The IEEE International Conference on Robotics and Automation. March, 1985. pp. 48-54.

Recognizing Partially Hidden Objects[1]

by

J. L. Turney, T. N. Mudge and R. A. Volz[2]

Abstract

In this paper, an approach is described for recognizing and locating partially hidden objects in an image. The method is based upon matching pairs of boundary segments of the template of an object with pairs of boundary segments in the image. Using a Bayesian based signal detection approach, pairs of segments are selected from the template of the object such that the probability of correctly identifying the object given that the pair is matched in the image is close to one. Assuming that models of all objects which might appear in the scene (a reasonable assumption for industrial applications) are known a priori, suitable pairs of segments can be determined a priori. Preliminary investigation suggests that the technique is robust and that subsecond recognition time can be achieved.

INTRODUCTION

A problem of great practical interest in machine vision is the recognition of objects that are partially hidden. For example, consider assembling a kit from parts dumped on a table, or extracting a part from a bin of parts. In both of these cases it is likely that some of the objects will be partially hidden from the view of the camera because others are lying on top of them. In this paper we will present a method for recognizing objects that are partially hidden. The discussion will be limited to recognizing flat two dimensional untilted objects. Each object will then have only two possible views, one for each flat side.

The method uses overlapping boundary segments and requires that all the segments of each view of all the objects that can occur in a scene be known a priori. The method is particularly suitable for an industrial environment where the objects and their geometry are known beforehand. An off-line training procedure selects, using the a priori information, configuration pairs of segments that can uniquely define the position and orientation (pose) of an object given the other objects that can occur in an image. A configuration pair is simply a pair of segments and their pose with respect to each other. Clearly, a segment that occurs in a large number of the configuration pairs of a particular object is more likely to be useful in recognizing that object, assuming that partial occlusion occurs randomly. We adopt the terminology of [TMV84] and refer to such segments as *salient* segments. The degree of saliency of a segment s can be quantified by counting the number of configuration pairs that s appears in. The set of segments that appear together with s in configuration pairs is termed the *coset* of s; the degree of saliency of s is the cardinality of coset(s). It is important to note that the degree of saliency of the segment of an object is dependent on the other objects in the a priori set of objects which can occur in an image.

[1]This work described in this paper was supported in part by Air Force Contract No. F49620-82-C-0089 and Army Research Office Contract No. DAAG29-84-K-0070.

[2]The authors are with the Robot Systems Division, Center for Robotics and Integrated Manufacturing, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, (313) 764-0203.

The recognition procedure searches for a object in a image of partially occluded objects by first trying to locate the most salient segment of that object in the image. If the most salient segment is found, configuration pairs containing it are sought. As soon as a configuration pair is located the pose of the object is assumed to be completely recognized, since the training procedure produces only those pairs that can uniquely identify an object. If the most salient segment cannot be found, or if none of the configuration pairs containing the segment can be found, the recognition procedure continues with the next most salient segment.

Many of the concepts developed in this paper are based on earlier work by the authors and others. A brief summary of some of this work follows.

Underlying the segment matching that is required as part of our overall recognition procedure is a generalization of the Hough transform. Work on this transform goes back a number of years. Duda and Hart [DuH72] used a version of the Hough transform to locate portions of the boundary of an object from edge points in an image. In the Hough transform an image boundary is located by constructing a parameterized version of the boundary and determining which parameters are most consistent with the image points. The set of parameters that describe a boundary can be regarded as a point in a multidimensional parameter space. For each image point the locus of all the parameter points that correspond to the image point (i.e., those parameters of boundaries that pass through the image point) is recorded. This is repeated for all image points. Parameter space is partitioned into discrete regions and a histogram is constructed that counts the number of loci passing through each region. The location of the peak in the histogram yields a region whose associated parameter point is considered to correspond to the boundary most consistent with the image. This work inspired considerable subsequent work. One of the main themes of this later work was to speed up the procedure and reduce the number of false peaks that could occur in the histogram by reducing the number of loci through the use of constraints (in the original form of the transform every image point is assumed to belong to every possible loci that can pass through it).

