VISUAL RECOGNITION FROM SPATIAL CORRESPONDENCE AND PERCEPTUAL ORGANIZATION

David G. Lowe
Courant Institute of Mathematical Sciences
New York University
251 Mercer St., New York, NY 10012

Abstract

Depth reconstruction from the two-dimensional image plays an important role in certain visual tasks and has been a major focus of of computer vision research. However, in this paper we argue that most instances of recognition in human and machine vision can best be performed without the preliminary reconstruction of depth. Three other mechanisms are described that can be used to bridge the gap between the two-dimensional image and knowledge of three-dimensional objects. First, a process of perceptual organization can be used to form groupings and structures in the image that are likely to be invariant over a wide range of viewpoints. Secondly, evidential reasoning can be used to combine evidence from these groupings and other sources of information to reduce the size of the search-space during model-based matching. Finally, a process of spatial correspondence can be used to bring the projections of three-dimensional models into direct correspondence with the image by solving for unknown viewpoint and model parameters. These methods have been combined in an experimental computer vision system named SCERPO. This system has demonstrated the use of these methods for the recognition of objects from unknown viewpoints in single gray-scale images.

Introduction

The standard model for much recent research in computer vision has been based on the reconstruction of depth information from the image prior to recognition. However, in this paper we will argue that this is not the primary pathway used for most instances of recognition in human vision. Although depth measurement has an important role in certain visual problems, it is often not available and is not needed for typical instances of recognizing familiar objects. Instead, we will propose that the primary bottom-up descriptive analysis of the image can best be performed by a process of perceptual organization. This process leads to the formation of significant groupings and structures directly from the two-dimensional image data. These groupings are partially invariant to viewpoint and can be matched directly against three-dimensional object models. The verification of these matches can be performed by spatially mapping the projection of three-dimensional object models onto the image data, through a process of viewpoint and model-parameter determination.

These methods have been combined in a vision system named SCERPO (for Spatial Correspondence, Evidential Reasoning, and Perceptual Organization). While seemingly solving a more difficult problem—the direct recognition of objects from unknown viewpoints in two-dimensional images—the approach is shown to be apparently simpler and more flexible than those that rely upon depth reconstruction. While it is true that the appearance of a three-dimensional object can change completely as it is viewed from different viewpoints, it is also true that many aspects of an object's projection remain invariant over large ranges of viewpoints (examples include instances of connectivity, collinearity, parallelism, repetitive textures, and certain symmetries). It is the role of perceptual organization to detect those image groupings that are unlikely to have arisen by accident of viewpoint or position. Once detected, these groupings can be matched to corresponding structures in the objects through a knowledge-based process of evidential reasoning. These methods for evidential reasoning were initially developed for combining probabilistic information in diagnostic expert systems, but they can be readily adapted to combining information regarding probabilistic associations between particular image features and object models. This probabilistic information is used to order the search strategy so that the most reliable and informative information is tested first.

The reliability of the search process depends upon the the final verification of each hypothesized interpretation. SCERPO uses a quantitative method to simultaneously determine the best viewpoint and object parameter values for fitting the projection of a three-dimensional model to given two-dimensional features. It allows a few initial hypothesized matches to be extended by making exact predictions for the locations of other object features in the image. This provides a highly reliable method for verifying the presence of a particular object, since it can make use of the spatial information in the image to the full degree of available resolution.

The role of depth recovery in human vision

A substantial fraction of recent computer vision research has been aimed at the bottom-up derivation of depth or surface orientation from image data, using information such as stereo, motion, shading or texture. This has come to be known as the "Shape from X" paradigm. Marr [12] suggested that these sources of information could be combined in a representation known as the $2\frac{1}{2}$-D sketch that would allow one source of information to compensate for the absence of another. The depth representation would then be used to determine correspondence with three-dimensional object representations, the assumption being that it would be easier to match a three-dimensional model to a depth representation than to two-dimensional image data.

