PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The forms of knowledge mobilized in some machine vision systems

Michael Brady

Phil. Trans. R. Soc. Lond. B 1997 **352**, 1241-1248
doi: 10.1098/rstb.1997.0106

The forms of knowledge mobilized in some machine vision systems

MICHAEL BRADY†

Robotics Research Group, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK

SUMMARY

This paper describes a number of computer vision systems that we have constructed, and which are firmly based on knowledge of diverse sorts. However, that knowledge is often represented in a way that is only accessible to a limited set of processes, that make limited use of it, and though the knowledge is amenable to change, in practice it can only be changed in rather simple ways. The rest of the paper addresses the questions: (i) what knowledge is mobilized in the furtherance of a perceptual task?; (ii) how is that knowledge represented?; and (iii) how is that knowledge mobilized? First we review some cases of early visual processing where the mobilization of knowledge seems to be a key contributor to success yet where the knowledge is deliberately represented in a quite inflexible way. After considering the knowledge that is involved in overcoming the projective nature of images, we move the discussion to the knowledge that was required in programs to match, register, and recognize shapes in a range of applications. Finally, we discuss the current state of process architectures for knowledge mobilization.

1. INTRODUCTION

Machine vision systems are deployed to perform a wide range of tasks in an equally wide range of applications. The knowledge that must be mobilized depends fundamentally on the requirements of the task, hence on the application. For example, consider an aerial imaging program that is required to register a newly acquired image to those obtained previously, notwithstanding changes in cloud or ground cover, and which may be required to detect significant changes in the environment, and perhaps to interpret what such changes are. To do so requires that the program have, and be able to mobilize, knowledge about the expected appearance of aerial images in the particular part of the world under investigation, about the spatial resolution at which the images are taken, and about the kinds of image noise and geometrical distortions that are expected to arise with the current imaging device. It needs to embody knowledge of the appearance of clouds, sand storms, seasonal changes in ground cover, and other 'non significant' changes; conversely, it needs to have some idea about the kinds of changes that should be considered significant (e.g. the appearance of a new building of a certain size). Some of this knowledge may be represented in a way that facilitates only a limited number of processes. For example, the expected appearance of ground cover may be represented as a set of textural descriptors to enable robust, automatic, region segmentation (Xie & Brady 1996). Conversely, some knowledge may be represented in a way that enables it to adapt quickly, even automatically, to changing goals.

Another machine vision system may be required to recognize an object in an image, a problem that is considerably more difficult when the imaged scene is three-dimensional (3D), and when the object may be partly occluded by others, and the ambient imaging conditions are not completely known in advance. Localizing an object in a 3D scene is easier than inspecting the instance or computing how to grasp it with a robot hand. This is because localization may be based on information integrated over the totality of the visible object, while inspection and grasp-planning additionally require information about its local geometry. Likewise, controlling a robot vehicle to navigate in an environment where clearances between obstacles are relatively large can be accomplished with representations that are crude approximations to the environment, say that it is composed of idealized geometric shapes. However, as the clearances become tighter, the more this will not do and more precise, local representations are needed.

A robot stereovision platform may be required to track an object; often 2D trackers suffice even though the object moves in 3D, a point which I will return to. However, it is occasionally required to build up a representation of the object that is being tracked. A smart security camera may not only be required to track a person, but to build up an adequate representation of the person's face, and perhaps even recognize him. An 'active vision' system is required to control, in real time, the motions of a device on the basis of visual

† Present address: Projet Epidaure, INRIA, Unité de Recherche INRIA Sophia Antipolis, 2004, route des Lucioles, B.P. 93-06902 Sophia Antipolis Cedex, France.

information, and deliberate movements of the device may be made to elicit further relevant information visually.

All the applications referred to above have been worked on in our laboratory. The systems we have constructed are firmly based on knowledge of diverse sorts, but that knowledge is often represented in a way that is only accessible to a limited set of processes that make limited use of it; and though the knowledge is amenable to change, in practice it can only be changed in rather simple ways (e.g. the adaptive control of a robot head).

Vision may also be the principal source of information for a system that is required to reason about a scene. Given a motion sequence of a roundabout or road junction, the goal may be to interpret the motions of vehicles in terms of prototypical behaviours such as lane changing, overtaking, or joining or leaving the roundabout (Howarth & Buxton 1996). From a video sequence of a football match, the task may be to interpret the motions of the players in the arcane language of football coaches. In both these cases, one might be interested in interpreting motions as 'abnormal': as dangerous driving or foul play. In the same way, a security system may be interested in people who are 'behaving suspiciously', while from a computed tomographic sequence of a beating heart one may hypothesize abnormal motion as ischaemia (Bardinet *et al.* 1995).

