critical rôle when we went to address the issue of computational power. None of the workstations on which *VIS* ran (and there were least five) could offer the power that was required. Parallelism, it seemed, was the only option. And having *VIS* configured as a complete, programmable, environment meant that it was an ideal candidate for coarse-grained parallelism. Quickly, *VIS* was ported to a multi-transputer environment, VISICL was enhanced and Remote Procedure Calls (RPCs) were added, and *VIS* $\vec{a}$ *VIS* was born: a name which is evocative of the idea of parallelism with two (or more) *VIS* systems concurrently operating and communicating with one another, sharing image data and processing.

VIS $\vec{a}$ *VIS* evolved also in functionality, beginning as an early-vision system, until now it comprises grouping processes, 3-D model construction and matching modules, and a Cartesian frame-of-reference robot programming language. It will continue to grow in functionality, absorbing, perhaps, manipulation strategies and capabilities for geometric reasoning. But that is in the future. What is important now is that *VIS* $\vec{a}$ *VIS* works and is providing a useful platform with which to engage in directed, and undirected, research.

D.V. and G.S.
Brussels 11/11/91

Chapter 1

An Introduction to Computer Vision.

David Vernon

The eye sees only what the mind is prepared to comprehend

Robertson Davies

1.1 What is computer vision: image processing or artificial intelligence?

Computer vision is *both* image processing and artificial intelligence, for the true goal of such a vision system is to exhibit adaptive anticipatory — so-called intelligent — behaviour through the processing of image data. Whether or not this behaviour is manifested by a roving robot or by a surveillance system in a high-security warehouse is immaterial; the behaviour exhibited by the system must encompass the assessment of what is happening in its local environment *and* the subsequent activation of appropriate events. For example, the robot might take evasive action while the surveillance system might alert a security guard to the presence of an intruder. More often than not, the effect of these actions will result in a change in the environment which is being monitored by the vision system, requiring in a re-assessment by the system. While this is more obviously true with a mobile robot, where the field of view is continually changing as the robot navigates its terrain, it is equally true of the surveillance system: panning the camera, zooming the lens, or sounding an alarm, will produce equally dramatic changes in the scene presented to the vision system.

We see then that computer vision systems are concerned with the interaction and manipulation of their environment in a coherent and stable manner. This interaction is facilitated by on-going intelligent interplay between perception, on the one hand, and action, be it navigation, manipulation, or communication, on the other.

If the computer vision system is concerned with its environment, it must somehow abstract relevant information about the environment so that it can 'reason' with it. It is these two aspects of vision, the process of abstraction and the representation of information, which form the kernel of current vision systems. Vision systems deal with the physical structure of a three-dimensional world and they do so by by the automatic analysis of images of that world. Since the images are two-dimensional, there is an inevitable loss of information and the recovery of this lost information is a central problem in computer vision. The processes of abstraction, then, must include many techniques, e.g. image processing (which is concerned with the transformation and encoding of images), pattern recognition, the useful description of shape and of volume, geometric modelling, and reasoning.

It is perhaps worth remarking here that the complexity of the environment is quite often a controlled variable in the overall process. That is, vision systems of varying levels of sophistication are developed for appropriately constrained visual domains. We are interested here in the more advanced (or adventurous) systems which address partially-constrained three-dimensional scenes. This type of computer vision is often referred to as image understanding or scene analysis. These scenes can be viewed from one (and often several) unconstrained viewpoints. The illumination conditions may be known but usually one will have to contend with shadows and occlusion, i.e. partially hidden objects. Occasionally, colour information is incorporated but not nearly as often as it should be. Range data is sometimes explicitly available from active range sensing devices, but a central theme of image understanding is the automatic extraction of both range data and local surface orientation from several two-dimensional images using, e.g., stereopsis, motion, shading, occlusion, texture gradients, or focussing.

One of the significant aspects of image understanding is that it utilises several mutually-relevant information representations (e.g. based on the object edges or boundaries, the disparity between objects in two stereo images, and the shading of the objects surface) and it also incorporates different levels of representation in order to organise the information being made explicit in the representation in an increasingly powerful and meaningful manner. For example, an image understanding system would endeavour to model the scene with some form of parameterised three-dimensional object models built from several low-level processes based on distinct visual cues.

It is these topics of visual representations and visual processes to which we now turn.

1.2 Organisation of visual processes

It is clear from the previous introduction that the tenets of conventional computer vision systems fall soundly in the domain of representationalism, i.e. the philosophy that perception is a mechanism by which the entity apprehends the world in which it finds itself, learns its structure and models it, and modifies its behaviour on the basis of what it learns. If we accept the validity of this approach, then a number of questions arise: What are these representations? What processes are necessary to generate them? How are these processes organised? To answer these questions, we must first acknowledge that, in proceeding from raw 2-D images of the world to explicit 3-D structural representations, we are making a significant leap across widely divided levels of representation; the information inherent in the former is implicit and iconic; that in the latter is explicit and predominantly symbolic. To traverse this gap, we must accept that no single process nor representation is going to be generally adequate. Consequently, a central theme which runs through the current, conventional, approach to image understanding is that intermediate representations are required to bridge this gap between raw images and the abstracted structural model. This realisation owes much to the work of David Marr (see [3]) who exerted a major influence on the development of the computational approach to vision. Marr modelled the vision process as an information processing task in which the visual information undergoes different hierarchical transformations at and between levels, generating representations which successively make more and more three-dimensional features explicit.

