Object recognition based on geometry: progress over three decades

The Royal Society

Phil. Trans. R. Soc. Lond. A 1998 **356**, 1213-1231
doi: 10.1098/rsta.1998.0218

Object recognition based on geometry: progress over three decades

By Joseph L. Mundy

*GE Corporate Research and Development, 1 Research Circle,
Niskayuna, New York, NY 12309, USA*

The evolution of object recognition systems over the last three decades has featured the use of geometric representations. In this paper, progress in object recognition will be reviewed and significant advances due to our deeper understanding of geometric relationships will be highlighted. An example of such progress is the development of projective and affine reconstruction from multiple uncalibrated camera views. This new understanding of the image projection of object geometry has had considerable impact on object representation and on grouping algorithms for recognition.

Keywords: recognition; classification; matching; invariants; grouping; appearance

1. Introduction

The central task for computer vision is to extract a description of the world based on images. Descriptions derived from images are essential to diverse applications of computers, such as virtual reality and human–machine interfaces. An important element of a description is the assertion that a specific individual object has been previously observed or that an object is similar to a set of objects seen in the past. This process of *recognition*, literally to RE-cognize, permits an aggregation of experience and the evolution of relationships between objects based on a series of observations. The ability to recognize objects in a cluttered scene with complex illumination and shadows has proven to be one of the most difficult challenges for computer vision. Progress has been gradual, but with significant advances since visual recognition became part of research in pattern recognition and artificial intelligence in the 1950s.

A great deal is now understood about the relationship between a geometric structure in three dimensions and its image projection. The process of constructing geometric descriptions from images has received considerable attention, and many approaches for indexing and classifying objects, based on geometric attributes, have been developed and implemented. There has also been progress in modelling the appearance of objects empirically from a set of training images. This empirical approach can provide a representation for objects for which a formal description is not yet known. Advances have also been achieved by integrating contemporary ideas about recognition into complete recognition systems, which provide benchmarks of progress.

2. What is recognition?

Not a small part of the difficulty of object recognition by computer is the illusiveness of the concept of recognition itself. Philosophers and psychologists have struggled to

Phil. Trans. R. Soc. Lond. A (1998) **356**, 1213–1231

Printed in Great Britain

TEX Paper

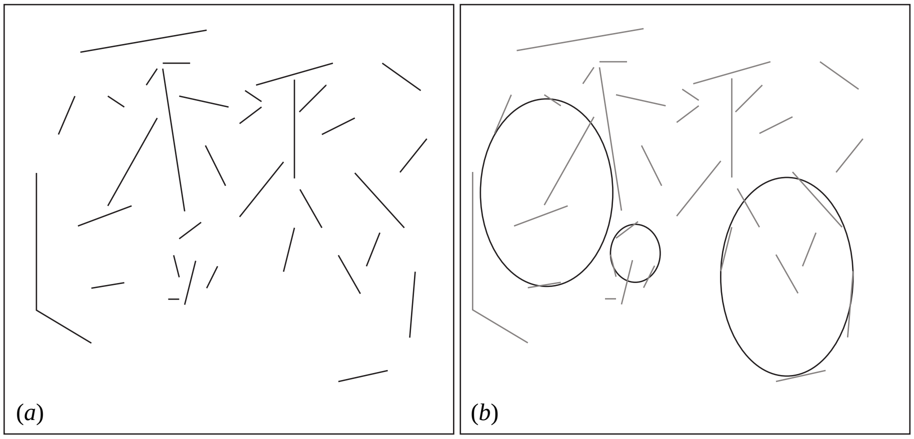

Figure 1. (*a*) An image has been segmented to form a set of fragmentary line segments. (*b*) The segments are seen to represent the partial boundaries of a bicycle, indicated by adding the wheels and drive sprocket. (Adapted from Lowe (1985).)

define an individual object and how objects are organized into classes. The philosophical state of affairs today is summarized by the following two principles: only individual objects exist; a class is defined by its individual members, which resemble each other.

For our purposes, it will be assumed that two objects are the same if a sufficient number of *significant* visual attributes are matched. Further, the definition of object classes is based on visual similarity and that the main purpose of classes is to enable *effective* recognition. The definition of what is significant and effective depends on the application, but is centred on the processes of image segmentation and spatial organization.

The process of recognition is composed of two parts: perception and classification. Perception is the process of assembling the features of an object in the image. A famous example by Lowe (1985) illustrates the perceptual grouping of line segments to form a bicycle. It is difficult to assemble the features of objects when seen against a complex background and with only partial extraction of the boundaries, as shown in figure 1. This process is also known as figure–ground separation.

Classification is the assignment of the set of assembled object features to an individual instance or perhaps to a larger class of objects. The problem of classification is illustrated by figure 2, which shows two examples of the common garden pepper. Classification has proven difficult for many concepts which are natural for human perception, such as *chair* or *table*. These classes are perhaps better treated by an analysis of function, as demonstrated by Stark & Bowyer (1991). However, the extraction of functional attributes, such as centre of gravity, from a single perspective image remains difficult.

These two parts of recognition, perception and classification, are independent since it is possible to separate an object from the background as an entity without being able to classify it. For example, consider a coloured bird against snow, or a moving object where figure–ground separation is achieved by motion differentiation. In many situations, however, an important role of class for computer-based object recognition is to define image constraints which can guide the perception process.

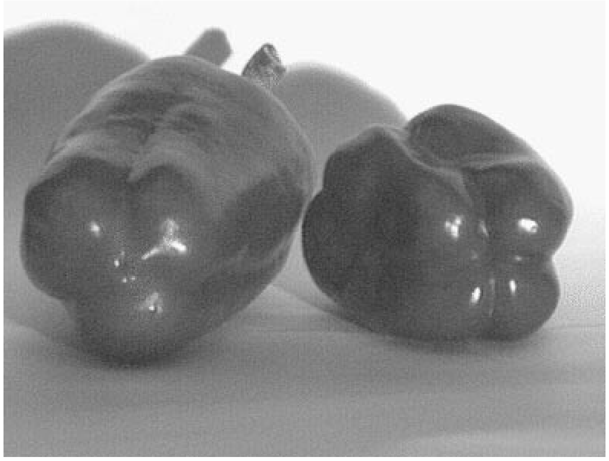

Figure 2. Two examples of peppers taken from the same plant. This amount of variation must be accepted in an object's shape while making the same classification for each individual.

