Computers See the Light

Will robots ever pick fruit? Ours is a multi-dimensional world, full of clutter and moving objects yet, all too often, sight is taken for granted. Not so, however, in the area of computer vision, as *David Vernon* explains. For computers may now be able to 'see' certain images — but they have yet to understand them.

Computer vision systems, machines which can process visual information generated by television cameras, have been around for over 25 years. From the early research systems to the current industrial inspection systems, the architectures have changed little: a typical system can be simply configured using a standard TV camera sensor, an image acquisition board to perform the analogue-to-digital conversion, and a micro-computer. Dedicated hardware is freely available to perform much of the simpler processing at real-time rates. The processing software has, on the other hand, matured significantly.

We have now recovered from the disappointments of the late '70s and early '80s when the enthusiasm of the computer vision scientist (and sales staff) far exceeded the functional capabilities of the systems, and we find ourselves designing and implementing systems which are indeed truly useful. From computer-based enhancement of Landsat satellite images, through automated visual inspection of printed circuit boards, to vision systems for guiding robot arms, computer vision technology has finally become a tool worth exploiting.

But is it really as useful as we normally make it out to be? Just how robust and flexible are current computer vision systems? And are we falling into the same trap as the vision scientists and engineers of the '70s, believing our own propaganda? Not quite. Current vision techniques are based on a much sounder foundation than those of a decade ago, but it is still not clear that the functional capabilities are any more advanced. Indeed, this article will suggest that the road to the future, when we can truthfully claim to have developed robust vision, is not well signposted.

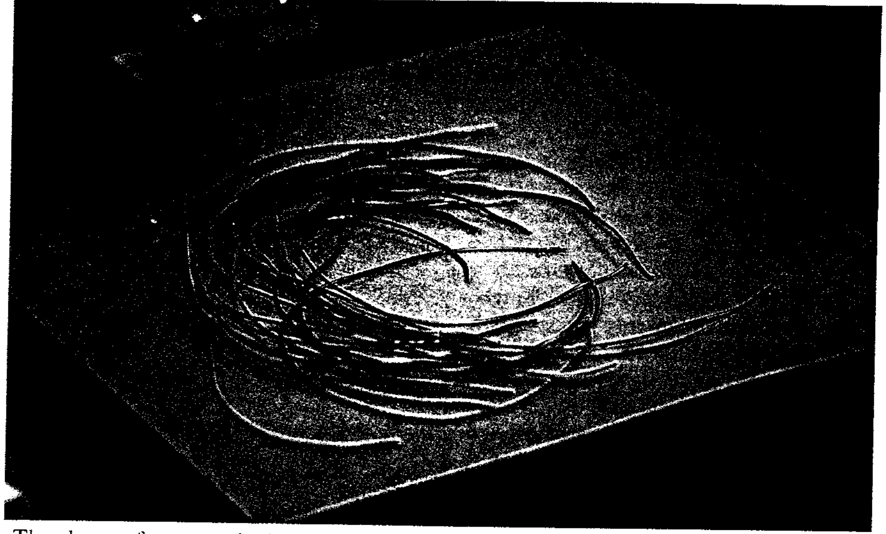
Let us begin with a brief overview of vision in the '80s. The term 'computer vision' is a very general one, covering a multitude of techniques, applications, and technologies. But there is a discernible distinction between three endeavours in computer vision: image processing, image analysis, and image understanding.

Image processing is an activity which is wholly concerned with the transformation of images to images. That is, image processing systems take images as their input and provide enhanced images as their output. The information which is processed is *iconic* in nature, as is the output; certainly the images mean something, but the meaning is, and remains, implicit. They require interpretation on the part of a human observer; indeed, the

