Active Object Recognition Integrating Attention and Viewpoint Control*

Sven J Dickinson[1], Henrik I Christensen[2],
John Tsotsos[1], and Göran Olofsson[3]

[1] Dept. of Computer Science, University of Toronto
6 King's College Rd., Toronto, Ontario, Canada M5S 1A4
[2] Laboratory of Image Analysis, IES
Aalborg University, DK-9220 Aalborg, Denmark
[3] Computational Vision and Active Perception Laboratory
Royal Institute of Technology, S-100 44 Stockholm, Sweden

Abstract. We present an active object recognition strategy which combines the use of an attention mechanism for focusing the search for a 3-D object in a 2-D image, with a viewpoint control strategy for disambiguating recovered object features. The attention mechanism consists of a probabilistic search through a hierarchy of predicted feature observations, taking objects into a set of regions classified according to the shapes of their bounding contours. If the features recovered during the attention phase do not provide a unique mapping to the 3-D object being searched, the probabilistic feature hierarchy can be used to guide the camera to a new viewpoint from where the object can be disambiguated.

1 Introduction

An important aspect of active vision is the use of an attentional mechanism to decide where in the image to search for a particular object [14]. Template matching schemes which move an object template throughout the image offer no attention mechanism since all positions in the image are treated equally. However, any recognition scheme that preprocesses the image to extract some set of features provides a basis for an attention mechanism. Assuming that the recovered image features correspond to model features, object search can be performed at those locations in the image where the features are recovered.

For an attention mechanism to be effective, the features must be distinguishing, i e., have low entropy. If the recovered features are common to every object

* The authors gratefully acknowledge the assistance of Lars Olsson, Gene Amdur, James Maclean, Sean Culhane, Winky Wai, Yiming Ye, and Feng Lu in the implementation of this work. Henrik Christensen and Göran Olofsson would like to acknowledge funding support from EP-7108-VAP "Vision as Process". The PLAYBOT project, based at the University of Toronto, is part of the Institute for Robotics ad Intelligent Systems, a Network of Centers of Exellence funded by the Government of Canada. Tsotsos is the CP-Unitel Fellow of the Canadian Institute for Advanced Research.

Lecture Notes in Computer Science, Vol. 801
Jan-Olof Eklundh (Ed.)
Computer Vision – ECCV '94

being searched, they offer little in the way of focusing a search for an object. This is typical in object recognition systems which match simple image features like corners or zeroes of curvature to model features, e.g., [10, 5, 8]. Although invariant to viewpoint, there may be an abundance of such features in the image, leading to a combinatorial explosion in the number of possible correspondences between image and model features that must be verified. In the first part of this paper, we will argue that regions, characterized by the shapes of their bounding contours, provide a more effective attention mechanism than simple linear features. We go on to present a Bayesian attention mechanism which maps objects into volumetric parts, maps volumetric parts into aspects, and maps aspects to component faces. Face predictions are then matched to recovered regions with a goodness-of-fit providing an ordering of the search locations.

In the second part of this paper, we extend our object representation for attention to support active viewpoint control. We will introduce a representation, called the *aspect prediction graph*, which is based on the aspect graph. Given an ambiguous view of an object, the representation will first tell us if there is a view of the object which is more discriminating. If so, the representation will tell us in which direction we should move the camera to encounter that view. Finally, the representation will tell us what visual events (the appearance or disappearance of features on the object) we can expect to encounter while moving the camera to the new viewpoint.

2 Attention

2.1 Review of the Object Representation

To demonstrate our approach to attention, we have selected an object representation similar to that used by Biederman [1], in which the Cartesian product of contrastive shape properties gives rise to a set of volumetric primitives called geons. For our investigation, we have chosen three properties including cross-section shape, axis shape, and cross-section size variation (Dickinson, Pentland, and Rosenfeld [3]). The cartesian product of the dichotomous and trichotomous values of these properties give rise to the set of ten volumes illustrated in Figure 1; to construct objects, the volumes are simply attached to one another.