It seems that the kind of applications referred to in the previous paragraph require knowledge to be represented in a way that facilitates reasoning. While processing of signals has traditionally been the domain of the (image processing) engineer, processing symbols has been the central concern of artificial intelligence (AI). It is in this sense that one refers to the transformation from 'signal to symbol' in the development of 'smart' vision systems. AI focuses on a number of questions about knowledge representation that usefully serve as the basis for discussion in this article.

1. What knowledge is mobilized in the furtherance of a perceptual task? As we noted above, the knowledge mobilized is application specific. How are the needs and constraints of a task specified?

2. How is the knowledge represented? A fundamental result of computer science is that there is an essential linkage between the representation of information and the processes such as matching that can effectively manipulate it. The way in which knowledge is represented in a system determines what David Marr (Marr 1982) called the accessibility, scope, and sensitivity of the representation.

3. How is the knowledge mobilized? Only a subset of the available knowledge may be mobilized in furtherance of any given task. What kind of process architecture enables opportunistic, dynamically changing perceptual processes?

Our current level of understanding of computer vision enables only preliminary answers to these questions. In the next section I review some cases of early visual processing where the mobilization of knowledge seems to be a key contributor to success, but where the knowledge is deliberately represented in a quite inflexible way. Section 3 reviews the knowledge that is involved in overcoming the projective nature of images. Sections 4 and 5 move the discussion to the knowledge that was required in programs to match, register, and recognize shapes in a range of applications. Section 6 simply contributes two observations that we have learned about learning. Finally, §7 discusses the current state of process architectures for knowledge mobilization. Necessarily, the discussion in each section is brief.

2. EARLY VISION

Humans, and nowadays computers, can deal effectively with many different kinds of image, each with very different characteristics. Examples include far infrared, synthetic aperture radar, (X-ray) mammograms, magnetic resonance imaging (MRI) and contrast-enhanced MRI. It has been found in practice that early vision processes, such as edge detection, that work well for certain classes of visual imagery, give very poor results when applied to other classes. It has further been discovered that reliable results can be obtained if one mobilizes knowledge of the physics of image formation. I will recall a number of examples developed in our laboratory, then draw some conclusions relevant to the subject of the meeting.

Far infrared (8–12 μm) imagery has many applications in night vision, not least in developing systems that contribute to safe driving. In comparison with the visual waveband, such images are very noisy, exhibit no shading, while relatively poor lenses and wide angle imagery lead to significant intensity variations (a process known technically as 'vignetting') (Highnam & Brady 1997). We have developed a model of infrared imagery from which we deduce that a retinex-like lightness computation algorithm, which uses relative brightnesses, enables reliable image enhancement, segmentation, and object tracking. X-ray mammograms also exhibit poor signal-to-noise ratio (SNR), while scattered illumination, radiation glare, the inevitable nonlinear variation of X-ray intensity across the film, film speed and exposure time all contribute to the resulting poor image quality. We have developed a model of the image formation process (Highnam *et al.* 1994) that models and corrects for all of these image degradations. Removal of the scatter component of the irradiation enables us to construct a representation as a surface of the non-adipose tissue in the breast. The importance of this representation is that it is based on anatomical information that is intrinsic to the breast, that is, it is invariant to image-specific parameters such as exposure time, or the particular X-ray machine's spectral characteristics. This enables the computation of information that is normalized across a patient group, which in turn enables a neural network to learn which masses are 'abnormal' (see §6). It also enables us to compute such 3D information as the separation of the compression plates, and to match images of the same breast over time. Mammograms are highly textured in appearance, with a high frequency texture composed of milk ducts, stroma, and larger blood vessels. Extracting these 'curvilinear structures' from a mammogram not only facilitates matching over time (a process that is easily distracted by high frequency information), but enables diagnostic signs such as microcalcifications to be interpreted more reliably as benign or malignant. We have modelled (Cerneaz & Brady 1995) the passage of X-rays through a compressed vessel. This knowledge is then embedded in a program to extract the curvilinear structures from mammograms. Note that the application of techniques that smooth an image (e.g. DOG (difference-of-Gaussians), Gabor, and wavelet filters or anisotropic diffusion) are ineffective at recovering curvilinear structure.

Two further examples (of several) suffice for our purposes, and they both concern MRI. MRI images are three-dimensional datasets, comprising a series of planar slices, much as one might slice a potato: there may be 256 slices, each having the same thickness. Each planar slice is further 'diced' into an array of samples, again typically 256 by 256 per slice. The individual samples are called 'voxels'. (This neologism has a simple etymology. When digital images were first produced the individual picture elements were called 'pixels'. For volumetric data the volume 'picture' elements were thus called voxels.) The first example of interest here is that brain MRI is subject to a low-frequency 'bias field' that greatly affects such classification tasks as the estimation of white matter. It is possible to exploit knowledge of the expected appearance of brain tissue in MRI volumes taken with a particular sequence and to mobilize this knowledge in an expectation-maximization algorithm that simultaneously estimates the bias field and reclassifies the brain voxels (Guillemaud & Brady 1997). The results are robust over time and patient head position. Finally, it is possible to model the uptake of a contrast agent by breast tissue to aid the radiologist in diagnosing breast cancer in women for whom mammography is ineffective. It is also possible to mobilize knowledge of the kinds of breast motions expected over an examination (typically 10 min), and then to solve simultaneously for both the slight motions of the breast and the uptake of the contrast agent. This leads to greatly improved assessment and localization of cancers (Hayton *et al*. 1997).