Human vision contains many of these components for recovering depth, and they presumably have important functions. However, biological visual systems have many objectives, so it does not follow that these components are central to the problem of visual recognition. In fact, the available evidence would seem to indicate the opposite. The first problem with these methods is that depth information is often unavailable or requires an unacceptably long interval of time to obtain. Stereo vision is only useful for objects within a restricted portion of the visual field and range of depths for any given degree of eye vergence, and is never useful for distant objects. Motion information is available only when there is sufficient relative motion between observer and object, which in practice is also usually limited to nearby objects. Recognition times are usually so short that it seems unlikely that the appropriate eye vergence movements or elapsed time measurements could be taken prior to recognition even for those cases in which they may be useful. Depth measurements from shading or texture are apparently restricted to special cases such as regions of approximately uniform reflectance or regular texture, and they lack the quantitative accuracy or completeness of stereo or motion.

Secondly, human vision exhibits an excellent level of performance in recognizing images—such as line drawings—in which there is very little potential for the bottom-up derivation of depth information. Whatever mechanisms are being used for line-drawing recognition have presumably developed from their use in recognizing three-dimensional scenes. The common assumption that line-drawing recognition is a learned or cultural phenomena is not supported by the evidence. In a seemingly definitive experiment, Hochberg and Brooks [6] describe the case of a 19-month-old human baby who had had no previous exposure to any kinds of two-dimensional images, yet was immediately able to recognize ordinary line drawings of known objects.

Finally, there has been no clear demonstration of the value of depth information for performing recognition, even when it is available. The recognition of objects from complete depth images, such as those produced by a laser scanner, has not been shown to be much easier than for systems that begin only with the two-dimensional image. This paper will describe methods for directly comparing the projection of three-dimensional representations to the two-dimensional image without the need for any prior depth information.

Of course, none of this is meant to imply that depth recovery is an unimportant problem or lacks a significant role in human vision. Depth information may be crucial for the initial stages of visual learning or for acquiring certain types of knowledge about unfamiliar structures. It is also clearly useful for making precise measurements as an aid to manipulation or obstacle avoidance. However, it seems likely that the role of depth recovery in common instances of recognition has been overstated.

Matching 3-D knowledge to the image

Although knowledge of object shape, context, and surface properties must naturally be represented in three-dimensional form, this knowledge can be matched directly against the two-dimensional image through the use of projection. A major practical difficulty is in using image measurements to determine the unknown projection parameters. Six parameters are needed to specify an arbitrary position and orientation of an object with respect to the camera, and there may be other unknown parameters internal to the object. However, each match between a point in the image and a point on the object allows us to solve for two parameters. Therefore, only three or four hypothesized matches between the image and an object model are typically needed to solve for the projection parameters. Once these parameters have been determined, it is straightforward to carry out the projection and extend the match by making accurate predictions for the locations of other model features in the image. These further matches may be used to solve for any remaining model parameters, but their most important function is to provide reliable confirmation for the correctness of an interpretation.

The author has previously presented a mathematical technique [7, 9] for solving for viewpoint and model parameters given some matches between image and model. Briefly, this method linearizes the projection equations and uses Newton-Raphson iteration to solve simultaneously for the unknown parameters. Since the projection equations are very smooth (consisting of linear combinations of *sin* and *cos* functions of viewpoint), the method has quadratic convergence and typically requires only 3 iterations to achieve high accuracy. This basic technique has been extended to perform least-squares solution of over-determined systems, and to allow matching of image lines to model lines (without concern for the location of line terminations). Given these methods, the problem of verification is largely solved for well-specified objects, and the remaining problems of recognition are those of reducing the size of the search space to produce the few initial matches.

There is experimental evidence that human recognition also relies upon the determination of viewpoint parameters for projecting a three-dimensional object description onto the image. Cooper & Shepard [4] describe experiments in which subjects are asked to compare images at varying orientations to previously memorized shapes. They found that the recognition time varied linearly in the angle of rotation between the image and the orientation of the original memorized shape. In conjunction with their other work on mental rotation, this would seem to indicate that recognition is performed by bringing a prior representation into spatial correspondence with image data by manipulating viewpoint parameters.

