PERCEPTUAL ORGANIZATION AS A BASIS FOR VISUAL RECOGNITION

David G. Lowe and Thomas O. Binford

Computer Science Department
Stanford University, Stanford, California 94305

Abstract

Evidence is presented showing that bottom-up grouping of image features is usually prerequisite to the recognition and interpretation of images. We describe three functions of these groupings: 1) segmentation, 2) three-dimensional interpretation, and 3) stable descriptions for accessing object models. Several principles are hypothesized for determining which image relations should be formed: relations are significant to the extent that they are unlikely to have arisen by accident from the surrounding distribution of features, relations can only be formed where there are few alternatives within the same proximity, and relations must be based on properties which are invariant over a range of imaging conditions. Using these principles we develop an algorithm for curve segmentation which detects significant structure at multiple resolutions, including the linking of segments on the basis of curvilinearity. The algorithm is able to detect structures which no single-resolution algorithm could detect. Its performance is demonstrated on synthetic and natural image data.

Introduction

A major goal of computer vision research is to relate visual images to prior knowledge of their constituents, and thereby label and interpret them. However, current model-based vision systems have been demonstrated only in tightly-constrained environments with a few well-specified models to compare to the image [2, 8, 11]. The difficulty in expanding performance to more general domains is not one of ambiguity—it is very unlikely that two different models will fully fit the same image data. Rather, the problem is one of searching for potential correspondences between models and the image, since increasing the number and generality of the models results in an excessively large space of possible matches. Continued research into recovering three-dimensional shape from images—using stereo, motion, shading, and texture—promises to reduce the size of this search space considerably. However, the problem of matching is far from solved even when given full three-dimensional information, and these methods fail to explain the excellent level of human performance in such simple domains as line drawings.

In order to interpret images about which we have little prior knowledge, it is necessary to use effective bottom-up techniques to structure and describe the image in a form that can be used to selectively index into a large body of world knowledge. In this paper we will describe methods for detecting and evaluating the significance of relations between image elements in a way that can be applied uniformly to all images before we have any knowledge of their contents. Previous research on this and related topics has gone under such names as image segmentation, perceptual organization, figure/ground phenomena, texture description, and Gestalt perception. There have been many efforts to develop algorithms for specific segmentation problems, such as the detection of collinearity or connectivity, but these have not been integrated and have often lacked general applicability. Marr's initial primal sketch formulation [7] was intended to make some of these relations explicit, but this aspect of it was never fully developed. Recently, Witkin and Tenenbaum [12] have argued for the importance of detecting regularities and imposing structure on the image for many of the same reasons given here. They describe a unified treatment of inference based on the assumption that regularities detected in the image are non-accidental. In this paper we will describe the role that this form of inference plays in model-based recognition, develop some underlying principles for this level of interpretation, and present new segmentation methods based upon these principles.

There are three valuable sources of information which the bottom-up organization of image features can provide, all of which simplify the problem of matching against world knowledge:

1) A major reduction in the search space is achieved by segmentation—the division of the image into sets of related features. This has long been recognized as a crucial problem in image interpretation. We do not want to match models against all possible combinations of features in an image, so good segmentation is crucial for reducing the combinatorics of this search.
2) Two-dimensional relations lead to specific three-dimensional interpretations, as we have described in previous papers [1, 5]. For example, collinear lines in the image must be collinear in 3-space, barring an accident of viewpoint. A corollary of this is that these image rela-

tions are normally invariant with respect to viewpoint, which greatly simplifies the problem of matching to three-dimensional objects of unknown orientation.

3) To the extent that these relations are stable under different imaging conditions and viewpoints, they can be used as reliable index terms to access a body of world knowledge. Not only can the names of the relations be used, but in addition each relation will have several parameters of variation whose relative values in the image can be used. For example, collinear line segments can be characterized by the relative sizes of the segments and gaps, which provides a viewpoint-invariant description that can be used to select a model for attempted matching.

Note that all three of these points assume that the relations found in the image are a result of regularities in the objects being viewed. This means that any relations which happen to arise accidentally from independent features will only confuse the interpretations. This distinction between significant and accidental relations is a point to which we will return.

