An Analysis of Lowe's Model-based Vision System

Alan M. M^{c}Ivor
Robotics Research Group,
Department of Engineering Science,
University of Oxford,
Oxford OX1 3PJ,
U.K.

This paper reports on the experiences gained with a re-implementation of Lowe's model-based vision system. This scheme was found to be robust and efficient. However, a number of deficiencies were encountered. In particular, the initial viewpoint estimation procedure is weak and needs improvement.

INTRODUCTION

In a series of papers [1,2,5], David G. Lowe has reported on the development of the SCERPO model-based vision system. This vision system is able to identify and locate instances of one or more predefined models in a single image. In this paper, we present and discuss results obtained from a re-implementation of this model-based vision system.

BRIEF OVERVIEW OF LOWE'S METHOD

Like several other model-based vision systems, the initial phase of Lowe's method involves applying an edge detector to the input image and extracting line segments from the result. This gives a set of k observed segments. These segments have to be matched to the n segments in the model. This matching process can be viewed as a search through an "interpretation tree" containing $(n+1)^k$ nodes [7]. For any reasonable values of k and n, this tree contains a very large number of nodes. A sophisticated search strategy is needed to complete this search in a reasonable time.

The search strategy developed by Lowe starts by grouping the observed line segments into significant multi-segment structures called perceptual groups. These "observed" groups are then matched to groups known to exist in the model. By initially matching groups instead of individual segments, the interpretation tree is pruned to a huge extent. Furthermore, since perceptual groups containing three or more segments are used to initialize the search, at each stage of the search it is possible to solve for the object's pose parameters (i.e. location and orientation). Using this estimate of pose, the position of each model segment's image can be predicted. Subbranches of the interpretation tree are only considered if the observed segments are in close proximity to the predicted image of their assigned model segment. These two factors lead to a very efficient search strategy.

Perceptual Organization

The technique of perceptual grouping is based on observations of the human visual system [1]. It involves grouping segments in the image into structures that are perceptually significant. The significance of an observed structure is measured by the probability that the structure arises by an accidental alignment of randomly located segments. As the probability of accidental alignment increases, the perceptual significance decreases. The primary perceptual groups that are found in SCERPO are segments which have endpoints in close proximity, segments that are collinear and segments that are parallel. These groups are themselves grouped into the more complex structures of parallelograms, trapezoids and skew-symmetries.

Not all pairings of observed segments are evaluated as possible primary perceptual groups. It was noted [2] that the measure of accidental matching for all types of primary group is directly proportional to the proximity of the segments. Hence, only the segments in a finite neighbourhood of each segment need be considered. Doing so reduces the search space. The search is also made more

AVC 1988 doi:10.5244/C.2.12

efficient by indexing all segments by endpoint location.

The calculation of the probability of an accidental match can be illustrated by considering the case of endpoint proximity. Given an endpoint, the expected number of endpoints in a circular neighbourhood of radius r centred on this endpoint is $E = \pi r^2 d$, where d is the density of endpoints. For E small, this approximates the probability, P, of finding a non-significant endpoint in the neighbourhood of radius r. Lowe [2] argues that $d = 2D/l^2$, where D is a scale invariant density parameter and l is the edge's length. Hence, for two segments of lengths l_1 and l_2 respectively, with $l_1 < l_2$, the probability that their endpoints separated by a distance r form an accidental grouping is

$$P = 2D\pi r^2/l_1^2 \tag{1}$$

The probabilities of accidental collinearity and parallelism are calculated in similar ways. The probability of accidental match for more complex groups is calculated by multiplying the probabilities of accidental match for each of the constituent primary perceptual groups. This is claimed to correspond to each constituent perceptual group having an independent probability of accidental match [6].

Once the probability of accidental match for a perceptual group has been calculated, the grouping is only considered significant if this probability is below a threshold. The perceptual groups containing 3 or more segments are used in the model matching stage.