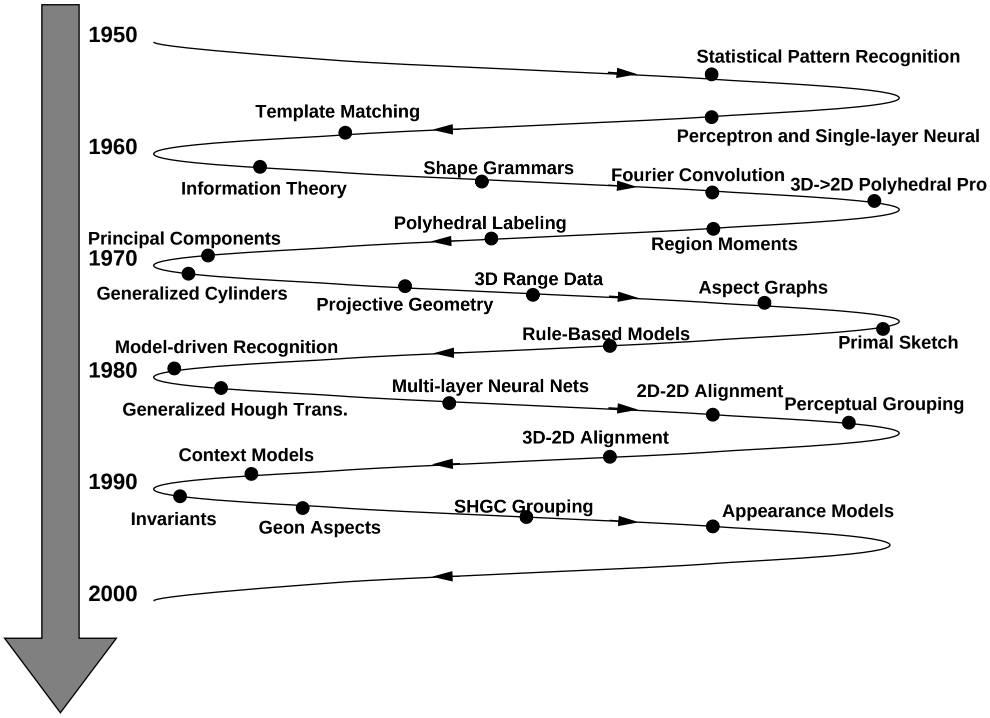

Figure 3. A history of some of the key ideas and paradigms for object recognition by computer.

3. A survey of ideas about recognition

The history of object recognition by computer vision extends back into the 1950s. The development has involved ideas or concepts which define various frameworks for carrying out object recognition. A time-line of some of these ideas is shown in figure 3. The key ideas are a mixture of representations, architectures and algorithms which

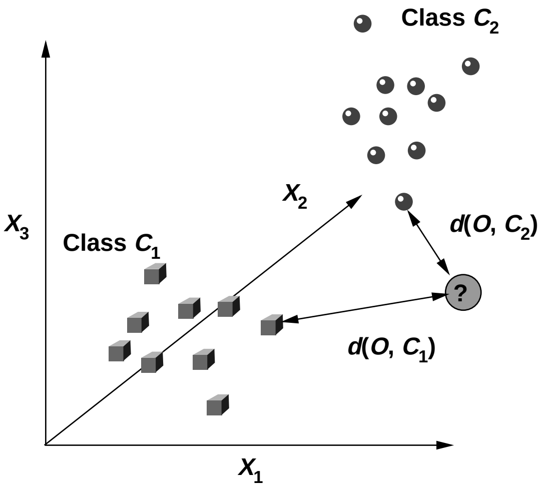

Figure 4. Object classes are defined by proximity in attribute space. Shown are a number of samples of two different classes. In this case, an object is described by three attributes, or features, X_1, X_2 and X_3. A distance measure, $d(X,Y)$ is defined on the space which determines the proximity of an unknown point to existing class samples.

have motivated extensive research activities. At many stages of this evolution, one approach or another was optimistically proposed to be a full solution to the problem of recognition. It now seems likely that no general solution exists. A competent recognition system will most certainly embody a multitude of representations and associated algorithms. Still, many important principles have been discovered which form the basis for designing and implementing recognition systems. These principles have emerged at various stages over the past 30 years and are often reinterpreted in terms of new capabilities and insights. This historical review will summarize the key ideas and their chain of development.

(*a*) *Object attributes as a geometric space*

The definition of class membership is proximity in a space of object attributes. If an object has properties which are *similar* to another object, then they are in the same class. In this case, similarity is equivalent to distance in the geometric attribute space. The concept is depicted in figure 4. The construction of this attribute space is dependent on the existence of a mapping of the attributes of an object, such as colour, intensity, texture, onto a set of numerical coordinates. If such a mapping can be defined, then a particular instance of an object can be represented as a point in the n-dimensional space of attributes. Object instances which belong to the same class are then *near* one another and form clusters. The classification process then becomes a problem of determining the distance from a point representing an unknown sample to the nearest cluster.

It is difficult to pinpoint the first use of this object representation and classification scheme, commonly known as image pattern recognition. However, certainly by the 1950s image pattern recognition systems based on n-dimensional feature vector classification were under widespread development (Chow 1957).

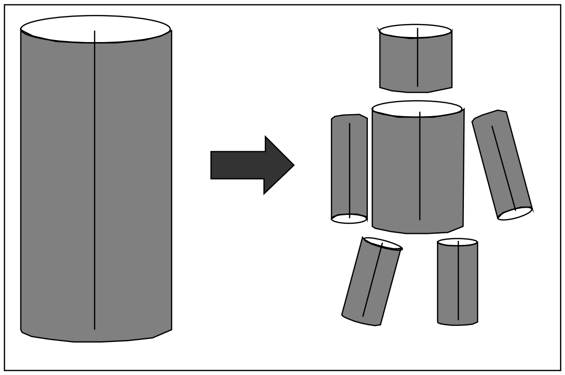

Figure 5. A three-dimensional structure, such as the human form, can be decomposed into cylindrical primitives. The configuration is characterized by the relative positions and orientations of the axes of symmetry. (After Marr & Nishihara (1978).)

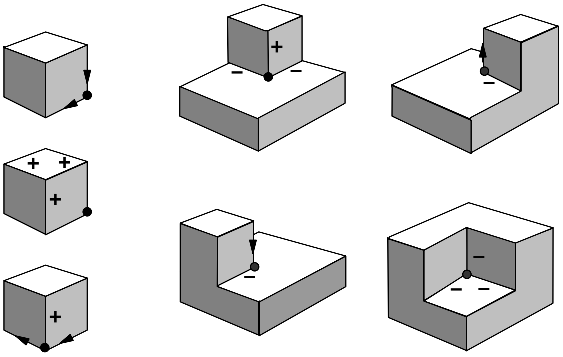

Figure 6. There are only a small number of possible views of tri-hedral junction. The possible edge labels are: convex, +; concave, −; and occluded, →. The figure illustrates a subset of the 16 feasible junction configurations, out of 208 possible combinations of edge labels. (After Raphael (1976).)

(*b*) *Structural decomposition*

Another key idea in the evolution of object recognition is the structural representation and decomposition of an object into primitive components. These components are then aggregated to form the overall object by a network of relationships among the components. In many structural representation schemes, these relations are geometric or topological. This idea of structural decomposition was used quite early in the representation of characters in support of optical character recognition (OCR) systems.

This approach evolved into a general approach called structural pattern recognition, or pattern grammars. The evolution reached a culmination in the mid-1970s (Fu 1974; Pavlidis 1977), when a full theory of pattern syntax and parsing was developed. Structures can be decomposed hierarchically into intermediate symbols and finally into so-called *terminal* symbols which are the actual primitives. The purely syntactic approach waned because many geometric relationships are difficult to express with simple formal grammars and full expressiveness is gained at the cost of intractable parsing complexity.