Traditional aspect graph representations of 3-D objects model an entire object with a set of aspects, each defining a topologically distinct view of an object in terms of its visible surfaces (Koenderink and van Doorn [7]). Our approach differs in that we use aspects to represent the (typically small) set of volumetric part classes from which each object in our database is constructed, rather than representing an entire object directly. The representation, called the *aspect hierarchy*, consists of three levels, including of the set of *aspects* that model the chosen volumes, the set of component *faces* of the aspects, and the set of *boundary groups* representing all subsets of contours bounding the faces. The ambiguous mappings between the levels of the aspect hierarchy were originally captured in a set of upward conditional probabilities (Dickinson et al. [3]), mapping boundary

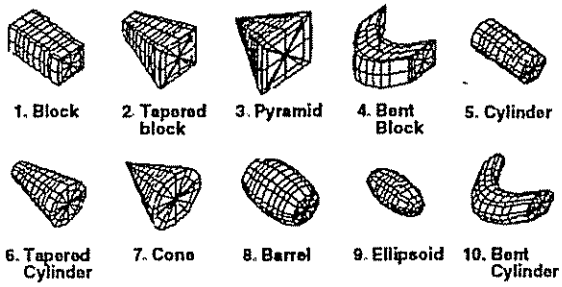

Fig. 1. The Ten Modeling Primitives

groups to faces, faces to aspects, and aspects to volumes.[4] However, for the attention mechanism described in this paper, the aspect hierarchy was augmented to include the downward conditional probabilities mapping volumes to aspects, aspects to faces, and faces to boundary groups. Figure 2 illustrates a portion of the augmented aspect hierarchy.

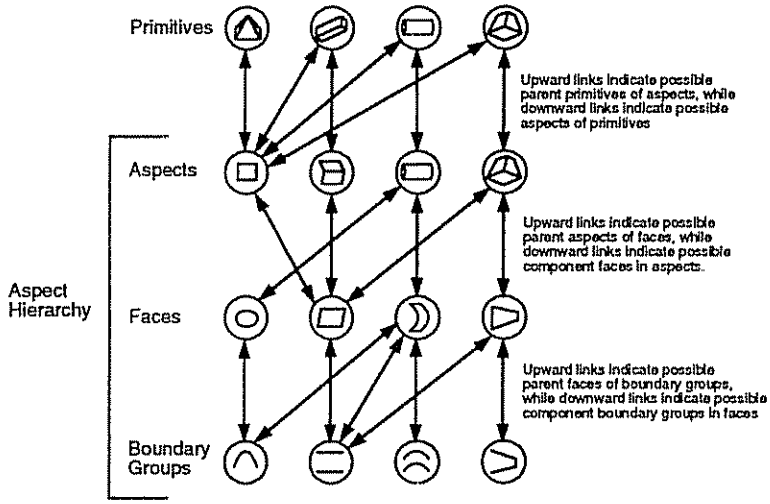

Fig. 2. The Augmented Aspect Hierarchy

[4] The probabilities are estimated from a frequency analysis of features viewed over a sampled viewing sphere centered on each of the ten volume classes.

2.2 A Case for Focusing on Regions

Given the various levels of the augmented aspect hierarchy, the question arises: At which recovered features from the image do we focus our search for a particular object? Many CAD-based recognition systems advocate extracting simple features like corners, high curvature points, or zeroes of curvature. Although robustly recoverable from the image, there may be many such features in the image offering marginal utility for directing a search. Such features are analogous to the boundary group level of features in the augmented aspect hierarchy. By examining the conditional probabilities in the augmented aspect hierarchy, we can compare the relative utility of boundary groups and faces in inferring the identity of a volumetric part.[5]

