Introduction

David Forsyth and Joe Mundy

Computer Science Division, U.C. Berkeley, Berkeley, CA 94720, USA
daf@cs.berkeley.edu,
WWW home page: http://www.cs.berkeley.edu/~daf

Abstract. Our understanding of object recognition can address the needs of only the most stylised applications. There is no prospect of the automated motorcars of Dickmanns *et al.* knowing what is in front of them anytime soon; searchers for pictures of the pope kissing a baby must search on a combination of text, guesswork and patience; current vision based HCI research relies on highly structured backgrounds; and we may safely guess that the intelligence community is unlikely to be able to dispense with image analysts anytime soon. This volume contains a series of contributions that attack important problems in recognition.

1 What we do well

We can solve some problems rather well. Unfortunately, these problems seem to be connected only rather tenuously with potential applications of object recognition. This is because all current algorithms for object recognition all model recognition as direct exploration of correspondence. Each belongs to a small number of types, each with characteristic individual misbehaviours. Much of this material is (or should be!) common cause, and will be reviewed only very briefly.

1.1 Geometric detail for point-like primitives

Point-like primitives project like points; points project to points, lines to lines, conics to conics, etc., with no more complex geometric behaviour than occlusion. This means that, though objects can look different from different views, these changes are highly structured. For point-like primitives, correspondences between image and object features can be searched in various ways for detailed geometric models (a fair sample of this literature includes [2–4, 6, 7, 9–11, 18, 19, 23–25]; there are hundreds of papers dealing with variant algorithms or specific technical issues). Typical algorithms of this class can usually find instances drawn from a small number of object models (sometimes from parametric families) against a background of moderate clutter and despite the effects of occlusion. Problems include: the restriction to exact geometry; the limited number of primitives; the difficulty of building sufficient appropriate models; the general unreliability of verification and segmentation; and the restriction to a small number of models.

1.2 Some cases of curved primitives

The difficulty with curved surfaces is that, though the change of outline with viewpoint is highly structured, this structure is generally complicated — so complicated that a detailed record is basically unmanageable. A great deal is known about particular cases, none particularly practical at present; all this information tends to suggest that *given a geometric class*, outline is a powerful constraint on geometry and viewpoint. There appear to be class-specific constraints on outlines for all cases, though only very few useful cases are known. This picture is complicated by the absence of any kind of covariant approximation theorem. We know no useful geometric theorems about the outlines of surfaces that "almost" belong to a particular class, and all indications are that such theorems are difficult to get.

1.3 Template matching

Template matching works rather well on some kinds of recognition problem (finding frontal faces is a good example; while some details are being worked out, the problem is quite well understood [1, 8, 14, 20, 21, 16]). Crude template matching behaves poorly under change of aspect and of illumination, but this can be resolved by adopting parametric families of templates (e.g. [12, 13, 15]). This approach requires objects to appear on their own, and becomes unwieldy for all but a few degrees of freedom. For more complex cases, models can consist of composites of primitives, which are themselves from parametric families of templates (e.g. [1, 5, 8, 14, 17, 22]).

2 What this volume describes

The authors of this piece see two quite different agendas for future research on object recognition. These agendas have been sketched in the next two papers. In the first, Forsyth argues that the main current difficulty in building acceptable recognition systems is the poor management of uncertainty within those systems; some sort of Bayesian reform is long overdue. He takes the position that many difficult issues — for example, what is an appropriate representation for a particular set of objects? or how should distinct cues be used to come up with a single recognition strategy? — are essentially empirical and statistical in nature, and that the community should be attempting to master and apply various statistical methods. In the second, Mundy argues that empirical methods can offer no fundamental breakthrough, but that better understanding of appropriate physical and formal models — for example, an understanding of the relationship between the geometry of objects, their surface properties and their image appearance — might. The rest of this volume consists of contributions from leading researchers, dealing with shape, shading, grouping, cue integration and recognition. Each section contains both research papers and papers with review and introductory material.

2.1 Shape

For the vast majority of recognition problems, shape is an important cue. In "*Shape models and object recognition,*" Jean Ponce and colleagues describe the current state of the art in shape modelling. The paper demonstrates a variety of representations for recognition, including their reworking of the idea of a generalised cylinder. They describe representations in terms of "parts", and show one way to get a canonical decomposition of an object into parts from an image. Finally, the paper discusses how the effects of viewing direction are themselves affected by image scale.

