Advanced Image Understanding and Autonomous Systems *

David Vernon
Department of Computer Science
Trinity College Dublin
Ireland

Abstract

The ultimate goal of most image understanding systems is to produce an unambiguous 3-D representation of the local visual environment. This representation can then be employed by robotic systems to effect some meaningful action. A great deal of research effort is concerned with the development of visual and 'manipulative' representations, and their generative processes, which allow for the effective linking of such visual perception and robotic action. Part of the motivation for this effort is the desire to develop autonomous systems. It is argued in this paper that the requirements for the development of autonomous systems are not fully compatible with the current representation-based A.I. paradigm. While this approach is ideal for the construction of goal-oriented systems which function in environments that can be specified *a priori*, it does not, and cannot, address the problems encountered when adaptive, self-determining, autonomous systems are required. It is argued that such autonomous systems must be self-organizing. The problem which then arises is how such autonomous systems can be imbued with a goal-oriented behaviour which reflects the requirements of its designer. This remains an open question.

1 Current Approaches to Image Understanding

Image understanding systems, in general, and robot vision systems, in particular, are concerned with the automatic interaction between computer-based machines and their environment. This interaction is facilitated by on-going intelligent interplay between perception, on the one hand, and action, be it navigation or manipulation, on the other. This is perhaps best characterized by the currently-popular paradigm of active vision where the sensor is actively moved through the local environment to validate and refine the interpretation the system has formed of the scene before it.

Current approaches to image understanding are based, for the most part, on the assumption that, if the image understanding system is concerned with its environment, it must somehow abstract relevant information about the environment so that it can 'reason' with it. These two aspects of vision, the representation of information and the processes which facilitate the abstraction of that information, form the kernel of current

*This work was supported by the Hitachi Dublin Laboratory; their assistance is gratefully acknowledged.

vision systems. Most advanced image understanding systems utilise several mutually-relevant information representations (based, e.g., on the object edges or boundaries, the disparity between objects in two stereo images, and the shading of the objects surface) and incorporate different levels of representation in order to organise the information being made explicit in the representation in an increasingly powerful and meaningful manner. Typically, an image understanding system will endeavour to model the scene with some form of parameterised three-dimensional object model built from several low-level processes based on distinct visual cues.

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, for the moment, 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? The answers to these questions form the body of research in advanced image understanding.

There is not sufficient space in this paper to include an exhaustive review of each of these processes and representations. Indeed, to do so would distract us from the central argument in the paper which concerns the relationshop between image understanding systems and autonomous systems. A summary can be found in [1] and in [2] and we will instead provide a brief thumbnail sketch of a few representative visual representations and processes.

1.1 Organisation of visual processes

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.

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.*

1.2 Visual Representations.

1.2.1 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. 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.

1.2.2 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.

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. The outcome of such grouping, the *full primal sketch*, makes explicit the region boundaries, object contours, and primitive shapes.

1.2.3 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.

1.2.4 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.

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 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. 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. There are two types of surface primitive (or surface patch): planar patches and 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.

1.3 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, and the computation of local surface orientation. We will summarise some of the main issues here.

1.3.1 Isolation of intensity discontinuities and detection of edges.

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*).

1.3.2 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 such as we suggested in section 1.2.2 which collect together non-trivial image based entities into groups which correspond to a single physical object.

1.3.3 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 image understanding systems, this is often 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: the identification of corresponding points in the two stereo images and 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.

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.

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. 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.

2 The Duality of Perception and Action

If we have managed to build an unambiguous 3-D model of the viewed scene, is the process then complete? The answer to this question must be no. Computer vision systems, *per se*, are only part of a more embracing 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. Let us now proceed to the topic of autonomy to see what is its relationships with image understanding.

3 Autonomous Systems

The word autonomy conjures up images of self-reliance and independence, self-sufficiency and isolation. The difficulty is that such autonomy never exists. There is no such thing as a wholly-closed independent self-sufficient isolated autonomous entity. If there were, then we could, by definition, have no knowledge of it. Autonomy certainly involves these issues, but there is more to it than that. This *more* is the object of the autonomy; the thing, the entity, that is autonomous. For the idea of autonomy seems to be meaningful only when we look at the concrete basis on which it is founded. So we will speak of automous systems and not of autonomy, for thus it is explicitly clear that we are dealing with the concept of a system which displays the characteristics of autonomy: independence, self-government, self-definition. And in this way, the study of autonomy becomes more objective, for the reference point of autonomy is thus a system and not the descriptive domain of our discussions.

Systems, whether autonomous or not, are defined in a context, in a universe of discourse; a domain of instantiation, if you will. That is, a system exists as part of something. Autonomous systems too exist, not in complete isolation, but in some reference domain. On the one hand, when we speak of an autonomous system, we speak of a system which has an identity and is capable of maintaining that identity. But, on the other hand, we must ask of what is it autonomous. We must look at the relationship between the autonomous entity and its universe. At first sight, this may appear to be contradictory: autonomous systems are independent and self-determining while the idea of their having an explicit relationship with their local universe implies the opposite. The resolution of this apparent paradox is the key to discussing autonomous systems. You cannot discuss

the one without making some statement on the other. It is the mutual relevance of the two and the mutual specification of the two by which autonomous systems arise. We differentiate this form of system from non-autonomous systems where the system is an aspect of, a direct component of, its environment: controlled by it and used by it.