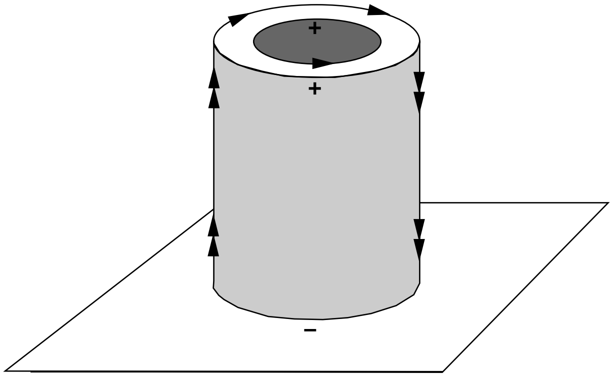

Figure 7. The labelling process can be extended to curved surfaces.

The idea of hierarchical reduction into primitives has remained an attractive concept because it seems the only way to control the complexity of object descriptions. The idea was extended to three-dimensional (3D) polyhedral structures by Roberts (1965) and to curved shapes by Binford (1971), who proposed the use of generalized cylinders as a generic 3D shape primitive. This idea was promoted also by Marr & Nishihara (1978) as a general representation of 3D shape. Figure 5 illustrates that the hierarchical decomposition provides a mechanism for levels of detail. The overall axis of symmetry summarizes the orientation of the body as a whole, the next level of detail gives the position and orientation of the symmetry axes of major components.

(*c*) *3D constraints induce 2D constraints*

(i) *Edge and junction labelling*

A key idea that has motivated many advances in object recognition is that constraints inherent in the configuration of an object in 3D space induces 2D image constraints which can be exploited in all stages of recognition processing. An early example of this concept is the idea of polyhedral labelling. As shown in figure 6 the image appearance of a polyhedron is highly constrained by the limited configurations of junction types and edge labels over all image viewpoints. These labels characterize all possible views of trihedral junctions. It is straightforward to enumerate the possible labellings by considering the volume occupancy of the underlying surface and for the discrete set of viewpoints which lead to a change in labelling.

Since these constraints were discovered by Huffman (1971) and Clowes (1971), the concept has been extended to curved surfaces by Chakravarty (1981), Shapira & Freeman (1978) and Malik (1987). An example of curved surface labelling is shown in figure 7. A new label, →→ is required for curved surfaces to represent the occlusion of the visible curved surface by itself. This occluding edge is called a *limb*. Polyhedral incidence constraints can be used in the reconstruction of the 3D geometry of an object from a single image. Sugihara (1986) showed that the requirement that the 3D planar polyhedral faces must intersect to produce the observable edges and vertices provides nearly enough constraints to reconstruct the solid geometry of the polyhedron from a single image.

These constructions established an algebraic basis for both the labelling and reconstruction of polyhedra. These ideas were extended by Rothwell *et al.* (1993) to

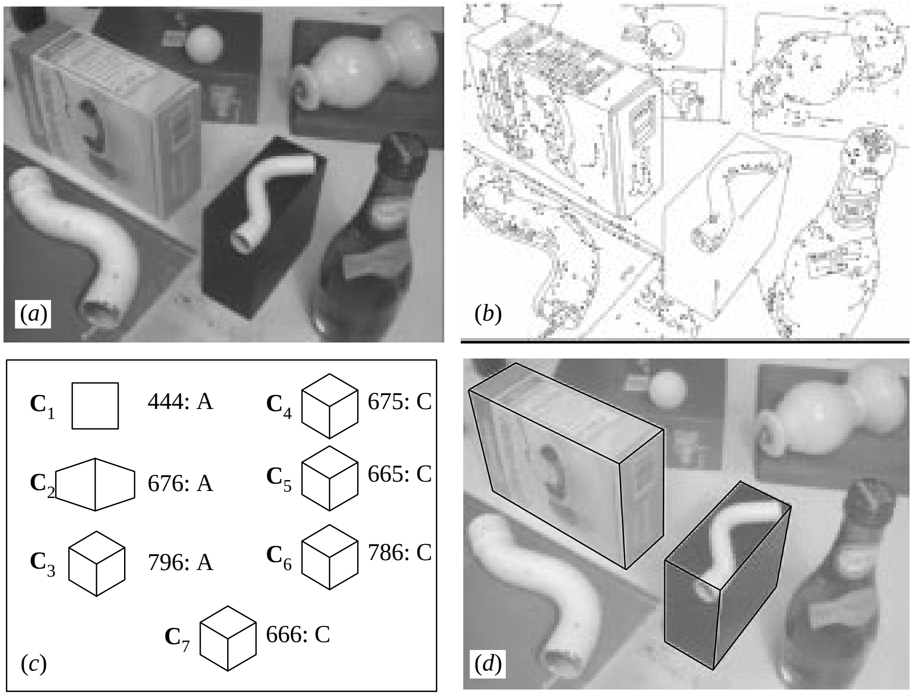

Figure 8. (*a*) Two rectangular prisms in a cluttered scene. (*b*) Edgel chains and junctions derived from intensity contrast boundaries. (*d*) The topological constraints for the prism model. A feasible view is defined by the recovered edges and vertices. Note that it is not necessary to recover a complete structure. (*c*) The matched prism. The projected 3D geometry is defined by the vanishing points of the parallel boundaries of the prism.

show that a viewpoint invariant description of a polyhedron can be derived in terms of 3D projective coordinates, based on the observed 2D image features. This use of polyhedral class constraints is demonstrated in the recognition of the two rectangular prisms shown in figure 8.

(ii) *Viewpoint consistency*

Another widely exploited constraint is that all points on a rigid object project with a single perspective transformation into the image. The principle of viewpoint consistency holds that all points on an object will project to their corresponding image positions for the same projection parameters.

This notion was initially used by Roberts (1965), who constructed a remarkably comprehensive recognition system for composite polyhedra. A key step in his system was the projection of the model into the image with a camera transform based on an initial set of image-to-model correspondences. In his system, a model hypothesis is considered valid if the projected model boundaries are in correspondence with those extracted from the image. The basic elements of his approach, i.e. bottom-up grouping followed by model verification based on viewpoint consistency, are still the basic paradigm for geometry-based recognition systems after three decades of ongoing research.

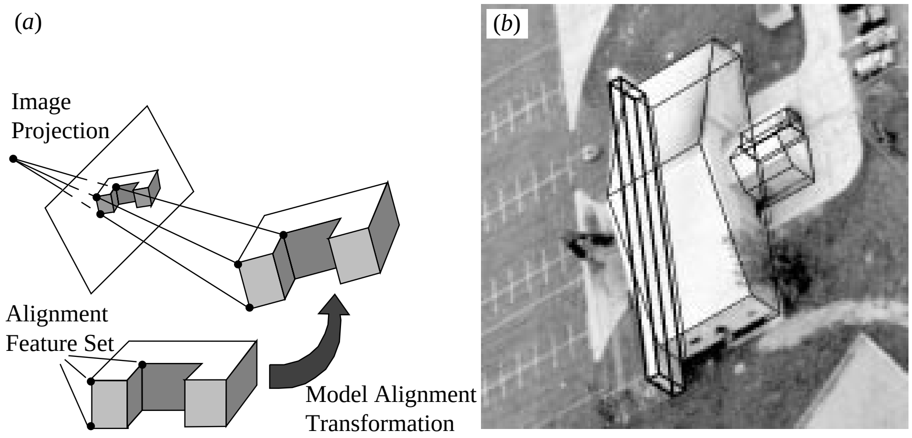

Figure 9. (*a*) A 3D model is transformed to align a set of features with an image projection of an object. When the camera is far enough from the object to eliminate strong perspective effects, the alignment of three object points is sufficient to compute the pose of the object with respect to the camera. (*b*) An example of the alignment of a model with a 2D image projection.