Allowing for variations in object models

The capability for recognizing objects from their two-dimensional projections is possible only because of previous knowledge regarding the objects. However, recognition does not imply that we must know every aspect of an object's appearance prior to recognition. Object models may be parameterized with variable sizes, angles, or articulations between components, with expected bounds given for each parameter. As already mentioned, it is possible to backsolve for these parameters using the same methods as when solving for viewpoint. Just as important is the fact that there is no precise boundary between what is an object and what is a component. It is possible to recognize commonly-occurring components, such as cylinders, rectangular solids, or repeated patterns, as parameterized objects in their own right. The only requirement is that there be fewer unknown parameters to the description than there are useful measurements to be made from the image data. These recognized components—even if the identification is only tentative—can then be used to suggest the identity of the more specific structure of which they are a part. If the identification of the components is quite certain, then they can even be combined into previously unknown or very loosely parameterized relationships. Most objects can be represented both in terms of their overall shape and in terms of a combination of components, and different images can best make use of each type of description depending upon such variables as image resolution, viewpoint, and occlusion.

Previous work on model-based vision

There is a considerable body of previous research in model-based vision. The remarkable early work of Roberts [13] demonstrated the recognition of certain polyhedral objects by precisely solving for viewpoint and object parameters. Unfortunately, this work was poorly incorporated in later vision research, which tended to emphasize less quantitative methods. The ACRONYM system of Brooks [1] used a general symbolic constraint solver to calculate bounds on viewpoint and model parameters from image measurements. These bounds could then be used to check the consistency of interpretations produced by general matching operations, and were capable of handling wide classes of generic object descriptions. Goad [5] describes the use of automatic programming methods to precompute a highly efficient search path and viewpoint-solving technique for each object to be recognized. This research has been incorporated in an industrial computer vision system by Silma Inc. which has the capability of performing all aspects of recognition within as little as 1 second. Because of their runtime efficiency, these precomputation techniques are likely to remain the method of choice for industrial systems dealing with small numbers of objects. Other closely related research on model-based vision has been performed by Shirai [14] and Walter & Tropf [16].

Perceptual organization in SCERPO

Unlike previous model-based systems, SCERPO makes use of perceptual organization as the central process for bottom-up analysis of an image. Perceptual organization refers to a basic capability of the human visual system to derive relevant groupings and structures from an image without prior knowledge of its contents. For example, people will immediately detect clustering, connectivity, collinearity, parallelism, and repetitive textures when shown an otherwise randomly distributed set of image elements. This grouping capability of human vision was studied by the early Gestalt psychologists [17] and is related to research in texture description [10]. A major function of perceptual organization is to distinguish non-accidental groupings from the background of groupings that arise through accident of viewpoint or random positioning [18, 8]. Those groupings that are non-accidental in origin will also be partially invariant with respect to viewpoint and be most suited to model-based recognition (see [9] for a much more detailed discussion).

In order to provide image features for input to perceptual organization, the first few levels of image analysis in SCERPO use established methods of edge detection, as shown in Figures 1–3. The 512-by-512-pixel image shown in Figure 1 was convolved with a Laplacian of Gaussian function ($\sigma = 1.8$ pixels) as suggested by the Marr-Hildreth [11] theory of edge detection. The zero-crossings of this function are shown in Figure 3. Of course, many of these zero-crossings do not correspond to significant edges in the image. We remove those corresponding to insignificant intensity changes by applying the Sobel gradient operator to the $\nabla^2 G$ convolution. Only those points that are above a chosen gradient threshold and lie on a zero crossing are retained in Figure 5. These remaining zero-crossings are linked into lists of points on the basis of connectivity.

The first stage of perceptual organization is to group the linked lists of points into perceptually significant curve segments. The author has previously described a method for finding straight-line and constant-curvature segmentations at multiple scales and for measuring their significance [9, Chap. 4]. However, here we use a simplified method that

selects only the single highest-significance line representation at each point along the curve. The significance of a straight line fit to a list of points is measured as the ratio of its length divided by the maximum deviation of a point from the line. This provides a scale-independent measure of significance that places no prior bounds on the allowable deviations. This is then used in a modified version of the recursive endpoint subdivision method. A segment is subdivided at the point with maximum deviation from a line connecting its endpoints. If the maximum significance of any of the subsegments is greater than the significance of the complete segment, then the subsegments are returned. Otherwise the single segment is returned. This procedure is applied recursively until each segment contains fewer than 3 points. The procedure will return a segment covering every point along the curve, but those with a length-to-deviation ratio less than 4 are discarded. This method is implemented in only 40 lines of Lisp code, yet does a reasonable job of detecting the most perceptually significant straight line groupings in the linked point data. The results are shown in Figure 4.