We draw the following conclusions from these, and a number of similar, examples. First, early vision algorithms that work well in some cases are completely ineffective in others. Second, mobilizing knowledge of the physics of image formation can give greatly enhanced results. Nevertheless, the results returned by any early vision algorithm are intrinsically uncertain, hence subsequent processing must be able to deal with uncertainties. Third, it is necessary to carry out early visual processing simultaneously with other more interpretive processes. Fourth, a great deal of work in 'knowledge-based' machine vision was predicated on the belief that adequate representations of intensity changes, textures, etc., were possible only with the deployment of 'higher level' knowledge. There continues to be an underestimation of what can be achieved with the mobilization of appropriate 'low level' knowledge.

3. IMAGES AS PROJECTIONS

Images are projections of the 3D world. A central goal of computer vision has been to understand how to reconstruct the 3D world from one or more images taken from one or more vantage points as in stereo structure from motion (SFM), and shape from shading or texture. In this section I make three observations concerning (i) the need in computer vision to have a number of geometrical models of projection; (ii) the information that is necessary to operate successfully in a 3D world; and (iii) tracking with an active stereo system.

First, a number of different mathematical definitions of projection from the three-dimensional world to the two-dimensional image (Mundy & Zisserman 1992) have proved useful in computer vision. These include perspective projection, affine projection (which for present purposes can be regarded as a simplification of perspective projection in which parallel lines project to parallel lines), and image 'aspectation' (which is simply a translation and rotation in the image plane together with a magnification of an object). Each of these mathematical definitions of projection constitutes a more or less crude approximation to the way a camera (or an eye) works. For example, the affine approximation works well when the range of depths surrounding the point of interest is less than one-tenth of the distance between that point and the camera (Shapiro 1995; Shapiro *et al*. 1995; Koenderinck & van Doorn 1991). Theoretically, the most general (perspective) definition of projection suffices. However, under certain frequently-occurring imaging conditions and camera motions, the solution to SFM may become underconstrained or ill-conditioned (Wiles & Brady 1996*b*). There are four main causes for this: (i) degeneracy of structure, e.g. viewing planar surfaces; (ii) certain degenerate motions; (iii) degenerate spatial positioning of features; and (iv) poor image preconditioning. In such cases, using a less general definition of projection, such as affine, is not only computationally more efficient, it is necessary. A system has been constructed that automatically determines the most appropriate definition of projection for tracking road vehicles (Wiles & Brady 1996*a*).

Second, it has generally been supposed that the goal of processes such as stereo and SFM is to generate an accurate 3D representation of the environment, as if vision were the same as photogrammetry. Usually, to compute depth it is necessary first to compute an image-based quantity such as stereo disparity, from which depth can be computed. This latter relies on the results of a complex process that estimates a number of internal parameters of the cameras (e.g. focal length of the lens, and the direction of the normal to the sensor surface) and the geometric transformation between the cameras. This process is called 'camera calibration'. Unfortunately, camera calibration is a complex nonlinear process that is often numerically ill-conditioned, particularly if lens distortions are taken into account. However, accurate 3D reconstruction of the environment is often not necessary. For example, the ground plane and ground plane obstacles can be estimated on the basis of disparity without calibration (see

Wiles & Brady 1996*a*, and the references therein). Similarly, Cipolla & Blake (1992*a*,*b*) have shown how deliberate motions enable the time-to-contact of an object to be computed without calculating depth, while relative motions of surface features enable an observer to distinguish bounding contours of an object (where the surface normal turns smoothly away from the viewer) from edges fixed in space (e.g. surface creases or reflectance changes). In a similar fashion, (Shapiro 1995) has shown how the instantaneous axis of 3D rotation can be computed with an uncalibrated camera. Even more remarkably, Zisserman and his colleagues (Beardsley *et al*. 1996) have shown that in certain practically important cases an accurate 3D model of an object can be acquired from an extended image sequence without knowledge of the camera's internal parameters and without knowledge of the camera's motion.

Finally, a number of 'active' stereo camera platforms have recently been built and used to track objects moving in 3D. The control of such platforms is complex, especially as it needs to be effected in real time. A key problem in tracking an object is to determine a subset of the image feature points that belong to the object, and to determine how the set moves over time. This is deeply related to the different mathematical definitions of projection referred to above.

