shown a series of slides, with each being exposed for 2000 msec or until the subject responded, whichever came first. The subjects were told in advance that some of the polygons would be blurred and/or have visual noise over them.

In each block the subjects were shown a series of 32 slides, 16 of which were the target polygon and 16 of which were different polygons. One target and one distractor was shown at each combination of each level of blurriness and variability; each nontarget polygon appeared twice with each target. Each version of each polygon was shown once as a target and once as a distractor throughout the experiment. The order of presentation of the stimuli within each block was randomized, subject to the constraint that no more than two targets or two distractors occurred in a row. The order of the blocks was different for each subject.

The subjects were asked to press the button labeled "same" if the figure shown was the same shape as the target, and the button labeled "different" if it was not. The subjects were told to respond as quickly as possible while remaining as accurate as possible. Each hand rested on a response button, and half of the subjects used the dominant hand to respond "same" and the nondominant hand to respond "different," and vice versa for the other half; thus, hand of response was counterbalanced with response, removing the possible effects of handedness from the results. The stimuli were presented by two random-access projectors, which were controlled by an Apple II+ computer; this computer also recorded the subjects' response times and decisions.

The data were analyzed as follows. First, the mean of the response times was computed for each subject, considering only trials on which the correct judgment was made. These means ranged from 490 to 1009 msec, with a grand mean of 678 msec. Each subject's times were then scaled to make their mean equal to the grand mean. Following this, response times that were greater than 2.5 times the mean for the remaining times in that cell (i.e., combination of level of blur, noise, and response) were discarded as outliers. This procedure resulted in 1.4% of the response times being trimmed. The mean times for each combination of blur and noise level were then considered in an analysis of variance.

The results allow us to discriminate unequivocably among the various alternative classes of algorithms. As is illustrated in Fig. 2, response times increased progressively as edge blur [$F(3,51) = 56.51, P < 0.0001$] and variability [$F(3,51) = 50.42, P < 0.0001$] increased. Note that the points in Fig. 2 generally are upwardly concave; however, the variability and blur scales were not designed to be precisely psychophysically linear. But more interestingly, the two variables interacted: the effects of increased blur were exacerbated by increased variability [$F(9,153) = 12.38, P < 0.00001$, for the interaction of blur and variability]. These results are what we would expect if both algorithms are at work and mutually interact during processing. In addition, there was a marginal tendency for subjects to respond more quickly for "same" figures than "different" ones [$F(1,17) = 3.17, P < 0.1$] and the effects of variability were more pronounced for "same" judgments [$F(3,51) = 4.30, P < 0.01$]. Finally, there was a three-way interaction between variability, blur, and response type, indicating that the interaction between variability and blur was more pronounced for "same" judgments [$F(9,153) =$

§We sought a self-scaling noise that obscured large- and small-scale features to an "equal" degree. White noise concentrates its power in high frequencies and is largely eliminated by low-pass filtering; "$1/f$" noise creates large-scale features that would compete with the shape of a polygon in figure/ground separation, even at low noise levels. The above power law is halfway between these two.