Ballard [Bal81] developed a restricted form of the generalized Hough transform that can be used for recognizing partially hidden objects. In the generalized Hough transform, a template of the object is parameterized by its location and orientation. A Hough approach is then used to determine these parameters. The generalized Hough has been shown to be an efficient form of template matching. Ballard restricted the generalized Hough approach by histograming only those parameters which allow the template to pass through the image point with the same slope as the image point. This work represented a significant improvement on earlier Hough transform based techniques. However, incorrect determination of the location and orientation of an object still occurs when the degree of occlusion is high [TMV84].

Perkins [Per78] developed a method for hidden part recognition that sought matches for straight line and circular arc segments that he referred to as "concurves." The matching was done in the slope angle-arclength representation of the boundaries. (The slope angle-arclength representation is discussed in [Bal82] and in the section on matching.) The concurves were determined from templates of the

objects during training. This approach allowed objects to be located that were partially hidden provided the degree of occlusion was not high. A high degree of occlusion hinders concurve matching, since concurves are generally large segments of the boundary.

Bolles and Cain [BoC82] developed an approach for hidden object recognition referred to as "local feature focus." During training a set of easily identifiable features, referred to as "focus features," were located in an object. A list of neighboring features which distinguish a focus feature from other features was compiled for each focus feature. During run-time a set of correspondences was established between image features and object features in the neighborhood of a focus feature. A graph was formed with the set of correspondences as nodes and the consistencies between correspondences as edges. A graph matching technique, the maximal clique algorithm, was used to locate the largest cluster of mutually consistent correspondences of the object features to the image features. Once a focus feature was located the orientation and location of the object could be determined. As noted in [BoC82] the inherent weakness of this approach is its reliance on detecting local neighborhoods of features; the local neighborhoods must be non-occluded to be useful in recognition. Our work extends this approach by relying on configurations of features that are more general than local neighborhoods of focus features, moreover we avoid the need for consistency checking.

The approach presented in [Seg83] matched extrema in curvature in the boundary of the image to extrema in the boundaries of templates of the objects. The approach works with global information in the following sense. A global orientation of the object was first determined by histograming the difference in orientation of each extrema in the image boundary with respect similar extrema in the template boundary, then a global translation along the x axis was determine by histograming differences in the x location of extrema in the image with respect to similar extrema in the template, and finally a global y location was determined in a similar fashion to the x location. The method achieved significant speed at the expense of accuracy. If similar objects or objects having similar extrema in different configurations appear in the image, the technique breaks down. Accurately determining the second derivative of the boundary to identify extrema is also a source of error.

The approach presented in [ABB84] approximated the boundary of an object by polygons. A linear segment taken from an approximation of the object is matched to a linear segment generated from an approximation of the image boundary. In order to reduce the number of such matches "preferred" segments were chosen that would occur with low frequency in the image. These preferred segments were generally those of longer length. If a match occurred, a hypothesis was generated for the possible location of the object from a comparison of the difference in pose of the object and image segments. Once a match between segments was found, neighboring segments of the object were compared to neighboring segments of the image to determined consistency. If they were consistent, a Kalman filtering technique was used to update the estimated pose of the object from the information gathered from the comparison of the neighboring segments. This technique combines good accuracy with speed, however, two drawbacks are its reliance on the polygon approximation and the use of preferred segments. If an image boundary is noisy, the polygon derived from it may differ significantly from that derived from the objects template. If the preferred segments are occluded, which is highly likely because they were chosen as the longest segments, the number of matches that must be perform grows rapidly.

Turney et al [TMV83], [TMV84] matched fixed length template contour segments to image boundary segments of the same length in a space where the slope angle of a contour is parameterized by its arclength. Templates segments were weighted according to their "saliency." The algorithm was able to recognize objects even when they were heavily occluded, but required a large amount of off-line computation.