The straight line segments are indexed according to endpoint locations and orientation. Then a sequence of procedures is executed to detect instances of collinearity, endpoint proximity (connectivity), and parallelism. A region around each endpoint or segment is examined to determine candidates for grouping. Each potential grouping is assigned a significance value that is roughly inversely proportional to the likelihood that it is accidental in origin. This is done in a scale-independent manner (i.e., measurements of endpoint proximity or separation of parallel lines are divided by the length of the shortest of the two line segments). After the execution of this grouping process, the many groupings are ranked in order of significance. Unfortunately, it is difficult to display the results of this grouping process without showing a separate image for each grouping that has been detected. Although several hundred significant groupings were detected in the line segments of Figure 4, we show in Figure 5 only the two sets of highly-ranked groupings that were actually used for successful recognition.

Evidential reasoning

Evidential reasoning refers to the combination of different sources of information or evidence in order to reach a conclusion with a specified level of certainty. This form of reasoning has been developed for use in diagnostic expert systems, among other applications. It can be used, for example, to calculate the likelihood that a particular disease is present given a number of symptoms. We are faced with a very similar problem in vision when we wish to calculate the likelihood that a particular object is present in an image given a number of detected features and other sources of information. The performance requirements for evidential reasoning in vision are much less stringent than in medical expert systems, since we have a reliable procedure for final verification and only need to use the evidential reasoning to suggest the most efficient ordering for our search.

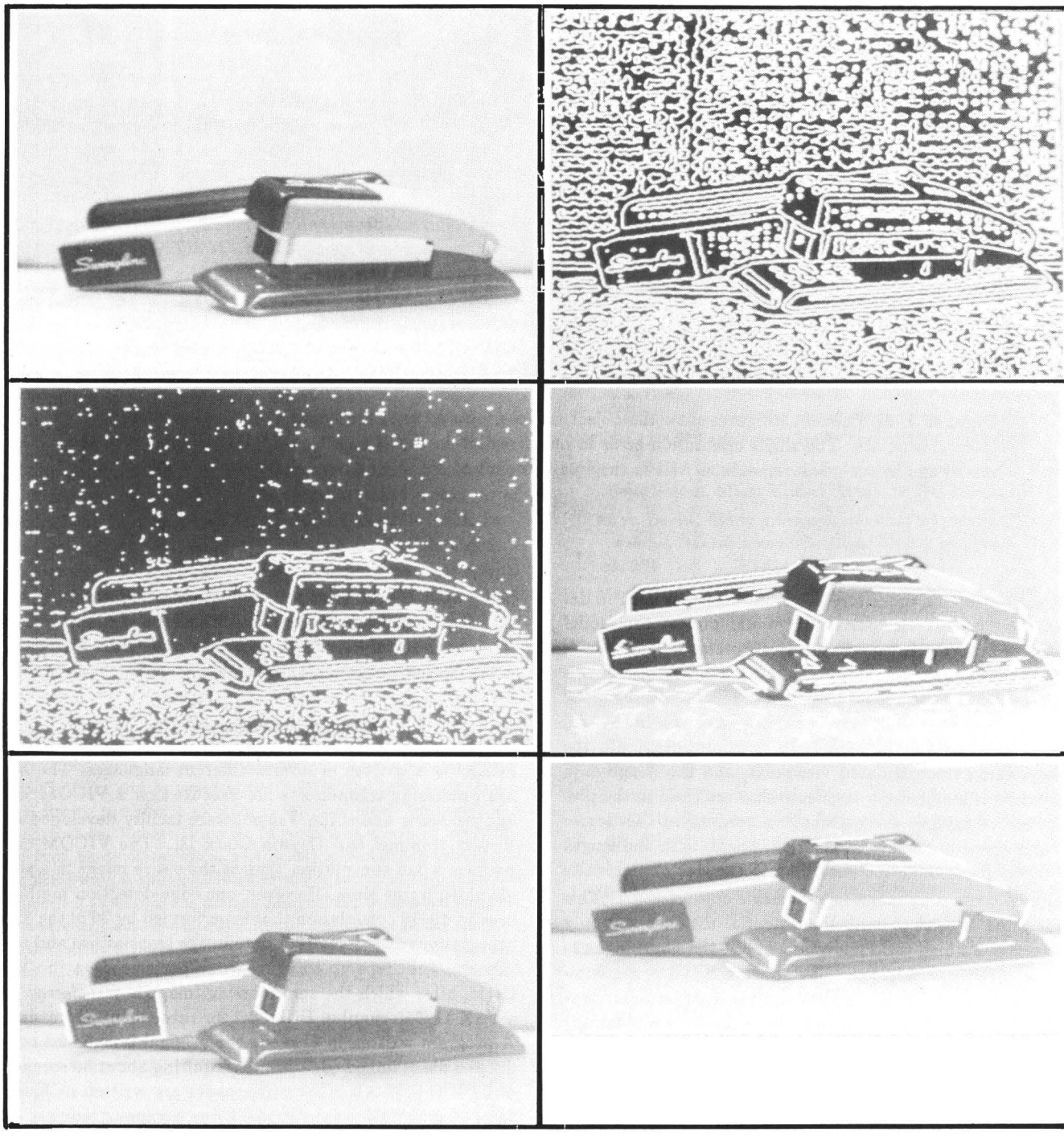
In order to minimize the search time, we would like to order our consideration of hypotheses according to decreasing values of P_k/W_k, where P_k is the probability that a particular hypothesis for the presence of object k is correct, and W_k is the amount of work required to verify or refute it. Evidence can come from many sources: we may have initial expectations for the presence of certain objects, contextual expectations resulting from the presence of already-detected objects, and information from many forms of image data such as perceptual groupings, texture, color, or metric measurements. The initial researchers in medical expert systems rejected the use of Bayesian methods for combining evidence [15], since they assumed that it would either require unrealistic independence assumptions or an impossibly large number of known statistical parameters. However, recent work by Charniak [2] has shown that it is possible to formalize the previous apparently ad-hoc methods within a Bayesian framework. The application of Charniak's methods to ordering search during recognition is discussed in [9, Chap. 6]. An important aspect of evidential reasoning is that it offers a strong basis for building learning systems in which the required statistical parameters are moved towards their correct values as the system gains experience.