Model Matching

The perceptual groups containing 3 or more segments are used in the model matching stage. The strategy used by Lowe is very simplistic. The list of observed perceptual groups is ordered in terms of increasing probability of accidental match. The first observed perceptual group from the list is matched to each model perceptual group of the same form in turn. For each observed perceptual group – model perceptual group pairing, an initial viewpoint estimate is obtained from the observed segment – model segment pairings defined by the observed perceptual group – model perceptual group pairing. This viewpoint estimate is refined by using a Newton-Raphson iteration. (See the following sections for details.) The outcome of the viewpoint refinement process is an estimate of the viewpoint, the number of observed segments supporting the viewpoint estimate, and the error of match between these observed segments and the assigned model segment images. If the "goodness" of this viewpoint estimate, based on the number of observed segments supporting the viewpoint estimate and the fit error, is acceptable then the search is stopped. If all possible model matches for the first observed perceptual group have been considered without finding an acceptable viewpoint estimate, then the second observed perceptual group is used. The search continues in this way until an acceptable viewpoint estimate is found or the list of observed perceptual groups is exhausted.

The viewpoint is specified by a 3x3 orthogonal matrix R and a 3-D vector D, such that the transform between the model based coordinate system P and the image plane coordinate system U is

$$\begin{aligned} x &= Rp & (2) \\ u_i &= \frac{f x_i}{x_3 + D_3} + D_i, \qquad i = 1,2 & (3) \end{aligned}$$

where the coordinate system X has the same origin as P, with the x_1, x_2 coordinate axes parallel to the image plane coordinate axes and x_3 pointing away from the image plane.

The list of model perceptual groups is also ordered in terms of decreasing likelihood of observation. It should be noted that it is necessary to include all permutations of a model perceptual group in the list. That is, each perceptual group containing n segments should appear n times, once with each segment being first in the ordered list of members of the perceptual group.

Initial Viewpoint Estimate

The initial viewpoint estimate is obtained from the observed segment – model segment pairings provided by the observed perceptual group – model perceptual group pairing. The rotation can be divided into two parts: rotation in depth and rotation in the image plane. An initial guess for the rotation in depth can apriori be made from the viewing directions that the surface containing the model perceptual group is visible from [2]. The rotation in the image plane can be estimated by causing one model segment to project onto the image plane with the same orientation as its assigned observed segment. Scale (i.e. D_3) can be estimated by finding the observed segment – model segment pair with the minimum ratio of model line length to observed segment length. This model segment should be closest to fronto-parallel and hence the ratio gives the scale. The other two components of D can be estimated by aligning the image of one model segment endpoint with the corresponding observed segment endpoint.

Viewpoint Refinement

The viewpoint estimate is refined by a Newton-Raphson iteration algorithm [2]. At each iteration, this involves linearizing the fit error between model segment - observed segment pairs about the current estimate, solving for the corrections by least squares since the number of data points (twice the number of segment pairs) is greater than the number of parameters (6), and updating the estimate by the corrections. The error between an observed segment and a model segment is measured by the perpendicular distance of each end of the model segment's image from the observed segment. The iterations are stopped when the correction factors fall below a tolerance level.

SYSTEM PERFORMANCE

Our implementation of Lowe's model-based vision system follows his descriptions in [1,2], except for the following differences:

- All the code was written in C on a SUN workstation.
- The Canny edge detector [8] and the line fitting algorithm of Pavlidis [9] were used to generate the list of observed segments in the input picture.
- The search for skew-symmetric perceptual groups was not implemented.
- The scale was estimated differently. From the estimate of R it is possible to calculate the x_3 component of each model segment. The model segment for which the difference between the x_3 values of its endpoints is smallest is the model segment which is closest to being fronto-parallel. The ratio of model segment length to assigned observed segment length for this model segment then gives the scale.