The importance of perceptual organization for recognition: A demonstration

The importance of these grouping operations as a stage in the processing of images by the human visual system can be demonstrated by a straightforward psychophysical experiment. In Figure 1(a) we have constructed a partial line drawing of a bicycle in such a way that most opportunities for bottom-up segmentation are eliminated (e.g., we have eliminated most cases of significant collinearity, endpoint proximity, parallelism, and symmetry). In informal experiments with 10 subjects who were told nothing about the identity of the object, this drawing proved to be remarkably difficult to recognize. Nine out of 10 subjects were unable to recognize the object within a 60 second time limit, and the tenth subject took 45 seconds. Note that this is in spite of the fact that the object level segmentation has already been performed—the task would be even harder if the bicycle were embedded in a normal scene containing many surrounding features.

Figure 1(b) is the same drawing as in 1(a) with only a single segment added. The added segment was placed in a strategic location which would allow it to be combined with other segments in a curvilinear grouping. The center of this circular grouping would then be coincident with the termination of another segment, leading to further groupings. As might be expected if we assume that bottom-up groupings play an important role in recognition, the recognition times for this second figure were dramatically lower than for the first, with 3 out of 10 subjects recognizing it within 5 seconds and with 7 out of 10 subjects recognizing it within the 60 second time limit. Presumably, if the added segment had been placed at some location which did not lend itself to perceptual groupings the change in recognition times would have been negligible.

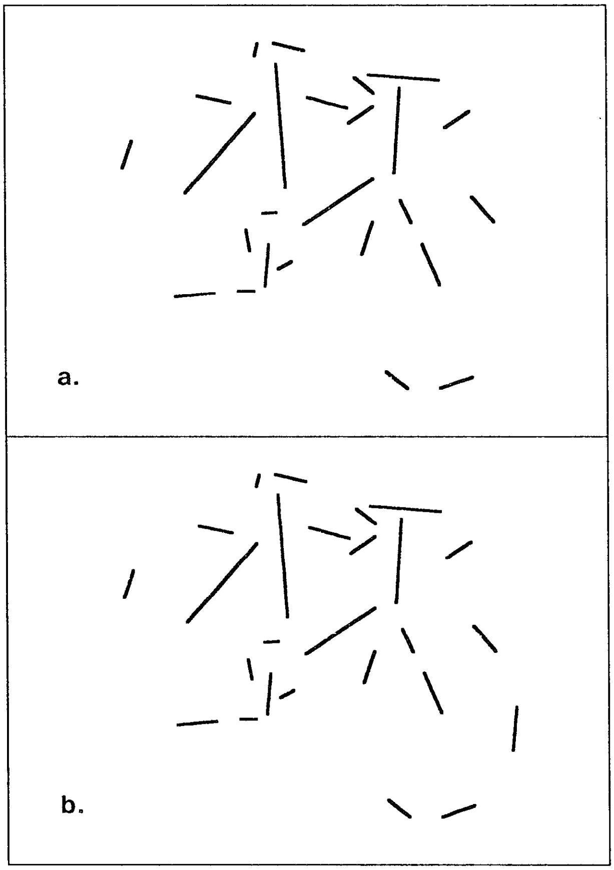


Figure 1: When opportunities for bottom-up grouping of image features have been removed, as was done for the line drawing of a bicycle in (a), the drawing is remarkably difficult to recognize. The average recognition time for (a) was over one minute when the subjects had no prior knowledge of the object's identity. When a single line segment was added in (b), which provided local evidence for a curvilinear grouping, the recognition times were greatly reduced.

These figures can also be used to demonstrate the human capability to make use of top-down contextual information to limit the search space for forming a match. As was demonstrated in experiments performed as early as 1935 [3], verbal clues naming even vague non-visual object classes can greatly reduce the recognition time. Subjects can usually interpret Figure 1(a) immediately upon being told that it is a bicycle. Thus this figure is on an interesting borderline where either bottom-up or top-down information can suddenly reduce the search space and lead to recognition. One can imagine a series of experiments that would systematically explore this search space and the reduction in its size created by different bottom-up or top-down clues. These figures can also be used to demonstrate the human equivalent of a back-projection algorithm [4] followed by image-level matching, where certain hypothesized partial matches can be used to solve for the position, orientation, and internal parameters of the model, which in turn lead to accurate predictions for further matches at specific locations in the image.