The evidential reasoning component of SCERPO has not yet been developed as fully as other parts of the system. Since the system has only been used with a single object under consideration, the performance requirements for minimizing search have not been great. The system makes use of a list of perceptual groupings and the model features that could give rise to them. This list is entered by the user at the same time as model specification. For example, the groupings shown in Figure 5 consist of particular combinations of parallelism and endpoint proximity that could be matched to various parts of the object model. The probabilities of non-accidentalness for the image relations that make up a grouping are multiplied together to calculate the probability for the grouping as a whole. This is multiplied by an estimate of the likelihood of correctness for the match (assuming a non-accidental grouping) that has been entered for each element of the association list, and these final values are used to order the search. We plan to explore methods for incrementally learning the required probability values in future research.

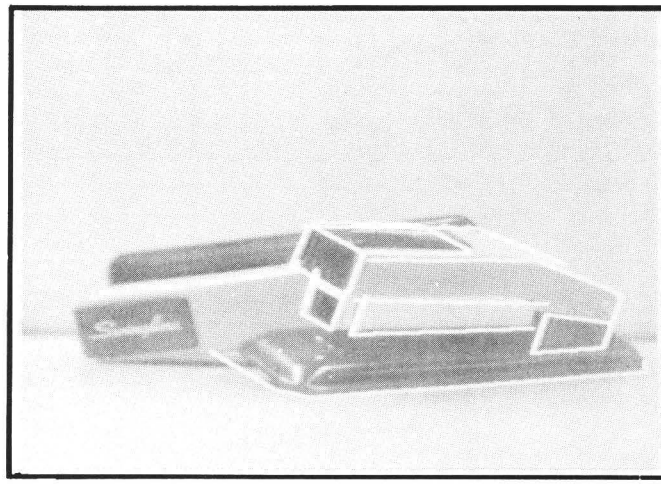
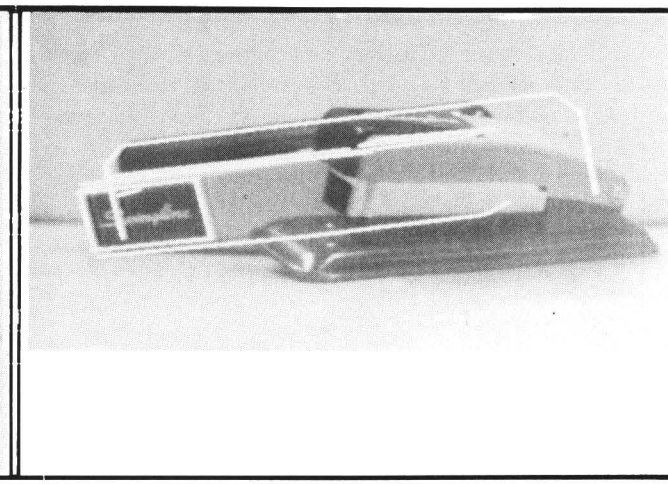
Verification of interpretations