Our implementation of Lowe's model-based vision system has been tested on views of the polyhedral object Widget obtained from Plessey Research Roke Manor [11]. See Figure 1 for a typical view. Out of 16 views of the Widget, the pose was calculated correctly in 10. In none was the pose calculated incorrectly. Of the 6 views for which there was no pose calculated, only three contained observed perceptual groups containing three or more members. Failure to find a model perceptual group match for one of these occurred because the corresponding object face was far from fronto-parallel. The typical viewpoint for this model perceptual group was fronto-parallel. Hence the initial viewpoint estimate was not close enough to the true viewpoint for the Newton-Raphson iteration scheme to converge to the correct answer. There is no apriori reason why the pose shouldn't have been found in the other two cases. The search passed over correct observed perceptual group - model perceptual group matches without recognizing them as such. The other three views for which no pose was found contained no 4-sided perceptual groups. Mainly because the edges were not found as continuous segments by the edge detector. To work in these cases, the vision system needs to build up sides out of collinear segments. It does not do this at present.

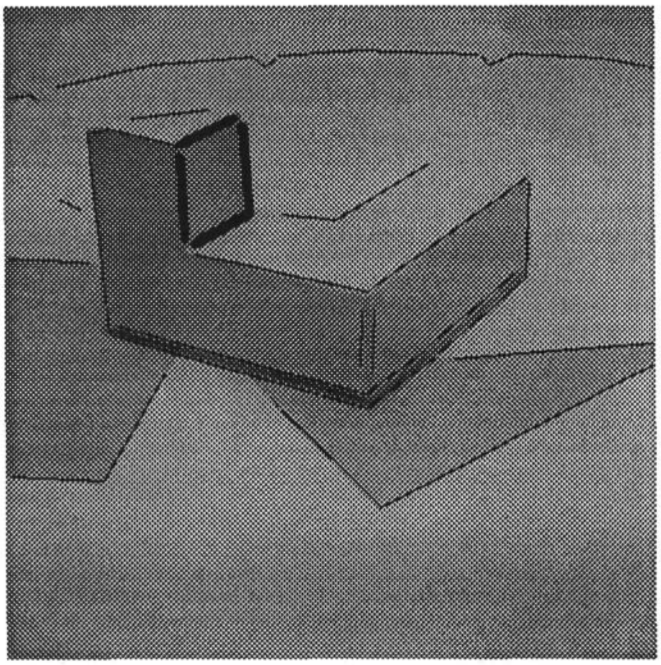

Figure 1: Image and its segments. Those in bold form the perceptual group used to get the viewpoint estimate.

Statistic	Min	Ave	Max
Perceptual grouping time (s)	3	4	5
Matching time (s)	1	20	78
Combined time (s)	4	23	82
Number of perceptual groups	99	156	317
Number of 4-sided groups	0	3	6

Table 1: Performance on "Widget" examples (on SUN 3/160)

Some performance statistics for the 16 Widget views are given in Table 1. In all these views, there was only one instance of the model. The model contained 16 vertices, 24 edges and 6 4-sided perceptual groups (there being 4 permutations of each). The results are illustrated in Figures 1 and 2. Figure 1 shows the input picture, the detected segments and the observed perceptual group used to produce the viewpoint estimate. Figure 2 shows the predicted model image based on the viewpoint estimate.