The work presented in [BhF84] used a two stage hierarchical stochastic labeling method for matching a object templates to the image boundary. They approximated the template and the image boundary by polygons. Associated with each image polygon segment were two probability vectors: one vector whose elements were the probabilities that the segment could be labeled as each of the segments of the template, and a second vector whose elements represented the compatibility of the neighbors of the image segment with the neighbors of each of the template segments. A two stage optimization technique was used to maximize a global criterion which maximized consistency while minimizing ambiguity. The algorithm is computationally intensive.

In [BoC82], [TMV83], and [TMV84] emphasis was placed on the context in which an object is found. Bolles and Cain chose local features which were unique to an object and its pose to disambiguate the object from other objects and other poses of the same object. Turney et al used extended features, i.e., the boundary segments of the object, to provide unique identification of the object of interest. In industrial applications one generally knows the number and exact shape of the objects that are to appear in a scene, and it is advantageous to use this information to distinguish objects. This paper presents preliminary results from an algorithm that also uses this contextual information. The next section explains the hidden object recognition method in term of signal detection theory. The subsequent sections discuss matching template segments of objects to image segments, training and recognition.

OBJECT RECOGNITION AS SIGNAL DETECTION

In this section we present a Bayesian based signal detection view of the recognition procedure, based on configuration pairs, that was outlined in the Introduction. Assume that the boundary of an object, object 1, appears in a scene with its location and orientation about an origin as shown in Fig. 1a. In the terms of signal detection theory this boundary represents a signal, $S^1(x_0,y_0,\phi_0)$, derived from object 1 when the object is placed at location (x_0,y_0) with orientation ϕ_0. This signal is transmitted to the camera and forms part of the image. Any rotation or shift of this boundary represents a different signal $S^1(x,y,\phi)$, where x, y, and ϕ represent a different pose from x_0, y_0, and ϕ_0. The rotations and shifts of other boundaries of other objects that can appear in the image correspond to different signals, $S^i(x,y,\phi)$, where $i \neq 1$.

Assume that R, shown in Fig. 1b, represents two boundary segments that have been extracted from an image. Treating R as a received signal we seek the probability that it identifies $S^1(x_0,y_0,\phi_0)$ as the signal sent, i.e., $Pr\ [\ S^1(x_0,y_0,\phi_0)$ sent $|\ R$ received $]$. This can be determined from

$$Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ ] = \frac{Pr\ [\ R\ |\ S^1(x_0,y_0,\phi_0)\ ]Pr\ [\ S^1(x_0,y_0,\phi_0)\ ]}{Pr\ [\ R\ ]} \quad (1)$$

R can be produced in many ways. In particular, R can be produced by an accidental alignment of segments of different objects (see Fig. 2). In this paper the possibility of accidental alignment is ignored, then

$$Pr\ [\ R\ ] \approx \sum_{i,x,y,\phi} Pr\ [\ R\ |\ S^i(x,y,\phi)\ ]Pr\ [\ S^i(x,y,\phi)\ ] \quad (2)$$

where $S^i(x,y,\phi)$, for all i, x, y and ϕ, represents all the possible signals that can be sent, i.e., all the boundaries of all the segments in all poses. Assume that R consists of two segments of fixed arclength, a_0. Call these r_1 and r_2 (see Fig. 1b), then (1) can be rewritten as

$$Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ ] \approx \frac{Pr\ [\ r_1r_2\ |\ S^1(x_0,y_0,\phi_0)\ ]Pr\ [\ S^1(x_0,y_0,\phi_0)\ ]}{\sum_{i,x,y,\phi} Pr\ [\ r_1r_2\ |\ S^i(x,y,\phi)\ ]Pr\ [\ S^i(x,y,\phi)\ ]} \quad (3)$$