These representations make different kinds of knowledge explicit and should expose various kinds of constraint upon subsequent interpretations of the scene. It is the progressive integration of these representations and their mutual constraint to facilitate an unambiguous interpretation of the scene that most characterises this approach to vision. It is interesting that most of the progress that has been made in the past few years has not, in fact, been in this area of representation integration, or 'data fusion' as it is commonly known, but rather in the development of formal and well-founded computational models for the generation of these representations in the first place and there remains a great deal of work to be done in the area of data fusion.

We can characterise image understanding, then, as a sequence of processes concerned with successively extracting visual information from one representation (beginning with digital images), organising it, and making it explicit in the representation to be used by other processes. From this perspective, *vision is computationally modular and sequential.*

At present, it is not clear how information in one representation should influence the acquisition and generation of information in another representation but some possibilities include:

- A bottom-up flow of data in which information is made explicit without recourse to *a priori* knowledge. Thus, we form our structural representation

purely on the basis of the data implicit in the original images in a context-free manner.

- Heterarchical Constraint Propagation. This is similar to the bottom-up approach but we now have the additional contraint that cues, i.e. a given information representation, at any one level of the hierarchical organisation of representations can mutually interact to delimit and reduce the possible forms of interpretation at that level and, hence, the generation of the information representations at the next level of the hierarchy. Perhaps one of the simplest examples is the integration of depth values generated by two independent processes such as binocular stereo and the parallax caused by a moving camera.

- A top-down, model driven, information flow whereby early vision is guided by expectations (or models) of what is to be seen.

This book is primarily concerned with the bottom-up approach and with heterarchical constraint propagation. However, it should be noted well that this emphasis is not intended to imply relative merit; it merely reflects our chosen approach. Model-based vision is currently a very popular paradigm and there is a wealth of interesting work being done in the area which would require a volume to itself to do it justice. We chose the bottom-up and heterarchical approachs because they can serve as a foundation of fundamental visual representations and processes upon which we can build in the future.

1.3 Representations.

1.3.1 Digital images.

The initial representation of a scene is the digital image: a two-dimensional, sampled and quantised, representation of the scene's reflectance function. Thus, the digital image represents a projection of the structure of the world, as encoded in the light reflected from each point on the surface of each object, onto the image plane of a camera or sensor system. Each point in the image, a pixel, is a sample of that reflectance function and typically represents the intensity, or grey-level, of the reflected light. Obviously, all information is coded implicitly in this iconic representation. Since the images are generated by projection there is no explicit information about the distance between the sensor system and the relevant points in the scene.

1.3.2 Primal sketches.

Taking as input a grey-level image, Marr proposed the generation of a *Raw Primal Sketch*, a representation which consists of primitives of edges, terminations, blobs, and bars at different spatial scales. Edge primitives are, effectively, local line segment approximations of discontinuities in intensity in an image; curves comprise a

sequence of edges, delimited at either end by the termination primitives. Instances of local parallelism of these edges are represented by bars, while blobs represent the discontinuities which are *not* manifested at several spatial scales. Each primitive has certain associated properties: orientation, width, length, position, and strength.

The computation of the raw primal sketch requires both the measurement of intensity gradients of different spatial scale and the accurate measurement of the location of these changes. In effect, the generation of the Raw Primal Sketch requires the prior detection of edges in the grey-scale images. We will return to the issue of detection of intensity discontinuities in chapter 3 and to the generation of the Raw Primal Sketch in Chapter 6.

As the information made explicit in the raw primal sketch is still local and spatially restricted, i.e. it does not convey any global information about shape in an explicit manner, we may now wish to group these primitives so that the groups correspond to physically meaningful objects. In this sense, the grouping process is exactly what is commonly meant by the term *segmentation*. Many of the more advanced segmentation techniques are based on Gestalt 'figural grouping principles', named after the Gestalt school of psychology formed in the early part of this century. For example, primitives can be grouped according to three criteria: continuity, proximity, and similarity. In the first case, for instance, lines bounded by terminations which are co-linear would be grouped. Primitives which are spatially proximate are also candidates for the formation of groups while so too are primitives which belong to the same class, i.e., which are similar. These criteria are not independent and the final grouping will be based on the relative weights attached to each criterion. The outcome of such grouping, the *full primal sketch*, makes explicit the region boundaries, object contours, and primitive shapes. Chapter 7 deals with these grouping processes in detail.

1.3.3 The $2\frac{1}{2}$-D sketch.

The next level of representation is the $2\frac{1}{2}$-D (two-and-a-half dimensional) sketch. This is derived both from the full primal sketch and from the grey-level image by using many visual cues, including stereopsis, apparent motion, shading, shape, and texture. The $2\frac{1}{2}$-D sketch is a viewer-centred representation of the scene, i.e. all metrics are defined in the viewer or image frame of reference, and it contains not only primitives of spatial organisation and surface discontinuity but also of the local surface orientation at each point and an estimate of the distance from the viewer. Thus, the $2\frac{1}{2}$-D sketch can be thought of as a 2-D array of 3-valued entities, representing the distance from the camera to the surface and the two angles specifying the surface normal vectors, in addition to the grouping and edge information made explicit in the primal sketch. The computation of depth information is dealt with in chapters 4 and 5 while local surface orientation is addressed in chapter 8.

1.3.4 3-D models.

The final stage of this information processing organisation of visual processes lies in the analysis of the $2\frac{1}{2}$-D sketch and the production of an explicit 3-D representation. There are two issues which must be addressed here:

1. The conversion from a viewer-centred representation to an object-centred representation. This is, in effect, the transformation between a camera co-ordinate system and the real-world co-ordinate system. This relationship is commonly referred to as the camera model.