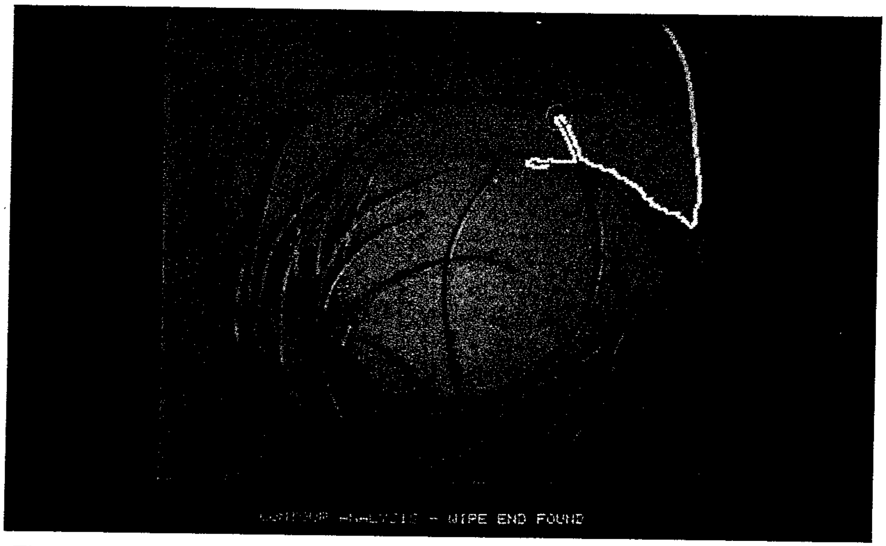
Image Analysis

Image analysis is concerned with extracting useful information about an imaged scene, information which, for instance, can be used to subsequently control an industrial process.

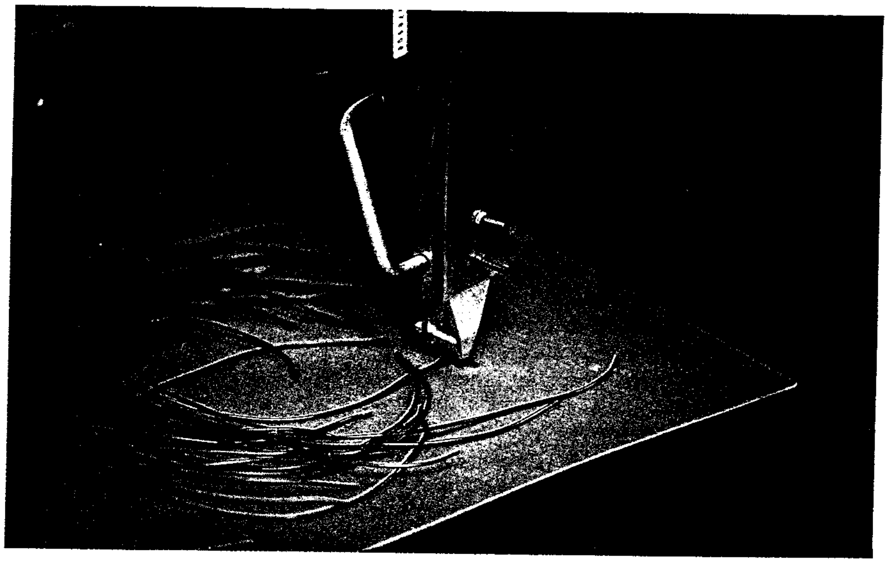
For example here, an image of a tray of wires is acquired:



The shape of extracted edge contours is analysed, and the end of a wire is characterised by a hair-pin bend.'



The position and orientation of the wire end can be used in controlling a robot to grasp and manipulate the wire.



objective of much image processing is to simplify this interpretation.

Image analysis contrasts sharply with image processing. Here, the effort is to perform some computation using image data and to produce useful explicit information about the image, perhaps to identify the position and orientation of objects being viewed by the TV camera and to estimate the size of each object. Thus, the transformation is from the iconic to the symbolic: information implicit in an image is made explicit and represented in some symbolic form which can be subsequently manipulated.

On completion of image analysis, one is in a position to make some decisions regarding the content of the scene viewed by the camera and use this information to control, for example, an industrial manufacturing process. What is significant is that the techniques that one employs in performing the analysis are tailored specifically to a particular application or visual environment. Furthermore, the success of the endeavour will depend heavily on the use of *a priori* knowledge: knowing what to expect and how to analyse it.

Many image analysis techniques address only two dimensional problems and perform relatively simple 2-D pattern recognition; applications which require analysis of three-dimensional objects are usually simplified (by appropriately engineering the environment) and reducing them to two-dimensional problems.

Unfortunately, the world is three dimensional (at least) and it is a cluttered and unstructured place. Objects are never where you expect them and, even if they are, they may be partially hidden. The average industrial or agricultural environment is also afflicted by extremely annoying objects which insist on moving: people. How, then, is a computer vision system to cope?

The search for a solution to this problem is the subject matter of a branch of computer vision called **image understanding,** a discipline which attempts to employ images of general 3-D scenes in automatically generating a symbolic description of the local environment. Image understanding includes image processing and analysis techniques in its armoury of computational techniques, but it also employs sophisticated algorithms for isolating partially hidden objects, for inferring 3-D structure, and for reasoning about 3-D structure.