Assume that all signals have equal a priori probability. Then (3) becomes

$$Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ ] \approx \frac{Pr\ [\ r_1 r_2\ |\ S^1(x_0,y_0,\phi_0)\ ]}{\sum_{i,x,y,\phi} Pr\ [\ r_1 r_2\ |\ S^i(x,y,\phi)\ ]} \quad (4)$$

Let s_j^i represent a segment of $S^i(x,y,\phi)$ of arclength a_0, and let $\sigma^i(x,y,\phi)$ represent the set of all configuration pairs (ordered pairs) of segments, $<s_j^i,s_k^i>$, that can be produced from $S^i(x,y,\phi)$. The probability that $<s_{j_1}^i,s_{j_2}^i> \in \sigma^i(x,y,\phi)$ is received as the pair of segments $<r_1,r_2>$ is the probability that noise resulting from sampling and quantization distorts $s_{j_1}^i$ and $s_{j_2}^i$ into r_1 and r_2 respectively. Let δ be a metric which measures the "distance" between configuration pairs of segments. We approximate the probability $Pr\ [\ r_1 r_2\ |\ S^i(x,y,\phi)\ ]$ by the value 1 if there exists a configuration pair of segments, $<s_{j_1}^i,s_{j_2}^i> \in \sigma^i(x,y,\phi)$, such that $\delta(<s_{j_1}^i,s_{j_2}^i>,<r_1,r_2>) < d_0$ (d_0 is a fixed threshold distance), and by 0 otherwise. Further assume that the reception of $<r_1,r_2>$ depends only on the configuration pair $<s_{j_1}^i,s_{j_2}^i>$. Then (4) becomes

$$Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ ] \approx \frac{Pr\ [\ r_1\ r_2\ |\ s_{j_1}^1\ s_{j_2}^1\ ]}{\sum_{i,x,y,\phi} Pr\ [\ r_1\ r_2\ |\ s_{j_1}^i\ s_{j_2}^i\ ]}. \quad (5)$$

Upon reception of R, it would be possible to estimate the the probability that $S^1(x_0,y_0,\phi_0)$ was sent during run-time. However, given the large number of possible signals, this would take a significant amount of computation. Instead, we adopt a simpler but less accurate approach which allows the bulk of the computation to be performed in an off-line training phase.

In order to eliminate run-time calculations it is necessary to eliminate the dependence of $Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ ]$ on the received signal R. From our approximation the numerator of (5) is zero unless $\delta(<s_{j_1}^1,s_{j_2}^1>,<r_1,r_2>) < d_0$. Terms containing the pair $<s_{j_1}^i,s_{j_2}^i>$ in the denominator will be zero unless $\delta(<s_{j_1}^i,s_{j_2}^i>,<r_1,r_2>) < d_0$. Thus, if $Pr\ [\ r_1\ r_2\ |\ s_{j_1}^i\ s_{j_2}^i\ ]$ is to contribute to the denominator, the largest distance that $<s_{j_1}^i,s_{j_2}^i>$ can be from $<s_{j_1}^1,s_{j_2}^1>$ is $2d_0$. We will include only terms in the denominator that have configuration pairs within $2d_0$ of $<s_{j_1}^i,s_{j_2}^i>$, and set their values to 1. This approximation is made, noting that its effect is to sometimes increase the value of the denominator above the value that would be estimated at run-time. This simply lowers the estimate of the probability to a more pessimistic value. In practice it has been found to work well. With this approximation the denominator no longer depends upon R. Rather it depends upon $s_{j_1}^1$ and $s_{j_2}^1$, which in turn depend upon x_0, y_0, and ϕ_0. Denote this approximate denominator by $D_{12}^1(x_0,y_0,\phi_0)$.