To compare the utility of boundary groups versus faces in recovering volumes, we will use the conditional probabilities captured in the augmented aspect hierarchy to define a measure of *average inferencing uncertainty*, or the degree to which uncertainty remains in volume identity given a recovered boundary group or face. More formally, we define average inferencing uncertainty for boundary groups, U^{BG}_{Avg}, and for recovered faces, U^{F}_{Avg}, as follows:[6]

$$U^{BG}_{Avg} = -\frac{1}{N_{BG}} \sum_{i=1}^{N_{BG}} \sum_{j=1}^{N_V} Pr(V_j \mid BG_i) \log Pr(V_j \mid BG_i) \tag{1}$$

$$U^{F}_{Avg} = -\frac{1}{N_{FA}} \sum_{i=1}^{N_F} \sum_{j=1}^{N_V} Pr(V_j \mid F_i) \log Pr(V_j \mid F_i) \tag{2}$$

where:

N_{BG} = number of boundary groups in the augmented aspect hierarchy
N_{FA} = number of faces in the augmented aspect hierarchy
N_V = number of volumes in the augmented aspect hierarchy

The average inferencing uncertainty for faces is 0.23 while that for boundary groups is 0.74. Clearly, faces offer a more powerful focus feature for the recovery of volumetric parts than do the simpler features that make up the boundary groups. However, this advantage is only realizable if the cost of extracting the two types of features is comparable. By using simple region segmentation techniques whose complexity is comparable to common edge detection techniques, we can avoid the complexity of grouping lines into faces. We can accommodate the segmentation errors associated with a cheap region grower by using partial information to intelligently guide viewpoint control to improve the interpretation.

[5] Since aspect recovery first requires the recovery of component faces, we will examine the choice between recovering simple contour-based features (boundary groups) and regions (faces)

[6] We have suppressed the zero-probability terms in this and remaining expressions for notational simplicity.

2.3 Focus of Attention

Our attention mechanism will exploit the augmented aspect hierarchy to map target objects down to target faces which, in turn, will be compared to image faces recovered during preprocessing. In selecting which recovered face to focus our attention on, we utilize a decision theoretic approach using a Baysian framework. A similar approach was reported by Levitt et al. [9], who use Baysian networks for both model representation and description of recovered image features. Specifically, they use Baysian networks for both data aggregation and selection of actions and feature detectors based on expected utility. The approach is thus centered around the use of a Baysian approach to integration and control. Similar techniques have also been reported by Rimey and Brown [13], and Jensen et al. [6], where both regions of interest and feature detectors are selected according to utility/cost strategies.

To select a region of interest, i.e., attend to a particular face, the augmented aspect hierarchy may be considered as a Baysian network, allowing us to utilize decision theory as described, for example, by Pearl [12]. To apply such a strategy, it is necessary to define both utility and cost measures. The utility function, U, specifies the power of a given feature at one level of the augmented aspect hierarchy, e.g., volumes, aspects, faces, and boundary groups, to discriminate a feature at a higher level. The cost function, C, specifies the cost of extracting a particular feature. The subsequent planning is then aimed at optimizing the benefit, $\max B(U, C)$; profit, e.g., $utility - cost$, is often maximized in this step.

For the system described in this paper, the face recovery algorithm was chosen to support a simple implementation on a Datacube image processor. From an input image, our bottom-up processing step yields a *face topology graph*, in which nodes encode the possible face interpretations of segmented image regions, and arcs encode region adjacency. For a given node, the face interpretations are ranked in decreasing order of probability; for details on face recovery, see [11]. Since the cost of face recovery is assumed to be constant and equal for all types of faces, the selection of which face to consider next should simply optimize the utility function.

Given a target object, $object_T$, the first step is to choose a target volume, $volume_T$, to search for. Next, given a target volume, $volume_T$, we choose a target aspect, $aspect_T$, to search for. Finally, given a target aspect, $aspect_T$, we choose a target face, $face_T$, to search for. Given a target face, $face_T$, we then examine the face topology graph for labeled faces which match $face_T$. If there is more than one, they are ranked in descending order according to their probabilities.