Since objects that are tracked are often quite small relative to their distance from the platform, the affine approximation to projection is often used. In this vein, Fairley *et al*. (1995) have shown how four points in an image establish an affine frame of reference; then affine transfer (Reid & Murray 1996) can be extended to stereo processing to control the vergence between the cameras (analogous to the vergence of the human eyes), track the object, and build an affine representation of the structure of the moving object.

4. MATCHING AND RECOGNITION

A fundamental task of computer vision is: given that a representation of an object is available inside a computer, and given an image of a scene that contains the object, find where it is in the image. A closely related problem is registration: find the 'closest' match between two instances of an object class in two separate images. An example might be to match up the heads of two people given MRI volumes, notwithstanding their different head shapes and the different relative sizes of their jaws and noses. There are several variants to the basic theme: (i) whether the object being represented is 2D or 3D and whether the image/scene is 2D or 3D – each of the cases (model–scene) 2D2D, 2D3D, 3D2D and 3D3D are important; (ii) whether the representation is parametric, so that by changing the values of the parameters one can specify a range of objects of the same class. In such a case, one wants to determine the parameter values as a side-effect of the process of finding the object in the image. In any case, it is necessary to localize the object in the image even when it is partly occluded by other objects. This has the important consequence that one cannot guarantee that local 'salient' features of the object will be visible in the image/scene. Localization is often treated as a hypothesis generation step, to be followed by a verification phase that attempts to test that each part of the predicted model is present.

An early attack on the problems 2D2D and 3D3D was made by Grimson (Grimson 1990). The representations of objects are piecewise linear (or planar in 3D), the localization algorithm amounts to searching a tree of interpretations subject to a set of constraints that are used to prune the tree as soon as possible. Despite the relatively inflexible representation of knowledge about the object, the program performs remarkably well. However, it cannot deal with representations that are parameterized. Also, in the case of a set of representations of different objects in which one (say A) refines another (say B), a quite common occurrence in practice, the interpretation tree process will always report instances of B when there are instances of A, because it cannot take account of negative evidence (parts of A are inconsistent with B). Reid & Brady (1996) have developed an alternative approach to 3D3D recognition that overcomes these two difficulties. Interestingly, their approach replaces the interpretation tree/fixed object representation used by Grimson *et al*. by a constraint propagation algorithm that was originally developed for computational reasoning/logic. A further refinement of the approach of Grimson for the case 3D(2D + *t*) was developed by the late Professor G. Sullivan and his colleagues (Sullivan 1993). The emphasis of that project was to track a model instance over several frames using a Kalman filter. Impressive though these programs are, they have the fundamental limitation that they are essentially based on polyhedral representations of objects.

Non-polyhedral representations of shape have naturally been explored in medical image processing. One idea has been to represent the bounding contours of objects as being 'elastic': they are able to stretch and bend, but always try to attain the shape that minimizes their 'internal energy', subject to (nearly) satisfying a set of positional constraints. Such curves are called 'snakes', and there have been numerous variations to the general theme used to model deformable objects (see Kass *et al*. 1987; Guillemaud *et al*. 1997, and the references therein). Similarly, Feldmar (Feldmar & Ayache 1996) has developed a scheme based on the iterated closest point algorithm for finding successively more accurate matches of two shapes such as faces, first by assuming a rigid transform, then an affine transform, and finally local affine transforms. Mostly, the schemes developed to date that follow both of these approaches assume prior segmentation of the data. This seems not to be the case for the principal component analysis approach of Taylor and his colleagues (Taylor, this volume). It is also not the case for the algorithm developed by Kok-Wiles (1997) for matching a pair of mammograms (over time or leftright). The latter approach is based on removing the curvilinear structures, as outlined in § 2, and then developing a representation of breast anatomy that captures knowledge of the way in which structures are embedded in others.

5. LOCAL SYMMETRY REPRESENTATIONS OF SHAPE

Representations of natural shapes are sometimes based on global information such as Fourier (or wavelet) descriptors. Another global approach is to imagine the shape to be made of a uniformly dense material and to compute its mechanical moments (area, principal axes, etc.). However, all such representations cope poorly with occlusion. At the other extreme, local descriptors are typically based on 'salient' (often curvature) properties of the bounding contour (2D shapes) or surface (3D). The latter representations have a number of interesting properties (Hoffman & Richards 1986; Thirion 1993), but fail to encode information about the positions of salient features relative to one another. For example, such a local representation would not make explicit such important attributes of a shape as its symmetries.