2. The type of 3-D representation we choose to model our objects. There are three main types of 3-D representation based on volumetric, skeletal, and surface primitives.

While chapter 8 discusses 3-D model construction in some depth, it might be useful to summarise the main approaches here. Details on how to compute the camera model can be found in the references cited in the bibliography at the end of the chapter.

Volumetric representations work on the basis of spatial occupancy, delineating the segments of a 3-D workspace which are, or are not, occupied by an object. The simplest representation utilises the concept of a voxel image (the word voxel derives from the phrase *volumetric element*) which is a 3-D extension of a conventional 2-D binary image. Thus, it is typically a uniformly-sampled 3-D array of cells, each one belonging either to an object or to the free space surrounding the object.

The oct-tree is another volumetric representation. However, in this instance, the volumetric primitives are not uniform in size and the spatial occupancy of a work-space can be represented to arbitrary resolution in quite an efficient manner. The oct-tree describes the occupancy of a 3-D image by representing large homogeneous volumes as single nodes in a tree, the position in the tree governing the size of the volume. When constructing the oct-tree, the work-space is initially represented by a single cubic volume. If the work-space is completely occupied by an object (a very unlikely situation), then it is represented by a single root node in the oct-tree. In the more likely situation that the volume is not completely occupied, the cube is divided into eight sub-cubes of equal volume and represented by eight nodes in the tree, each of which are offspring of the node which was sub-divided. Again, if any of these sub-cubes are completely occupied then that volume is represented by a node at the second level in the tree; alternatively, the sub-cube is further sub-divided into another eight cubes and the same test for complete spatial occupancy is applied. This process is re-iterated until we reach the required resolution for spatial occupancy, i.e., the smallest required cubic volume: this is equivalent to the voxel in the previous representation.

The generalised cylinder, also referred to as the generalised cone, is among the most common skeletal 3-D object representations. A generalised cylinder is defined as the surface created by moving a cross-section along an axis. The cross-section can vary in size, getting larger or smaller, but the shape remains the same and the axis can trace out any arbitrary three-dimensional curvilinear path. Thus, a single generalised cylinder can represent, e.g., a cone (circular cross-section; linear decrease in diameter; linear axis), a sphere (circular cross-section; sinusoidal variation in diameter; linear axis), or a washer (rectangular cross-section; constant area; circular axis). However, a general 3-D model comprises several generalised cones and is organised in a modular manner with each component comprising its own generalised cylinder based model. Thus, the 3-D model is a hierarchy of generalised cylinders.

Finally, we come to the third type of 3-D model which is based on surface representations. We immediately have a choice to make regarding the type of surface primitives (or surface patches) we will allow: planar patches or curved patches. Although there is no universal agreement about which is the best, the planar patch approach is quite popular and yields polyhedral approximations of the object being modelled. This is quite an appropriate representation for man-made objects which tend predominantly to comprise planar surfaces. It is not, however, a panacea for 3-D representational problems and it would appear that many of the subtleties of 3-D shape description cannot be addressed with simplistic first-order planar representations. Nevertheless, it does have its uses and, even for naturally curved objects, it can provide quite a good approximation to the true shape, if an appropriate patch size is used.

1.4 Early visual processes.

From the preceding discussion, it is clear that we require several visual processes in order to generate each representation. Amongst those we mentioned are:

- The detection of intensity discontinuities.
- Grouping processes and segmentation.
- The computation of depth information.
- The computation of local surface orientation.

As before, although each of these is dealt with in the following chapters, we will summarise some of the main issues here.

1.4.1 Isolation of intensity discontinuities and detection of edges.

As might be expected when dealing with a process which is fundamental to image processing, the literature concerning edge detection is large. Our interest here is not in reviewing all detectors in detail but in identifying the broad thrust of the approach and, specifically, in briefly introducing the Marr-Hildreth edge detection scheme which is used in *VIS à VIS*, the parallel computer vision system described in this book. We will return to deal with this technique in more detail in Chapter 3.

If we define a local edge in an image to be a transition between two regions of significantly different intensities, i.e. an intensity discontinuity, then the spatial first derivatives of the image, which measures the rate of change of intensity, will have large values in these transitional boundary areas. Thus first-derivative, or gradient, based edge detectors enhance the image by estimating the partial derivatives and then signal that an edge is present if these derivatives, or combinations of the derivatives, exceed some defined threshold. On the other hand, second derivatives too can be used to detect intensity discontinuities. In this instance, however, we seek not local maxima of the image gradient but instances where the function crosses from a positive to a negative value (or *vice versa*).

In more detail, if $\partial/\partial x$ and $\partial/\partial y$ represent the rates of change of a 2-D function $f(x, y)$ in the x and y directions respectively, then the Laplacian operator:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

(i.e.,the sum of second-order, unmixed, partial derivatives) responds on either side of the edge, once with a positive sign and once with a negative sign. Thus in order to detect edges, the image is enhanced by evaluating the digital Laplacian and isolating the points at which the resultant image goes from positive to negative, i.e., at which it crosses zero. Unfortunately, the Laplacian has one significant disadvantage: it responds very strongly to noise.

A different and much more successful application of the Laplacian to edge detection was proposed by Marr and Hildreth in 1980. This approach first smooths the image by convolving it with a two-dimensional Gaussian function, and subsequently isolates the zero-crossings of the Laplacian of this image:

$$\nabla^2 I(x, y) * G(x, y)$$

where $I(x, y)$ represents the image intensity at a point (x, y) and $G(x, y)$ is the 2-D Gaussian function, of a given standard deviation σ, defined by :

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

Despite some criticism of this technique, it is widely used. The operator possesses a number of useful properties which we will discuss in Chapter 3.