Apparently, autonomous systems do two things: they act and they perceive. They may do other things also, such as *think* but it is not clear whether thinking is a characteristic of autonomous systems or a description from the point of view of an observer of the action of the autonomous system. Autonomous systems may also be conscious but again this is a word which is, in the current context at least, ill-founded. For the present, we can be satisfied with observing that autonomous systems exist, they act and interact with their environment; they perceive — in the sense that the are capable of adapting to a dynamic environment — and they exhibit, through these activities, a behaviour which could on occasion be construed as intelligent. So, in discussing autonomous systems, and in developing a formalism for treating autonomy, we must inevitably be drawn into considering many secondary topics such as robotics, cybernetics, perception, intelligence, and, indeed, ontology. But, apart from ontology, these issues are secondary. It is the purpose of this paper[1] to argue that these considerations are in fact contingent upon the autonomy of the system and that they can be explained in terms of the autonomous activity of the system.

Autonomous systems derive their autonomy from their intrinsic self-organization. This self-organization implies a dynamic relationship between the autonomous system and the universe of which it is a part. In this dynamic relationship, the components which constitute the system are continually organized and re-organized, new components entering the sytem, and components leaving. What remains constant is the identity of the system in that an entity with a given self-organization endures. In a strong sense, the autonomous system distinguishes itself from the environment and maintains that distinction. Specifically, it maintains that distinction in the face of (despite) the independently fluid and changing nature of the universe of which it is a part and in self-distinction from the local universe. This formulation of autonomy owes a great deal to the work of Maturana and Varela (see, e.g., [6, 7, 8]) who introduced the concept of autopoietic, self-producing, systems.

Let us now interpret the terms perception and cognition in the context of what we have been discussing. Given that autonomous systems exhibit a critical level of (self-)organization, and given that this implies a dynamic relationship between the environment (or universe) and a dynamic relationship between the components which constitute the system, then we might search for a phrase which describes this autonmous activity. We could, perhaps, use the phrase dynamic self-organization and self-specification and self-distinction. But this would be a little clumsy. We more commonly use the terms perception and cognition to refer to what an autonomous systems does and I would like to argue that the concepts of perception and cognition are identical to the self-organizing, self-determining, self-specifying systemic activity which is a part of and is required of an autonomous system as it distinguishes itself from the environment. Perception is the term which we could use to emphasise the aspects which pertain to accommodating the perturbations on the system by the environment while cognition is a term which we could use to emphasise the aspects which are internal to the system and which could be interpreted as the 'making sense of' the world (*c.f.* creating representations). Perception then corresponds to the apprehension of the external 'reality' while cognition corresponds

[1] In fact, due to the limited amount of space available, I will not argue the point in depth here; rather, I will state the argument and leave its detailed discussion to another paper [5] which begins the development of autonomous systems with a treatment of ontological issues.

to the construction of sense of the universe. The study of autonomous systems is, in effect, the study of spontaneously self-organizing, self-determining, systems which exhibit a structural plasticity but an organizational constancy. By this means, autonomous systems manage to preserve a fixed — enduring — identity, despite the fluid and dynamic nature of the universe of which they are a part. The organizational principles which facilitate this dynamical relationship constitute the subject matter of autonomous systems.

4 The Problem of Representations in Autonomous Systems

Clearly, computer vision addresses many of the issues which are relevant to the design of artificial autonomous, adaptive and anticipatory, systems. Image understanding systems attempt to understand the physical structure of the local environment with the express purpose of allowing robotic systems to interact with that environment. The central problem in trying to design autonmous systems which are based on the type of representational vision systems described above is that the designer is acting as an implicit *homunculus*, an interpreter, at the perception/action interface: he or she decides on the representations which will be used and on the rules which will be invoked in response to certain perceptual (representational) stimuli. This is at variance with the the structural, organizational, and behavioural plasticity which is fundamental to autonomous systems and, I would argue, so strongly prejudices the structure of the system that the likelihood of developing a truly autonomous system is quite low. However, by addressing directly the organizational principles by which autonomous systems arise — rather than the representations which we as system designers believe to be appropriate to autonomous systems — the likelihood of creating truly autonomous systems is increased. The problem which then arises is how such autonomous systems can be imbued with a goal-oriented behaviour which reflects the requirements of its designer. This remains an open question. Our task now is to develop a formal science of autonomous systems and self-organization.

References

[1] Vernon, D. and Sandini, G. 1992. *Parallel Computer Vision — The* VIS $\vec{a}$ VIS *System.* , Ellis Horwood, London.

[2] Vernon, D. 1991. *Machine Vision*, Prentice-Hall International, London.

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

[4] Vernon, D. and Tistarelli, M. 1990. 'Using Camera Motion to Estimate Range for Robotic Parts Manipulation', IEEE Transaction on Robotics and Automation, Vol. 6, No. 5, pp. 509-521.

[5] Vernon, D. and Furlong, D. 1992. *Limitations of Scientific Ontology*, Technical Report, Department of Computer Science, Trinity College, Dublin, Ireland.

[6] Maturana, H. R. and Varela, F. 1980. *Autopoiesis and Cognition — The Realization of the Living*, D. Reidel Publishing Company, Dordrecht, Holland.

[7] Varela, F.J. 1979. *Principles of Biological Autonomy*, Elsevier North Holland, New York.

[8] Maturana, H. R. and Varela, F. 1988. *The Tree of Knowledge*, New Science Library, Boston.