The relationship between image measurements and object shape is complicated. This relationship has combinatorial, as well as geometric, aspects as Stefan Carlson shows in "*Order structure, correspondence and shape based categories.*" The order structure of groups of points does not change arbitrarily as the groups are projected to an image; this fact can be used to reason about the object in view.

Plane curves are often subject to some form of group action before they are seen in an image. A standard mechanism for discounting this group action is to represent a curve by a plot of invariant properties against an invariant parameter. This approach usually requires that one be able to measure an inconvenient number of derivatives. In "*Quasi-invariant parametrisations and their applications in computer vision*", Sato and Cipolla show how to use a quasi-invariant parametrisation for representing curves. This has the advantage that fewer derivatives need be measured if one is willing to accept some variation in the representation.

2.2 Shading

Illumination is an important source of the variation in object appearance. In "*Representations for recognition under variable illumination,*" Kriegman, Belhumeur and Georghiades describe a device called the *illumination cone*, which records the appearance of an object under all possible illuminants. They show that this convex cone can be used to recognise faces, despite substantial shading variations which confound the usual strategies. This theory does not treat shadows, which are dealt with in "*Shadows, shading, and projective ambiguity,*" by Belhumeur, Kriegman and Yuille. Objects that give the same shadow pattern in a fixed view are geometrically equivalent up to a *generalised* bas-relief *ambiguity*; furthermore, objects can be reconstructed up to this ambiguity from multiple images under multiple, unknown light sources.

2.3 Grouping

Grouping is the process of assembling image components that appear to belong together. There are a variety of reasons that components may belong together — in "*Grouping in the normalized cut framework*" Malik and colleagues describe a mechanism that segments an image into components that satisfy a global

goodness criterion. Their mechanism is attractive because (as they show) it can be used for a variety of different grouping cues, including intensity, texture, motion and contour.

Another reason that image components may belong together is that they lie on the same plane in the world. In "*Geometric grouping of repeated elements within images,*" Schaffalitzky and Zisserman show how to determine when a pattern consists of plane elements, repeated according to a variety of rules. They determine the basic element of the pattern, and can then reconstruct the pattern because repetition rules on the plane lead to repetition rules in the image in a well defined way.

In satellite image applications, both the pose of the ground plane and the calibration of the camera are usually known. This means that one can tell whether objects lying on the ground plane appear to have a symmetry, as Curwen and Mundy demonstrate in "*Constrained symmetry for change detection.*" This cue makes it possible to group together interesting edge fragments, which are likely to have come from, for example, human artifacts.

In "*Grouping based on coupled diffusion maps,*" Proesmans and Van Gool describe the use of anisotropic diffusion processes for grouping. In this approach, pixels are connected by a diffusion process that is modified to prevent smoothing over large gradients. They show examples where this process is used for segmentation, for detecting symmetries, for stereoscopic reconstruction and for motion reconstruction.

2.4 Representation and recognition

In "*Integrating geometric and photometric information for image retrieval,*" Schmid, Zisserman and Mohr describe a local representation of images in terms of *interest points*. These interest points are defined using photometric information; collections of interest points yield a representation that incorporates information about shape and about photometry. This representation can be used to match query images against an image database. They show how the representation can be extended to 3D curves, using the osculating plane of the curve, to obtain a matching process that can match 3D curves between small baseline stereo pairs.

Mundy and Saxena describe another approach for integrating photometric and geometric information in "*Towards the integration of geometric and appearance-based object recognition.*" They propose using facet models of surface brightness to index object identity, and compare renderings using the Oren-Nayar model of surface reflectance to real data. This approach is extended to colour images in "*Recognising objects using color annotated adjacency graphs,*" by Tu, Saxena and Hartley. In this paper, objects are represented by adjacency graphs of coloured faces, which are matched using a graph matcher. Their approach uses a method from linear algebra for computing graph compatibilities that is somewhat reminiscent of the method of Malik *et al.*.

In "*A cooperating strategy for object recognition,*" Chella, Di Gesù, Infantino, Intravaia and Valenti describe a complete recognition system. Objects are represented using either interest points or a discrete symmetry transform; their

system uses cooperating agents to mediate the crucially important relationship between top-down and bottom-up information flow.