A symmetry is an invariant to a particular transformation of the shape, e.g. a rotation, reflection, translation, or scaling. The familiar mathematical notion of symmetry (an element of a transformation group) is too strict—it implies a global symmetry of the shape. Blum's (1973) insight was to develop a definition of a local symmetry, and to realize that the locus of such local symmetries often coincided with what one might regard as the 'skeleton' of the shape. Mathematically, Blum defined a local symmetry (in the case of a 2D shape) as the centre of any of the largest discs that could be fitted inside the shape, normally touching (tangentially) the bounding contour of the shape at two points. He called the locus of all such centres (and the radius of the bitangent circle at each point) the symmetric axis transform (SAT) and noted that it is an information preserving representation—given the SAT, one can uniquely reconstruct the shape. Based as it is on centres of bitangency, the SAT naturally favours (local) reflectional symmetries. In the case of elongated shapes, the SAT coincides with the skeleton, which Blum related to growth processes, as discussed by D'Arcy Thompson (Thompson 1952). Blum also discussed the 'brushfire' algorithm to generate the SAT, which computes the locus of centres as the steady state solution to the diffusion equation given the bounding contour as the initial condition. More recently, McAuliffe *et al.* (1996) extended the SAT to provide some invariance to scaling transformations. To do this, they analysed the way that shapes change as they are blurred increasingly, and showed how one can select automatically the scales that are, in a certain mathematical sense, 'natural' for analysing the shape. This has already been of considerable value in segmenting and analysing medical images. Finally, Blum & Nagel (Blum & Nagel 1978) showed how one might generate a symbolic description of a shape; but their technique seems not to have been used.

Some years ago, we noted the advantage of making explicit all the local symmetries of a shape, not just the subset defined by maximal bitangencies. The most compelling aspect of the symmetry set (SS) scheme is the embedding property: suppose that a shape A is contained inside another B, then the SS (but not the SAT) is composed of (i) the SS of A; (ii) the SS of B; and (iii) additional local symmetries that detail how A is embedded in B. This makes the SS considerably more robust than the SAT, and hence more useful in practice, but rather more difficult to compute. To this end, we have developed a parallel wave-diffusion algorithm to compute the SS, and have implemented it on a parallel, distributed processor (Mukherjee & Brady 1996). The SS computed in this way favours rotational and reflectional symmetries equally. We have also developed a system to compute hierarchical symbolic representations of shapes of objects, and shown how such representations might be learnt and generalized.

The 2D SS is, however, not invariant to skews as might occur in projection of the three-dimensional world onto a two-dimensional image. Succinctly, the SS of a skewed shape is not the same as the skewed version of the SS of the original shape. This appeared for some time to lessen its usefulness for recognizing occluded shapes in non-frontoparallel planes. Recall, however, that there are several useful camera models, corresponding to different definitions of projection. Among these, the affine camera is particularly useful, and it has the important property of preserving parallelism, which is a strong local cue for parallelism. This means that an image parallelism may be used to infer scene parallelism. Mukherjee *et al.* (1995) have developed an algorithm that can detect and verify local symmetries for affine skew transformations, then unskew the shape for recognition.

The 3D SAT is relatively undeveloped, not least because there are many ways to generalize the maximal bitangent circle definition. On the other hand, it is straightforward to generalize the SS to give local symmetry faces. Intersections of such local symmetry surfaces often correspond to skeletons or 'axes' of 3D shapes, as explored in computer vision in 'generalized cylinders'. Szekely's recent habilitationsschrifft (written in English) (Szekely 1996) is a thorough summary of the state of the art of local symmetry representations.

6. LEARNING

Relative to studies in human perception, learning has had remarkably little impact on the development of computer vision. Rarely has the constraint that a representation (of shape, image formation, or whatever) be learnable been a consideration at the outset of a computer vision project. Conversely, work in machine learning has either worked exclusively with symbolic information (rarely with information computed from signals) or has utilized simplistic representations of visual information. Typical of the latter are vectors of 'features' in the context of a pattern recognition or neural network system (Bishop 1996). Recent developments in machine learning, not least advances in neural networks, encourage the thought that it is timely to attempt a synthesis of the more promising ideas in computer vision and machine learning. Brady & Connell (1987) made an early attempt at such a synthesis, in their case, on the one hand of local symmetry representations of shape, and on the other of an algorithm developed for learning and reasoning by analogy. More recently, Blake and his colleagues (Blake *et al.* 1995; Reynard *et al.* 1996) have shown how one can represent, track, and learn prototypical motions such as lip movements.

A second point is illustrated by our work in mammography. Mammography is typical of many problems in medicine: the class of real interest (i.e. abnormalities, in our case masses) is under-represented in the database of available examples, and hence its prior probability will be very low—for every one thousand women who are screened, only five, on average, go on to develop cancer. Lack of standardization in equipment and in data acquisition (breast compression, film exposure time, etc.) has so far made it impossible to assemble multi-centre databases. As a result of this, there are very few examples of abnormalities in any of the existing databases. If a neural network classifier is trained using the standard approach of minimizing the mean-squared error at the output, the under-represented class will be ignored. We have been exploring (Tarassenko *et al.* 1995) an alternative approach in which we attempt to learn a description of normality using the large number of available mammograms that do not show any evidence of mass-like structures. The idea is then to test for novelty against this description in order to try and identify candidate masses in previously unseen images.

