Using Global Consistency to Recognise Euclidean Objects with an Uncalibrated Camera

D.A. Forsyth
Computer Science
University of Iowa
Iowa City, IA 52242

J.L. Mundy
General Electric
Center for Research and Development
Schenectady, NY 12345

A. Zisserman
Robotics Research Group
Oxford University
Oxford, UK

C.A. Rothwell
Robotics Research Group
Oxford University
Oxford, UK

Abstract

A recognition strategy consisting of a mixture of indexing on invariants and search, allows objects to be recognised up to a Euclidean ambiguity with an uncalibrated camera. The approach works by using projective invariants to determine all the possible projectively equivalent models for a particular imaged object; then a system of global consistency constraints is used to determine which of these projectively equivalent, but Euclidean distinct, models corresponds to the objects viewed. These constraints follow from properties of the imaging geometry. In particular, a recognition hypothesis is equivalent to an assertion about, among other things, viewing conditions and geometric relationships between objects, and these assertions must be consistent for hypotheses to be correct. The approach is demonstrated to work on images of real scenes consisting of polygonal objects and polyhedra. **Keywords:** *Recognition, Computer Vision, Invariant Theory, Indexing, Model-based Vision*

1 Introduction

Many recent object recognition systems model viewing with an uncalibrated camera or using an uncalibrated stereo pair as inducing either an affine or a projective transformation on figure. This approach allows invariants of the appropriate transformation to be used to index models to produce a selection of recognition hypotheses. These hypotheses are combined as appropriate, and the result is back-projected into the image, and verified by inspecting relationships between the back-projected outline and image edges [3, 5, 9, 12, 14]. Indexing using projective invariants has been demonstrated for plane objects and simple polyhedral objects, and has been extended with varying success to certain types of surfaces [1, 6, 8, 13, 15]. One main disadvantage of this approach is that objects are identified only up to either an affine or a projective ambiguity. This paper argues that this ambiguity is a consequence of considering recognition hypotheses in isolation, and is not intrinsic to the approach.

Systems based on indexing using projective invariants have not, to date, been able to distinguish between objects that are projectively equivalent, but not Euclidean equivalent, because such objects have the same projective invariants. In this paper, we show that a view of two or more coplanar objects, or polyhedra is enough to allow the objects to be recognised up to only a Euclidean ambiguity, if the objects can be recognised at all and if Euclidean models are available.

1.1 Frames and terminology

Much of the work we describe consists of reconciling different assertions about coordinate frames. As a result, the discussion can become confusing without an established terminology. The paper uses the following terms:

- **object:** an actual thing in the world.
- **model:** a collection of known measurements of the projective and Euclidean geometry of an object, which is stored in the system. A model could consist of a mixture of points, lines, planes, conics and more complicated curves or surfaces.
- **model frame:** the frame of reference in which the model measurements are taken; the reference points of an object in the world are within a Euclidean transformation of the reference points in this frame.
- **world frame:** a global frame of reference, in which objects exist. If the world consists of coplanar plane objects, then the world frame is the frame of reference within this plane; otherwise, the world frame is three-dimensional.
- **image frame:** a frame of reference constructed in the image plane, usually by reference to the pixel positions in the camera. For a view of a plane world, the image frame is within an unknown projective transformation of the world frame.
- **Euclidean transformation:** a projective transformation, equivalent to a rigid motion (rotation and translation), expressed in homogenous coordinates.

In this paper, the relationships between frames are emphasized; these relationships are usually determined by computing transformations between image features and model features. Such a transformation, although computed using

1063-6919/94 $3.00 © 1994 IEEE

some specific set of features, *expresses the transformation between the model frame and the image frame.*

2 Plane objects

Consider a scene consisting of a set of distinct, coplanar plane objects, many of which are represented in a model-base. It is well known that any view of this scene with an uncalibrated camera can be obtained by applying an appropriate plane projective transformation to the scene. The goal of a recognition algorithm is from an image of the scene, label each object correctly up to Euclidean equivalence.

Indexing using projective invariants (as in [9]) associates with each group of image features a collection of object models (labels), which are projectively equivalent, but Euclidean inequivalent.