1.4.2 Grouping and segmentation.

In the previous section, we dealt with one element of the process of edge detection: the isolation of discontinuities in intensity. However, edge detection as a whole is a complex procedure as it requires not only the the isolation of the image features we call edges, but it is also concerned with the inference of the physical cause of those features. Edge detection techniques belong to a generic class of image processing and analysis operations, a class normally referred to as segmentation, which effect the

isolation of specific image regions corresponding, unambiguously, to given objects. Such object isolation is often achieved quite simply by edge detection alone but this is the case only if the scene is extremely simple. For more realistic natural scenes, the segmentation process requires more sophisticated grouping techniques which collect together non-trivial image based entities into groups which correspond to a single physical object.

In *VIS à VIS*, we deal with grouping in two stages, exactly in the manner suggested by Marr in his seminal work in this area. The first concerns the formation of the raw primal sketch which make explicit preliminary groups of intensity discontinuities into 'locally-straight' line segments and assigns to the groups attributes of, e.g., size, contrast, position, orientation, and local parallelism. The formation of the raw primal sketch also involves, to a small extent, the inference of the physical cause of the edge, in that the line segment is only formed if the constituent intensity discontinuities appear at several spatial scales. This means that the intensity discontinuities are present, irrespective of whether we are looking at that local region in detail (small spatial scale) or at its gross, smoothed, appearance (large spatial scale). Thus, the raw primal sketch involves a simple type of edge authentication by requiring the spatial coincidence of intensity discontinuities derived with different levels of smoothing, typically achieved by the convolution of the image with Gaussian filters of different standard deviation. We will return again to the generation of the raw primal sketch in chapter 6.

The second grouping procedure is concerned with the global spatial distribution of these attributed local line segments. This grouped representation is referred to as the full primal sketch. Many grouping principles have been suggested but those implemented in *VIS à VIS* are based on the figural grouping principles of Gestalt psychology. These rules attempt to capture some of the perceptual behaviour of humans, forming groups of entities on the basis of, e.g., their similarity (e.g., in shape, in orientation, in size) of their spatial distribution (e.g. proximity, co-linearity, curvilinearity). Significantly, Gestalt grouping principles depend quite heavily on the ability to compound the groupings, i.e. to group the groups, and in chapter 7, we will discuss the effect of this type of recursive process.

1.4.3 Stereopsis and visual motion

The distance, or depth, from a viewer to any given point in the observed scene is required to construct the $2\frac{1}{2}$-D sketch. In passive computer vision, which rely solely on the analysis of the reflected light of a naturally illuminated scene rather than on contrived illumination, this is normally achieved by triangulation. This involves the use of two (or more) views of a scene to recover the distance of objects in the scene from the observer (the cameras). The camera model (or, more accurately, an algebraic variant, the *inverse perspective transformation*), allows us to construct a line describing all of the points in the 3-D world which could have been projected onto a given image point. If we have two images acquired at different positions in the world, i.e. a stereo pair, then for the two image co-ordinates which correspond

to a single point in 3-D space, we can construct two lines, the intersection of which identifies the 3-D position of the point in question. Thus, there are two aspects to stereo imaging:

1. The identification of corresponding points in the two stereo images.
2. The computation of the 3-D coordinates of the world point which gives rise to these two corresponding image points.

The main problem in stereo is to find the corresponding points in the left and right images; this is commonly referred to as the correspondence problem.

In characterising a stereo system, we must address the kind of visual entities on which the stereo system works and the mechanism by which the system matches visual entities in one image with corresponding entities on the other.

Typically, the possible visual entities from which we can choose include, at an iconic level, points of intensity discontinuity, patches (small areas) in the intensity image, and patches in filtered images; or, at a more symbolic level, line segment tokens, such as are made explicit in the raw primal sketch.

The matching mechanism which establishes the correspondence will depend on the type of visual entities which we have chosen: iconic entities will normally exploit some template matching paradigm, such as normalised cross-correlation, while token entities can be used with more heuristic search strategies. The stereo matching system in *VIS á VIS* is described in chapter 4.

While stereopsis involves the analysis of two images for binocular stereo, or three images in the case of trinocular stereo, it is possible to exploit many more images if either the object or the observer is moving. This analysis of object motion in sequences of digital images, or of apparent motion in the case of a moving observer, to provide information about the structure of the imaged scene is an extremely topical and important aspect of current image understanding research. In this book, we confine our attention to the problem of camera motion and, in particular, to the study of apparent motion of objects in a scene arising from the changing vantage point of a moving camera. This restriction is not necessarily a limitation on the usefulness of the technique; on the contrary, the concept of a camera mounted on the end-effector of a robot, providing hand-eye coordination, or on a moving autonomously guided vehicle (AGV), providing navigation information, is both appealing and plausible.

From an intuitive point of view, camera motion is identical to the stereo process in that we are identifying points in the image (e.g. characteristic features on an object) and then tracking them as they appear to move due to the changing position (and, perhaps, attitude) of the camera system. At the end of the sequence of images, we then have two sets of corresponding points, connected by optic flow vectors, in the first and last images of the sequence. Typically, we will also have a sequence of vectors which track the trajectory of the point throughout the sequence of images. The depth, or distance, of the point in the world can then be computed using the inverse perspective transformation.