The verification component of SCERPO is able to take a tentative match between a couple of image features and model features and return a reliable answer as to whether the match is correct. If the object is present, this module will extend the match as much as possible and determine the precise viewpoint.

Given the initial set of correspondences, the iterative



Figures 1–6: The original image of some desk staplers is shown in Fig. 1. This image was convolved with a $\nabla^2 G$ function ($\sigma = 1.8$ pixels). The zero-crossings of this function are shown in Fig. 2. The gradiant of the convolved image was measured, and Fig. 3 shows only those zero-crossings at locations where the function had a gradient above a selected threshold value. Fig. 4 shows the segments that resulted from linking of zero-crossings and selection of the most significant straight-line segmentations (shown superimposed on the original image). Fig. 5 shows the two perceptual groupings that were actually used to initiate successful recognition. After solving for model viewpoint, selecting new segments most consistent with model predictions, and iterating, the segments shown in Fig. 6 were selected as being consistent with one viewpoint.

7.   8.

Figures 7–8: These final figures show the object model projected onto the image from the two final calculated viewpoints. The slight orientation error in one direction in Fig. 8 is due to small inaccuracies in the model and image measurements as well as the small amount of data being used to determine viewpoint.

viewpoint-solving procedure described earlier is used to determine the best viewpoint that would project the model features onto the image features. The current implementation solves only for viewpoint and does not allow variable model parameters. If large errors remain following the least-squares fit, the solution is rejected as inconsistent. All edge features from the model are then projected onto the image using the calculated viewpoint, and the image data structure is searched for segments that are close to the predictions. Matches are evaluated according to the degree of agreement in transverse location, orientation, and length with the prediction, and according to the lack of ambiguity between competing matches for a single object feature. This evaluation is used to rank the potential matches and only those above a high threshold value, or else the single highest-ranked match, is selected. The selected matches are combined with the original matches and the least-squares viewpoint determination is repeated. An estimate is maintained of the error bounds, based upon the number of matches and the least-squares deviations, so that instances of ambiguity become less likely as the viewpoint estimate improves. The set of matches is repeatedly extended until no more can be found. The final result of this process is the selection of a set of segments, as shown in Figure 6, that are consistent with a single viewpoint of the model, as shown in Figure 7.

The current verification process in SCERPO could clearly be extended to include many other aspects of verification than just the matching of line segments. For example, the viewpoint determination for the model instance shown in Figure 8 has a small error in orientation, due to errors in image measurements and the small number of segments being used for the least-squares matching. However, given this degree of recognition, it would now be straightforward to go back to the original image data or zero-crossings and make further image measurements.

Implementation details

SCERPO is written in several different languages. The image processing components are executed on a VICOM image processor under the *Vsh* software facility developed by Robert Hummel and Dayton Clark [3]. The VICOM can perform a 3x3 convolution against the entire image in a single video frame time. However, our edge detection method uses an 18x18 convolution that is performed by 36 of the 3x3 convolutions and the appropriate image translations and additions. The steps up to figure 5 are performed on the VICOM, after which the zero-crossing image is transferred to a VAX 11/750 running UNIX 4.2 for subsequent processing. A program written in C reads the original image and produces a file of linked edge points (requiring about 30 seconds of CPU time). All other components are written in Franz Lisp. Segmentation into straight line segments requires 40 seconds, indexing and grouping operations require about 1 minute and the later stages of matching and verification took 40 seconds for this example.

The object models used by the system consists merely of a set of straight 3-D line segments. Each segment has a simple visibility specification, listing viewpoint ranges over which it is visible. A full hidden-line algorithm and more complete object models would improve the performance of the system.