If only one known object is present, the task is possible only if there is just one possible label for that object. If two or more labels apply, the task can be considered in terms of constructing the largest possible consistent labelling, because implicit in each recognition hypothesis is information about the frame in which the objects lie. This information can be formalised to obtain possible contradictions between recognition hypotheses. The details of the idea appear below; an example that illustrates the reasoning appears in section 2.1.1.

2.1 Theory

Consider two coplanar plane objects, o_1 and o_2, for which we have models m_1 and m_2. Write the transformation from m_k to o_k as E_k, and the transformation from m_k to the image frame as P_k. E_k is Euclidean, and moves the model into its position in the world frame. Write the projective transformation from the world frame to the image frame as Q. Then $P_k = QE_k$, so that $P_1^{-1}P_2 = E_1^{-1}Q^{-1}QE_2 = E_1^{-1}E_2$ which is Euclidean. Since labelling an image curve with a particular model name determines the transformation from that model's frame to the image frame, the pairwise consistency of labellings can be checked by forming a system of matrices $P_1^{-1}P_2$, and checking whether they are Euclidean.

Objects can consist of points, or of some mixture of points, lines, conics, and other curves, as long as a projective transformation can be computed from the object frame to the image frame. This observation also justifies our emphasis on coordinate frames, rather than on particular geometric configurations. Section 2.2 details the ambiguities implicit in this scheme.

If a labelling is consistent it is possible to reconstruct the whole plane, in the frame of one given model, since $P_j^{-1}P_i$ gives the transformation from the configuration in m_i's frame to that in m_j's frame. To reconstruct the plane in, say, m_1's frame, for all m_k compute $P_1^{-1}P_k$ and then apply this map to m_k; this will give a collection of objects in m_1's frame

2.1.1 Example

Given an image of three objects, o_1, o_2 and o_3, which are plane and coplanar, and which are instances of known models, the recognition system would proceed as follows:

1. **Determine projective equivalence classes** by indexing the model-base using appropriate projective invariants. For each object, the indexing stage returns a collection of possible Euclidean models to which it might correspond. Assume that the response is: $o_1 \rightarrow (m_1, m_4, m_7)$, $o_2 \rightarrow (m_2, m_5, m_8)$ and $o_3 \rightarrow (m_3, m_6, m_9)$.
2. **Determine all image-model transformations** for every image-model correspondence using a least squares process. Call the transformations between o_i and m_j, P_{ij}. There is a total of nine transformations.
3. **Test consistency** between model hypotheses for each pair of objects. Thus, for o_1 and o_2, form the matrices:

$$P_{11}^{-1}P_{22}, P_{14}^{-1}P_{22}, P_{17}^{-1}P_{22}, P_{11}^{-1}P_{25}, P_{14}^{-1}P_{25},$$

$$P_{17}^{-1}P_{25}, P_{11}^{-1}P_{28}, P_{14}^{-1}P_{28}, P_{17}^{-1}P_{28}$$

   if, say, $P_{11}^{-1}P_{22}$ is very close to being a Euclidean matrix, then accept the pairing (o_1m_1, o_2m_2). For this example, assume that the pairs: (o_1m_1, o_2m_2), (o_1m_1, o_3m_3), (o_2m_2, o_3m_3), (o_1m_7, o_2m_8) are consistent.
4. **Form the longest possible consistent hypothesis** by merging consistent pairings. Thus, in this example, the longest consistent hypothesis is (o_1m_1, o_2m_2, o_3m_3). This is accepted as the correct labelling for the image; consistency is defined by ensuring that each object has at most one label, so that two pairings are consistent if they refer to distinct objects, or if they assign the same labels to objects that they share. The other possible consistent labelling is (o_1m_7, o_2m_8), which is the result of an ambiguity.

2.2 Ambiguities

An ambiguous image supports two or more consistent labellings that are indistinguishable, one of which will be correct. Ambiguities arise from quite complex interactions between properties of the image and of the modelbase; some modelbases may not admit ambiguities. We assume that projectively distinct objects receive distinct labels, and study inherent ambiguities in the consistency process.

> **Definition:** Two pairs of models, say $\{m_1, m_2\}$, $\{m_1', m_2'\}$, admit an ambiguous labelling if there is some image containing objects $\{o_1, o_2\}$ so that $\{o_1m_1, o_2m_2\}$, and $\{o_1m_1', o_2m_2'\}$ are both consistent labellings of the objects.