However, there are a number of differences. First, the tracking is achieved quite often, not by a correlation technique or by a token matching technique, but by differentiating the image sequence with respect to time to see how it changes from one image to the next. There is often a subsequent matching process to ensure the accuracy of the computed image change, and sometimes it is not the grey-scale image which is differentiated but, rather, a filtered version of it. Nevertheless, the information about change is derived from a derivative (or, more accurately, a first difference) of the image sequence. Second, the 'correspondence' between points is established incrementally, from image to image, over an extended sequence of images. Thus, we can often generate accurate and faithful maps of point correspondence which are made explicit by a 2-D array of flow vectors which describe the trajectory of a point over the image sequence.

Chapter 5 describes two simple types of camera motion which are exploited in *VIS á VIS*.

1.4.4 Other visual cues.

The construction of the $2\frac{1}{2}$-D sketch and 3-D models require one further element: the computation of the local orientation of a point, i.e. the surface normal vector. The analysis of the shading of a surface, based on assumed models of the reflectivity of the surface material, is sometimes used to compute this information.

The amount of light reflected from an object depends on :

1. The surface material
2. The emergent angle, e, between the surface normal and the viewer angle.
3. The incident angle, i, between the surface normal and light source direction.

There are several models of surface reflectance, the simplest of which is the Lambertian model. A Lambertian surface is a surface that looks equally bright from all view points, i.e. the brightness of a particular point does not change as the viewpoint changes. It is a perfect diffuser: the observed brightness depends only on the direction to the light source, i.e., the incident angle i.

Let E be the observed brightness, then for a Lambertian surface:

$$E = r \cos i$$

where r is a constant called the surface albedo and is peculiar to the surface material under analysis.

There will, in general, be many surface orientations which satisfy $E = r \cos i$ for a given brightness E_1. However, the solution set is well-formed and can be derived either by analytic means or by empirical calibration. Nonetheless, an extra constraint is required to uniquely determine the orientation of the imaged point from the brightness: this is supplied by assumptions of surface smoothness (or continuity),

i.e. that the surface should not vary much from the surface direction at neighbouring parts. If we have some points on the surface of the object to anchor the process, we can infer the orientation of neighbouring points.

On occluding boundaries of objects without sharp edges, i.e. on extremal boundaries, surface direction is perpendicular both to the viewer's line of sight and to the local orientation of the contour. Thus, we know immediately the surface orientation of every occluding contour point. Since the surface orientation changes smoothly, all points on the surface close to the occluding boundary must have an orientation which is not significantly different from that of the occluding boundary. The surface orientation of each point adjacent to the occluding boundary can now be computed by measuring the intensity value and identifying the corresponding orientation in the solution set, knowing that its orientation is similar to that of the adjacent occluding boundary anchor point. This scheme of local constraint is re-iterated using these newly computed orientations as constraints, until the orientation of all points on the surface have been computed.

This technique has been studied in depth in the computer vision literature and it should be emphasised that this description is intuitive and tutorial in nature; you are referred to the appropriate texts cited in the bibliography at the end of the chapter. As we have noted, however, there are a number of assumptions which must be made in order for the technique to work successfully, e.g. the surface orientation must vary smoothly and, in particular, it must do so at the occluding boundary (the boundary of the object at which the surface disappears from sight). Look around the room you are in at present. How many objects do you see which fulfil this requirement? Probably very few. Allied to this are the requirements that the reflective surface has a known albedo and that we can model its reflective properties, or alternatively, that we can calibrate for a given reflective material. Finally, we assume that we know the incident angle of light. The necessity of these several assumptions limits the usefulness of the technique for general image understanding.

There are other ways of estimating the local surface orientation. As an example of one coarse approach, consider the situation where we have a 3-D raw primal sketch, i.e. a raw primal sketch in which we know the depth to each point on the edge segments. If these raw primal sketch segments are sufficiently close, we can compute the surface normal by interpolating between the edges, generating a succession of planar patches, and effectively constructing a polyhedral model of the object. The surface normal is easily computed by forming the vector cross product of two vectors in the plane of the patch (typically two non-parallel patch sides).

1.5 Vision: an end in itself?

If we have managed to build an unambiguous 3-D model of the viewed scene, is that enough? Is that the end of the vision process? The answer must be no. It is important to realise that computer vision systems, *per se*, are only part of a more embracing system, a larger system which is simultaneously concerned with making

sense of the environment and interacting with the environment. Without action, perception is futile; without perception, action is futile. Both are complementary, but strongly-related, activities and any intelligent action in which the system engages in the environment, i.e. anything it does, it does with an understanding of its action, and it gains this quite often by on-going visual perception. Computer vision, then, is not an end in itself; that is, while the task of constructing an unambiguous explicit 3-D representation of the world is a large part of its function, there is more to vision than 'just' the explication of structural organisation. In essence, computer vision systems, or image understanding systems, are as concerned with cause and effect, with purpose, with action and reaction as they are with structural organisation.

Computer vision, then, truly lies in the domain of artificial intelligence since it is intimately concerned with the design of artificial autonomous, adaptive and anticipatory, systems. That being said, however, we have not advanced greatly in these aspects of image understanding and computer vision and a great deal of attention is being focussed on the development of formal and well-founded bases of visual processes so that the discipline of computer vision is developed just as that, a discipline, and not an *ad hoc* collection of techniques and algorithms. For the most part, then, current activity in computer vision *is* concerned with the explication of structural organisation and it is this aspect of vision which is the subject matter of this book. However, the issues we have just raised, in effect the temporal semantics of vision in contribution to and in participation with physically interactive systems, are critically important to the long-term development of intelligent, autonomous, visual systems.

Bibliography

[1] Ayache, N. and Faugeras, O.D. 1986 'HYPER: A new approach for the recognition and positioning of two-dimensional objects' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAMI-8 No 1, pp. 44–54.