Conclusions and future research

The current capabilities of SCERPO provide a framework that could be used to incorporate numerous additional capabilities, each of which would improve the generality of the system or its level of performance. A brief list of these possible extensions might include the following: incorporation of a wider range of perceptual grouping operations, the ability to handle variable model parameters, the recognition of object components and their subsequent combination, more complete modeling with surface information and hidden-line algorithms, the use of color and texture information, the expanded use of evidential reasoning, the incremental learning of associations and probabilities, the detection of curve segments as well as straight lines, and more detailed verification in terms of the original image data.

Perceptual organization and the methods for achieving spatial correspondence offer an alternative to the use of depth reconstruction and matching in three-dimensions. It has been argued in this paper that most instances of recognition in human vision also work directly from two-dimensional data. It should be possible to provide a definitive answer to this question by designing psychophysical experiments that test human recognition capabilities with different combinations of available information. A final answer to this question would carry many implications for the future design of knowledge-based vision systems.

Acknowledgments

Implementation of the SCERPO system relied upon the extensive facilities and software of the NYU vision laboratory, which are due to the efforts of Robert Hummel, Jack Schwartz, and many others. Robert Hummel, in particular, provided many important kinds of technical and practical assistance during the implementation process. Mike Overton provided help with the numerical aspects of the design. Much of the theoretical basis for this research was developed while the author was at the Stanford Artificial Intelligence Laboratory, with the help of Tom Binford, Rod Brooks, Chris Goad, David Marimont, Andy Witkin, and many others.

References

[1] Brooks, Rodney A., "Symbolic reasoning among 3-D models and 2-D images," *Artificial Intelligence,* **17** (1981), 285-348.

[2] Charniak, Eugene, "The Bayesian basis of common sense medical diagnosis," *Proceedings of AAAI-83* (Washington, D.C., August, 1983), 70-73.

[3] Clark, Dayton and Robert Hummel, "VSH user's manual: an image processing environment," *Robotics Research Technical Report,* Courant Institute, New York University (September 1984).

[4] Cooper, Lynn A., and Roger N. Shepard, "Turning something over in the mind," *Scientific American,* **251**, 6 (December 1984), 106–114.

[5] Goad, Chris, "Special purpose automatic programming for 3D model-based vision," *Proceedings ARPA Image Understanding Workshop* (1983).

[6] Hochberg, Julian E. and Virginia Brooks, "Pictorial recognition as an unlearned ability: A study of one child's performance," *American Journal of Psychology,* **75** (1962), 624-628.

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

[8] Lowe, David G. and Thomas O. Binford, "Perceptual organization as a basis for visual recognition," *Proceedings of AAAI-83* (Washington, D.C., August 1983), 255–260.

[9] Lowe, David G., *Perceptual Organization and Visual Recognition* (Boston, Mass: Kluwer Academic Publishers, 1985).

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

[11] Marr, David, and Ellen Hildreth, "Theory of edge detection," *Proc. Royal Society of London, B,* **207** (1980), 187-217.

[12] Marr, David, *Vision* (San Francisco: W.H. Freeman and Co., 1982).

[13] Roberts, L.G., "Machine perception of three-dimensional objects," in *Optical and Electro-optical Information Processing,* Tippet *et al.,* Eds. (Cambridge, Mass.: MIT Press, 1966), 159-197.

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

[15] Shortliffe, Edward H. and Bruce G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences,* **23** (1975), 355-356.

[16] Walter, I. and H. Tropf, "3-D recognition of randomly oriented parts," *Proceedings of the Third International Conf. on Robot Vision and Sensory Controls* (November, 1983, Cambridge, Mass.), 193–200.

[17] Wertheimer, Max, "Untersuchungen zur Lehe von der Gestalt II," *Psychol. Forsch.,* **4** (1923). Translated as "Principles of perceptual organization" in *Readings in Perception,* David Beardslee and Michael Wertheimer, Eds., (Princeton, N.J.: 1958), 115-135.

[18] Witkin, Andrew P. and Jay M. Tenenbaum, "On the role of structure in vision," in *Human and Machine Vision,* Beck, Hope & Rosenfeld, Eds. (New York: Academic Press, 1983), 481-543.