[28] Matching Geometrical Descriptions in Three-Space

Stephen B Pollard, John Porrill, John E W Mayhew and John P Frisby

AI Vision Research Unit
University of Sheffield, Sheffield S10 2TN, UK

Reprinted, with permission of Butterworth Scientific Ltd, from *Image and Vision Computing*, 1987, **5**, 73-78.

Abstract

A matching strategy for combining two or more three space descriptions, obtained here from edge based binocular stereo, of a scene is discussed. The scheme combines features of a number of recent model matching algorithms with heuristics aimed to reduce the space of potential rigid transformations that relate scene descriptions.

1. Introduction

The topic addressed by this paper is the matching of stereo-based 3D edge descriptions obtained from two or more different views of a scene. We describe a matching algorithm with wide potential application in the temporal aggregation of scene descriptions for scene model evolution and autonomous vehicle guidance.

The algorithm is for example of use as a precursor to the combination of data about the 3D geometry of a given edge into an improved estimate (Porrill et al[1]). It is also useful for obtaining an accurate estimate of the location of the viewpoint with respect to the scene. Moreover, the algorithm is shown to be useful as a primitive model matcher in which a visual description of a known model can be matched to an instance of the modelled object in a cluttered scene.

The algorithm can also be used for a purely visual method of creating an object model description. Given descriptions of the object-to-be-modelled from multiple viewpoints, the algorithm can combine these into a single description which can then serve as a model of the object-to-be-recognised in subsequent views of cluttered scenes containing the object. For this and other model matching applications it will clearly be desirable to incorporate into the model description some organisation of its view potential[2–4] including the flagging of non-rigid relationships (eg hinges). However, whilst we have obtained some success in this regard it remains a topic beyond the scope of the current paper.

The reasonable assumption that the geometry of the scene remains constant between views provides a powerful matching constraint. It allows our goal to be defined as identifying the best set of matches that is consistent with a single rigid transformation. The rigidity constraint can either be explicit, requiring each primitive to undergo the same global transformation, or implicit, requiring that local geometrical relationships between primitives be preserved. Tree searching strategies based upon each have been exploited by Faugeras *et al*[5, 6] and Grimson *et al*[7–9] respectively. Constraints based upon local geometrical relationships have the computational advantage that such relationships can be precompiled for each scene and stored in look up tables, with the result that simple pairwise comparisons are all that is necessary to exploit rigidity. Under the alternative strategy, each potential global transformation requires both computation and then application to each descriptive primitive in one scene to locate matching descriptive primitives in the other.

The matching strategy described here differs from that of Grimson et al in that tree search is replaced by a hypothesis and test strategy based upon a number of focus features[10, 11]. Exhaustive tree searching strategies are only computationally efficient where all the data in one scene is known to be present in the other[12, 13], otherwise, despite the powerful constraint provided by rigidity they tend to be combinatorially explosive. Here, unfortunately, due to the difference in viewpoint, the vagaries of the image formation process, and the imaging process itself the mapping between two geometrical descriptions of the same scene is generally many-to-many. Hence major computational problems will arise with simple tree search. Concentrating initial attention to the matches of a relatively small number of heuristically determined focus features allows, at the expense of some generality, the search space to be reduced considerablely.

2. Geometrical Description

The task of matching three space descriptions of well carpentered scenes is discussed. For brevity and simplicity we shall restrict our attention to scenes, or regions of scenes, that are amenable to characterisation by their straight surface discontinuities. However extensions to include both circular and space curve descriptions in the matching process are currently under investigation. All such descriptions are obtained here from edge based binocular stereo triangulation[14–16]. Matched edge points are aggregated into extended edge structures[3] and described by a process of recursive segmentation and description as either straight, circular, planar or space curves[17–19]. The grouping of edge descriptions in this way provides an initial, though impoverished, viewer centred scene description called the Geometric Descriptive Base[20] (GDB).

In the GDB straight lines are represented (in an overdetermined fashion) by the triple $(\mathbf{v},\mathbf{p}_1,\mathbf{p}_2)$, that is, their two end points $\mathbf{p}_1$ and $\mathbf{p}_2$ and the direction vector between them $\mathbf{v}$. The centroid of a line (its midpoint $(\mathbf{p}_1+\mathbf{p}_2)/2$) shall be denoted $\mathbf{c}$. Where the actual physical occupancy of a line is not important it is sometimes helpful to represent them by the vector pair $(\mathbf{v},\mathbf{c})$.