The above top-down sequence of predictions represents a best-first search of a tree defined by each object; the root of the tree represents the target object, while the leaf nodes of the tree represent target faces. The target volume subtrees for each object tree are independent of the object database and can be specified at compile time. The branching factor at a given node in any object tree can be reduced by specifying a probability (or utility) threshold on a prediction. The heuristic we use to guide the search is based on the power of an object's features, e.g., volumes, aspects, and faces, to identify the object. For example,

to determine how discriminative a particular volume, $volume_i$, is in identifying the target object, $object_T$, we use the following function:

$$D(volume_i, object_T) = \frac{Pr(object_T|volume_i)}{\sum_j Pr(object_j|volume_i)} * Pr(volume_i) \quad (3)$$

The numerator specifies how discriminative $volume_i$ is for $object_T$, while the ratio specifies the "voting power" of $volume_i$ for the object of interest. $Pr(object_i|volume_j)$, for any given i and j, is computed directly from the contents of the object database. The last term specifies the likelihood of finding the volume, and is included to discourage the selection of a volume which is highly discriminative but very unlikely. The $Pr(volume)$ may be calculated as follows:

$$Pr(volume_i) = \sum_k (Pr(volume_i|object_k) * Pr(object_k)) \quad (4)$$

where $Pr(volume_i|object_k)$, for any given i and k, is computed directly from the object database, and $Pr(object_k)$ represents a priori knowledge of scene content. $D(aspect_i, volume_T)$ and $D(face_i, aspect_T)$ are defined in an analagous fashion.

When we descend the search tree to a given target face, we search for matching face candidates in the face topology graph. We focus our attention on the best face matching the target face, and proceed to verify the object. If a target face, target aspect, or target volume cannot be verified, the search algorithm backtracks, applying the above utility functions to remaining faces, aspects, and volumes in the search tree.

2.4 Verification

Verification is the process by which we move from a matched target face node in the search tree back up to an object. Once we have a matched face leaf node, our next step is to verify its parent (target) aspect [3]. This entails searching the vicinity of the target face for faces whose labels and configuration match the target aspect using an interpretation tree search (Grimson and Lozano-Pérez [4]). Note that the resulting verified aspect has a score associated with it which can be compared to a score threshold to terminate the search from a particular target face. The score of a recovered aspect is calculated as follows:

$$AspectScore = \frac{1}{N} \sum_{k=1}^{N} Pr(Face_k) * \frac{Length(BG_k)}{Length(Region_k)} \quad (5)$$

where: N is the number of faces in model aspect, $Length(BG_k)$ is the length of boundary group, and $Length(Region_k)$ is the perimeter of the region. Note that if the region boundary graph recovered for the shape *exactly* matches some face in the augmented aspect hierarchy, its probability will be 1.0 and the length of its boundary group will be the perimeter of the entire region. The score of a volume is calculated as follows:

$$VolumeScore = AspectScore * Pr(ModelVolume|ModelAspect) \quad (6)$$

where:

$Pr(ModelVolume|ModelAspect)$ = probability of volume given aspect (from the augmented aspect hierarchy)

Once a target aspect is found, we then proceed up the tree one level to the target volume, defining a mapping between the faces in the target aspect and the surfaces on the target volume. Moving back one level to the object, we must then decide whether or not we have enough information confirming the target object. If so, the recognition process is complete. If not, we must then decide which volume to search for next. If we choose a volume which is connected to a volume we have already verified, we can move back down its branch in the tree and, when matching its target faces to image faces, consider only those image faces that are topologically adjacent to the faces belonging to the verified volume.