Viewpoint consistency was later exploited as a mechanism for perceptual organization where a small set of image features are grouped together corresponding to a subset of the features of a single object. These feature sets are selected to be sufficient to compute the camera transformation parameters. Therefore, for two such feature groups to be consistent, the camera parameters which they define must be the same. An early example of this approach is the work of Stockman (1987).

The view-consistency principle was also the basis of a series of recognition systems based on a hypothesize–verify search of the space of possible camera viewpoints and object models. In this approach, a small set of correspondences are used to project the model into the image and then the projected model is used to verify the correspondence set. If a restricted form of camera projection, called weak perspective, is used, three point correspondences are sufficient to compute the camera transformation as shown in figure 9. These search methods are now widely known as *alignment*, since the projection of the model is aligned with the image features (Huttenlocher & Ullman 1987). The hypothesis and verification process can be based on more extended sets of features which are grouped on generic relations which all objects satisfy. For example, Lowe used approximate constraints, such as *parallel lines in 3D space are parallel in the image*, to form larger object-feature group hypotheses before verification.

The hypothesize–verify process can proceed incrementally as in the work of Grimson & Lozano-Pérez (1984), who treated the process as a branch-and-bound search space, called the *interpretation tree*. Alternatively, the search can proceed in a a parallel fashion as in the vertex-pair algorithm of Thompson & Mundy (1987), where the computed model projection parameters are clustered in a six-dimensional space.

(*d*) *View-centred representation*

Another approach to describing objects for recognition is called a *view-centred* representation. In this approach the 3D object is represented by a number of 2D geometric image projections of the object, or even by actual intensity images. Perhaps

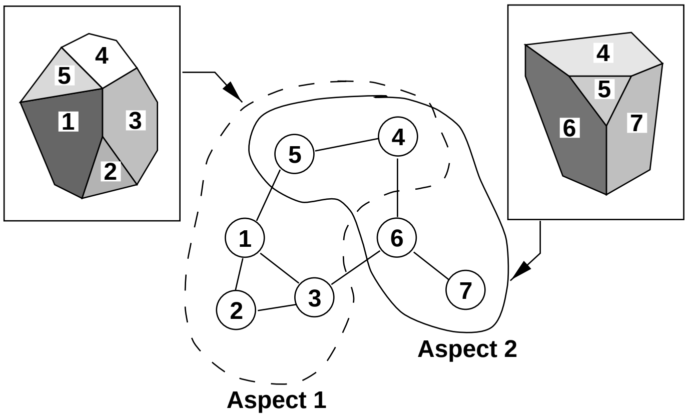

Figure 10. Two views, or aspects of a polyhedral solid are linked by common topological structures, i.e. faces 4 and 5. The projected topological structure is captured in the face-edge graph, where faces are indicated by labelled circles, joined by edges according to common edges on the 3D solid. (After Underwood & Coates (1975).)

the earliest mention of this idea was by Underwood & Coates (1975). They proposed that the description of an object can be learned by acquiring multiple, overlapping views of the object. The idea is illustrated in figure 10. The nodes and edges of the graph are faces and adjacency relations between the faces, respectively. As more views are acquired, the graph is extended until a complete view structure is obtained. The variation from one view to the next is defined by the topological structure of the view. The projected features in a new view are matched to a subgraph of the view structure to achieve recognition.

(iii) *Visual events*

The set of views of an object which are defined by changes in the topology of the image projection of the object is called an *aspect graph*. The concept of aspects was generalized to arbitrary 3D surfaces by Koenderink and van Doorn in 1979. A set of critical events is defined that arise from a change in the structure of the image projection of the surface. One of these critical events can be illustrated by the torus as shown in figure 11. The computation of critical visual events entails difficult problems in symbolic manipulation, for example, the outline curve of the torus for view (b) in figure 11 is defined implicitly by

$$\begin{aligned}
&13u^4r_1^4 - 8r_1^4v^4 + 24u^6v^2 - 6u^2r_1^6 + 40v^2r_2^4u^2\\
&\quad - 40v^2r_2^2u^4 - 88v^4u^2r_2^2 - 40r_1^2v^4u^2 - 40r_1^2v^4r_2^2\\
&\quad - 20r_1^4v^2r_2^2 - 24v^2r_2^6 + 52v^4r_2^4 - 48v^6r_2^2\\
&\quad + 48v^6u^2 + 52v^4u^4 - 20r_1^2r_2^4u^2 + 44r_1^2v^2r_2^4\\
&\quad - 8r_2^4u^4 + r_1^8 - 44u^4v^2r_1^2 + 13r_2^4r_1^4 - 12r_2^6r_1^2\\
&\quad - 6r_2^2r_1^6 + 22r_2^2u^2r_1^4 - 20r_2^2u^4r_1^2 + 20u^2v^2r_1^4\\
&\quad + 4r_2^8 + 4u^8 + 16v^8 - 12u^6r_1^2 = 0, \qquad (3.1)
\end{aligned}$$

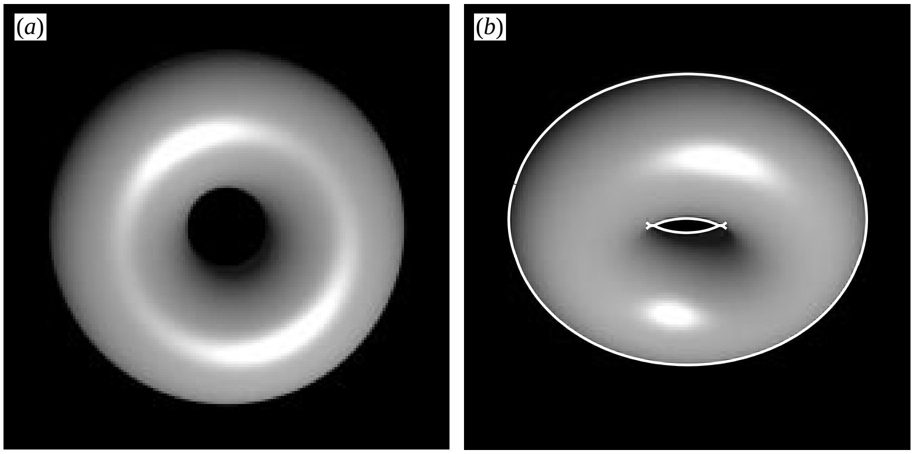

Figure 11. Critical viewing events occur when the viewpoint of the torus changes. When viewed from directly above in (*a*), the torus is bounded by two circular limbs. With a shift in viewpoint in (*b*), a cusp forms at each end of the hole and defines a new view structure of the torus.

where $(r_1 + r_2)$ and $(r_1 - r_2)$ are the outer and inner radii of the torus, respectively. In spite of this complexity, the recognition of curved surfaces requires an analytical representation for the appearance of objects like the torus. Ponce & Kriegman (1992) used polynomial expressions like equation (3.1) to align object outlines in images to recognize classes of algebraic surfaces.