7. ARCHITECTURES FOR KNOWLEDGE MOBILIZATION

The previous sections have outlined the diverse kinds of knowledge required to complete a sensor-based task. Equally, there is need for an architecture that can effectively mobilize the knowledge. Such an architecture needs to: (i) support opportunistic selection and use of the knowledge that a system has, depending on the task and the situation; (ii) be able to manipulate uncertain information at multiple levels of representation and constraint; and (iii) be robust and efficient. Many architectures have been experimented step by step with the development of computer science and technology. Initially, technology forced software architectures to consist of a relatively small number of processes (Allen Newell referred to such systems as 'coarse grain') designed to be executed on a single serial processor. Recent work has explored 'finer' grain software architectures, involving tens or hundreds of individual collaborating processes, as well as genuinely parallel and/or distributed processors. It should be recognized that there remains a limited stock of ideas about the most appropriate software and hardware architectures to support perceptual systems, and that they are all in a state of rapid evolution. The following remarks are primarily historical.

Very early in the development of computer vision it was realized that a single sequential thread of processing was inadequate. This was clear even in early experiments aimed at developing printed character recognition programs: the recognition step depended on perfect segmentation and enhancement of the printed characters; yet these two processes could benefit from each other's intermediate results, and those of the recognition step. The fundamental reason why a single sequential thread does not work well in practice is that such a program resembles a house of cards: failure of the early processes inevitably foils later processes. Since the early processes proved to be (and still are) difficult (if not impossible) to make absolutely reliable, early programs only worked in carefully contrived situations. It was, and remains, tempting to imagine that simply by making arrows that indicate the flow of control between two processes point in both directions, understanding is somehow advanced.

Nevertheless, this 'insight' has been the basis for a great deal of work in computer vision, leading to systems that are admirably baroque but mostly do not work, or, if they do, are quite sensitive to changes in their environment. One popular variant to this theme was the so-called blackboard architecture (Erman *et al.* 1980), in which a number of independent processes 'interact' by reading the intermediate results of other processes (about which they have limited understanding) from a single central datastructure called a 'blackboard', and writing their own results on to the blackboard. Another early approach was to exploit developing ideas about interacting process models from computer science, such as coroutines (Brady & Wielinga 1979).

An idea, borrowed from control theory, is to make explicit the constraints of a task and then attempt to maximize agreement with the particular image data while satisfying (mostly) the constraints. Such variational models led ultimately to the active contour model ('snakes') described earlier in this article, and to the system of Blake *et al.* referred to in § 6. This is a powerful framework, but one that is limited to representing knowledge as constraints. Further, optimization of nonlinear sets of constraints encourages programs to become trapped in local minima as they search for the global minimum, and this in turn encourages algorithms that attempt to escape from local minima, such as simulated annealing (Murray *et al.* 1986), genetic algorithms, and graduated non-convexity (Blake & Zisserman 1987). Most schemes for avoiding local minima involve an occasional random movement that (hopefully) leads away from the local minimum to a more promising part of the search space. For example, simulated annealing is an optimization scheme in which 'normally' the program takes what appears to be the most promising step in searching for the global optimum value, but from time to time, with low probability, takes what appears to be a less favoured step. The idea can best be understood by analogy with hill-walking: in attempting to descend to the deepest valley (the global optimum), it is occasionally better to climb higher to cross a ridge. The name 'simulated annealing' is by analogy with treatment of metals in which the heat is gradually reduced: in this case the height of the upwards step one can occasionally take is gradually reduced. The major problem with simulated annealing is that though in certain circumstances it is guaranteed to find the global minimum, in practice to do so can take an unacceptably long time. Genetic algorithms are another approach to optimization that involve a random jump to another part of the search space. They explore many different parts of the search space, often to quite a shallow depth, whereas simulated annealing explores a relatively localized part of the search space deeply. Graduated non-convexity also explores a relatively localized part of the search space, and does so by approximating the search space at different scales by blurring it at each scale to convert it to the form of a convex well whose minimum is easy to find; then the search space around that minimum is blurred less and the process recommences.

Recently, there has been considerable interest in perceptual psychology in parallel distributed processes (PDP) (McClelland & Rumelhart 1986) and in computational models of brain operation called 'neural networks' (Bishop 1996). However, to date, neither PDP nor neural networks have had much impact on computer vision. One reason is the need for normalization (as in our work on mammography) which is often hard to attain. A second related problem is the difficulty of learning 'abnormality' discussed in the previous section. Mostly, however, the problem seems to be that current neural techniques work from impoverished representations of knowledge. This may not be insuperable. PDP encourages thinking about fine-grain truly parallel systems, though the vast majority of implemented systems only run in simulation. Genuinely real-time, parallel (often distributed) systems (Fairley *et al*. 1995; Sharkey *et al*. 1993; Rygol *et al*. 1992; Li *et al*. 1995) are hard to construct and control, but as we have found in the case of active vision, once they are available they may change fundamentally the way we think about seeing.