3 Viewpoint Control

Through segmentation errors, occlusion, or "accidental viewpoint"[7], a recovered aspect may be ambiguous. By extending our object representation, we can use the recovered aspect to drive the sensor to a new position from which the object's part can be disambiguated. The extended representation, called the *aspect prediction graph*, tells us which of the volume's aspects represents a "better" view of the volume, how the camera should be moved in order to achieve this view, and what visual events can be expected as the camera is moved.

The aspect prediction graph (APG) is derived from two sources: an aspect graph and the augmented aspect hierarchy. The APG is a more efficient version of the aspect graph in which topologically equivalent nodes are grouped regardless of whether their faces map to different surfaces on the object. For example, the APG encodes 3 aspects for a block (volume 1 in Figure 1) while an aspect graph encodes 26 aspects. Next, the APG specifies the visual events in terms of which faces appear/disappear when moving from one aspect to another. Furthermore, the position of such a face appearance/disappearance from a source aspect to a target aspect is specified with respect to particular contours of faces in the source aspect (event contours) Moreover, the transition between two nodes (aspects) encodes the direction(s) relative to the event contours that one must move in the image plane in order to observe the visual event. Finally, the APG borrows from the augmented aspect hierarchy both the $Pr(volume|aspect)$ and $Pr(aspect|volume)$ conditional probabilities, and assigns them to the nodes in the APG.

To illustrate the above concepts, Figure 3 presents the APG for the block volume, illustrating the three possible aspects of the block. Between every two nodes (aspects) in the aspect prediction graph are a pair of directed arcs. The directed arc between aspect 1 and aspect 2 in Figure 3(a) is expanded in Figure 3(b). From aspect 1 in Figure 3(a), there are three ways to move to a view

[7] The probability of an "accidental viewpoint" is actually quite significant, as was shown in [15].

in which aspect 2 will be visible. Movement relative to contours 0 and 1 on face 2 will cause a visual event in which face 2 disappears at contour 1 on face 0 and at contour 3 on face 1. Or, movement relative to contours 0 and 1 on face 0 will cause a visual event in which face 0 will disappear at contour 0 on face 1 and contour 0 on face 2. Finally, movement relative to contours 0 and 3 on face 1 will cause a visual event in which face 1 will disappear at contour 0 on face 0 and contour 1 on face 2.

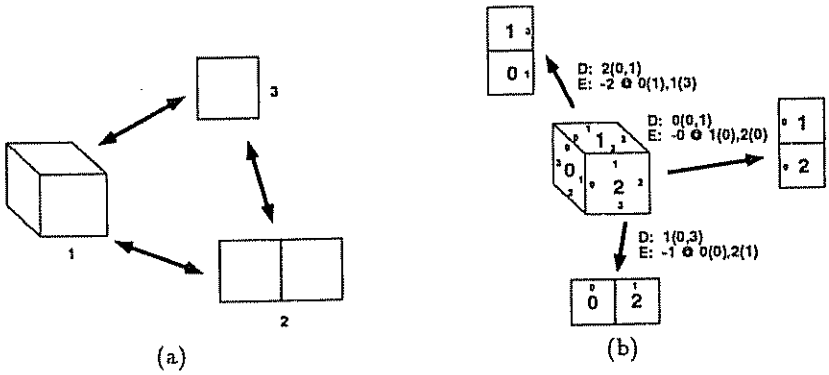

Fig. 3. (a) Aspect Prediction Graph (APG) for Volume 1 (Block) (b) APG Transitions from Aspect 1 to Aspect 2

It should be noted that in the augmented aspect hierarchy, each aspect has an indexing of its component faces, and each component face has a similar indexing of its bounding contours. By referring to the normals of such well-defined contours in a recovered aspect, we can qualitatively specify direction rules with respect to an aspect-centered coordinate system. The direction of view change is hence specified as a vector sum of the normals to particular contours of the recovered aspect corresponding to the current APG aspect.[8] The face events are also defined with respect to these specified contours. For example, we can predict along which contour in the current aspect a new face will appear or disappear when moving towards a new aspect.