Principles of segmentation

There are virtually an infinite number of relations that could be formed between the elements of any image. What general principles can we derive for selecting those relations which are worth forming and for measuring their significance? As was mentioned earlier, segmentations are useful only to the extent that they represent actual structure of the scene rather than accidental alignments. Therefore, a central function of the segmentation process must be to distinguish, as accurately as possible, significant structures from those which have arisen by accident. All of the relations we have considered can arise from accidents of viewpoint or random positioning as well as from structure in the image. However, by examining the accuracy of each relation and the surrounding distribution of features in the image, it is possible to give probabilistic measures of the likelihood that any given relation is accidental. These nonrandomness measures can then be used as the basic test for significance during the segmentation process.

If there were a significant level of prior knowledge regarding the expected distributions of features and relations, this could be used for judging the significance of segmentations. However, the range of common images seems to be so wide that any prior knowledge at this level must be very weak. We have chosen to carry out our computation of significance with respect to the null hypothesis that features are independent with respect to orientation, location, and scale. Significance is then inversely proportional to the probability that the relation would have arisen from such a set of independent features. It is a matter for psychological experimentation to see whether the human visual system is biased in any direction from this independence assumption. But since a scene typically contains many independently positioned objects (leading to independence with respect to orientation, location, and scale in the image), the discrimination of relations with respect to this background seems like a reasonable criterion for judging significance.

A second major principle of segmentation is that each operation must have limited computational complexity. It is obviously impossible to test all combinations of features in an image, so the relations can only be formed over distances that do not include too many false candidates of the particular type being examined. Figure 2 shows an example in which a highly significant grouping of five equally-spaced collinear dots is not apparent to human vision when there are enough surrounding false targets. It would presumably be useful for the purposes of interpretation and recognition to detect such a statistically significant grouping, so this failure must be attributable to a lack of computational resources. This does not mean that groupings are diameter-limited in any absolute sense, since groupings can be attempted at many different scales; however, if there are more than a few false candidates at some scale, then no groupings can be formed at that scale of description.

The principles above describe which groupings will be formed and how they will be evaluated for a given class of relations, but they do not specify which classes of relations will be attempted. There are several factors which influence this choice. One important factor is the same imaging-invariance condition that was mentioned earlier—it is only worth looking for image relations which do not depend on a specific viewpoint, light-source position, or other image-formation parameter. For example, collinearity is useful because it is present in the image over all viewpoints of collinearity in the scene. But it would be pointless to detect lines at right-angles in the image, since even if right-angles are common in the scene the angle in the image would change with almost any change in viewpoint. Witkin and Tenenbaum [12] argue that prior probabilities play a role in selecting which relations are the easiest to distinguish from accidentals, and should therefore be attempted. If some relation arises only very rarely from the structure of typical scenes, then it is more likely that some instance of the relation in an image is accidental (although it would still be possible to distinguish the relation from accidentals given accurate-enough image measurements). Of course, it is also less productive to devote resources searching for properties which seldom arise than for those which are common.

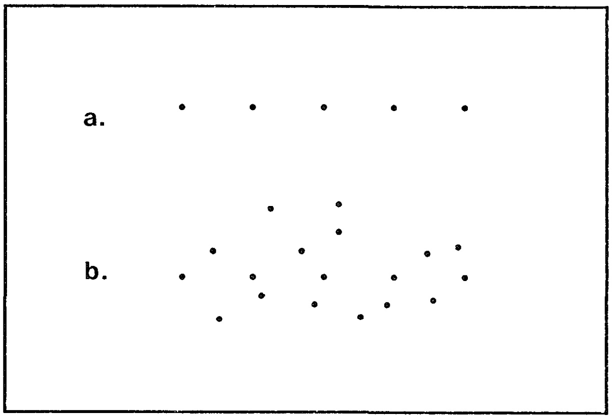


Figure 2: The pattern of five equally-spaced collinear dots in (a) is not detected spontaneously by human vision if it is surrounded by enough competing candidates for grouping within the same proximity, as in (b). This occurs even though the relation remains highly significant in the statistical sense and would therefore likely be of use for recognition.