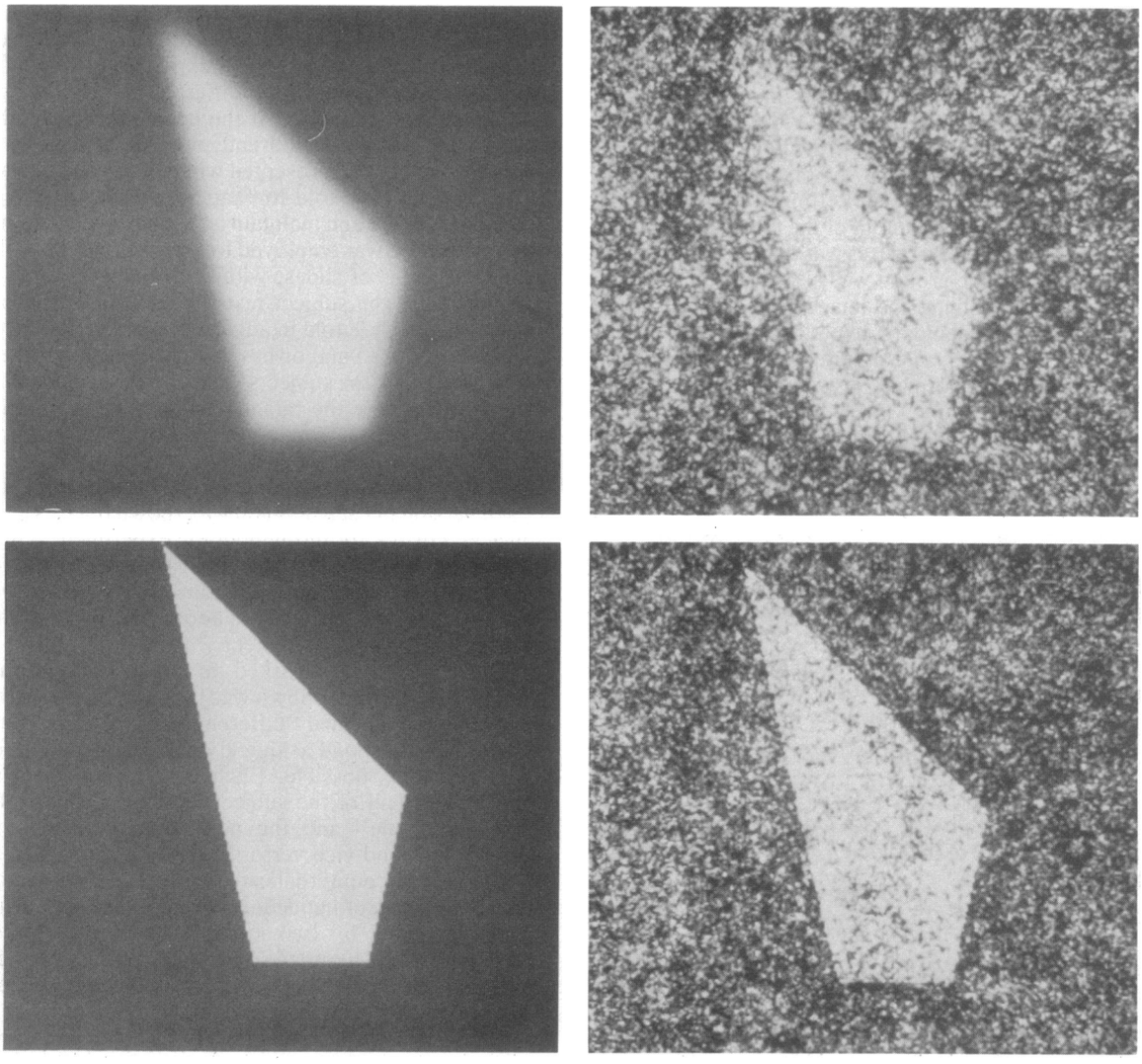

FIG. 1. Examples of test stimuli that had to be compared to a memorized standard. Values of blur and variability are 0,0 (*Lower Left*), 0,2 (*Lower Right*), 2,2 (*Upper Right*), and 2,0 (*Upper Left*). Four levels of blur and variability were used, with the additional two levels roughly dividing the scale between 0 and 3 into two subjectively equal increments. Note that the addition of noise seems to blur the edge, which is due to masking of high spatial frequencies; if the slides are blurred, thereby filtering out high spatial frequencies, the edges appear equally sharp in the 0,0 and 0,2 cases. The subjective impression of a blurred edge in the 0,2 case can be taken as further evidence that region-growing processes are used in figure/ground segregation.

2.44, $P < 0.02$]; this interaction probably reflects the fact that it was relatively easy to evaluate "different" stimuli that were very dissimilar to the target, hence they need not be processed as thoroughly.