Using the attention mechanism described earlier in section 2.3, the search for an object includes a search for its component volumes. Each recovered volume is characterized by the aspect in which it is viewed. For a given aspect of a volume, we can use the volume to aspect mappings in the augmented aspect hierarchy to determine which aspects (if any) are more probable (or stable) than the current one, by maintaining an ordered list of aspects for each volume, ranked in decreasing order of their downward conditional probabilities. Conversely, if we have an ambiguous aspect whose mapping to the hypothesized volume is weak, we can use the aspect to volume mappings in the augmented aspect hierarchy to determine which aspects offer a less ambiguous mapping to that volume. These

[8] For concave and convex curve segments, the normal at the midpoint is used.

aspects, ranked in decreasing order of their upward conditional probabilities, offer an effective means of disambiguating a given view of a volume.

When we want to move the camera in a direction to get a "better" view, we first check the APG to see which aspects (neighboring nodes) can be reached from the current aspect (node). The probabilities associated with the APG nodes tell us to which aspect to move in order to achieve a more likely view of the volume or to disambiguate it. The arc to this "best" neighbor node encodes the view change direction (in the image plane) in terms of a function of the normals of selected aspect contours. We calculate the values of these normals in the image and get a direction for camera movement with respect to the current aspect

While moving the camera, we must track the aspect from one frame to the next so that we can verify the visual events as specified in the APG. For the experiments described in this paper, we will assume a fixation mechanism which can track a region through successive frames. Our visual event verification strategy will therefore consist of focusing the attention mechanism at the tracked region and searching for predicted aspect. The recovered aspect can then be compared to the original aspect to verify the expected visual events. Tracking the object between frames is beyond the scope of this paper and is described in [2].

4 Results

We test the attention and viewpoint control strategies in the context of a multi-disciplinary research effort exploring active vision in the domain of robotic aids for the handicapped (PLAYBOT). Through a touch-screen interface, a child can instruct a mobile robot vision system to identify, localize, and manipulate 3-D objects in its environment. To support simple manipulation of the objects, the domain of objects that the system can visually identify consists of the ten volumetric shapes outlined in Figure 1[9]; more complex objects, modeled as constructions of the ten shapes, will be supported in the future. In the following results, the images were acquired using the stereo head at the CVAP Laboratory at KTH, Stockholm. Only one camera was used to acquire images and during viewpoint control, the camera was fixated on the object.

In Figure 4, we present the results of applying the attention mechanism to a scene containing single-volume objects. Moving top to bottom and left to right, the first image shows the results of the region segmentation step; recall that the face topology graph constructed from the region topology graph is the input to the attention mechanism. The next three images show the three best instances of the block viewed in its most likely aspect containing three faces. The faces in the aspect are highlighted in the image; only those contours (boundary group) used in defining the face are highlighted in the face.

Using this measure, the first three volumes received the score of 1.0, 1.0, and 0.86, respectively. The next three images show the best three instances of the

[9] Each of the ten objects is assumed to be equally likely

second most likely aspect containing two faces; each recovered aspect received the score of 0.48. Continuing, the next two images show the best two instances of the least likely aspect containing one face; each recovered aspect received the score of 0.22. The last three images show the highest-scoring instances of the tapered block (0.79), the cylinder (0.27), and the barrel (0.46), respectively. Due to noise and occlusion, only certain portions of each shape were recovered. In the case of the barrel, region undersegmentation results in the recovered aspect being incorrectly oriented with the visible end assumed to be occluded at the bottom.

In Figure 5, we show the result of the attention mechanism as it searches for a block (first image) and a cylinder (third image). The block is recovered in its second most likely aspect which is ambiguous (common to volumes 1 and 4), while the cylinder is recovered in its second most likely aspect which is ambiguous (common to volumes 1, 2, 3, 4, 5, and 10). Guided by the aspect prediction graph, the camera is moved to the left in each case and the attention scheme is guided to disambiguate the volume by searching for its most likely aspect, as is shown in the second and fourth images.