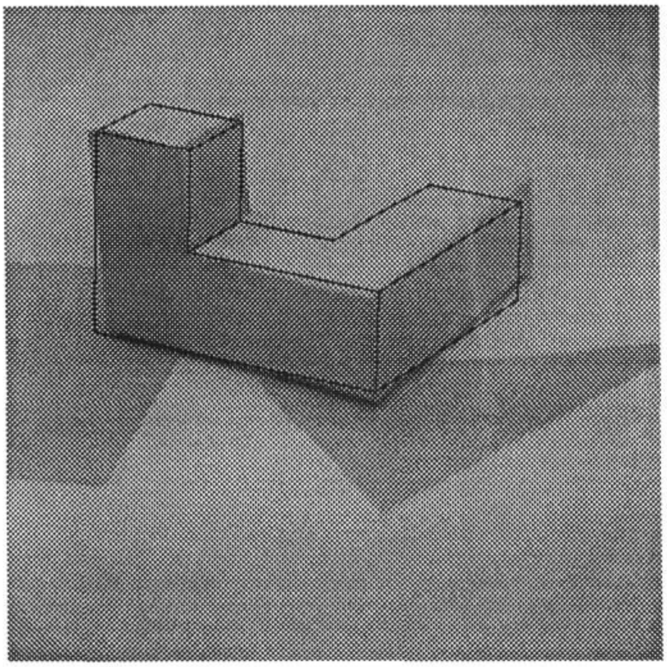

Figure 2: Model superimposed over image from viewpoint estimate.

DISCUSSION

From experience with our implementation of Lowe's model-based vision system, a number of points about it have been raised. These are discussed below.

Perceptual Grouping

The following points were noted with respect to the perceptual grouping phase:

1. The hypothesis that the density d of segments with length greater than or equal to l is given by $d = D/l^2$ has not been tested [6]. A regression test of this hypothesis was performed using a sample of 50 images [10]. It was found that this hypothesis is incorrect over the whole range of lengths but is reasonable when very short and long edges are excluded. From images where the hypothesis was accepted, of which there were 38, the estimated value of D is 0.33, which is much less than the predicted value of 1 [2].

2. There are two measures of segment density used by Lowe. These are the density, $\tilde{d}$, of endpoints in the neighbourhood of a segment of length l, and the density , d, of segments with length greater than or equal to l. The latter, d, is defined by $d = D/l^2$ as discussed in the previous paragraph. The former is claimed to be given by $\tilde{d} = 2d$ [2]. Although it is plausible that $\tilde{d} = \tilde{D}/l^2$, it does not seem likely that $\tilde{D} = D$.

3. The probability measure of accidental match seems to be very insensitive to the angle between segments when considering parallelism and collinearity. Segments of similar size and large orientation difference can have a lower probability of accidental match than segments of dissimilar size and smaller orientation difference.

4. The collinearity of two image segments can be due to two effects. It could be due to their alignment as Lowe discusses. It could also be due to the fact that they are both images of parts of the same object edge. That is, the edge detector missed a section of the edge and reported two segments instead of one. The probability of accidental match needs to be modified to reflect the latter case. This could be done by considering a single segment approximating both, such as the one joining the most distant endpoints, and measuring how well this fits the two segments, such as in a line fitting routine, e.g. [9]. The probability that the edge detector missed a segment of the size implied by the "gap" between the segments would also have to be incorporated.

5. The probability of accidental match for a complex perceptual group is calculated by multiplying the probabilities of accidental match of its constituent primary perceptual groups. This is claimed to follow from an independence assumption [2,6]. However, this is incorrect. Independence of the primary perceptual groups implies that the probability of a complex group should be calculated by multiplying the probabilities of non-accidental match (i.e. 1 - Prob(accidental match)) for the primary groups. This leads to the property that the probability of accidental match increases with complexity, which is not what is intuitively expected. This shows that it is necessary to incorporate dependence and apriori probabilities into the calculation of the probability of accidental match [6].

Model Matching

The following points have been noted with respect to the model matching phase:

1. The projection transform (2–3) is an affine approximation, not true perspective. To see this, consider true perspective

$$\tilde{x} = R(p-t) \tag{4}$$

$$u_i = \frac{f\tilde{x}_i}{\tilde{x}_3}, \qquad i = 1,2 \tag{5}$$

If we define $\tilde{t} \triangleq Rt$, then from (2) and (4), $\tilde{x} = x - \tilde{t}$. Hence from (3) and (5),

$$\begin{aligned} D_i &= -\frac{f\tilde{t}_i}{x_3 - \tilde{t}_3} \qquad i = 1,2 \qquad (6) \\ D_3 &= -\tilde{t}_3 \qquad (7) \end{aligned}$$

and it can be clearly seen that assuming that D_1 and D_2 are constants amounts to an affine approximation.