Error rates are another reflection of processing, and we also submitted these data to an analysis of variance. The percent errors for each cell are presented in Fig. 3. Errors increased with increases in edge blur [$F(3,51) = 12.76$, $P < 0.0001$] and variability [$F(3,51) = 19.66$, $P < 0.0001$]; however, as is evident, most of these effects are captured by the upper right four cells of Fig. 3. As before, the two variables interacted: the effects of increased blur were exacerbated by increased variability [$F(9,153) = 11.86$, $P < 0.0001$, for the interaction of blur and variability]. In addition, subjects committed more errors for "same" figures than "different" ones [$F(1,17) = 5.95$, $P < 0.05$], and the effects of blur and of variability were more pronounced for "same" judgments [$F(3,51) = 5.46$, $P < 0.01$, and $F(3,51) = 5.57$, $P < 0.01$, respectively, for the interaction of each variable with response type]. Finally, the interaction between variability and blur was more pronounced for "same" judgments [$F(9,153) = 5.49$, $P < 0.001$]. These results, then, dovetail nicely with those from the response times, with increases in times and errors both reflecting increases in underlying difficulty of processing; from inspection of Figs. 2 and 3, there is no hint of speed/accuracy trade-offs.

We also attempted to fit each class of model to the 4×4 table of mean response times (in these analyses, times from the two responses were pooled to decrease the noise before eliminating outliers). We modeled each type of algorithm in the following way.¶

First, the simplest algorithms, positing only an edge-based process or only a region-based process, were modeled by arbitrary functions of one of the variables, with four parameters in each model. The best function of blur accounted for only 35% of the variance, and the best function of variability accounted for 37% of the variance.

Second, the algorithm in which there are two completely

¶These numerical models are not to be taken strictly like laws of physics, but rather as formulae that make concrete the possible qualitatively different interactions of the variables. Numerous specific formulations are possible within each qualitative class; we have taken the most straightforward examples we could find.

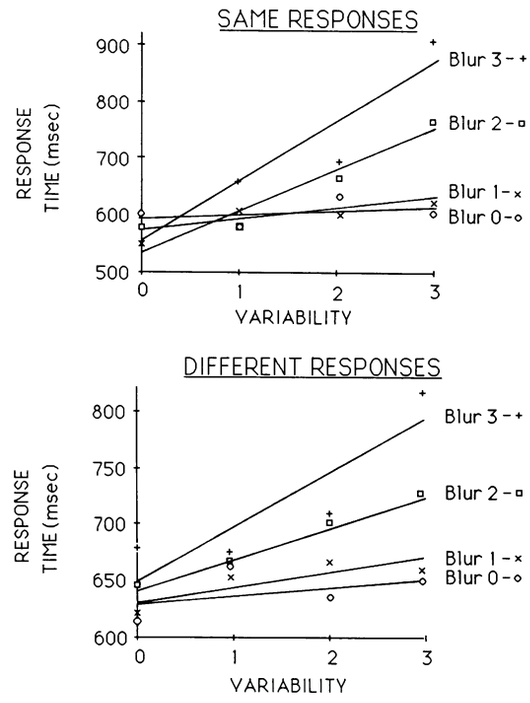

FIG. 2. The time to make "same" or "different" judgments to geometric forms that were degraded in two possible ways, by blurring the edge or by adding variability to the values of the figure and ground. Linear functions that best fit the points were computed using the least-squares method and are illustrated here. Precise values of the levels of blur and variability are provided in the text.