5 Conclusions

In examining the balance between recovery and verification, we have clearly moved towards recovery. Recovering more discriminating features facilitates an effective attention mechanism based on the viewing probabilities in the aspect hierarchy. However, there is a cost in attempting to recover more complex features (in our case, a set of regions and their bounding shapes). Our solution to this problem is to pass along this cost to a dynamic sensor. We assume that some relatively unoccluded, fronto-parallel surfaces will project into regions that can be quickly and cheaply extracted using simple region segmentation techniques. We use this recovered partial information to intelligently guide the sensor to a position where the object can be disambiguated. The aspect hierarchy, and its extension to the aspect prediction graph, provides a unifying representation for the active vision problems of attention and viewpoint control.

References

1. I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.
2. S. Dickinson, P. Jasiobedzki, H. Christensen, and G. Olofsson. Qualitative tracking of 3-D objects using active contour networks. In *Proceedings, CVPR '9* Seattle, June 1994.
3. S. Dickinson, A. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992.
4. W. Grimson and T. Lozano-Pérez. Model-based recognition and localization from sparse range or tactile data. *International Journal of Robotics Research*, 3(3):3–3 1984.

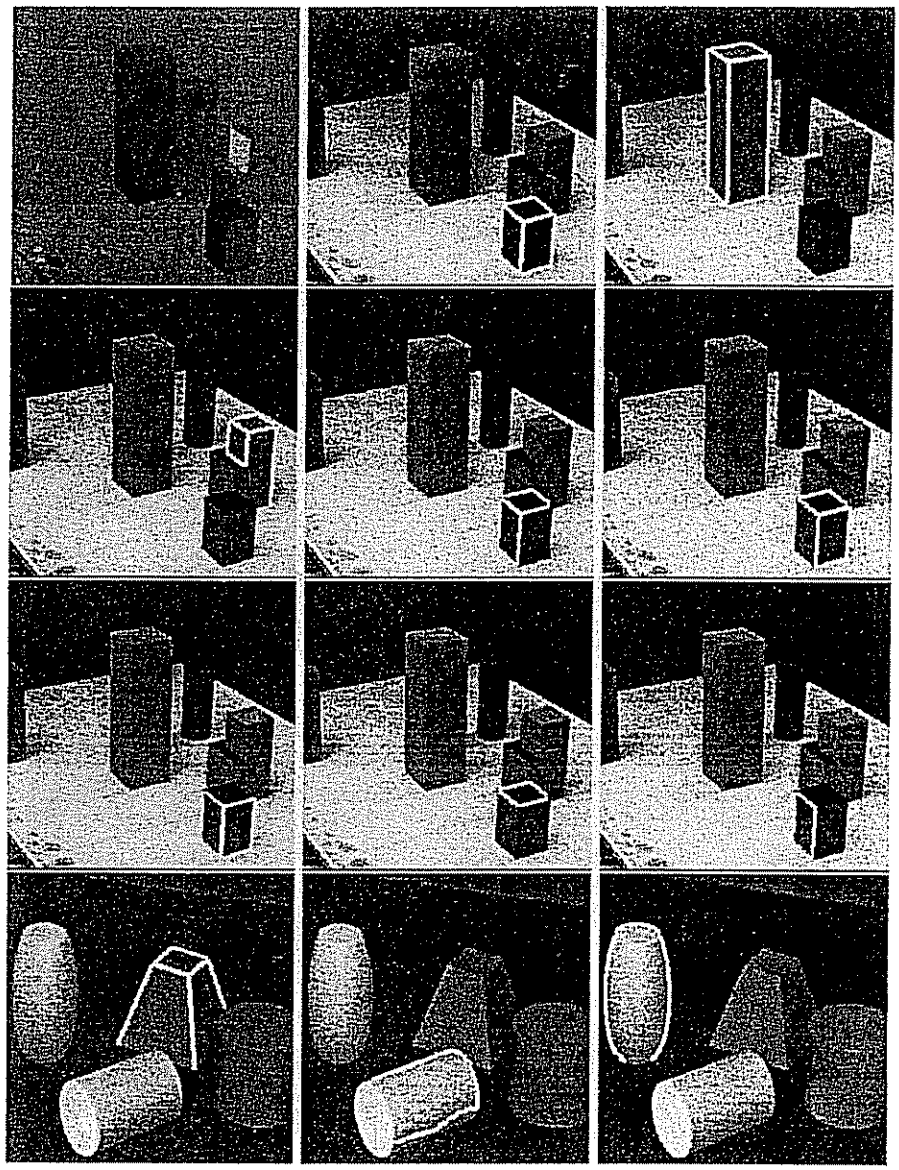

Fig. 4. Demonstration of the Attention Mechanism. The top three rows show the results of searching for Volume 1 in Figure 1. The last row shows the results of searching for Volumes 2, 5, and 8. (see text for explanation)

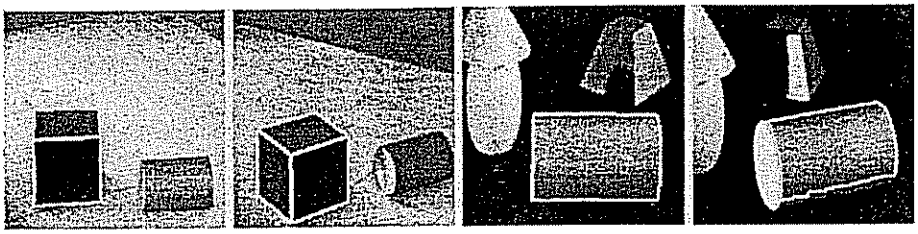

Fig. 5. Moving the Sensor to Disambiguate Recovered Volumes

5. D. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.