Critical visual events can produce a very complex aspect graph, even for a relatively simple object. Consider a surface with 3D texture, such as the dimples on a golf ball. The number of critical views expands intractably with the number of surface undulations. The problem of complex view structure was discussed by Binford (1981) who pointed out that most critical views occur with very low probability over the set of all viewpoints.

This observation defines a relatively small number of views which characterize the major features of the object projection but not necessarily all minor topological configurations. Therefore, one can assume that a view is generic and does not involve critical alignment of boundary features. This assumption frees a recognition algorithm from considering complex feature relationships during the perception of object boundaries. Still, the problem of determining the scale below which critical viewing events can be ignored is an unresolved problem.

(*e*) *Appearance models*

The approach of representing an object by a set of views has been used in recognizing a large library of isolated objects. A system, called SLAM, developed by Murase & Nayar (1995), is capable of recognizing an object in an arbitrary intensity view by comparing it against thousands of views stored in a library of 100 different objects. Each object is represented by a large number of views taken with respect to variations that are expected to occur during recognition, such as rotation about the vertical axis of the object and illumination direction. This representation of the object is called an *appearance model.*

An object is classified by comparing the current image with the set of stored views for each object. This comparison is carried out efficiently by interpolating between compressed, stored views. The image compression is carried out using principal components which capture the main variations between images. Principal components are the eigenvectors of the covariance matrix of the image samples. In this way, the space of 16 384 pixels for a 128 × 128 image is compressed to 15 or so principal components. A dense set of images, collected according to a systematic exploration of camera viewpoint and illumination direction, forms a manifold in this space for each object. A new image is then classified by its distance to the nearest point of a compressed manifold. This approach is similar to the classical nearest-neighbour classification algorithm widely used in pattern recognition (Duda & Hart 1973). Similar techniques have also been applied in the recognition of objects using coherent optical correlation and holographic pattern library storage (Casasent & Psaltis 1977).

If an unoccluded image view of an object is obtained, the process of finding the nearest manifold is very efficient and recognition proceeds without image segmentation. However, the appearance approach is sensitive to occlusion and it becomes impractical to collect the combinatorially large space of images representing various states of occlusion of one object by others. The appearance manifolds ultimately must be based on local object features, which leads back to the segmentation and structural representations discussed earlier.

This integration of model-based and appearance-based recognition approaches is now being vigorously pursued (Schmid *et al*. 1996; Pope & Lowe 1996). The great advantage of the appearance method is that it is not necessary to define a representation or model for a particular class of objects, since the class is implicitly defined by the selection of the test objects. On the other hand, a model or recognition classification theory is required in order to achieve generalization to the recognition of similar objects. It is not yet clear how to make use of empirically derived appearance models to achieve generalization.

(*f*) *Class-based recognition*

It is now widely accepted that interpretation of a complex scene cannot proceed in a purely bottom-up manner. That is, successful feature grouping is guided by general constraints associated with object classes. Thus the recognition process becomes an interleaved top-down bottom-up process. An example of this approach is provided by the MORSE recognition system developed at a number of research institutions (Zisserman *et al*. 1995). The recognition system is aimed at a set of object classes that induce strong grouping constraints in an image. Examples of these classes are surfaces of revolution, polyhedra, canal surfaces, i.e. surfaces swept by a sphere with varying radius along a space curve, and extruded surfaces where a planar cross-section is extruded to form a three-dimensional surface. A scene containing a number of objects representing these classes was shown in figure 8 for the discussion of recognition of polyhedra. The processing of a surface of revolution (SOR) is shown in figure 12. The SOR class is defined by an axis of rotational symmetry as displayed in the figure. Feature grouping is based on the image constraints imposed by the symmetry. This example represents the more general idea that a class can be identified from the consistency of image constraints defined by the class, i.e. it is not necessary to consider a specific SOR in carrying out the recognition.

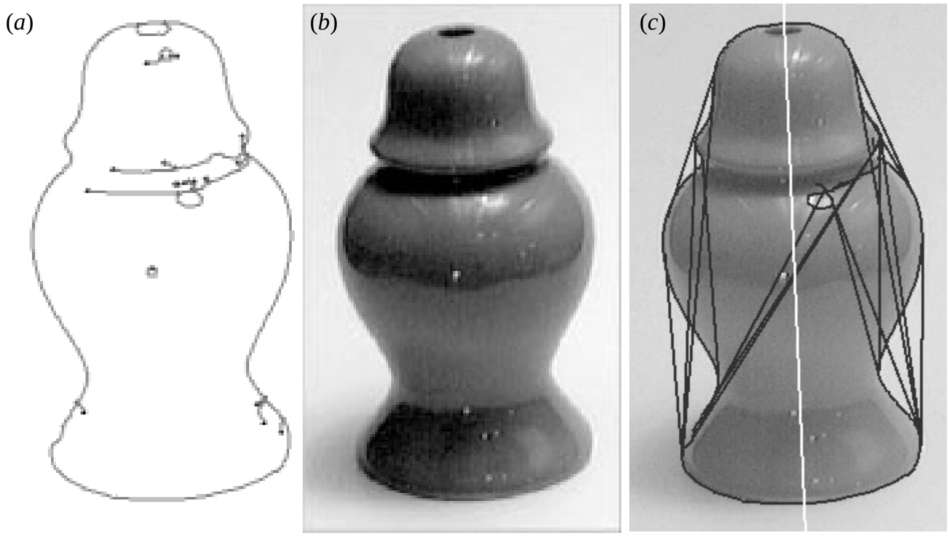

Figure 12. (*a*) A lamp base as an example of the class of objects represented by a surface of revolution. (*b*) Edgel chains and junctions derived from the intensity image in (*a*). (*c*) The computed rotational symmetry axis is shown as a white line. The symmetry axis was recovered by grouping bi-tangents, which are shown in black.

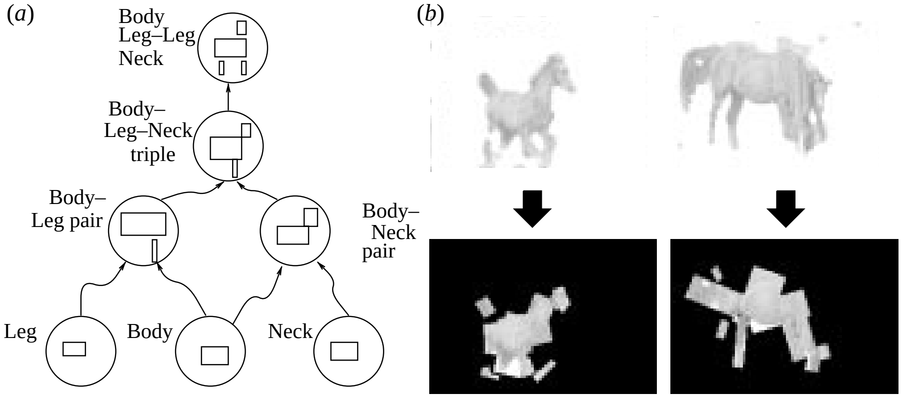

Figure 13. (*a*) A feature grouping hierarchy for the class of horse-like shapes. (*b*) Two examples of horse images with the extracted groupings. The assertion is that horse recognition is entailed by successful feature group hypothesis. (From Forsyth with permission (personal communication).)