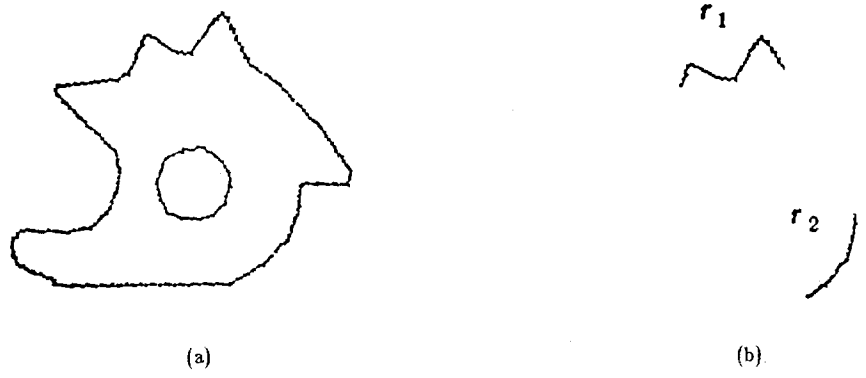


Figure 1. Signal sent and signal received.

Figure 2. Accidental alignment.

With these approximations $D_{12}^{1}(x_0,y_0,\phi_0)$ can be calculated off-line, and the probability $Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ ]$ can be estimated during run-time by

$$Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R\ = <r_1,r_2>\ ] \approx \frac{1}{D_{12}^{1}(x_0,y_0,\phi_0)} \quad (6)$$

In accordance with our approximation the numerator, $Pr\ [\ r_1\ r_2\ |\ s_{j_1}^{1}\ s_{j_2}^{1}\ ]$, takes on the value 1 in (6).

Since the calculation of $D_{12}^{1}(x,y,\phi)$ depends only upon a relative distance metric for pairs of objects on the same object, $D_{12}^{1}(x,y,\phi)$ is independent of the pose of the object. Therefore, in the following, the notation will reflect this and $D_{12}^{1}(x,y,\phi)$ will be shortened to D_{12}^{1}.

The counting method of determining D_{12}^{i} for all i is discussed in the section on training.

This Bayesian approach of estimating whether or not a signal has been sent given the reception of a pair of segments is used together with the matching approach to locate partially hidden objects. Matching is discussed in the next section.

MATCHING SEGMENTS

A critical phase of the procedure for locating an object in a scene of partially hidden objects involves matching segments from the template of the object to be located to segments in the image of the scene. The approach used in this work has been discussed in detail in [TMV83] and [TMV84], and is summarized here.

The template and image boundaries are represented in two spaces, in normal cartesian space and in slope angle-arclength space, or θ–a space (see Fig. 3). The template and image boundaries in both θ–a space and cartesian space are partitioned into segments of fixed arclength a_0.

Matching is performed in θ–a space since it is more efficient than matching in cartesian space. Rotations in cartesian space become offsets in θ–a space.

During matching a θ–a representation of the template segment (shown with a heavy line in Fig. 4) is moved along the s axis so that its center is aligned with the center of the image segment to which it is to be compared. The template segment is then shifted in the θ direction so that the mean θ value of the template segment has the same mean θ value as the image segment. This θ shift (see Fig. 4) measures the average slope angle difference between the template and image segments and will be referred to as the "angle of match." The difference in θ is found between corresponding points of the

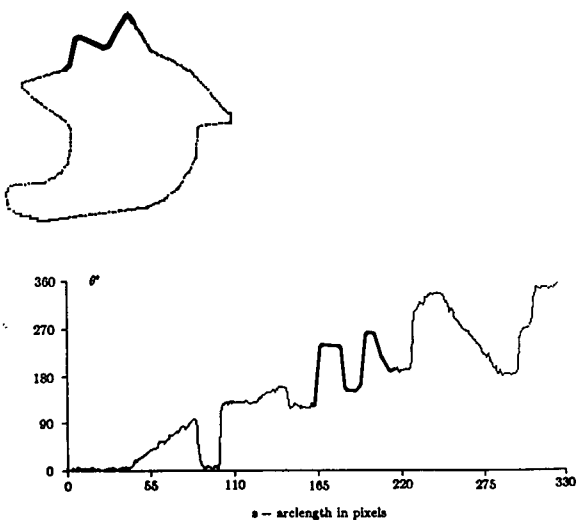


Figure 3. Cartesian and θ–a representation of an object.