Advanced image understanding systems have rudimentary knowledge of elementary, or naïve, physics (e.g. containment, support, and stability). Although complete image understanding systems are not yet commercially available, it is one of the most active areas of computer vision research at present and it is likely that some of the successful implementations will be reaching the marketplace in the near future.

That's all very well, you may say, but what exactly does this mean? Advanced computer vision can certainly take images of the world, transform them to other (nicer) images, analyse them and produce symbolic descriptions; it can even begin to do this for unstructured real 3-D environments. But do such systems really understand the symbols they produce? Who makes sense of them? To answer these questions it is necessary to look deeply at the word 'understand'.

To understand something is, fundamentally, to perceive its meaning: to apprehend all its significance, and to know how to deal with it. But meaning emerges from and changes with one's experiences; experiences occurring as one interacts with one's environment. Thus, to gain an understanding of something necessitates a continual ongoing interaction with it.

In this light, the term understanding is one which should have far-reaching implications when used in the context of computer vision: that the computer vision system should interact — perceive and (re)act — with its environment in a continual and circular manner. This, of course, is easier said than done.

If one allows an artificial system, such as a computer vision machine, to interact with the world, there are two possible outcomes: either the system will adapt to the changing nature of the environment, stabilise, and become in a sense self-preserving, or (as is much more likely) the interaction will be catastrophic and the vision system will unceremoniously crash. The latter is endemic among current computer vision systems, despite their claims to be capable of 'image understanding'.

What is important here is the realisation that the development of systems with faculties of understanding is deeply dependent on their ability to act and re-act; that continued understanding is achieved through perception and interaction with the world. It is axiomatic that such an ability requires autonomy on the part of the (artificial) system. It subsumes the properties of learning, self-organisation, and cognitive activity that are normally applied as labels or attributes to artificially-intelligent entities.

This is very heady stuff and unfortunately, due to our inepitude at constructing genuinely artificially intelligent systems at present, evidence to support these conjectures is a little thin on the ground. However, the arguments in the above paragraphs do merit serious considerations.

The problem of how to progress remains. Either we can all become experts in cognitive science (a branch of psychology concerned with quantitative and computational models of cognition) or we can re-apply our technology and attempt to build real autonomous systems.

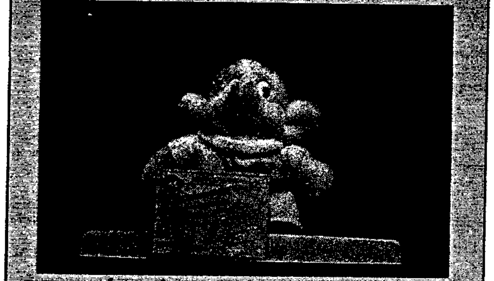
Ideally, we might adopt both options, but there is little doubt that an approach pulled by demanding applications, e.g. a mobile hospital library on an autonomously guided vehicle (AGV) or a general-purpose robotic fruit harvester, would be both invigorating and useful as long as the basic premises underpinning the application are embraced.

What technologies should we adopt? It seems pointless to discard the significant advances that have been made in computer vision in the last five years, especially in so-called *early vision* where techniques for edge detection, stereopsis, object and camera mo-

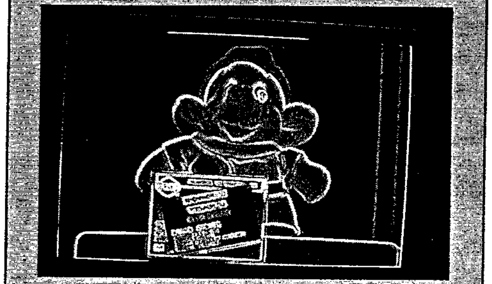
Image Processing

Image processing techniques transform images into images with accentuated features.

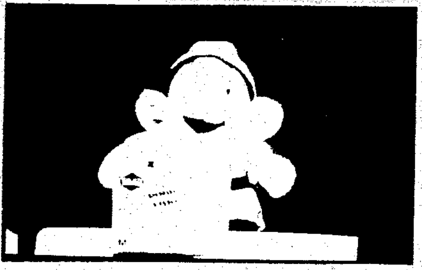
For example, a digital image of a soft toy is acquired,



the edges are enhanced (in this instance, using a simple Sobel operator):



Alternatively, the image might be thresholded, re-colouring bright regions as white and dark regions as black:



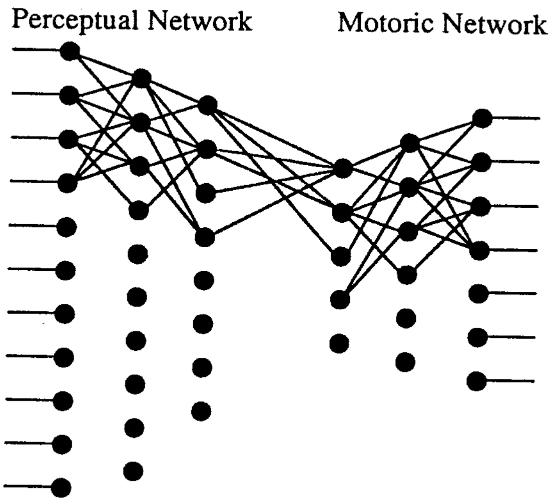
tion analysis (optical flow), texture analysis and (to an extent) shape analysis, are now well-understood and based on sound foundations. So these might be gainfully employed in constructing autonomous systems, remembering always that the loop between sensed and sensor (perceived and perceiver) must be closed, and that facilities for dynamically constructing and updating world models and representations must be incorporated.

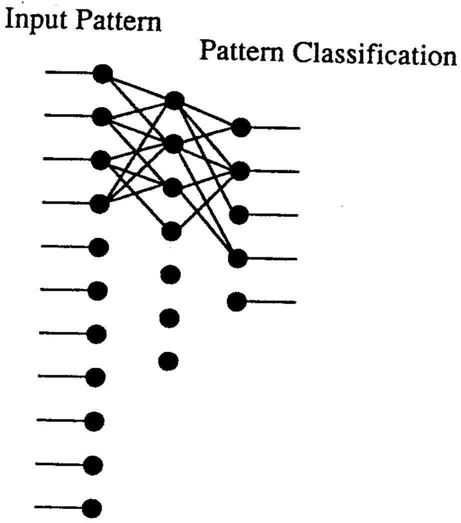
Alternatively, there has been a tremendous resurgence of interest in artificial neural networks in the last two or three years deriving from studies of the neural structures in the human brain. This neural modelling is concerned with building interconnected networks of 'neural processors' or (simplified) neurons where the interactions result in remarkable properties of self-organisation and learning. Unfortunately, many neural network systems and much work remains in investigating *why* these simple computational machines exhibit such complex (and useful) behaviour.

Whichever approach one adopts, artificial neural networks or conventional A.I., one needs always to be aware that a system which understands is interactive and participative, autonomous and stable, and that, if one neglects to address these issues, the machine will always fall short of being truly intelligent.

Such is the stuff of future research; for A.I., for vision, and, necessarily, for industry. The way is still not well signposted, but the direction is clear.

Dr David Vernon is director of the vision and Sensor Research Unit in the National AMT programme and a lecturer in computer science at Trinity College, Dublin.





Neural Networks

Artificial neural models utilise interconnected networks of 'neural processors' or (simplified) neurons to build systems which exhibit properties of self-organisation, learning, and pattern recognition.

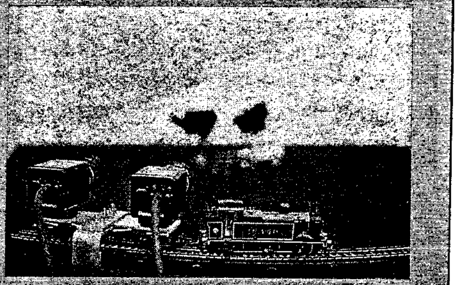
Conventional networks work in a feed-forward manner, taking patterns of activity on the input layer and producing a resultant output. Normally these outputs are mutually exclusive and, hence, the neural network can act as a pattern recogniser.

If neural networks for sensing and control are utilised together, it might be possible to construct an autonomous system.

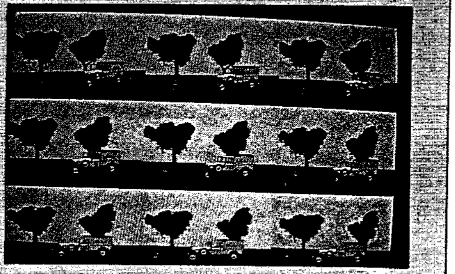
Image Understanding

Image understanding systems attempt to reconstruct the three-dimensional structure of real natural scenes and to develop reasoning facilities to allow, for example, autonomous navigation by un-manned vehicles.

In this example, a pair of cameras are mounted on a train track and nine images are acquired as the camera moves around the scene.



After extracting edge contours (top right), the apparent movement (or optical flow) between each image is computed (bottom left).



This movement, or parallax, is inversely proportional to the distance from the object from the camera (bottom right) — here distance is proportional to brightness.