Admitting an ambiguous labelling poses a stringent constraint on the models in the modelbase. If two pairs of models, say $\{m_1, m_2\}$, $\{m_1', m_2'\}$, admit an ambiguous labelling, then there exist projectivities $P_{11'}$ and $P_{22'}$ such that $m_1' = P_{11'}m_1$ and $m_2' = P_{22'}m_2$

It can be shown that, for the configurations to be ambiguous, there are Euclidean transformations E_a, E_b such that $P_{11'} = E_a P_{22'} E_b$. This is an action of two copies of the Euclidean group on the space of projective transformations, and invariants can be obtained for this action. For example, writing the i, j'th component of a matrix Q as q_{ij}, the expression $(q_{20}^2 + q_{21}^2)^3 / Det(Q)^2$ is an invariant under this action. This means that this expression must take the same value for $Q = P_{11'}$ as it does for $Q = P_{22'}$. Thus, for a modelbase to admit ambiguities, it must contain at least two pairs of projectively equivalent models, where the projectivities between the ambiguous models *have special properties*. In turn, this statement suggests that ambiguities are unlikely. However, there is some reason to believe that modelbases containing man-made objects are likely to contain ambiguities; for example, a sequence of scaled versions of several different objects will certainly give rise to ambiguities.

2.3 Implementation details and experiments

A system implementing the approach described has been demonstrated on real images of simple scenes, using a stripped-down version of the system described in detail in [9] to perform early vision and fitting. Indexing, though performed in a reimplementation of that system, follows essentially the same pattern, but in the present system a successful indexing attempt returns a collection of Euclidean models. To focus attention on the Euclidean labelling properties of the system, the model base contains only one projective equivalence class of models, consisting of five projectively equivalent but Euclidean distinct models, so that for all known models the projective invariants are the same. The models all consist of polygons with five sides; objects were obtained by cutting these polygons out of black cardboard.

The success of this approach can be measured both by determining its effectiveness in labelling the scene, and by looking at Euclidean invariants of an unknown object, coplanar with the known objects in the scene, and reconstructed using the techniques described; stability in these invariants means that the Euclidean labelling was sufficiently successful to allow the Euclidean structure of other objects to be determined from the labelling. Some results are shown in figures 1 and 3.

3 3D objects

The situation is more difficult when the objects are three-dimensional. It is known that the projective geometry of a range of polyhedra can be recovered partially or completely from a single perspective view with an uncalibrated camera (see [10, 11]). In turn, projective invariants

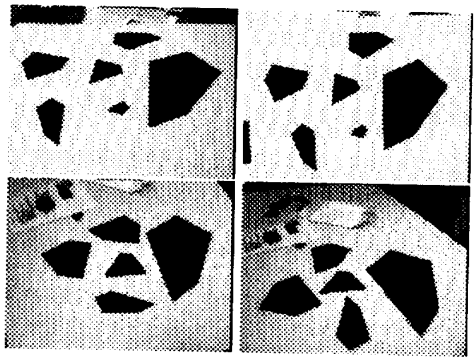

Figure 1: Six examples of scenes containing known coplanar plane objects (five sides), and an unknown object, imaged with a projective camera.

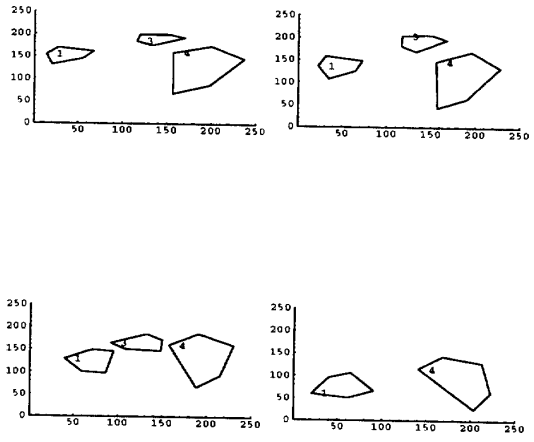

Figure 2: The Euclidean labels chosen by a global consistency analysis of the corresponding scenes in figure 1, superimposed on the backprojected outlines of the objects, which are five-sided plane polygons. Although the labellings are correct, the system consistently ignores one object (apparently as a result of a segmentation difficulty with one poorly cut corner).