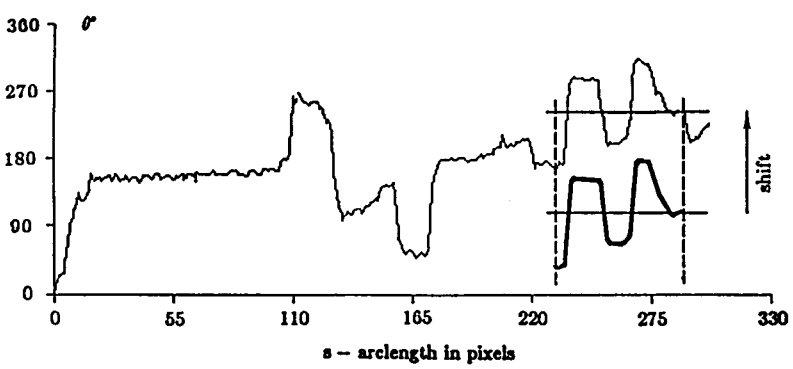


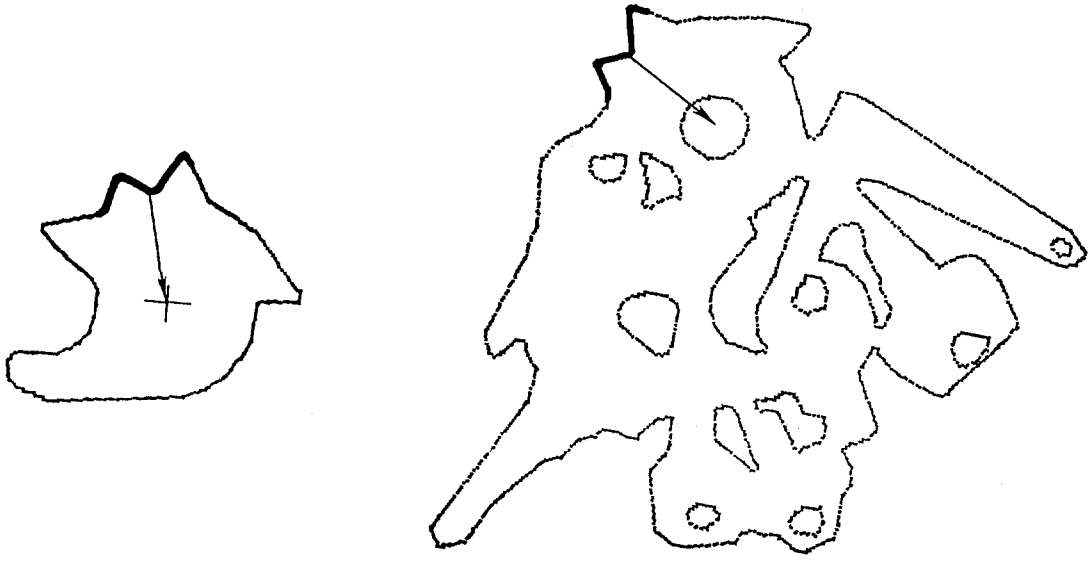
Figure 4. Matching in θ–a space.

template and image segment. The inverse of sum of the squares of these differences is used to measure the similarity of the two segments. If they are similar, as determined by a function of the threshold d_0, they are assumed to match.