3. The Matching Problem

In general it is not possible to place a restriction on the allowable transformations that take primitives from one scene description into another unless an a priori estimate of the difference in their viewpoint is available. Of course in the domain of autonomous navigation and scene evolution it is likely that just such an estimate exists. If for instance the temporal delay that separates scene descrip-

tions obtained by an autonomous vehicle is sufficiently small (ie the frame rate is sufficiently high) the transformation that takes one view point into another will be of a limited magnitude. Alternatively if a less intensive rate of low level image processing is to be preferred an estimate of the trajectory could be used to approximate the geometry that relates successive viewpoints. The design of our matching algorithm is general in this regard. An estimate of the global transformation is not required to achieve successful matching; if however such information is readily available it can be exploited to reduce the set of potential matches and thus also the computational requirements.

Given the somewhat impoverished nature of our descriptive basis and the potential for unfortunate, and unforeseen, occlusion relationships to arise, it seems prudent not to restrict potential matches on the basis of the descriptive properties of the line primitives (eg length, contrast etc). Hence in principle, each primitive extracted from one scene is able to match with each of the primitives in the other. Furthermore we do not feel, at the present time, able reliably to obtain higher level features and topological relationships (eg vertices or connected edge segments describing a polyhedral face) and focus initial matching about these. Frequently relationships of this kind will not be preserved between views. It has not even assumed that the locations of the end points of lines remains constant as continuous lines in one image may appear broken in the other. Lines are however expected to overlap significantly.

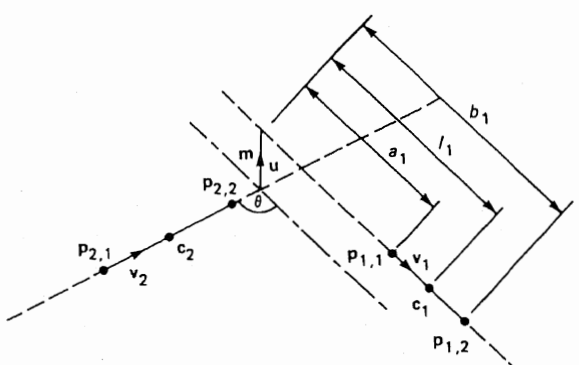

Figure 1. Geometrical relationships are illustrated for a pair of lines. These are: their orientation difference θ*, the distance m between their extensions, and the distances* a_1 *and* b_1 *from the ends of the physical line to the point of minimum separation.*

3.1. Exploiting Rigidity

Matches for two non-parallel line segments are sufficient to constrain all six degrees of freedom that constitute a putative transformation between scene descriptions[6]. Once a transformation is hypothesised, rigidity provides a powerful constraint upon other consistent matches (subject to tolerance errors; the details of which are not discussed here for reasons of brevity). As discussed above rigidity can be exploited more cheaply (though less strongly) if expressed in terms of the consistency in a number of pairwise relationships. Here we adopt just three, they are (illustrated also in figure 1):

(i) orientation differences, given by $\theta = \cos^{-1}(\mathbf{v}_1.\mathbf{v}_2)$.

(ii) minimum separations between (extended) lines. The unit vector in the direction of closest approach (normal to each line) is given by $\mathbf{u} = (\mathbf{v}_1 \times \mathbf{v}_2)/|\mathbf{v}_1 \times \mathbf{v}_2|$ and component of the vector difference between the lines in that direction by $m = (\mathbf{c}_2 - \mathbf{c}_1).\mathbf{u}$. However if the lines are close to parallel it is more sensible to simply measure the perpendicular distance between the lines $m = |(\mathbf{c}_2 - \mathbf{c}_1) - ((\mathbf{c}_2 - \mathbf{c}_1).\mathbf{v}_1)\mathbf{v}_1|$.