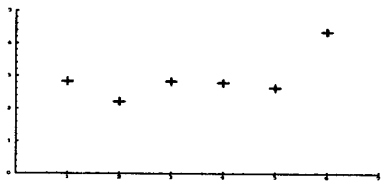

Figure 3: The graph shows the area of the unknown quadrilateral, measured in the six images shown above by computing a backprojection based on the Euclidean recognition hypotheses; note that the area is relatively stable.

can be computed and used to index the polyhedron in a model-base. Appropriate polyhedra are position-free in views, and tend to contain many faces with four or more vertices. However, for most situations, the resulting projective ambiguity is too great. To proceed, it is necessary to assume that that, for each generic view of every polyhedron in the modelbase, a distinctive projective structure can be recovered (details appear in [10]).

The perspective projection from 3D projective space, P^3, to the image plane, P^2, is modelled by a 3×4 projection matrix, $\mathbf{P}$, so that

$$\mathbf{x} \approx \mathbf{PX} \tag{1}$$

where homogeneous coordinates are used, $\mathbf{X} = (X, Y, Z, 1)^t$, $\mathbf{x} = (x, y, 1)^t$ and $\approx$ indicates equality up to a non-zero scale factor. Following Hartley [4], we partition $\mathbf{P}$ as

$$\mathbf{P} = (\mathbf{M} | -\mathbf{Mt}) \tag{2}$$

where $\mathbf{t}$ is the focal point (since the focal point projects as $\mathbf{PX} = \mathbf{0}$). Provided the first 3×3 matrix, $\mathbf{M}$, is not singular (i.e. the focal point is not on the plane at infinity), $\mathbf{P}$ can always be partitioned in this way.

To determine the position of the focal point in the model frame proceed as follows:

1. Compute the projection matrix $\mathbf{P}$ from the known model vertices and their corresponding image positions.
2. Partition $\mathbf{P}$ as above. This determines $\mathbf{t}$, which is the focal point in the object's frame.
3. The rays passing through other image outline points are given as the pre-image in $\mathbf{P}$ of the image points.

An alternative construction, due to Mohr, is also possible [7]. Labelling an image with a consistent Euclidean labelling proceeds as follows:

- Determine a set of projectively equivalent, Euclidean inequivalent labels for each polyhedron visible, using the indexing methods of [10].
- For each labelling of each item, compute the focal point and an appropriate cone of rays in the object's frame.
- Construct the largest pairwise consistent labelling of the scene, where pairwise consistency is checked by determining that the focal point and cone of rays constructed by assuming one object, is Euclidean equivalent to that constructed by assuming a second object.

As in the case of coplanar plane objects, although a correct labelling must be consistent in the sense given, a consistent labelling may not be correct. Thus, for particular scenes and particular model-bases, a unique labelling may not, in fact, be possible.

3.1 Ambiguities

The range of possible ambiguities in the case of polyhedral objects is wider than in the case of plane objects. Problems arise both as a result of viewing and self-occlusion issues and because the reconstructed polyhedra do not share the same projective frame. Ambiguities resulting from self-occlusion are not treated here, as the nature of the ambiguities depends in a complicated way on the structure of the recognition system. If the projective class of the object has been correctly recovered, and the cones through the object vertices and the focal point have been constructed correctly, the following, tractable question remains; given an image, and the projective structure of the polyhedra represented in that image, what ambiguities exist in the Euclidean labelling process described?

There is now a second source of ambiguity; many distinct objects can produce the same cone of rays through the focal point. It can be shown that, for a modelbase to admit an ambiguity, it must contain four elements p_1, p_2, p_1', p_2', with p_i and p_i' projectively equivalent, and where the transformations between the model frames satisfies:

$$P_{p_2 p_2'} = ET^{-1}DTP_{p_1 p_1'}$$

for some arbitrary translation T, elation D and Euclidean transformation E. Note that, since the total number of arbitrary degrees of freedom is 13, not every pair of transformations $P_{p_2 p_2'}, P_{p_1 p_1'}$ will have this property. As a result, not every modelbase admits an ambiguity, and unambiguous labellings appear possible for at least some modelbases. Many man-made objects yield modelbases that admit ambiguities (for example, a modelbase containing only cubes of different sizes).