This class-based approach to recognition has also been used by Zarroug & Nevatia (1996) to recover descriptions of more complex objects composed of generalized cylinders. Their emphasis is on the decomposition of an object into parts. It is clear that the recognition of a large database of objects cannot be efficiently achieved without the ability to *divide and conquer* by representing an object in terms of components. At the current state of development, however, there is no universally accepted formal definition of what constitutes a *part* and no general approach for decomposing an object into parts. A key unsolved requirement is that the part decomposition can be robustly acquired from image constraints alone.

The idea of class-based recognition can be taken to a very general level. A very

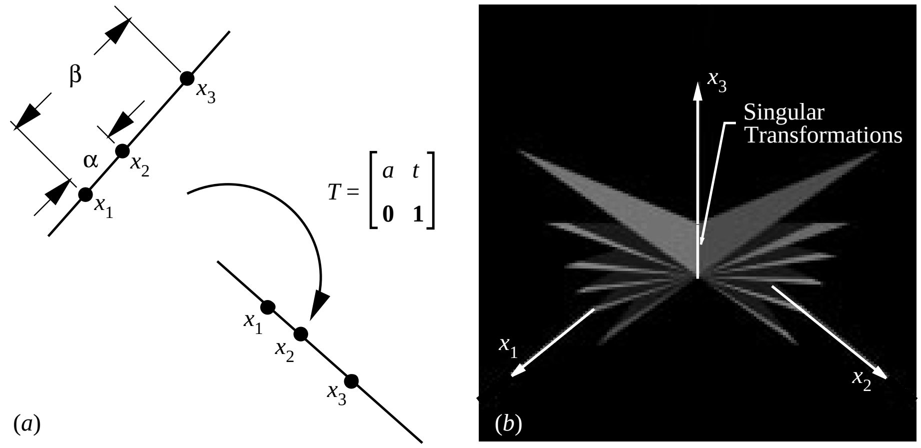

Figure 14. (*a*) The affine 1D mapping of three collinear points defines a 3D measurement space. (*b*) The orbits of three 1D points under an affine projection form a pencil of planes parametrized by the affine-invariant length ratio, β/α.

flexible form of class-based recognition processing is illustrated in figure 13. Here a set of constraints between the legs and body of a horse are used to extract horses from natural scenes. In this case the grouping of 2D symmetrical ribbons is achieved by applying statistically derived constraints which hold between the parts in a canonical frame. The relative geometric relationships in this canonical frame are approximately invariant to viewpoint.

4. Geometric appearance

At this time, the field of object recognition is experiencing a peak of rapid development, particularly in the exploitation of ideas from intensity appearance-based methods. This enthusiasm has left geometric approaches somewhat in the background, since recent demonstrations of intensity-based recognition have been so effective. In keeping with the theme of this meeting, the underlying role of geometry in the analysis of visual appearance is described, which complements the current successes of appearance-based methods and reasserts the importance of geometry in our understanding of object class.

(*a*) *Group orbits*

An important concept is the orbit of transformation actions. An *orbit* is a surface defined in a space of geometric measurements which results from variation of transform parameters. More formally,

> the orbit of a feature vector, $\boldsymbol{X}$, with respect to a group, $\mathcal{G}$, of transformations is the set, $\{x \mid \text{ where } \boldsymbol{x} = \mathcal{T}X \text{ for all } \mathcal{T} \text{ in } \mathcal{G}\}$.

Each orbit can be identified by a unique set of values which are the invariants to the group actions.

Consider the affine mapping from one line onto another. The direct coordinates of the three points define a three-dimensional observation space and the manifold is

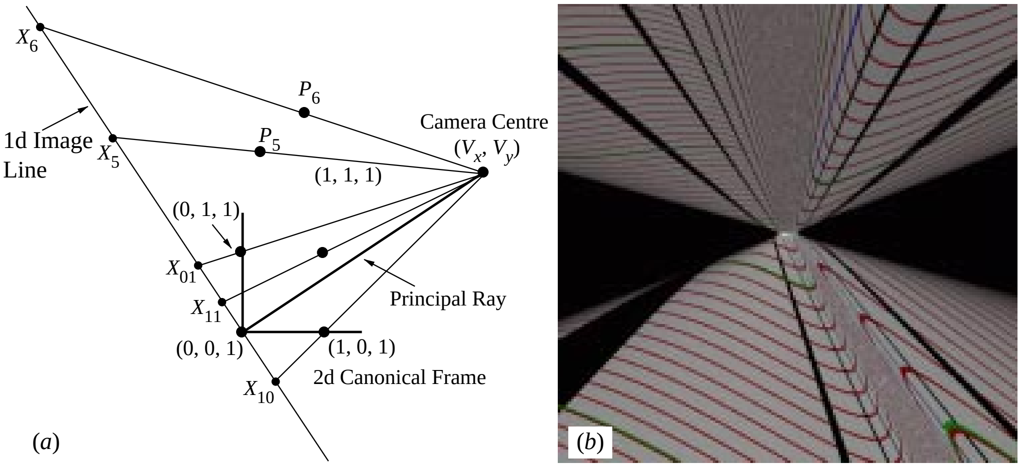

Figure 15. (*a*) A perspective camera constructed to converge to the limit as an affine camera. The point, $(0\ 0\ 1)^{\mathrm{T}}$ is fixed on the image line and the image line is always perpendicular to the principal ray. As the viewpoint, $(V_x\ V_y)$, recedes to infinity, the camera projection becomes affine. (*b*) The manifold of projected point coordinates. The parametric contours of constant viewpoint location, (V_x, V_y) are projected onto the manifold.

a planar surface. The orbits of this configuration are shown in figure 14. The orbit planes are generated by the parameter, β/α, which is the affine-invariant length ratio, $(x_3 - x_1)/(x_2 - x_1)$. The planes intersect on the line corresponding to $a = 0$, which is a set of singular 2×2 matrices and not included in the affine group.

(*b*) *The appearance of point-sets*

The set of perspective projections does not form a group, since any such projection is not invertable. However, it is still possible to consider the manifold of projected geometric features, in analogy to the group manifolds just discussed. In order to construct a manifold which can be visualized, a projection model is defined for the case of 2D → 1D perspective projection, as shown in figure 15.

Let V_x and V_y be the coordinates, i.e. centre of projection, of the camera. Let

$$s = \frac{1}{\sqrt{V_x^2 + V_y^2}}, \qquad t = \frac{V_x}{V_y}.$$

Then the 2D → 1D perspective projection, represented as a 3×2 matrix, parameterized by s and t:

$$C = \begin{bmatrix} -1/\sqrt{1+t^2} & t/\sqrt{1+t^2} & 0 \\ -ts/\sqrt{1+t^2} & -s/\sqrt{1+t^2} & 1 \end{bmatrix}.$$

In this representation of the camera, $\tan^{-1}(1/t)$ is the slope of the vector from the origin to the camera centre and $1/s$ is the distance of the camera centre to the origin. As s approaches 0, the camera centre goes to infinity, and the perspective projection approaches an affine projection.