If the template and image segments match in θ–a space, the match is recorded as follows. In cartesian space a vector from the center of the template segment to the template centroid is determined. This vector is is rotated by the "angle of match" and translated so that its tail is centered at the same location as the center of the image segment (see Fig. 5a). The location of the head of this vector represents a potential location of the centroid of the template in the image. Each pixel location in the image has an associated list. If the head of the vector falls on a particular pixel, a record containing the identity of the template segment and the angle of match is stored in the list at that pixel location (see Fig. 5b).

of the next object, generating new lists of match records. These lists are again analyzed, possibly resulting in D_{12}^1 being further incremented. All other objects are matched and all possible contributions to D_{12}^1 are counted.

In our preliminary implementation only configuration pairs with denominators equal to 1 were output in a table as part of the training phase. The table is termed the training table and is indexed by the subscripts of the configuration pair. Entries in the training table are considered to be the configuration pairs that can uniquely determine the pose of their associated object. The saliency of each segment can be determined from the table. It is the cardinality of the coset of each distinct segment that occurs in any of the configuration pairs in the table.



(a) (b)

Figure 5. Storing a record of the match.

TRAINING

The denominators for the conditional probabilities of an object are trained off-line. The template of the object to be trained is matched to templates of all of the objects (including itself) that can appear in the image. From this matching information one can determine the denominators of the conditional probabilities.

The object whose conditional probabilities are to be determined is termed the training object. As before let D_{12}^1 denote the denominator term for the conditional probability when $<s_{j_1}^1, s_{j_2}^1>$ form the configuration pair. The calculation proceeds as follows. The segments of the template of the training object are matched to the segments of the template of one of the objects. After matching, the list of records at each pixel location is examined. If the list at a pixel location or the list of any nearby pixel contains a record of a match by segment $s_{j_1}^1$ and a record of a match by $s_{j_2}^1$ at approximately the same angle of match, then D_{12}^1 is incremented by 1. After all lists generated for this match have been examined they are disposed of and the template of the training object is matched to the template

LOCATING PARTIALLY HIDDEN OBJECTS

When locating an object, for example object 1, a segment, $s_{j_1}^1$, of the template of the object is matched to the segments of the boundaries in the image using the approach discussed previously. The most salient segments are matched first according to the strategy outlined in the Introduction. When a template segment of the object matches a image segment, a record of the match is stored in a list associated with a pixel at the location of a possible centroid of the template. Then the list at that pixel and of all neighboring pixels are examined to see if there exists any previous record of a match with another template segment, say $s_{j_2}^1$, at the same match angle.

If such a record exits, the training table is examined using j_1 and j_2 as indices to find if this configuration pair is present. If this is the case then object 1 has been located.

RESULTS

Fig. 6 shows the boundaries of the set of objects that were used during training in our experiments. Fig. 7 shows an example of the joint conditional probability, $Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R = <r_1,r_2>\ ]$. The bullet indicates the center of segment $s^1_{j_1}$. A vertical line is drawn from each possible center of $s^1_{j_2}$. The length of the line is proportional to the joint conditional probability $Pr\ [\ S^1(x_0,y_0,\phi_0)\ |\ R = <r_1,r_2>\ ]$ that one would obtain if $s^1_{j_1}$ were centered about each of these possible locations.

Figure 8 illustrates the recognition of two objects from a pile of parts. Preliminary estimates indicate subsecond recognition times on an Apollo 660 workstation.

SUMMARY

In this paper it has been shown that a Bayesian approach, together with template segment matching in θ–α space can be used as an effective approach to locate partially hidden objects. A key assumption was that the received pair of segments, $<r_1,r_2>$ were not an accidentally alignment of two segments each from a different object, but were received from a single object. In a real industrial scene, particularly a bin of parts situation where there are many copies of the same part, this assumption is likely to be violated. In these cases some configurations may require more than a pair of segments.