6. F. Jensen, H. Christensen, and J. Nielsen. Baysian methods for interpretation and control in multiagent vision systems. In K. Bowyer, editor, *SPIE Applications of AI X: Machine Vision and Robotics*, volume 1708, pages 536–548, Orlando, FL, April 1992.
7. J. Koenderink and A. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216, 1979.
8. Y. Lamdan, J. Schwartz, and H. Wolfson. On recognition of 3-D objects from 2-D images. In *Proceedings, IEEE International Conference on Robotics and Automation*, pages 1407–1413, Philadelphia, PA, 1988.
9. T. Levitt, J. Agosta, and T. Binford. Model based influence diagrams for machine vision. In M. Herion, R. Shacter, L. Kanal, and J. lemmer, editors, *Uncertainty in Artificial Intelligence 5*, volume 10 of *Machine Intelligence and Pattern Recognition Series*, pages 371–388. North Holland, 1990.
10. D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, 1985.
11. D. Metaxas and S. Dickinson. Integration of quantitative and qualitative techniques for deformable model fitting from orthographic, perspective, and stereo projections. In *Proceedings, Fourth International Conference on Computer Vision (ICCV)*, Berlin, May 1993.
12. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, Inc., 1987.
13. R. Rimey and C. Brown. Where to look next using a bayes net: Incorporating geometric relations. In G. Sandini, editor, *European Conference on Computer Vision (ECCV)*, volume 588 of *Lecture Notes in Computer Science*, pages 542–550. Springer Verlag, May 1992.
14. J. Tsotsos. On the relative complexity of active vs passive visual search. *International Journal of Computer Vision*, 7(2):127–141, 1992.
15. D. Wilkes, S. Dickinson, and J. Tsotsos. Quantitative modeling of view degeneracy. In *Proceedings, 8th Scandinavian Conference on Image Analysis*, University of Tromsø, Norway, May 1993.