(iii) distance to the beginning and end of each physical line with respect to the point of minimum separation and in the direction of the line. This relationship is only applicable for non-parallel lines. The vector between the points of closest approach is given by $\mathbf{m} = ((\mathbf{c}_2 - \mathbf{c}_1).\mathbf{u})\mathbf{u}$. Adding $\mathbf{m}$ to $\mathbf{c}_2$ gives $\mathbf{c}'_2$, where lines $(\mathbf{v}_1, \mathbf{c}_1)$ and $(\mathbf{v}_2, \mathbf{c}'_2)$ are coplanar and meet at the point of closest approach on $(\mathbf{v}_1, \mathbf{c}_1)$. The signed distance to that point from $\mathbf{c}_1$ in the direction $\mathbf{v}_1$ is given by $l_1 = ((\mathbf{v}_2 \times \mathbf{v}_1).(\mathbf{v}_2 \times (\mathbf{c}'_2 - \mathbf{c}_1)))/|\mathbf{v}_2 \times \mathbf{v}_1|^2$. Hence the distance from $\mathbf{p}_{1_1}$ and $\mathbf{p}_{1_2}$ to that point are given by $a_1 = l_1 + (\mathbf{c}_1 - \mathbf{p}_{1_1}).\mathbf{v}_1$ and $b_1 = l_1 + (\mathbf{c}_1 - \mathbf{p}_{1_2}).\mathbf{v}_1$ respectively. Similarly for distances to the point of closest separation on the other line $a_2 = l_2 + (\mathbf{c}_2 - \mathbf{p}_{2_1}).\mathbf{v}_2$ and $b_2 = l_2 + (\mathbf{c}_2 - \mathbf{p}_{2_2}).\mathbf{v}_2$.

Potential matches for each pair of descriptive elements from one scene description can be checked for geometrical consistency in the other. Rigidity implies that each of the pairwise relationships will be preserved between scene descriptions, hence any measured discrepancies must lie within a range predicted by the magnitude of allowable errors. Furthermore a pair of consistent non-parallel matches provides a powerful constraint upon the remaining matches. Hence they can be thought to represent, implicitly and weakly, a global transformation. The representation is weak because it is possible, on occasion, for matches that are not consistent with a single global transformation to satisfy the pairwise relationships. In practice such problems are not major. Furthermore if the basis of the implicit transforms is raised from a pair to a triple, quadruple or even a quintuple of matches, such inconsistencies are far less likely (additionally the margin of allowable error on each new match will be reduced).

3.2. Look Up Tables

The pairwise geometrical relationships, upon which local constraints are based, have the advantage that they can be precomputed for each pair of lines independently for each scene description and stored as look up tables [as with [7-9, 12]]. Each relational property is stored as a range of values consistent with the allowable error. It is these ranges that must overlap for a pair of matches to be considered geometrically consistent. Errors in centroid location and orientation are considered separately and combined in a conservative fashion that simply adds their contributions resulting in the largest feasible range of pairwise geometrical relationships.

Given a pair of lines with allowable errors $|\varepsilon_1| < \alpha_1$ and $|\varepsilon_2| < \alpha_2$ on the location of their centroid and solid angles ϕ_1 and ϕ_2 on their direction vector the following ranges can be derived:

(i) on orientation differences: the interval $[\max(\theta-\phi_1-\phi_2,0),\min(\theta+\phi_1+\phi_2,\pi)]$.

(ii) on minimum separations between (extended) lines: the interval

$$m +/- (\alpha_1+\alpha_2+|l_1|\tan\phi_1+|l_2|\tan\phi_2)$$

(iii) on the distances to the beginning and end of each physical line with respect to the point of minimum separation: the approximate intervals

$$a_1 +/- (\alpha_1+\alpha_2+|l_2|\tan\phi_2/\sin\theta)$$

$$b_1 +/- (\alpha_1+\alpha_2+|l_2|\tan\phi_2/\sin\theta)$$

$$a_2 +/- (\alpha_1+\alpha_2+|l_1|\tan\phi_1/\sin\theta)$$

$$b_2 +/- (\alpha_1+\alpha_2+|l_1|\tan\phi_1/\sin\theta)$$

3.3. Feature Focus

Our current approach to matching is to apply heuristics similar to those of feature focus[10, 11] in order to avoid unbounded search. The strategy is to concentrate initial attention upon a number of salient features. Only matches associated with these features are subsequently entitled to *grow* hypothetical transformations. Currently processing terminates only after all focus features have been considered. As an alternative it could be possible to complete computation once a *sufficiently good* match has been located. However, at the present time, a suitable definition of *sufficiently good* is not available. The feature focus strategy adopted here differs from those considered previously as familiarity with the scene it is not assumed. As a result focus features and matching strategies are not an integral component of our scene description: they must be generated at run time.