[2] Ballard, D.H., and Brown, C.M. 1982. *Computer Vision* Prentice-Hall, New Jersey.

[3] Ballard, D. H. and Sabbah, D. 1983. 'Viewer Independent Shape Recognition' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 6, pp. 653–660.

[4] Asada, H. and Brady, M. 1986 'The Curvature Primal Sketch' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 1, pp. 2–14.

[5] Bamieh, B. and De Figueiredo, R. J. P. 1986. 'A General Moment-Invariants/ Attributed-Graph Method for Three-Dimensional Object Recognition from a Single Image', IEEE Journal of Robotics and Automation, Vol. RA-2, No. 1, pp.31–41.

[6] Barnard, S.T., and Fischler, M.A. 1982.'Computational Stereo', SRI International, Technical Note No. 261.

[7] Besl, P.J. and Jain, R. 1985. "Three-Dimensional Object Recognition", ACM Computing Surveys, Vol. 17, No. 1, pp.75–145.

[8] Bhanu, B. 1982. 'Surface Representation and Shape Matching of 3-D Objects', Proceedings of IEEE Computer Society Conference on Pattern Recognition and Image Processing, Las Vagas, NV, USA, pp.349–354.

[9] Bhanu, B. 1984. 'Representation and Shape Matching of 3-D Objects', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, No.3, pp.340–351.

[10] Binford, T.O. 1982. 'Survey of Model-Based Image Analysis Systems', The International Journal of Robotics Research, Vol. 1, No. 1, pp.18–64.

[11] Boyle, R.D. and Thomas, R.C. 1988. *Computer Vision - A First Course* Blackwell Scientific Publications, Oxford.

[12] Brady, J.M. (ed.) 1981. *Computer Vision* Elsevier Science Publishers, The Netherlands.

[13] Brady, M. 1982. 'Computational Approaches to Image Understanding', ACM Computing Surveys, Vol. 14, No. 1, pp.3–71.

[14] Brady, M. 1982. 'Describing Visible Surfaces', Proceedings of COMPSAC 82, IEEE Computer society's Sixth International Conference on Computer Software and Applications, pp.385–391.

[15] Brady, M. and Asada, H. 1984. 'Smoothed Local Symmetries and their Implementation' The International Journal of Robotics Research, Vol. 3, No. 3, pp. 36–61.

[16] Brooks, R.A. 1981. 'Symbolic Reasoning Among 3-D Models and 2-D Images', Artifical Intelligence, Vol.17, pp.285–348.

[17] Brooks, R.A. 1983. 'Model-Based Three-Dimensional Interpretations of Two-Dimensional Images', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 2, pp.140–150.

[18] Canny, J. 1986 'A computational approach to edge detection' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAMI-8, No 6, pp. 679–698.

[19] Davis, L. S., Janos, L. and Dunn, S. M. 1983. 'Efficient Recovery of Shape from Texture' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 5, pp. 485–492.

[20] Dawson, K. and Vernon, D. 1990. 'Implicit Model Matching as an Approach to Three-Dimensional Object Recognition', Proceedings of the ESPRIT Basic Research Action Workshop on 'Advanced Matching in Vision and Artificial Intelligence', Munich, June 1990.

[21] Duda, R.O., and Hart, P.E. 1973.*Pattern Classification and Scene Analysis* Wiley.

[22] Fairhurst, M.C. 1988. *Computer Vision for Robotic Systems* Prentice-Hall International (UK), Hertfordshire.

[23] Fischler, M.A. 1981. 'Computational Structures of Machine Perception', SRI International, Technical Note No. 233.

[24] Fischler, M.A. and Bolles, R.C. 1986. 'Perceptual Organisation and Curve Partitioning', IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 1, pp.1–5.

[25] Frigato, C., Grosso, E., Sandini, G., Tistarelli, M., and Vernon, D. 1988. 'Integration of Motion and Stereo', Proceedings of the 5th. Annual ESPRIT Conference, Brussels, Edited by the Commission of the European Communities, Directorate-General Telecommunications, Information Industries and Innovation, North-Holland, Amsterdam, pp.616–627.

[26] Giordano, A., Maresca, M, Sandini, G., Vernazza, T., and Ferrari, D. 1987. 'VLSI-based Systolic Architecture for fast Gaussian Convolution', Optical Engineering, Vol. 26, No. 1, pp.63–68.

[27] Gonzalez, R.C. and Wintz, P. 1977. *Digital Image Processing* Addison-Wesley, Reading, Mass.

[28] Grimson, W. E. L. *From Images to Surfaces* MIT Press Cambridge, Mass.

[29] Grimson, W.E.L. and Hildreth, E.C. 1985. 'Comments on Digital Step Edges from Zero Crossings of Second Directional Derivatives', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, No. 1, pp.121–127.

[30] Guzman A. 1968. *Computer Recognition of Three-Dimensional Objects in a Visual Scene*, Ph.D. Thesis, MIT, Mass.

[31] Hall, E.L. 1979. *Computer Image Processing and Recognition* Academic Press, New York.

[32] Hanson, A.R. and Riseman, E.M. 1978. 'Segmentation of Natural Scenes', in *Computer Vision Systems* Hanson, A.R. and Riseman, E.M. (Eds.).

[33] Haralick, R.M., Watson, L.T., and Laffey, T.J. 1983. 'The Topographic Primal Sketch', The International Journal of Robotics Research, Vol. 2, No. 1, pp.50–72.

[34] Haralick, R.M. 1984. 'Digital Step Edges from Zero-Crossing of Second Directional Derivatives', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, No. 1, pp.58–68.