An algorithm for curve segmentation

A significant bottleneck in creating a computer program which can perform these bottom-up perceptual processes on natural images is the problem of creating appropriately segmented edge descriptions. The best current edge operators detect "edge points" which are then linked using nearest-neighbor algorithms into lists of points. Although there has been considerable research into the problem of fitting smooth curves to these lists of points [9, 10, 11], almost without exception these efforts have concentrated on a single pre-selected resolution of segmentation and have attempted merely to smooth out noise induced by the imaging process. (We use "resolution" in the context of curve segmentation to refer to the allowable transverse deviation from the smoothed curve description.) Although these smoothed results may appear reasonable to the naïve human eye, that is because the human visual system can still perform the lower resolution groupings even though they have not been detected and described by the program. Figure 3 illustrates the problem, where the segmentation in Figure 3(b) is adequate to recognize one instance of collinearity, but other groupings are only apparent when lower resolution structures are recognized as in Figure 3(c). We have developed a new algorithm, based on the principles of segmentation outlined earlier, which measures the degree of nonrandom structure in edge-point lists over a wide range of resolutions and selects the most significant structures for the curve description. In our implementation, we examine all groupings which are either linear or of constant curvature. These can be splined to represent arbitrary smooth curves, although it is possible that human vision includes the detection of more general primitive curve groupings, such as spirals.

Measuring the significance of a curve segmentation

The first task in developing a segmentation algorithm is to determine how we will measure the significance of each grouping. In this case, since the points were originally linked on the basis of proximity, we must be careful not to confuse nonrandomness in proximity with the measurement of nonrandomness in linearity. For example, if we start by looking at a set of only three points, we might measure the significance of their linearity by measuring the distance of one point from the line joining the other two. However, this would confuse the effects of proximity with those of linearity, since by being close to one of the other points the third point would automatically be close to the line on which they lie, as is shown in Figures 4(a) and 4(b). Therefore, we have chosen to define nonrandomness in linearity to be how unlikely a point is to be as close as it is to a curve given its distance from the closest defining point of the curve. This is equal to $2\theta/\pi$, where θ is the angle between the curve and the vector from the closest endpoint, as shown in 4(c). This can be extended to 4 or more points by recursively looking for the point which is farthest away from any of the points considered thus far and calculating the likelihood for that point in terms of its minimum proximity to these previously considered points. Since these likelihood values are independent, they can be multiplied together to produce an overall value for the curve.

Testing all possible groupings

Given this significance test for sets of points, we want to divide the initial linked list of points into segments which have the highest significance values. Previous methods of curve segmentation have usually attempted to search for corners (tangent discontinuities) on a curve, where the curve can be divided into different segments. Our approach is the dual—we look for segments of the curve which exhibit significant nonrandomness, and tangent and curvature discontinuities are assigned to the junctions between neighboring segments. In contrast to the earlier approaches, our method will fail to assign any segmentation where the curve

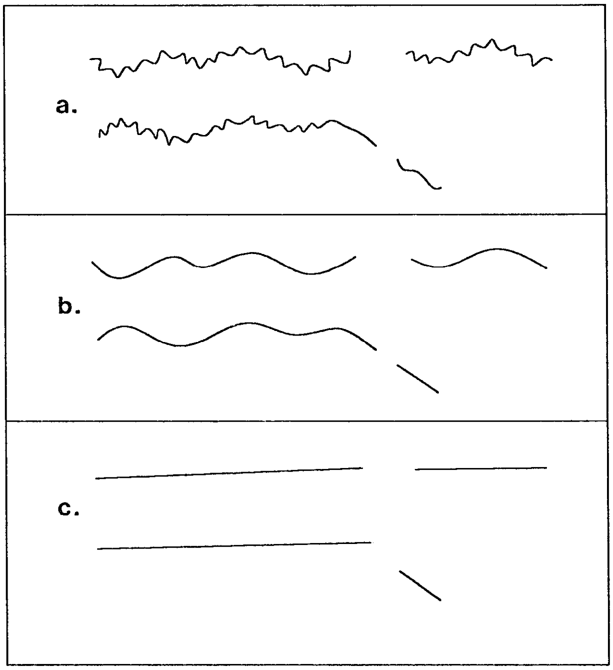


Figure 3: The data in (a) can be segmented at at least two different resolutions of description, as shown in (b) and (c). One instance of collinearity can only be detected in segmentation (b) while the other instance of collinearity and the parallelism can only be detected straightforwardly in (c).