Focus features are identified in a single scene description. Currently they are chosen simply on the basis of their length, a property associated with salience. Some effort is expended to ensure that all regions of the scene are represented by chosen features, ie a feature is identified as a focus if there are not more than a certain number of longer features within a predetermined radius of it.

Our matching strategy proceeds as follows

(1) a focus features is selected (in turn);

(2) the S closest features to it with length greater than L are identified;

(3) potential matches for the focus feature are considered, unlike matching in general, these are selected conservatively on the basis of length (which must lie within 30% of each other);

(4) consistent matches for each of the neighbouring primitives are located;

(5) this set of matches (including that of the focus feature) is searched for maximally consistent cliques of cardinality at least C, each of these can be thought of as a potential implicit transformation;

(6) each clique is extended by adding new matches for all other lines in the scene if they are consistent with each of the matches in the clique;

(7) mutually consistency can be ensured by some further (cheap) tree search;

(8) extended cliques are ranked on the basis of the sum of the length of their matched lines, the contribution from each match being the lesser of the lengths of its constituent lines;

Note that any/all of the focus features are potentially able to discover the implicit transformation (clique of correct matches) that takes one viewpoint into the other. Hence only allowing focus features to match conservatively does not greatly hinder the matching strategy. Furthermore some unnecessary computation can be avoided if consistent cliques arrived at via different focus features are identified and combined prior to their extension in step (6).

Insisting that at least one focus feature obtains a match places a bound upon depth of search to which current assignments are allowed to be all null. In a similar fashion, restricting the set of focus feature matches reduces the breadth of search. Furthermore requiring that mutually consistent matches be found for C of the S near neighbouring primitives further controls the search. Consider each matched focus feature to be the origin of an independent search tree; paths below the depth S are bounded unless at least C matches occur above them. In practice it is this constraint that provides the greatest prune, as very few incorrect transformations satisfy this requirement.

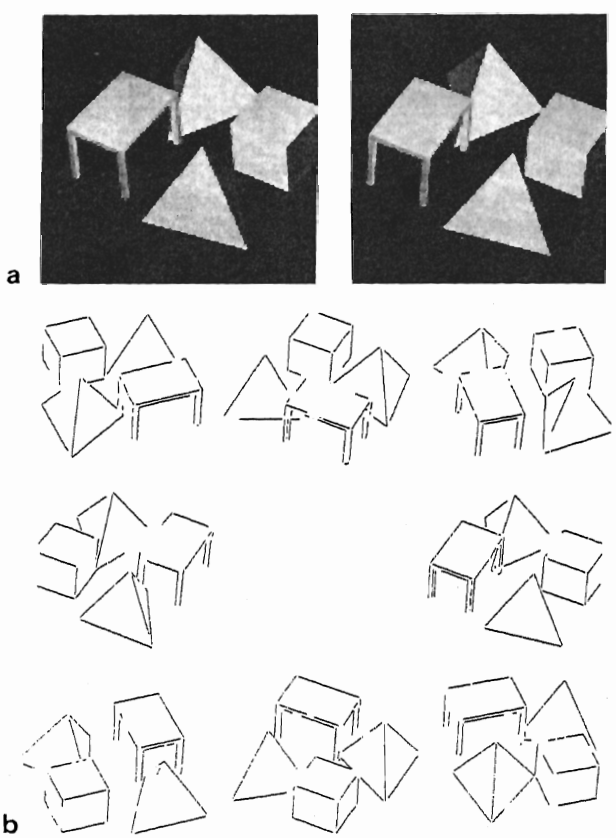

Figure 2. Part (a) is a stereogram of one view of a synthesised scene (arranged for cross eyed fusion). Part (b) shows GDB descriptions of this scene from eight equally spaced viewpoints.

The power of our focusing heuristics are dependent upon the choice of S and C; if C is too small many putative transformations will be explored at great expense, hence C must be large enough to impose considerable constraint upon potential transformations. Conversely, whilst S must be sufficiently large that C amongst them will locate consistent matches, if S becomes too large the time spent searching increases dramatically. In practice, as will be illustrated below, very few consistent sets of matches, beyond the correct one, are found if $C = 4$ and $S = 7$.