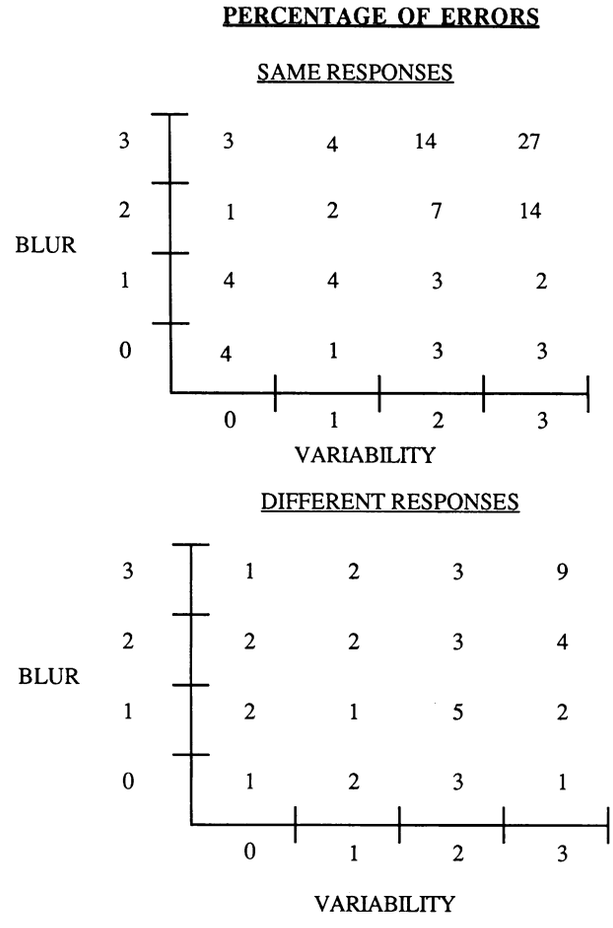

FIG. 3. Percent errors for the 16 presentation conditions, separately for "same" and "different" trials.

independent processes, with only the output from the fastest process being used, was modeled by $\min(a + b * \text{blur}, c + d * \text{variability})$. This model accounted for only 76% of the variance with four parameters.

Third, the algorithm in which both processes are used, but one follows the other, was modeled by the sum of $a + (b * \text{blur}) + (c * \text{variability})$. A model of this kind would follow, for instance, by the most narrow interpretation of the neurophysiological architecture, with area 17 acting as a edge-detection module and later visual areas acting as region growers. This model accounted for only 66% of the variance. Note also that the interaction between blur and variability observed in the analysis of variance serves to rule out this class of models, which predicts strictly additive functions of the two variables (11).

Fourth, the class of algorithm in which both processes are used simultaneously and interactively was divided into two subclasses. The most common subclass is a "feature plus blackboard" algorithm (e.g., ref. 12). In this process, an edge-based module and a region-based module independently post features on a single "blackboard"; the rate at which features are posted decreases linearly with increases in blur or variability (depending on the module). A decision is reached whenever the total number of features reaches a threshold. This algorithm was modeled by a harmonic mean of two linear functions, $a + [(b + c * \text{blur})^{-1} + (d + e * \text{variability})^{-1}]^{-1}$. (We restricted this model, and the previous ones, to the linear case in order to equate roughly the number of parameters in each of the models.) This five-parameter model accounted for only 72% of the variance in the data.

Finally, we considered a second subclass of the simultaneous interactive algorithms, which posits active processing in the "blackboard," not simple accumulation of features. In this model, a low-level feature-detection module operates in parallel on the whole image; this module reports, in constant time, local features, such as edge elements, blobs and bars, to a buffer [i.e., the structure in which Marr's *primal sketch* occurs (9)]. The problem is to organize these features into figure and ground. We assume that the local segments of a polygon's edges are reported with a strength decreasing with blur [strength of edge $= (a + b * \text{blur})^{-1}$] and that the variability produces extraneous features, such as small blobs and bars, that do not correspond to edge segments. The strength of extraneous features can be represented as $c * \text{variability}$. Then we assume that a combined edge-region algorithm finds the optimal figure/ground segregation in the buffer. This algorithm relies upon (*i*) the relative absence of distinguishing features in the interior and (*ii*) the coherence of the local edge elements surrounding the figure. Both sorts of information are used simultaneously, and the optimal figure/ground segregation is achieved by satisfying both sorts of constraints simultaneously. The time this process takes increases from some minimal time with the number of extraneous features but decreases with increasing strength of the edge elements. The simplest possible representation of this is $[(d + c * \text{variability})/(a + b * \text{blur})^{-1}]$. This gives us a model of response time with a bilinear function $a' + b' * \text{blur} + c' * \text{variability} + d' * \text{blur} * \text{variability}$. This four-parameter model accounts for 85% of the variance in the data.