2. The derivation of the differentials of u_i with respect to the three components of the rotation R can be made rigorous by noting that R is the exponential of a skew-symmetric matrix and that the derivatives obtained by Lowe [2] correspond to the differentials with respect to the three parameters of the skew-symmetric matrix.

3. Using true hidden line removal instead of the approximation used by Lowe would have eliminated some false matches in our example by reducing the length of a partially occluded model segment's image. This would reduce the search space for matches and increase the accidental match probabilities.

4. The calculation of the initial viewpoint estimate is the weakest part of this system. A more effective procedure needs to be developed.

5. The claim that there will always be a marked difference between the correct observed perceptual group – model perceptual group match and the others in terms of fit error and number of observed segment – model segment matches isn't true. Although it always seems to be true that the true match has the most number of model segment – observed segment matches and the smallest fit error, there can be a lot of support for a false match in the background clutter. This makes it hard to find a search stop criteria apriori. Perhaps a "semi-exhaustive" search will be needed for reliability [7].

FUTURE WORK

The major areas in which this model-based vision system can be improved are:

1. The procedure for the initial viewpoint estimate needs to be made more robust. For example, if the model perceptual group is a rectangle, it is possible to determine the rotation in depth exactly.

2. Perceptual groups containing only three member segments, e.g. a pair of parallel lines with another segment closing one end, need to be incorporated into the system.

3. The whole issue of probability measures needs to be studied further. Especially the calculation of the probability of accidental match for complex perceptual groups.

4. It is possible to determine the derivatives of image location for true perspective with no more difficulty than using Lowe's affine approximation. These could be used in the Newton-Raphson iteration scheme, improving the accuracy of the viewpoint estimate. Perhaps at the expense of decreased numerical stability.

5. At present the fit error between an observed segment and the matched model segment is measured by the perpendicular distance of the endpoints of the model segment's image from the observed segment. This could be reformulated in terms of the difference between the parameters of the supporting lines.

We propose to continue our study of this model-based vision system, improving it as outlined above.

References

[1] D. G. Lowe, Perceptual Organization and Visual Recognition, MA, USA: Kluwer Academic Publishers, 1985.

[2] D. G. Lowe, "Three Dimensional Object Recognition from Single Two Dimensional Images", Artificial Intelligence, Vol 31, 1985, pp 355–395.

[3] D. G. Lowe, "Solving for the Parameters of Object Models from Image Descriptions", Proceedings: Image Understanding Workshop, April 1980, pp 121–127

[4] D. G. Lowe, "Four Steps towards General-Purpose Robot Vision", Proceedings: Fourth International Symposium of Robotics Research, August 1987.

[5] D. G. Lowe, "The Viewpoint Consistency Constraint", International Journal of Computer Vision, Volume 1, Number 1, 1987, pp 57–72.

[6] D. G. Lowe, *private communication*, March 1988.

[7] W. E. L. Grimson and T. Lozano-Pérez, "Recognition and Localization of Overlapping Parts from Sparse Data", MIT AI Memo No 41, June 1985.

[8] J. Canny, "A Computational Approach to Edge Detection", IEEE Trans Pattern Recognition and Machine Intelligence, Vol PAMI-8, No. 6, November 1986, pp 679–698.

[9] T. Pavlidis, Algorithms for Graphics and Image Processing, MD, USA: Computer Science Press, 1982.

[10] A. M. M^{c}Ivor, "Test of Lowe's hypothesis on edge density", Oxford University Robotics Research Group Report, April 1988.

[11] C. G. Harris and J. M. Pike, "3D Positional Integration form Image Sequences", Proceedings, Alvey Vision Conference, September 1987, pp 233–236.