The set of S primitives are chosen to neighbour the focus feature in question for two reasons. First, the constraints provided by pairwise consistency are strongest over modest physical separations as the allowable error ranges are smaller. And second, in the absence of a more sophisticated scheme, neighbouring primitives are thought *more likely* to occupy similer view potentials and hence appear simultaneously in scene descriptions obtained from different views.

The number of focus features used for matching is increased with n (in the experiments below approximately $0.2 \times n$). Similarly the number of potential matches, the number of potential transformations, and the cost of extending each transformation all increase with n. However as S and C remain constant the the expense of exploring each focus match will on average also remain constant. Hence whilst computational expense is high (increasing with some multiple of n^4) combinatorial explosion is avoided. Furthermore if an appropriate computer architecture were available it may be possible to do some proportion of this work in parallel (for example the consideration of each match of each focus feature).

A similar matching strategy has been proposed recently by Ayache *et al*[21], except that they consider transformations for all consistent matches of pairs of privileged lines (equivalent to chosing C to be 2), of which there may be a great many. Furthermore each pair of such matches is used by them to compute an explicit transformation, rather than the implicit representation we prefer.

4. Matching Experiments

Performance of the matching algorithm is illustrated quantitatively for artificial stereo data provided by the IBM Winsom[22] body modeller and qualitatively for natural stereo data. The former is used to obviate the accurate stereo calibration problem which is a current research topic in the laboratory.

Consider first the synthesised scene depicted in the stereogram in figure 2a. GDB descriptions of this scene, obtained from eight equally spaced viewpoints (45 degrees apart) are shown in 2b. Each description consists of approximately 40 above-threshold GDB line primitives. These are to be matched between viewpoints to construct a more complete model of the scene. The results of the matching process between the first two views is illustrated in figure 3. The ten focus features chosen in view 1 obtained a total of 174 potential matches in view 2. Setting S to 7 and C to 4 only 13 independent implicit transformations result. After extension the best consistent transformation included 18 matched lines. The best rigid rotation and translation (in that order) that takes view 1 to view 2 is computed by the least squares method discussed by Faugeras *et al*[6] in which rotations are represented as quaternions (though for simplicity the optimal rotation is recovered before translation). In figure 3a view 1 is transformed into view 2 (the error in the computed rotation is 0.48 degrees) and matching lines are shown bold, many of the the unmatched lines are not visible in both