3.2 Experiments

A system implementing the approach described has been demonstrated on real images of simple scenes. The modelbase contains five polyhedral objects, all projectively equivalent. Polygon vertices are marked in the image by hand; all further processing is automatic. Figure 4 shows

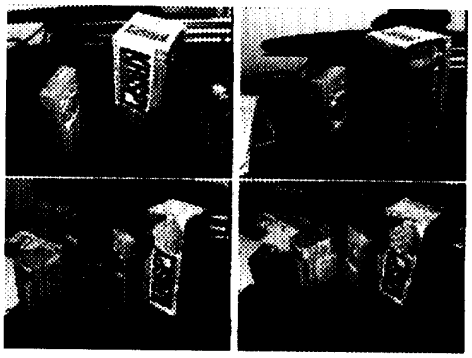

Figure 4: Four examples of scenes containing known polyhedral objects all of which are projectively equivalent, imaged using a perspective camera.

typical images; figure 5 shows the Euclidean labelling for corresponding images and figure 6 shows a reconstruction. The reconstruction techniques used at present can produce a reconstruction that is within an improper rotation (with negative determinant) of the original world; this appears to be an intrinsic ambiguity in the purely projective methods used, and may be overcome by considering the direction in which the camera is pointing. The techniques do not extend to reconstructing unknown objects in the way that the plane techniques do.

4 Discussion

We have shown methods for using geometric consistency in 2D from 2D recognition, and in 3D from 2D recognition; because the argument is based on frames and maps, the 2D from 2D argument carries over to the 3D from 3D case without modification (for example, on the output of "uncalibrated stereo"[2]).

Linking these ideas is the observation that *recognition hypotheses are frame hypotheses.* When a program asserts that some Euclidean model produced an image observation, it is making a statement about camera position and internal parameters. Such statements lead to global consistency constraints that must hold. If, in a world with many known objects, objects can be recognised effectively and unknown objects can be reconstructed up to a Euclidean ambiguity without camera calibration, there is no reason to calibrate the camera. Other constraints can be applied to the reconstruction (for example, using known objects, occlusion cues and up-vector estimates to bound the distance to the object). This paper has dealt with discrete modelbases; the case of parametrised systems of models bears further investigation.

More global consistency mechanisms are available - for example, the orderly and effective exploitation of the potential of t-junctions to explain occlusions. Many other sources of information, not necessarily primarily geometric, could be used to inform and strengthen recognition

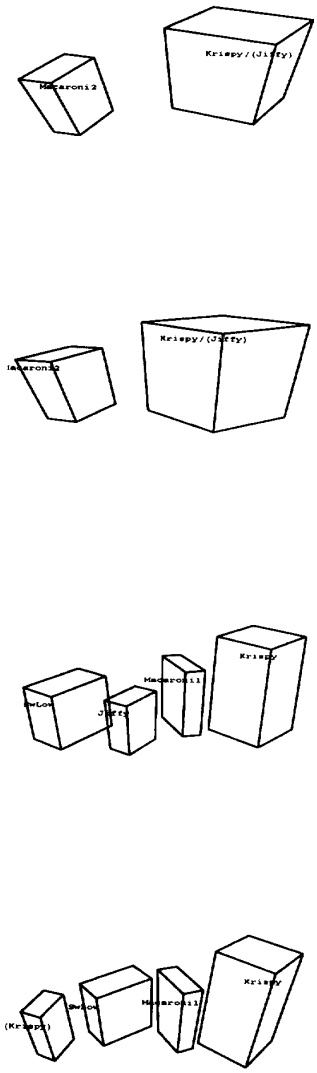

Figure 5: Euclidean labels chosen by a global consistency analysis of the scenes in figure 4, superimposed on the backprojected outlines of the objects. A label "i/j" means that there is a consistent interpretation where the object is either object i or object j. Incorrect labels are enclosed in parentheses.