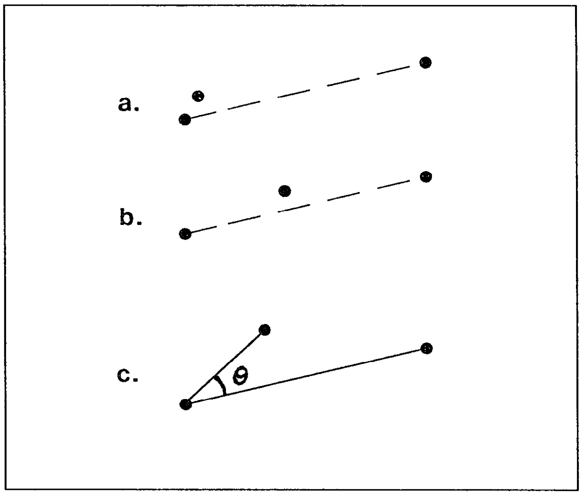
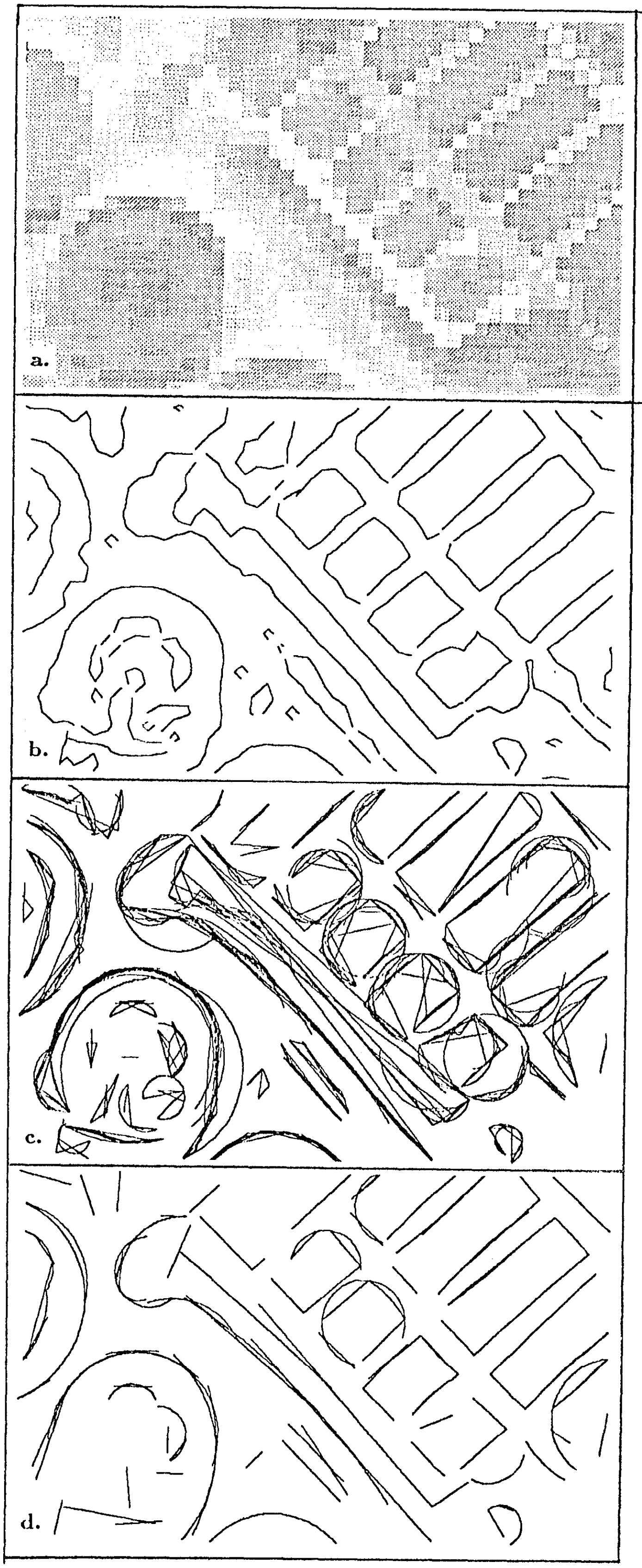


Figure 4: The middle dots in (a) and (b) are both the same distance from the dashed line joining the other two dots. Yet the three dots in (b) are much more significant in terms of their collinearity than those in (a), since the middle dot in (a) could be close to the line merely as a result of its proximity to the first endpoint. Therefore, we measure the probability of a point being within a given distance from a line in terms of its proximity to the closest endpoint defining the line, as shown in (c).