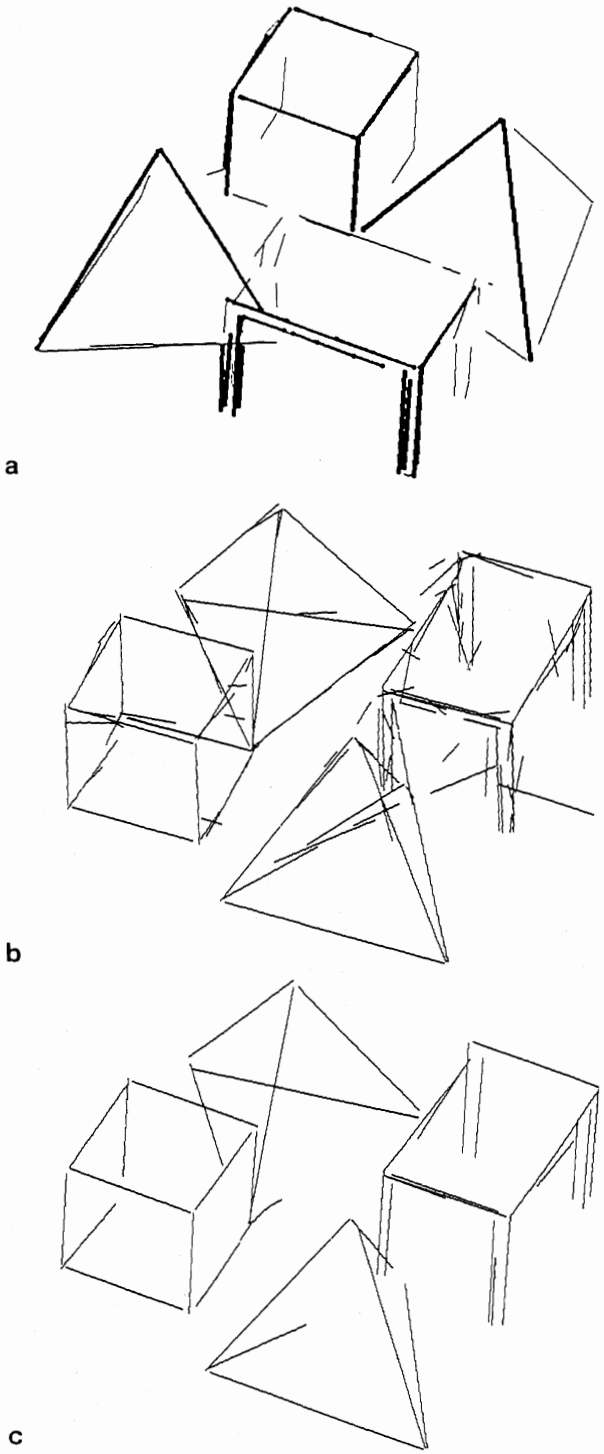

Figure 3. The results of matching the first two views in figure 2 are shown in (a); matched lines are shown bold. All views when combined by the matcher result in (b). Those lines that have been matched (and hence appear in more than one view) are shown in (c).

views. If the model is matched and updated with respect to each view in the sequence (chosing focus features only from the features that were matched in the previous cycle) the scene description in figure 3b results. This description contains a large quantity of noisy data that appeared in one or other view. A cleaner model can be obtained by filtering out primitives that have never been matched (see figure 3c).

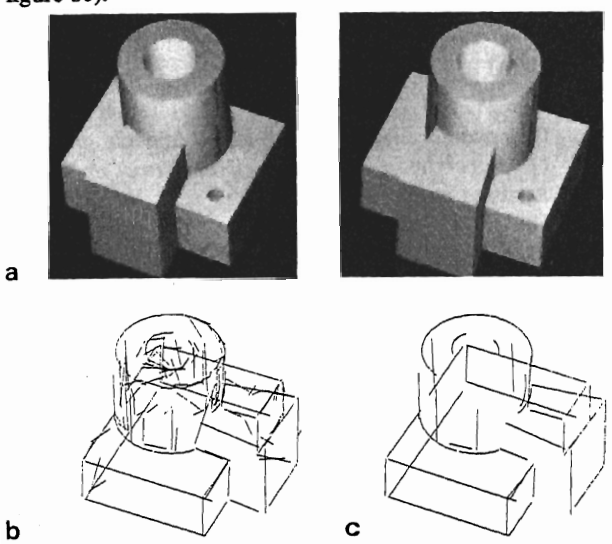

Figure 4. A single view of a synthesised test object is shown in (a). Noisy and clean models are shown in (b) and (c) respectively.

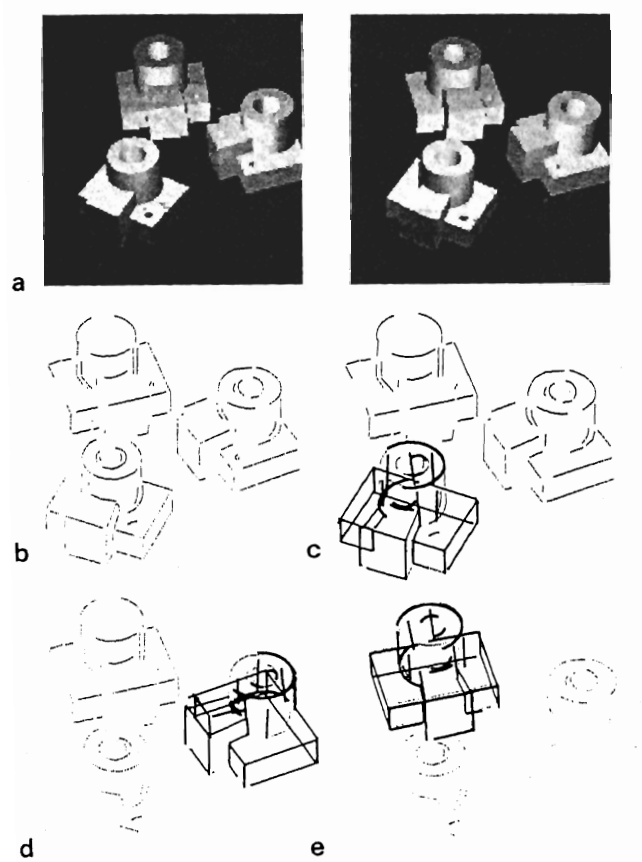

Figure 5. A synthesised bin picking scenario is depicted in (a), with corresponding GDB description in (b). Model instances are identified in (c), (d) and (e).