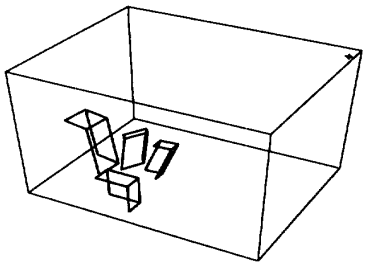

Figure 6: A Euclidean reconstructions of the polyhedral world shown in one image (bottom left image, all labels correct) taken from figure 4. The focal point is the marked point in the top right-hand corner of the figure. Note the distortions of the boxes, caused by mapping all boxes into the frame of one box; more sophisticated techniques might distribute error more evenly. The focal point is included so the reader can assess the effectiveness of the reconstruction by comparing with figure 4, which is seen from a different viewpoint; note that, for example, the bases of all boxes are near to coplanar.

hypotheses (and thereby shore up minor failures of the consistency mechanism).

Finally, consistency mechanisms of the type described are most effective in worlds well-populated with familiar objects. We believe that the tremendous potential power of a global consistency analysis will become most important in a system with a large modelbase, operating in a complex world.

Acknowledgements

DAF, JLM and AZ were all supported in part by a grant from United States Air Force Office of Scientific Research AFOSR-91-0361. DAF was supported in part by the National Science Foundation under award no. IRI-9209729, and in part by a National Science Foundation Young Investigator Award with matching funds from GE, Eugene Rikel, Rockwell International and Tektronix. JLM was supported in part by by General Electric. AZ was supported in part by ESPRIT BRA project "VIVA". CAR was supported by a grant from General Electric.

References

[1] Binford, T.O., Levitt, T.S., and Mann, W.B., "Bayesian inference in model-based machine vision," in Kanal, L.N., Levitt, T.S., and Lemmer, J.F., *Uncertainty in AI 3*, Elsevier, 1989.

[2] Faugeras, O.D. "What can be seen with an uncalibrated stereo rig?", Proc. European Conference on Computer Vision, 1992.

[3] Forsyth, D.A., Mundy, J.L., Zisserman, A.P., Coelho, C., Heller, A. and Rothwell, C.A. "Invariant Descriptors for 3-D Object Recognition and Pose," *PAMI*-13, No. 10, p.971-991, October 1991.

[4] Hartley, R.I. "Cheirality invariants," *Proc. DARPA Image Understanding Workshop*, pages 745–753, 1993.

[5] Lamdan, Y., Schwartz, J.T. and Wolfson, H.J. "Object Recognition by Affine Invariant Matching," Proceedings CVPR88, p.335-344, 1988.

[6] Liu J., Mundy J.L., Forsyth D.A., Zisserman A. and Rothwell C.A., "Efficient Recognition of Rotationally Symmetric Surfaces and Straight Homogeneous Generalized Cylinders", *CVPR*, 1993.

[7] Mohr, R., Morin, L. and Grosso, E., "Relative positioning with uncalibrated cameras," in Mundy, J.L and Zisserman, A. (eds) *Geometric Invariance in computer vision*, MIT press, 1992.

[8] Ponce, J. "Invariant properties of straight homogenous generalised cylinders," *IEEE Trans. Patt. Anal. Mach. Intelligence*, **11**, 9, 951-965, 1989.

[9] Rothwell, C.A., Zisserman, A., Mundy, J.L. and Forsyth, D.A. "Efficient Model Library Access by Projectively Invariant Indexing Functions", Proceedings CVPR92, p.109-114, 1992.

[10] Rothwell, C.A., Forsyth, D.A., Zisserman, A. and Mundy, J.L., "Extracting projective structure from single perspective views of 3D point sets," International Conference on Computer Vision, Berlin, 573-582, 1993.

[11] Sugihara, K., *Machine Interpretation of Line Drawings*, MIT Press, 1986.

[12] Taubin, G. and Cooper, D.B. "Recognition and Positioning of 3D Piecewise Algebraic Objects," Proceeding DARPA Image Understanding Workshop, p.508-514, September 1990.

[13] Ulupinar, F, and Nevatia, R. "Shape from Contour using SHGCs," *Proc. ICCV*, Osaka, 1990.

[14] Weiss, I. "Projective Invariants of Shapes," Proceedings DARPA Image Understanding Workshop, p.1125-1134, April 1988.

[15] Zerroug, M. and Nevatia, R., "Using invariance and quasi-invariance for the segmentation and recovery of curved objects," Proc 2'nd Joint Workshop on the Applications of Invariance in Computer Vision, Ponta Delgada, 1993.