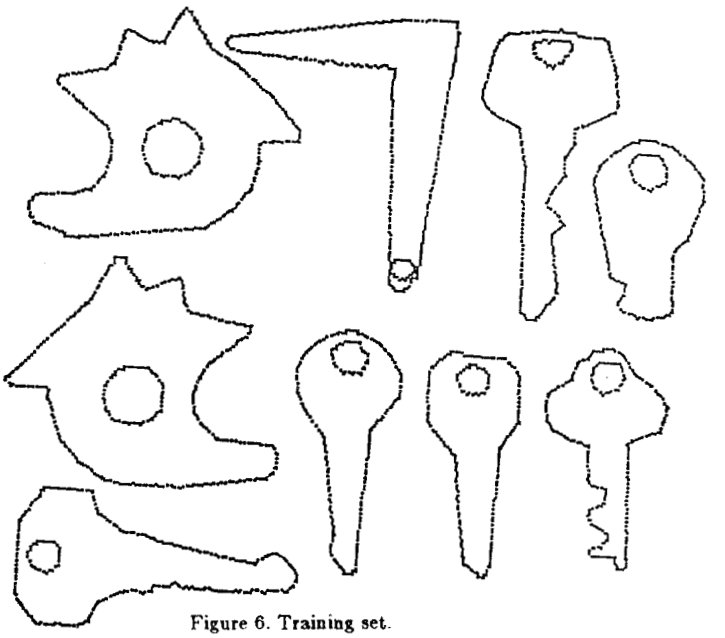


Figure 6. Training set.

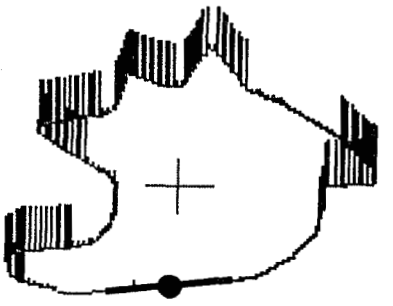


Figure 7. Joint conditional probability.

REFERENCES

[ABB84] N. Ayache, J. D. Boissonnat, B. Bollack, and B. Faverjon, "Automatic Handling of Overlapping Workpieces", *I.C.P.R.*, Montreal, 1984.

[Bal81] D. H. Ballard, "Generalizing the Hough Transform to Detect Arbitrary Shapes," *Pattern Recognition*, vol. 13, no. 2, 1981, pp. 111-122.

[Bal82] D. H. Ballard, C. M. Brown, *Computer Vision*, Englewood Cliffs, N.J: Prentice Hall, 1982, p. 237.

[BhF84] B. Bhanu and O. D. Faugeras, "Shape Matching of Two-Dimensional Objects," *IEEE Pattern Analysis and Image Processing*, vol. 6, March 1983, pp. 137-155.

[BoC82] R. C. Bolles, R. A. Cain "Recognizing and Locating Partially Visible Workpieces", *Proceedings IEEE Conference on Pattern Recognition and Image Processing*, June 1982, pp. 498-503.

[DuH72] R. O. Duda and P. E. Hart, "Use of the Hough Transform to Detect Lines and Curves in Pictures," *CACM*, vol. 15, Jan. 1972, pp. 11-15.

[Per78] W. A. Perkins, "A Model-based Vision System for Industrial Parts," *IEEE Transactions on Computers*, vol. C-27, Feb. 1978, pp. 126-143.

[TMV83] J. L. Turney, T. N. Mudge and R. A. Volz, "Experiments in Occluded Parts Recognition," *Proceedings of the Society of Photo-optical Instrumentation Engineers Cambridge Symposium on Optical and Electro-optical Engineering*, Cambridge, MA, Nov. 1983, pp. 719-725.

[TMV84] J. L. Turney, T. N. Mudge and R. A. Volz, "Recognizing Partially Occluded Parts," *IEEE Pattern Analysis and Image Processing*, (in review).

[Seg83] J. Segen, "Locating Randomly Oriented Objects from Partial View," *Proceedings of the Society of Photo-optical Instrumentation Engineers Cambridge Symposium on Optical and Electro-optical Engineering*, Cambridge, MA, Nov. 1983, pp. 676-684.

Figure 8. Objects located.