2.5 Statistics, learning and recognition

Methods from statistics and from statistical learning theory are starting to have a substantial impact on the practice of computer vision. A classical statistical problem that turns up in many different vision applications is *model selection* — from which of several models was the data set obtained? This issue must be dealt with in recognition, where it is often known as *verification* — the issue is *do the pixels in this region come from an object or the background?* — in structure from motion (*what kind of camera produced this scene?*), and in a variety of other areas of vision. Torr reviews this topic in "*Model selection for two view geometry: a review.*"

Forsyth, Haddon and Ioffe describe probabilistic algorithms for object recognition in "*Finding objects by grouping primitives.*" These algorithms are structured around the use of simple primitives: firstly, people and animals are represented as cylinders, and can then be found by a grouping process that assembles cylinders that together "look like" a person; secondly, folds in cloth are found using a classifier that recognizes their appearance, and then the Markov chain Monte Carlo method is used to group them into assemblies that look like buckle patterns in clothing. It is usually difficult to know what to use as a primitive in this sort of work; the paper argues that the criteria are statistical in nature, and that primitives could be learned from data.

A more detailed committment to learning appears in in "*Object Recognition with Gradient-Based Learning.*" by LeCun, Haffner, Bottou and Bengio. Images of hand-written characters are filtered by a sequence of filters at various scales, and the filter outputs passed to a neural net classifier. Not only the classifier, but the filters themselves are learned from example data, using a procedure known as *gradient based learning*. Convolutional neural networks are then rigged together, to yield a *space displacement network*, that can read a sequence of handwritten characters. Information about the probabilistic structure of handwritten numbers is incorporated using a *graph transformer network*.

References

1. M.C. Burl, T.K. Leung, and P. Perona. Face localisation via shape statistics. In *Int. Workshop on Automatic Face and Gesture Recognition*, 1995.
2. O.D. Faugeras and M. Hebert. The representation, recognition, and locating of 3-D objects. *International Journal of Robotics Research*, 5(3):27–52, Fall 1986.
3. D.A. Forsyth, J.L. Mundy, A.P. Zisserman, C. Coelho, A. Heller, and C.A. Rothwell. Invariant descriptors for 3d object recognition and pose. *PAMI*, 13(10):971–991, 1991.
4. W.E.L. Grimson and T. Lozano-Pérez. Localizing overlapping parts by searching the interpretation tree. *IEEE Trans. Patt. Anal. Mach. Intell.*, 9(4):469–482, 1987.

5. C-Y. Huang, O.T. Camps, and T. Kanungo. Object recognition using appearance-based parts and relations. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 877–83, 1997.
6. D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proc. Int. Conf. Comp. Vision*, pages 102–111, London, U.K., June 1987.
7. D.J. Kriegman and J. Ponce. On recognizing and positioning curved 3D objects from image contours. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12(12):1127–1137, December 1990.
8. T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. In *Int. Conf. on Computer Vision*, 1995.
9. D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.
10. J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT Press, Cambridge, Mass., 1992.
11. J.L. Mundy, A. Zisserman, and D. Forsyth. *Applications of Invariance in Computer Vision*, volume 825 of *Lecture Notes in Computer Science*. Springer-Verlag, 1994.
12. H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *Int. J. of Comp. Vision*, 14(1):5–24, 1995.
13. S.K. Nayar, S.A. Nene, and H. Murase. Real time 100 object recognition system. In *Int. Conf. on Robotics and Automation*, pages 2321–5, 1996.
14. M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet templates. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 193–9, 1997.
15. H. Plantinga and C. Dyer. Visibility, occlusion, and the aspect graph. *Int. J. of Comp. Vision*, 5(2):137–160, 1990.
16. T. Poggio and Kah-Kay Sung. Finding human faces with a gaussian mixture distribution-based face model. In *Asian Conf. on Computer Vision*, pages 435–440, 1995.
17. A.R. Pope and D.G. Lowe. Learning object recognition models from images. In *Int. Conf. on Computer Vision*, pages 296–301, 1993.
18. L.G. Roberts. Machine perception of three-dimensional solids. In J.T. Tippett et al., editor, *Optical and Electro-Optical Information Processing*, pages 159–197. MIT Press, Cambridge, 1965.
19. K. Rohr. Incremental recognition of pedestrians from image sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9–13, 1993.
20. H.A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing 8*, pages 875–881, 1996.
21. H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 203–8, 1996.
22. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.
23. D.W. Thompson and J.L. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *IEEE Int. Conf. on Robotics and Automation*, pages 208–220, Raleigh, NC, April 1987.
24. S. Ullman. *High-level Vision: Object Recognition and Visual Cognition*. MIT Press, 1996.
25. S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(10):992–1006, 1991.