[35] Henderson, T.C. 1983. 'Efficient 3-D Object Representations for Industrial Vision Sytems', IEEE Transactions on Pattern Analysis and Machine Intelligence', Vol. PAMI-5, No. 6, pp.609–618.

[36] Hildreth, E.C. 1983. 'The Measurement of Visual Motion', MIT Press, Cambridge, USA.

[37] Hildreth, E.C. 1985. 'Edge Detection', A.I. Memo No. 858, M.I.T. A.I. Lab.

[38] Horaud, P., and Bolles, R.C. 1984. '3DPO's strategy for Matching 3-D Objects in Range Data', International Conference on Robotics, Atlanta, GA, USA, pp.78-85.

[39] Horn B.K.P. and Schunck, B.G. 1981. 'Determining Optical Flow', Artificial-Intelligence 17, No.1–3, pp.185-204.

[40] Horn B.K.P. and Ikeuchi, K. 1983. 'Picking Parts out of a Bin', A.I. Memo No. 746, MIT A.I. Lab.

[41] Horn, B.K.P. 1986. *Robot Vision* The MIT Press, Cambridge, Mass.

[42] Huertas, A. and Medioni, G. 1986. 'Detection of Intensity Changes with Subpixel Accuracy Using Laplacian-Gaussian Masks', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 5, pp.651–664.

[43] Ikeuchi, K. 1983. 'Determining Attitude of Object From Needle Map Using Extended Gaussian Image', MIT. A.I. Memo No.714.

[44] Ikeuchi, K. 1984 'Determining Surface Orientations of Specular Surfaces by Using Photometric Stereo Method', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-3, No. 6, pp. 661–669.

[45] Ikeuchi, K., Nishihara, H.K., Horn, B.K., Sobalvarro, P., and Nagata, S. 1986. 'Determining Grasp Configurations using Photometric Stereo and the PRISM Binocular Stereo System', The International Journal of Robotics Research, Vol. 5, No. 1, pp.46–65.

[46] Ju, W. K., Yang, J. Y. and Huang, T. S. 1987. 'Matching Perspective Views of a Polyhedron Using Circuits', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-9, No. 3, pp. 390–400.

[47] Kanade, T. 1981. 'Recovery Of The Three-Dimensional Shape Of An Object From A Single View', Artificial Intelligence, Vol.17, pp.409–460.

[48] Kanade, T. 1983. 'Geometrical Aspects of Interpreting Images as a 3-D Scene', Proceedings of the IEEE, Vol.71, No.7, pp.789–802.

[49] Kanatani, K. 1986. 'The Constraints on Images of Rectangular Polyhedra', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 4, pp. 456–463 .

[50] Kim, Y.C. and Aggarwal, J.K. 1987. 'Positioning Three-Dimensional Objects using Stereo Images', IEEE Journal of Robotics and Automation, Vol. RA-3, No. 4, pp.361–373.

[51] Kuan, D.T. 1983. 'Three-Dimensional Vision System for Object Recognition', Proceedings of SPIE, Vol.449, pp.366–372.

[52] Laurendeau, D. and Poussart, D. 1987. 'Model Building of Three-Dimensional Polyhedral Objects Using 3-D Edge Information and Hemispheric Histogram', IEEE Journal of Robotics and Automation, Vol. RA-3, No. 5, pp. 459–470.

[53] Lawton, D.T. 1983. 'Processing Translational Motion Sequences', Computer Vision Graphics and Image Processing, Vol. 22, pp. 116–144.

[54] Lowe, D.G. and Binford, T.O. 1985. 'The Recovery of Three-Dimensional Structure from Image Curves', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, No. 3, pp.320–326.

[55] Lowe, D. G. 1985. *Perceptual Organisation and Visual Recognition* Kluwer Academic Publishers, Boston.

[56] Lowe, D. G. and Binford, T. O. 1985'The Recovery of Three-Dimensional Structure from Image Curves' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, No. 3, p.p 320–326.

[57] Lunscher, W. F. and Beddoes, M. P. 1986. 'Optimal edge detector design I: parameter selection and noise effects' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 2, pp. 164–177.

[58] Lunscher, W. F. and Beddoes, M. P. 'Optimal edge detector design II: coefficient quantization' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 2, pp. 178–187.

[59] Magee, M. J., Boyter, B. A., Chien, C. and Aggarwal, J. K. 1985. 'Experiments in Intensity Guided Range Sensing Recognition of Three-Dimensional Objects' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, No. 6, pp. 629–637.

[60] Marr, D. 1976. 'Early Processing of Visual Information', Philosophical Transactions of the Royal Society of London, B275, pp.483–524.

[61] Marr, D. and Poggio, T. 1979. 'A Computational Theory of Human Stereo Vision', Proceedings of the Royal Society of London, Vol. B204, pp.301–328.

[62] Marr, D. and Hildreth, E. 1980. 'Theory of Edge Detection' Proceedings of the Royal Society of London Vol. B207, pp. 187–217

[63] Marr, D. 1982. *Vision* W.H. Freeman and Co., San Francisco.

[64] Martin, W.N. and Aggarwal, J.K. 1983. 'Volumetric Descriptions of Objects from Multiple Views', IEEE Transactions on Pattern Analysis and Machine Intellegence, Vol. PAMI-5, No.2, pp.150–158.

[65] McFarland, W.D., and McLaren, R.W. 1983. 'Problem in Three Dimensional Imaging', Proceedings of SPIE, Vol.449, pp.148–157.

[66] McPherson, C.A., Tio, J.B.K., Sadjadi, F.A., and Hall, E.L. 1982. 'Curved Surface Representation for Image Recognition', Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing, Las Vegas, NV, USA, pp.363–369.