The resulting appearance manifold for six points is shown in figure 15*b*. The manifold is a quadric in three dimensions when expressed in terms of the projective coordinates of a 1D basis defined on three of the projected points. Four of the six points in the plane form a 2D projective basis. It is thus possible to determine one of the coordinates of the camera centre, since the 2D canonical basis does not depend on the specific configuration of the original six points, assuming general positions.

In the case of affine projection, $s = 0$, the 2D canonical frame is composed of three points and the 1D affine image basis is defined by two points (Mundy & Zisserman 1992). The single affine camera parameter, t, can be eliminated in terms of the affine-invariant image coordinates of a 2D canonical frame, independent of the configuration of the six object points.

These results hold more generally for 3D → 2D perspective and affine projection. It can be shown that the perspective appearance manifold is still a quadric but embedded in a higher-dimensional space. For example, for six points the perspective appearance manifold is a quadric in four dimensions (Jacobs 1996). The 3D affine basis is constructed from four points and the affine image basis requires three points. The two invariant image coordinates of the fourth 3D basis point can be used to eliminate the two unknown camera viewpoint parameters, as in the 1D example.

(*c*) *Implications for recognition*

The first important observation is that, for the affine case, it is possible to index point-sets from a single view without knowledge of the 3D structure and without initial information about the camera, as observed by Jacobs (1992). This affine appearance will depend on viewpoint, but since viewpoint is known, effective indexing is possible. This result has been derived for point features. An obvious investigation to consider is the derivation of affine geometric appearance indices for other types of features. Jacobs (1993) has already considered oriented points. However, the approach suggested here provides a framework for considering features other than points, i.e. curves and surfaces. This general line of investigation will yield a sound geometric basis for indexing the manifolds of intensity-based features.

A second key observation is that geometric appearance is the substrate for intensity-based appearance. The underlying geometry of a surface defines an irreducible minimum complexity of appearance which is further complicated by the variations due to illumination and shadows. As shown earlier, it is possible to generate complete perspective geometric point-set appearance manifolds from only one parameter. Another way of considering this result is that the entire appearance manifold can be generated from just two views, which is just a reinterpretation of the well-known projective reconstruction theorem (Hartley 1994). Efforts to understand the structure and intrinsic dimensionality of appearance manifolds can only benefit from a deeper understanding of these geometric bounds on complexity.

References

Binford, T. O. 1971 Visual perception by computer. *Proc. IEEE Conf. on Systems and Control.*

Binford, T. O. 1981 Inferring surfaces from images. *Artif. Intell. J.* **17**, 205–244. Special Edition on Computer Vision.

Casasent, D. & Psaltis, D. 1977 New optical transforms for pattern recognition. *Proc. IEEE* **65**, 77–84.

Chakravarty, I. 1981 A single-pass, chain generating algorithm for region boundaries. *Computer Graphics Image Processing* **15**, 182–193.

Chow, C. K. 1957 An optimum character recognition system using decision functions. *IRE Trans. Elec. Comp.* **EC-6**, 247–254.

Clowes M. B. 1971 On seeing things. *Artif. Intell. J.* **2**, 79–116.

Duda, R. O. & Hart, P. E. 1973 *Pattern classification and scene analysis.* Wiley.

Fu, K. S. 1974 *Syntactic methods in pattern recognition.* New York and London: Academic Press.

Grimson, W. E. L. & Lozano-Pérez, T. 1984 Model-based recognition and localization from sparse range or tactile data. *Int. J. Robotics Res.* **3**, 3–35.

Hartley, R. 1994 Projective reconstruction and invariants from multiple images. *IEEE Trans. Pattern Analysis Machine Intell.* **16**, 1036–1040.

Huffman, D. A. 1971 Impossible objects as nonsense sentences. In *Machine intelligence* (ed. B. Meltzer & D. Michie), vol. 6, pp. 295–324. Edinburgh University Press.

Huttenlocher, D. & Ullman, S. 1987 Object recognition using alignment. In *Proc. First Int. Conf. on Computer Vision, London*, pp. 102–111.

Jacobs, D. 1992 Space efficient 3D model indexing. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 439–444.

Jacobs, D. 1993 Generalizing invariants for 3D to 2D matching. In *Applications of invariance in computer vision* (ed. J. L. Mundy, A. Zissermann & D. Forsyth). Lecture Notes in Computer Science, vol. 825, pp. 415–435. Springer.

Jacobs, D. 1996 The space requirements of indexing under perspective projections. *IEEE Trans. Pattern Analysis Machine Intell.* **18**, 330–333.

Lowe, D. 1985 *Perceptual organization and visual recognition.* Kluwer Academic.

Malik, J. 1987 Interpreting line drawings of curved objects. *Int. J. Computer Vision* **1**, 73–103.

Marr, D. & Nishihara, H. K. 1978 Visual information processing: artificial intelligence and the sensorium of light. In *Readings in computer vision* (ed. M. Fischler & O. Firschein), pp. 616–637. Morgan Kauffman.

Mundy, J. L. & Zisserman, A. 1992 *Geometric invariance in computer vision.* MIT Press.

Murase, H. & Nayar, S. K. 1995 Visual learning and recognition of 3D objects from appearance. *Int. J. Computer Vision* **14**, 5–24.

Pavlidis, T. 1977 *Structural pattern recognition.* Berlin, Heidelberg, New York: Springer.

Ponce, J. & Kriegman, D. 1992 Curved object recognition from image contours In *Geometric invariance in computer vision* (ed. J. L. Mundy & A. Zisserman), pp. 408–439. MIT Press.

Pope, A. & Lowe, D. 1996 Learning appearance models for object recognition. In *Object representation in computer vision II* (ed. J. Ponce, M. Hebert & A. Zisserman). Lecture Notes in Computer Science, vol. 1144, pp. 201–219. Springer.

Raphael, B. 1976 *The thinking computer: mind inside matter.* New York: W. H. Freeman & Company.

Roberts, L. G. 1965 Machine perception of three-dimensional solids. In *Optical and electrooptical information processing* (ed. J. Tippett, D. Berkowitz, L. Clapp, C. Koester & A. Vanderburgh), pp. 159–197. MIT Press.

Rothwell, C., Forsyth, D., Zisserman, A. & Mundy, J. L. 1993 Extracting projective structure from single perspective views of 3D point sets. In *Proc. 4th Int. Conf. on Computer Vision*, pp. 573–582. IEEE Computer Society Press.

Schmid, C., Bobet, P. & Mohr, R. 1996 An image-oriented CAD approach. In *Object representation in computer vision II* (ed. J. Ponce, M. Hebert & A. Zisserman). Lecture Notes in Computer Science, vol. 1144, pp. 221–246. Springer.

Shapira, R. & Freeman, H. 1978 Computer description of bodies bounded by quadric surfaces from sets of imperfect projections. *IEEE Trans. Computers*, pp. 841–854.

Stark, L. & Bowyer, K. 1991 Generalized object recognition through reasoning about association of function to structure. *IEEE Trans. Pattern Analysis Machine Intell.* **13**, 1097–1104.

Stockman, G. 1987 Object recognition and localization via pose clustering. *Computer Vision Graphics Image Processing* **40**, 361–387.