The author thanks his many collaborators whose publications have been referred to in this paper. Nicholas Ayache kindly hosts my sabbatical at INRIA, a sabbatical made possible by grants from the EPSRC and NATO.

REFERENCES

Bardinet, E., Cohen, L. D. & Ayache, N. 1995 Superquadrics and free-form deformation: a global model to fit and track 3D medical data. In *First int. conf. on computer vision, virtual reality and robotics in medicine* (ed. N. Ayache). CVRMed'95, Nice, France, 1995. Lecture Notes in Computer Science. Springer.

Beardsley, P., Torr, P. & Zisserman, A. 1996 3D model acquisition from extended image sequences. In *Proc. ECCV'96*, pp. 683–695. Cambridge, England, 1996. Lecture Notes in Computer Science. Springer.

Bishop, C. M. 1996 *Neural networks for pattern recognition*. Oxford University Press.

Blake, A. & Zisserman, A. 1987 *Visual reconstruction*. The MIT Press Series in Artificial Intelligence. MA: MIT Press.

Blake, A., Isard, M. A. & Reynard, D. 1995 Learning to track the visual motion of contours. *Artif. Intell.* **78**, 101–134.

Blum, H. 1973 Biological shape and visual science. I. *J. Theor. Biol.* **38,** 205–287.

Blum, H. & Nagel, R. N. 1978 Shape description using weighted symmetric features. *Pattern Recognit.* **10**, 167–180.

Brady, M. & Connell, J. H. 1987 Generating and generalising models of visual objects. *Artif. Intell.* **31** (2).

Brady, M. & Wielinga, B. J. 1979 Reading the writing on the wall. In *Computer vision systems* (ed. A. Hanson & E. M. Riseman), pp. 283–301. Academic Press.

Cerneaz, N. J. & Brady, J. M. 1995 Finding curvilinear structures in mammograms. In *First int. conf. on computer vision, virtual reality and robotics in medicine* (ed. N. Ayache), pp. 372–382. CVRMed'95, Nice, France, 1995. Lecture Notes in Computer Science. Springer.

Cipolla, R. & Blake, A. 1992*a* Motion planning using image divergence and deformation. In *Active vision* (ed. A. Blake & A. Yuille), pp. 39–58. MA: MIT Press.

Cipolla, R. & Blake, A. 1992*b* Surface shape and the deformation of apparent contours. *Int. J. Computer Vision* **9** (2), 83–112.

Erman, L. D., Hayes-Roth, F., Lesser, V. R. & Reddy, D. R. 1980 The hearsay-ii speech understanding system: integrating knowledge to resolve uncertainty. *Computing Surveys*, **12** (2), 213–253.

Fairley, S. M., Reid, I. D. & Murray, D. W. 1995 Transfer of fixation for an active stereo platform via affine structure recovery. In *Proc. of the fifth int. conf. on computer vision*. Boston, MA, June 1995, pp. 1100–1105. Los Alamitos, CA: IEEE Computer Society Press.

Feldmar, J. & Ayache, N. 1996 Rigid, affine and locally affine registration of free-form surfaces. *Int. J. Computer Vision* **18**, 99–120.

Grimson, W. E. L. 1990 *Object recognition by computer: the role of geometric constraints*. MA: MIT Press.

Guillemaud, R. & Brady, M. 1997 Enhancement of MR images. *IEEE Trans. Med. Imaging*. (In the press.)

Guillemaud, R., Sakuma, M., Marais, P., Feldmar, J., Crow, T., deLisi, L., Zisserman, A. & Brady, M. 1997 Cerebral symmetry analysis from MRI scans. *Psychiatric Res. Neurosci*. (In the press.)

Hayton, P., Brady, M., Tarassenko, L. & Moore, N. 1997 Analysis of dynamic MR breast images using a model of contrast enhancement. *Med. Image Understand.* **1**, 207–224.

Highnam, R. P. & Brady, M. 1997 Model-based image enhancement of infra-red images. *IEEE Trans. Patt. Anal. Mach. Intell.* (In the press.)

Highnam, R. P., Brady, J. M. & Shepstone, B. J. 1994 Computing the scatter component of mammographic images. *IEEE Trans. Med. Imaging* **13**, 301–313.

Hoffman, D. D. & Richards, W. A. 1986 Parts of recognition. In *From pixels to predicates: recent advances in computational and robotic vision* (ed. A. P. Pentland), pp. 268–293. MA: MIT Press.