[67] McPherson, C.A. 1983. 'Three-Dimensional Robot Vision', Proceedings of SPIE, Vol.449, part 4, pp.116–126.

[68] Nalwa, V.S. and Binford T.O. 1986. 'On Detecting Edges', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 6, pp.699–714.

[69] Nackman, L. R.and Pizer, S. M. 'Three-Dimensional Shape Description using the Symmetric Axis Transform 1: Theory' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, No. 2, pp. 187–202.

[70] Nevatia, R. 1982. *Machine Perception* Prentice-Hall, New Jersey.

[71] Nishihara, H.K. 1983. 'PRISM: A Practical Realtime Imaging Stereo Matcher',Proceedings of SPIE, Vol.449, pp.134–142.

[72] Pentland, A. 1982. 'The Visual Inference of Shape: Computation from Local Features', Ph.D. Thesis, Massachusetts Institute of Technology.

[73] Poggio, T. 1981. 'Marr's Approach to Vision', MIT A.I. Lab., A.I. Memo No. 645.

[74] Pradzy, K. 1980. 'Egomotion and Relative Depth Map from Optical Flow', Biol. Cybernetics Vol. 36, pp.87–102.

[75] Pratt, W.K. 1978. *Digital Image Processing* Wiley, New York.

[76] Pugh, A. (ed.) 1983. *Robot Vision*, IFS (Publications) Ltd, U.K.

[77] Roberts, L.G. 1965. 'Machine Perception of Three-Dimensional Solids' in *Optical and Electro-Optical Information Processing*, J.T. Tippett *et al.* (Eds.), MIT Press, Cambridge, Massachusetts, pp.159–197.

[78] Rosenfeld, A. and Kak, A. 1982. *Digital Picture Processing* Academic Press, New York.

[79] Safranek, R.J. and Kak, A.C. 1983. 'Stereoscopic Depth Perception for Robot Vision: Algorithms and Architectures', Proceedings of IEEE International Conference on Computer Design: VLSI in Computers (ICCD '83), Port Chester, N.Y., USA, pp.76–79.

[80] Sandini, G. and Vernon. D. 1987. 'Tools for Integration of Perceptual Data', in *ESPRIT '86: Results and Achievements* Directorate General XIII, Commission of the European Communities (Eds.), Elsevier Science Publishers B.V. (North Holland), pp.855–865.

[81] Sandini, G., Tistarelli, M., and Vernon, D. 1988 'A Pyramid Based Environment for the Development of Computer Vision Applications', IEEE International Workshop on Intelligent Robots and Systems, Tokyo.

LLIS HORWOOD SERIES IN
RTIFICIAL INTELLIGENCE
eries Editor: Professor JOHN CAMPBELL, Department of Computer Science, University
ollege London

ARALLEL COMPUTER VISION
he VIS a VIS System
D. VERNON, Department of Computer Science, Trinity College Dublin, Ireland, and
ofessor G. SANDINI, DIST, University of Genoa, Italy

escribing a computer vision environment which supports coarse granularity parallelism on
nsputer-based systems, this book discusses a system which treats the organization of the
age data and its representation and interdependencies equally with the visual processes
hich operate on that data. As an environment, the system – VIS a VIS – provides dynamic
eation of images and attendant data structures, each of which are linked to provide an
plicit representation of the interdependence of the visual cues. As a vision system, VIS a
S offers all the required functionality of 3-D image understanding, from image acquisition
rough detection and analysis of intensity discontinuities, computation of depth from stereo
d camera motion, segmentation using raw and full primal sketches, through to 3-D model
nstruction and object recognition. Both geometric and functional parallelism are achieved
exploiting both the in-built vision control language – with its facility for remote execution
procedures – and the integrated system data-structures.

of these issues are dealt with in considerable detail, and the book's emphasis on system
ganization integration of information and computer architectures, rather than purely
eoretical and algorithmic issues, will make it a very welcome addition to the literature.

adership: Computer vision scientists, Industrial research and development engineers,
mputer scientists, Information technologists.

r. Vernon's interests stretch from robot vision to the philosophy of science. His recent research
vities have been concerned with the development of robot systems which display true autonomy. His
rk encompasses computational theories of perception, self-organization, and intelligence. Dr Vernon
Fellow of Trinity College Dublin, Ireland.
lio Sandini has been working on research in computer and biological vision for more than twelve
rs, his primary interests at present being the integration of many low-level visual cues in a robust
mputational framework. He has also been very active in exploiting the neurobiological principles of
y vision to postulate and implement artificial visual sensing at the 'retinal' level of computer vision.
fessor Sandini is currently at the University of Genoa in Italy.

es of related interest

RMAL SPECIFICATION OF ADVANCED AI ARCHITECTURES
N D. CRAIG, Department of Computer Science, University of Warwick.

TIFICIAL INTELLIGENCE AND INTELLIGENT SYSTEMS
Implications
VID ANDERSON, Division of Computer Science, Teesside Polytechnic.

TIFICIAL INTELLIGENCE TERMINOLOGY:
eference Guide
neral Editor: COLIN BEARDON, Department of Computer Science, University of Exeter.

ISBN 0-13-932716-9

LLIS HORWOOD

Vernon and Sandini **PARALLEL COMPUTER VISION** the VIS a VIS system

ELLIS HORWOOD

ELLIS HORWOOD SERIES IN ARTIFICIAL INTELLIGENCE

PARALLEL COMPUTER VISION
the VIS a VIS system

D. Vernon and G. Sandini