Sugihara, K. 1986 *Machine interpretation of line drawings.* MIT Press.

Thompson, D. & Mundy, J. L. 1987 Three-dimensional model matching from an unconstrained viewpoint. In *Proc. Int. Conf. on Robotics and Automation, Raleigh, NC*, pp. 208–220.

Underwood, S. A. & Coates, C. L. 1975 Visual learning from multiple views. *IEEE Trans. Computers* **C-24**, 651–661.

Zarroug, M. & Nevatia, R. 1996 From an intensity image to 3D segmented descriptions. In *Object representation in computer vision* (ed. J. Ponce, M. Hebert & A. Zisserman). Lecture Notes in Computer Science, vol. 1144, pp. 11–24. Springer.

Zisserman, A., Forsyth, D., Mundy, J. L., Rothwell, C., Liu, J. & Pillow, N. 1995 3D object recognition using invariance. *Artif. Intell. J.* **78**, 239–288.

Discussion

O. Faugeras (*INRIA, France*). Dr Mundy showed us this interesting example of the torus and the equation of the outline which was very complicated and then he said that if you go to higher degrees of surfaces, it's even worse, and we should drop this approach and look more into combinations of geometry and intensity. If you look at bidirectional reflectional distribution functions, they can be very, very complex as well, so wouldn't you bump into the same wall on that avenue, as you would on the geometry tack?

J. L. Mundy. This is a good point and actually I was trying to correct this impression in the last part of the talk by saying that I think the way forward is to combine the best theory we are able to do, which provides robust recoverable attributes, and what we have to acquire from the image empirically. I think the combination of these two models is really going to be the only possible way forward. There are certainly many examples over history where achievement has been in advance of the theory. Cathedrals stayed up, even though the knowledge of structural mechanics at that time was relatively meagre. We cannot escape the fact that we are going to have to allow some empiricism. On the other hand, I would hate to see the field dominated entirely by empiricism. There is quite a strong push in that direction today in the appearance-based vision realm where people have almost given up any theory whatsoever. Instead the focus is on learning everything about the object by taking thousands of pictures of it. I think that's wrong too, and that this approach will not really provide us with a great movement forward because we won't understand general object classes and we won't improve our ability to extract and group image features.

T. Kanade (*Robotics Institute, Carnegie Mellon University, Pittsburgh, USA*). Dr Mundy said that in discussing formal versus empirical methods we need language for describing objects. I'm not sure about that. I think we should not confuse recognition of an object whose geometry is either known or at least can be varied by a certain rule or relationship, such as the size versus length, and recognition of object class, such as chair, desk and so on. For the latter type of recognition, I don't think we have much of a clue yet how to do it. Yet, if we suddenly begin to say, humans seem to describe them by language, and therefore we need language as a tool, I think that's wrong. Now I do not have an answer to that either, but the answer seems to lie in the perceptual grouping process. If there is any theory here, somehow we have to develop sound mathematical theory for perceptual grouping that relates observable properties with the description of the object, not a linguistic theory that relates symbolically represented properties with objects. Simply saying that geometry is done and the language to describe functions is the next direction sounds like we are going back to the old days before geometry pattern recognition, when all sorts of soft AI-ish ideas were dominant.

J. L. Mundy. I partly agree with Professor Kanade. The point I was really trying to make is that this area down here (on the figure indicating subconscious perception) is terra incognita. In other words we do it, but we don't know what we do, and we have no model for what we're doing. Maybe it could be exposed by some behaviourist approach, you ask subjects questions, yes or no, that kind of idea, but we have no way of capturing the process in any formal way. If we're going to implement an algorithm we have to have a formal description; this is my belief anyway, which is a very mathematical point of view, if you will. But I believe that unless these concepts, which we can't even articulate, bubble up through human language and finally become formalized as a mathematical structure, only then can we write an algorithm to do grouping or an algorithm to classify. There's no other alternative, so we have to take that route. Indeed, over the last 2000 years this is what has been happening, that vague ideas, maybe at first that we didn't even have language for, became ideas that we could talk about informally. Eventually, these ideas became formalized to the point where we could write a computer algorithm. I don't think we can escape this long painful formalization process. The only way to escape it, which then doesn't leave us with particularly any understanding, is to rely totally on empirical measurements. In other words, if you say, this object is class A, I show it to a computer, I tell the computer it's class A and I try to record everything about it. I don't know what to record, so I must record everything, and then use that kind of behaviourist classification in order to empirically derive the class; but the machine would have no formal model of that class at all other than the set of attributes that were collected empirically. To me this is a last resort and does not represent significant progress in our understanding of recognition.

C. d'Souza (*Nottingham Trent University, UK*). In most of the literature I have read on the viewpoint independence of the curvature, the data used are mostly range data. Is work being done on the intensity data, since this would, in my view, represent more problems in different views of lighting?

J. L. Mundy. Yes, to my knowledge, some of the best work in this line was by Ponce, who used the Gaussian curvature as one attribute to restrict the choice of objects in recognition using the alignment of these algebraic surfaces. The sign of the curvature was an additional feature to the choice of possible object. With relatively good resolution and relatively simple scenes, approaches such as the fitting of splines to the surface outline in the image does allow decent enough measurement of the sign of curvature to be an important clue. I don't think anyone was proposing, including Ponce, that it is practical to estimate Gaussian curvature to the tenth decimal place, but Ponce was at least confident that it was positive rather than negative which was the main thrust of that work.

M. Sabin (*Numerical Geometry Ltd, Cambridge, UK*). This seems to be a bit of a take stock meeting. I'm from outside the field, and it's very interesting seeing the kind of different ideas which are being put together within the meeting. From what has been said about the last 30 years, it sounds as if it's a good time to take stock—many things have been tried and none of them has proved to be the breakthrough. Now, there are a couple of other related fields which have also felt the need to take stock around about this time. There's the computational geometry field, which is actually concerned with efficient algorithms for geometric computation, and they went as far as, last year, having a task force to say, 'what should we be doing now?'. And the second is the surface design field, which calls itself computer aided surface design, has more or less saturated the easy theory and there's a lot of tidying at the edges going on, and there are a few interesting looking avenues which are addressing nonlinear things rather than the linear formulations we've had in the past. Is there much contact with these other two groups?

J. L. Mundy. I would say, in the case of computational geometry, yes. I can mention one person who has a foot in both camps, Dan Huttenlocher at Cornell University. He has had a good influence by conveying many important ideas from computational geometry into this field. Keep in mind that we do not necessarily talk about the implementation of our algorithms at the level where ideas from computational geometry would be obvious. But, many computer vision algorithms and many of the results that I've shown, embody many key ideas from computational geometry, such as sorting, clever binning and line-sweeping methods.

I tried to give an honest portrayal of the progress, and actually it's been very good. If you look at Robert's thesis from 1963, the scenes were relatively uncluttered, the background was black, the objects were white. Today, there are many recognition systems where you can throw an object down in a cluttered scene and have half of the object occluded and still be able to perform successful recognition for certain classes of objects. So, taken in the large I would say that is a tremendous advance. Have we achieved the ability to recognize a general object in an arbitrary setting under totally variable illumination? No. Will we do it in another 30 years? Probably not. But I expect that it will continue to make steady progress.