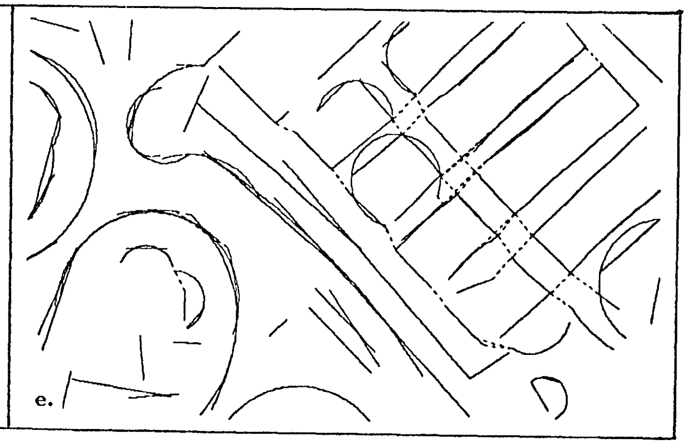


Figure 5: The small 30 by 45 pixel region of an aerial photograph shown in (a) was run through the Marimont edge detector to produce the linked edge points shown in (b). Figure (c) shows all the segments at different scales and locations which were tested for significance. After selecting only those segments which were locally *maximum* with respect to size of grouping, and thresholding out those which are not statistically significant, we are left with the segments shown in (d). It is then fairly simple to form collinearity and curvilinearity relations between these segments as shown by dotted lines in (e).

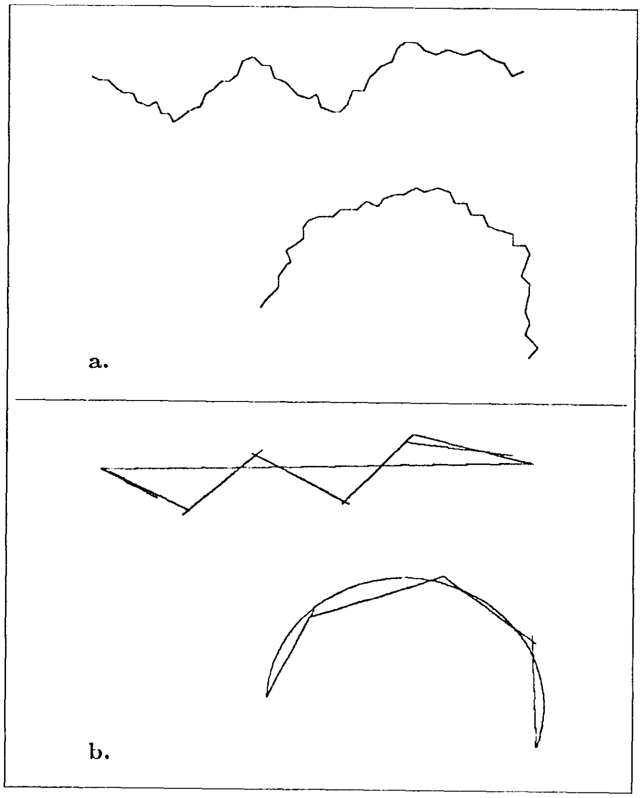


Figure 6: The hand-input curves in (a) have been created to exhibit significant structure at multiple resolutions. When these are given as data to the curve segmentation algorithm, it produces the results shown in (b), which makes these multiple levels of structure explicit.

appears to wander randomly at all resolutions, and will assign multiple segmentations where it exhibits different structure at different resolutions.

It would clearly be too costly to test every possible segment of the curve for nonrandomness. However, if we allow a reasonable margin of error, it is possible to cover all scales and locations with a relatively small number of groupings. We examine groupings at all scales differing by factors of two, from groupings of only three adjacent points up to groupings the size of the full length of the curve (amounting to 6 scales for a curve of 100 points). At each scale, we examine groupings at all locations along the curve, with adjacent groupings overlapping by 50%. This means that any given segment of the curve will have at least one grouping attempted which covers 50% of its length but does not extend outside its borders.

After measuring the significance of each grouping, a thinning procedure is executed which steps through the different resolutions at each location along the curve and selects only those segmentations which are locally maximum in their significance values. It is possible that there will be more than one local maximum if the curve exhibits different structures at different resolutions of grouping. There is also a threshold at the 0.05 significance level, below which groupings are not considered significant.