A similar sequence of processing has been performed for the artificial test object shown in figure 4a. Noisy and clean models obtained for it are shown in figure 4b and 4c respectively. Notice that few of the occluding contours that arise from the cylinder are ever matched. The matcher does not match the circular section of object, the unsophisticated update procedure simply passes through circles that were observed in the last known view. Matching was not hindered by the presence of circular data.

Once obtained this simplistic model (consisting of 41 line sections) can be matched in a bin picking scenario. Figures 5 and 6 illustrate this process for artificial and natural disparity data. The latter suffers camera calibration error; the resolution of our current calibration technique is suitable for epipolar stereo matching but not for accurate disparity interpretation. Figures 5a and 6a show scenes of a number of test objects, and figures 5b and 6b GDB data extracted from these. The best match of our model is superimposed, and shown bold, over each in 5c and 6c. Whilst the match for artificial data is near perfect, some geometrical distortion is visible in the real data. Removing matched portions of the GBD data allows the second (figures 5d and 6d) and third (figures 5e and 6e) best matches to be located. Unfortunately the third match of the real data results in mismatch.

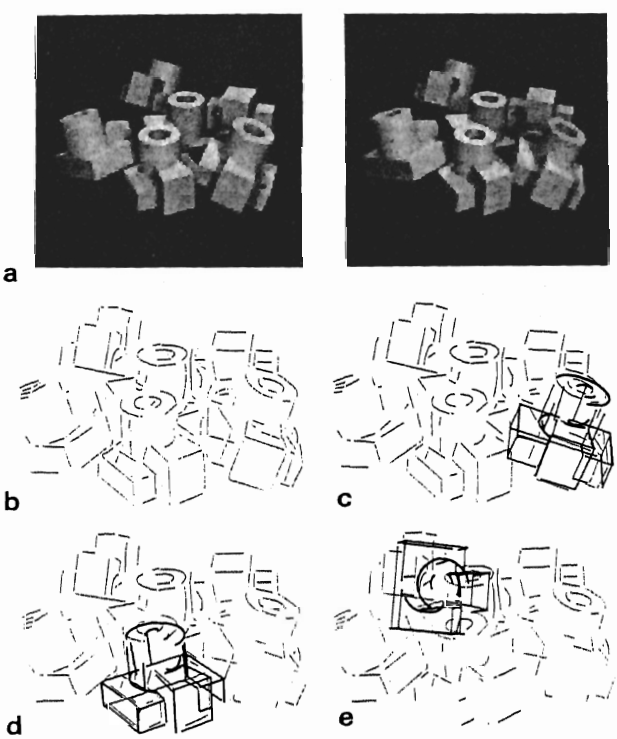

Figure 6. A real bin picking scenario is depicted in (a), with corresponding GDB description in (b) (current camera calibration is unreliable). Model instances are identified in (c) and (d). Mismatch results in (e).

5. Concluding Comments

A matching strategy for combining two or more three space descriptions of a scene has been discussed. It combines features of a number of algorithms that have appeared recently in the literature on three dimensional model matching. Its has two principal (almost novel)

features. First, a number of pairwise relationships are seen as implicitly specifying the rigid transformation that relates the scenes. And secondly, search has been controlled by requiring that local cliques of mutually consistent matches must be located in the vicinity of at least one of a number of focus features, with the result that very few hypothetical transformations require attention.

A number of extensions to this strategy are currently under investigation. These fall into two categories: (i) those concerned with description and model building, eg the primitive base, occluding contours, partial rigidity, and view potential (all discussed briefly above); (ii) and those concerning the matching process itself. Currently the pairwise relation table is computed exhaustively, relationships between every primitive are stored. It should be possible to exploit rigidity using only a subset of the pairwise relations. Furthermore the scheme could be expanded to include unforeseen non-rigidity (when acquiring a model of a scene with moving objects in it).