Howarth, R. & Buxton, H. 1996 Visual surveillance monitoring and watching. In *Proc. ECCV'96*, pp. 321–335. Cambridge, England, 1996. Lecture Notes in Computer Science. Springer.

Kass, M., Witkin, A. P. & Terzopoulos, D. 1987 Snake: active contour models. In *Proc. of the first int. conf. on computer vision*, vol.1, pp. 259–268. London: IEEE Press.

Koenderinck, J. J. & van Doorn, A. J. 1991 Affine structure from motion. *J. Opt. Soc. Am.* **8** (2), 377–385.

Kok-Wiles, S. 1997 Comparing mammogram pairs in the detection of mammographic lesions. Ph.D. thesis, University of Oxford, UK.

Li, F., Brady, M. & Hu, H. 1995 Visual guidance of an AGV. In *Int. symp. on robotics research 1995* (ed. G. Giralt & G. Hirzinger). Springer.

Marr, D. 1982 *Vision*. San Francisco: W. H. Freeman.

McAuliffe, M. J., Eberly, D., Fritsch, D. S., Chaney, E. L. & Pizer, S. M. 1996 Scale-space boundary evolution initialised by cores. In *Visualisation in biomedical computing* (ed. K. Heinz & R. Kikinis), pp. 173–182. Springer.

McClelland, J. L. & Rumelhart, D. E. 1986 *Parallel distributed processing. 1. Explorations in the microstructure of cognition. 2. Psychological and biological models*. MA: MIT Press.

Mukherjee, D. P. & Brady, M. 1996 Symmetry analysis through wave propagation. *Int. J. Pattern Recognit. Artif. Intell.* **10** (4), 291–306.

Mukherjee, D. P., Zisserman, A. & Brady, M. 1995 Shape from symmetry: detecting and exploiting symmetry in affine images. *Phil. Trans. R. Soc. Lond.* A **351**, 77–106.

Mundy, J. L. & Zisserman, A. P. 1992 *Geometrical invariance in computer vision*. MA: MIT Press.

Murray, D. W., Kashko, A. & Buxton, H. 1986 A parallel approach to the picture restoration algorithm of Geman and Geman on an SIMD machine. *Image Vision Computing* **4**(3), 133–142.

Reid, I. D. & Brady, M. 1996 Recognition of object classes from range data. *Artif. Intell*. **78**, 289–326.

Reid, I. D. & Murray. D. W. 1996 Active tracking of foveated feature clusters using affine structure. *Int. J. Computer Vision* **18**(1), 41–60.

Reynard, D., Wildenberg, A., Blake, A. & Marchant, J. 1996 Learning dynamics of complex motions from image sequences. In *Proc. of the fourth European conf. on computer vision*, pp. 357–368. Cambridge, England, April 1996.

Rygol, M., Pollard, S., Brown, C. & Mayhew, J. 1992 A parallel 3D vision system. In *Active vision* (ed. A. Blake & A. Yuille), pp. 239–262. MIT Press.

Shapiro, L. S. 1995 *Affine analysis of image sequences*. Cambridge University Press.

Shapiro, L. S., Zisserman, A. P. & Brady, M. 1995 3D motion from point matches via affine epipolar geometry. *Int. J. Computer Vision* **16**, 147–182.

Sharkey, P. M., Murray, D. W., Vandevelde, S., Reid, I. D. & McLauchlan, P. F. 1993 A modular head/eye platform for real-time reactive vision. *Mechatronics* **3**(4), 517–535.

Sullivan, G. D. 1993 Visual interpretation of known objects in constrained scenes. *Phil. Trans. R. Soc. Lond.* B **337**, 118–126.

Szekely, G. 1996 Shape characterization by local symmetries. Habilitationsschrift ETH. Zurich.

Tarassenko, L., Hayton, P., Cerneaz, N. & Brady, M. 1995 Novelty detection for the identification of masses in mammograms. In *Fourth int. conf. on artificial neural networks*, pp. 442–447. Cambridge, UK: Institute of Electrical Engineering.

Thirion, J.-P. 1994 The extremal mesh and the understanding of 3D surfaces. *Int. J. Computer Vision*.

Thompson, D'A. W. 1952 *On growth and form*. Cambridge University Press.

Wiles, C. S. & Brady, M. 1996*a* Ground plane motion camera models. In *Proc. ECCV'96*, II, pp. 238–250 (ed. B. Buxton & R. Cipolla). Cambridge, England, 1996. Lecture Notes in Computer Science. Springer.

Wiles, C. S. & Brady, M. 1996*b* On the appropriateness of camera models. In *Proc. ECCV'96* (ed. B. Buxton & R. Cipolla), II, pp. 228–237. Cambridge, England, 1996. Lecture Notes in Computer Science. Springer.

Xie, Z.-X. & Brady, M. 1996 Texture segmentation using local energy in wavelet scale space. *Image Understand*. (In the press.)