The algorithm in action

This algorithm have been implemented in MACLISP on a KL-10 computer and tested on synthetic data as well as edges derived from natural images. Figure 6(a) shows some hand-drawn curves which exhibit different structures at different resolutions, much as was shown in Figure 3. Figure 6(b) gives the output of the curve segmentation algorithm when given this data, and demonstrates the algorithm's ability to detect significant structure at multiple resolutions--results which no single-resolution algorithm could have produced.

Figure 5 shows the results of running the algorithm on a small 30 by 45 pixel region of an aerial photograph of an oil tank facility. The original digitized image is shown in 5(a). Figure 5(b) shows some linked edge data generated from this image by an edge detection program written by David Marimont [6], which detects edge points to subpixel accuracy and links them into lists. Figure 5(c) shows all the groupings at all resolutions, although the widely differing significance values are not apparent. Figure 5(d) shows the results after the thinning process which selects local maxima with respect to resolution. Given these segments, it is relatively easy to form collinearity and curvilinearity relations between them as shown by the dotted lines in Figure 5(e). It would also be fairly straightforward to detect endpoint proximity, parallelism, constant intervals, and other perceptual groupings.

Summary

We began this paper by demonstrating the importance of bottom-up perceptual organization for human vision. These image relations play a major role in limiting the size of the search space that must be considered when matching against world knowledge. The unifying principles of detecting nonrandom structure, avoiding combinatorial complexity, and looking for viewpoint-invariant relations were suggested. An algorithm for curve segmentation, based upon these principles, was developed and demonstrated. There are many other problems besides recognition in which these groupings would be useful. An example is the stereo correspondence problem, since to the extent that these image relations represent structure in the scene and are invariant with respect to viewpoint, they will be detected in images taken from different viewpoints.

The specific algorithms developed are preliminary implementations of the general methodology of segmenting perceptual data by looking at groupings over a wide range of scales and locations and retaining those which are the most unlikely to have arisen by accident from the background distribution. This same methodology could be applied to a wide range of other perceptual segmentation problems or signal analysis.

Acknowledgements

We would like to thank Peter Blicher, Chris Goad, David Marimont, Marty Tenenbaum, Brian Wandell, Andrew Witkin, and our other colleagues for stimulating discussions and suggestions. This work was supported under ARPA contract N00039-82-C-0250. The first author was also supported by a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Binford, Thomas O., "Inferring surfaces from images," *Artificial Intelligence*, **17** (1981), 205-244.

[2] Brooks, Rodney A., "Symbolic reasoning among 3-D models and 2-D images," *Artificial Intelligence*, **16** (1981).

[3] Leeper, R., "A study of a neglected portion of learning—the development of sensory organization," *Journal of Genetic Psychology*, **46** (1935), 41-75.

[4] Lowe, David G., "Solving for the parameters of object models from image descriptions" *Proceedings ARPA Image Understanding Workshop* (College Park, MD, April 1980), 121-127.

[5] Lowe, David G. and Thomas Binford, "The interpretation of three-dimensional structure from image curves," *Proceedings IJCAI-7* (Vancouver, Canada, August 1979), 613-618.

[6] Marimont, David, "Segmentation in ACRONYM," *Proceedings ARPA Image Understanding Workshop* (Stanford, California, September 1982).

[7] Marr, David, "Early processing of visual information," *Philosophical Transactions of the Royal Society of London, Series B*, **275** (1976), 483-524.

[8] Nevatia, R., and T.O. Binford, "Description and recognition of curved objects," *Artificial Intelligence*, **9** (1977).

[9] Pavlidis, T., *Structural Pattern Recognition* (New York, NY: Springer-Verlag, 1977).

[10] Rutkowski, W.S., and Azriel Rosenfeld, "A comparison of corner-detection techniques for chain-coded curves," TR-623, Computer Science Center, University of Maryland, 1978.

[11] Shirai, Y., "Recognition of man-made objects using edge cues," *Computer Vision Systems*, A. Hanson, E. Riseman, eds. (New York: Academic Press, 1978).

[12] Witkin, Andrew P. and Jay M. Tenenbaum, "On the role of structure in vision." To appear in *Human and Machine Vision*, Rosenfeld and Beck, eds. (New York: Academic Press, 1983).