Our conclusion is that the algorithm humans use to segregate figure from ground involves an interplay between the one-dimensional information given by edge-based processes and the two-dimensional information given by region-based processes. Some such hybrid algorithms have been proposed recently (e.g., refs. 13–17), but it is not clear whether these algorithms would predict the bilinear pattern observed here.

It is of interest to compare our results to those of Uttal (18), who has examined the detection of dot figures in the presence

of spurious dots. His results show that the presence of collinear dots that suggest extended lines in a figure was the principal factor that enabled observers to discern a figure in spite of noise. Athough Uttal interpreted this result as an argument for the use of the autocorrelation function in visual processing, it can also be interpreted as indicating that when region-based algorithms are disrupted by noise, edge detection-based algorithms become crucial, and these algorithms pick out collinear dots (e.g., even a simple cell with a central excitatory strip and inhibiting flanks would pick out such dots).

The present results lead to a clear prediction that can be tested by single-cell recordings—namely, that areas such a V2 should contain cells that respond to nonlocal configurations underlying region-growing and edge-grouping processes. The response of a cell involved in region growing should be influenced by the extent and shape of the region to which it belongs following segmentation, an area likely to extend outside the classical "receptive field" of the cell. For instance, one might seek cells that respond only when a curve outside the receptor field completely or nearly surrounds the field. It would be surprising if there were not direct neural correlates to the behavioral results found here.

We thank M. Van Kleeck for valuable observations and discussion. This work was supported by National Science Foundation Grant IST-8511606 and Office of Naval Research Contract N00014-85-K-0291.

1. Roberts, L. G. (1965) in *Optical and Electro-optical Information Processing*, ed. Tippett, J. P. (MIT Press, Cambridge, MA), pp. 159–197.
2. Canny, J. (1986) *IEEE Trans. Patterns Anal. Mach. Intel.* **8,** 679–698.
3. Marr, D. & Hildreth, E. (1980) *Proc. R. Soc. London B* **207,** 187–217.
4. Horowitz, S. & Pavlidis, T. (1974) *Proc. Int. Joint Conf. Pattern Recognition* **2,** 424–433.
5. Ohlander, R., Price, K. & Reddy, R. (1979) *Comput. Graph. Image Process.* **8,** 3.
6. Burt, P. J., Hong, T. H. & Rosenfeld, A. (1981) *IEEE Trans. Systems Man Cybern.* **12,** 802–809.
7. Haralick, R. M. & Shapiro, L. G. (1985) *Comp. Vision Graph. Image Process.* **29,** 100–132.
8. Hubel, D. & Wiesel, T. (1968) *J. Physiol. (London)* **195,** 215–243.
9. Marr, D. (1982) *Vision* (Freeman, San Francisco), chap. 2.
10. Von der Heydt, R., Peterhaus, E. & Baumgartner, G. (1984) *Science* **224,** 1260–1262.
11. Sternberg, S. (1969) *Acta Psychol.* **30,** 276–315.
12. Lindsay, P. & Norman, D. (1979) *Human Information Processing* (Freeman, San Francisco).
13. Geman, S. & Geman, D. (1984) *IEEE Trans. Patterns Anal. Mach. Intel.* **6,** 721–741.
14. Grossberg, S. & Mingolla, E. (1985) *Percept. Psychophys.* **38,** 141–171.
15. Grimson, W. E. L. & Pavlidis, T. (1985) *Comput. Vision Graph. Image Process.* **30,** 316–330.
16. Mumford, D. & Shah, J., in *Image Understanding 1986*, eds. Ullman, S. & Richards, W. (MIT Press, Cambridge, MA), in press.
17. Sejnowski, T. & Hinton, G., in *Vision, Brain, and Cooperative Computation*, eds. Arbib, M. & Hanson, A. R. (MIT Press, Cambridge, MA), in press.
18. Uttal, W. (1975) *An Autocorrelation Theory of Form Detection* (Erlbaum, Hillsdale, NJ).