Acknowledgements

We would like to thank Tony Pridmore and Ian Elsley for useful comments and advice and Chris Brown for his valuable technical assistance. This research was supported by SERC project grant no. GR/D/16796-IKBS/099 awarded under the Alvey programme. Stephen Pollard is an SERC IT Research Fellow.

References

1 Porrill J, SB Pollard and JEW Mayhew (1986) The optimal combination of multiple sensors including stereo vision, Alvey Computer Vision and Image Interpretation Meeting, Bristol, and submitted to Image and Vision Computing.

2 Chakravarty,I. and Freeman,H (1982) Characteristic Views as a Basis for Three-dimensional Object Recognition, SPIE Vol.336 Robot Vision.

3 Koenderink JJ (1985) The Internal Representation of Solid Shape Based on the Topological Properties of the Apparent Contour, Image Understanding.

4 Mayhew JEW (1986) Review of the YASA Project : May 1986, AIVRU Memo 014, University of Sheffield.

5 Faugeras OD, M Hebert, J Ponce and E Pauchon (1984) Object representation, identification, and positioning from range data, Proc. 1st Int. Symp. on Robotics Res, M Brady and R Paul (eds), MIT Press, 425-446.

6 Faugeras OD and M Hebert (1985) The representation, recognition and positioning of 3D shapes from range data, submitted to Int. J. Robotics Res.

7 Grimson WEL and T Lozano-Perez (1984) Model based recognition from sparse range or tactile data, Int. J. Robotics Res. 3(3): 3-35.

8 Grimson WEL and T Lozano-Perez (1985) Recognition and localisation of overlapping parts from sparse data in two and three dimensions, Proc IEEE Int. Conf. on Robotics and Automation, Silver Spring: IEEE Computer Society Press, 61-66.

9 Grimson WEL and T Lozano-Perez (1985) Search and sensing strategies for recognition and localization of two and three dimensional objects, Proc. Third Int. Symp. on Robotics Res.

10 Bolles RC and RA Cain (1982) Recognizing and locating partially visible objects, the local feature focus method, Int. J. of Robotics Res. 1(3): 57-82.

11 Bolles RC, P Horaud and MJ Hannah (1983) 3DPO: A three dimensional part orientation system, Proc. IJCAI 8, Karlshrue, West Germany, 116-120.

12 Gaston PC and T Lozano-Perez (1984) Tactile recognition and localization using object models: the case of the polyhedra on a plane, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol PAMI-6, No. 3, 257-266.

13 Grimson WEL (1984) The combinatorics of local constraints in model-based recognition and localization from sparse data, MIT AI Lab. Memo 763.

14 Pollard SB, JEW Mayhew and JP Frisby (1985) PMF: A stereo correspondence algorithm using a disparity gradient limit, Perception 14, 449-470.

15 Pollard SB, J Porrill, JEW Mayhew and JP Frisby (1985) Disparity gradient, Lipschitz continuity and computing binocular correspondences, Proc. Third Int. Symp. on Robotics Res.

16 Pollard SB (1985) Identifying correspondences in Binocular stereo, unpublished Phd thesis, Dept of Psychology, University of Sheffield.

17 Porrill J, TP Pridmore, JEW Mayhew and JP Frisby (1986) Fitting planes, lines and circles to stereo disparity data, AIVRU memo 017.

18 Pridmore TP, JEW Mayhew and JP Frisby (1985) Production rules for grouping edge-based disparity data, AIVRU memo 015, University of Sheffield.

19 Pridmore TP, J Porrill and JEW Mayhew (1986) Segmentation and description of binocularly viewed contours, Alvey Computer Vision and Image Interpretation Meeting, Bristol, and submitted to Image and Vision Computing.

20 Pridmore TP, JB Bowen and JEW Mayhew (1985) Geometrical description of the CONNECT graph #2, the geometrical descriptive base: a specification, AIVRU Memo, 012.

21 Ayache N, OD Faugeras, B Faverjon and G Toscani (1985) Matching depth maps obtained by passive stereo, Proc. Third Workshop on Computer Vision: Representation and Control, 197-204.

22 Quarendon P (1984) WINSOM user's guide, IBM Doc. No. UKSC 123.