NOTE: This paper was originally published (on pages 380—401) in
Physical and Biological Processing of Images, Editors: O. J. Braddick and A. C. Sleigh
Springer Verlag Berlin Heidelberg New York 1983
Proceedings of an International Symposium
Organised by The Rank Prize Funds
London, England, 27 - 29 September, 1982
Springer Series in Information Sciences
Editors: King-sun Fu, Thomas S. Huang, Manfred R. Schroeder
This was an invited contribution.

(I had completely forgotten I had written this until I found a reference to it in October 2006. I therefore decided to make it available online. This is the result of scanning+OCR and still incomplete post-editing. A sequel to this paper is: **A. Sloman. On designing a visual system (Towards a Gibsonian computational model of vision). Journal of Experimental and Theoretical AI, 1(4):289--337, 1989,** available online at http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7,
recently reformatted with figures reinstated. Some further developments, emphasising vision as perception of processes, can be found in http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk35
A presentation on evolution of vision and language is at
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk111

Image Interpretation: The Way Ahead?

Aaron Sloman
Cognitive Studies Programme,
University of Sussex,
Brighton, United Kingdom
(Now at the University of Birmingham:
http://www.cs.bham.ac.uk/~axs/)

Abstract
Some unsolved problems about vision are discussed in relation to the goal of understanding the space of possible mechanisms with the power of human vision. The following issues are addressed: What are the functions of vision? What needs to be represented? How should it be represented? What is a good global architecture for a human-like visual system? How should the visual sub-system relate to the rest of an intelligent system? It is argued that there is much we do not understand about the representation of visible structures, the functions of a visual system and its relation to the rest of the human mind. Some tentative positive suggestions are made, but more questions are posed than answers.

A.1. Introduction

The human visual system is the most powerful information-processing system known. It very rapidly processes a continuous stream of information, from millions of receptors. It copes with an enormous variety of formats, and many kinds and degrees of image degradation. It improves itself over time, and it can be used for purposes as varied as sight-reading music, diagnosing diseases, peeling a banana and enjoying a ballet. Explaining how it all works is a mammoth task, and no sane person should claim to be able to see the way ahead. But having been asked, I shall try to peer into the mists.

Assuming that a criterion for understanding is the ability to design a working model, there are many things we still don't understand, despite the progress of the last few years, mostly concerned with 'low Level' processing. (See HANSEN and RISEMAN [12], BRADY [6], MARR [25], recent Artificial Intelligence conference proceedings, e.g. IJCAI 1981, and this volume. A lot of this research was originally inspired by the work of HORN, e.g. [22], who showed that more information about scenes could be extracted from images in data-driven fashion than was previously thought possible. Some of the work is very technical, and I cannot claim to have understood it all.)

I take it that our aim is to understand not just human vision but general principles relevant to the design of visual systems, especially those which approximate the power of the human visual system. Science is not just the study of what is, but also of what is and is not possible and why. (See my [30] chapters 1 and 2.) This aim may not be important for special-purpose applications of image processing, but in the long term, the design of flexible and robust robots will require a deep understanding of general principles, including the geometry and physics of scenes and image formation, and also computational principles concerned with how best to represent information, how to cope with ambiguity and degraded information, how to combine multiple sources of information for a single task, how to maximise speed or trade space and time, how to improve the system's abilities over time, etc.
In the short run, for commercial reasons, most research and development funding is likely to be directed to special-purpose systems which are heavily model-driven, and cannot cope with arbitrary scenes, unlike human beings and many other animals, which are not restricted to seeing known sorts of objects.

There is a great discrepancy between the kinds of tasks that can be performed by existing computer models and the experienced richness and multiple uses of human vision. This is not merely a quantitative difference which might easily be overcome by the use of better hardware. There are too many limitations in our theoretical understanding for technological advances to make much immediate difference. Given computers many times faster and bigger than now, and much better TV cameras, we still would not know how to design the visual system for a robot which could bath the baby or clear away the dinner things, let alone enjoy a ballet.

A.2. The main problems

The literature cited includes descriptions of both achievements and detailed unsolved problems, especially problems of interpreting local image features in terms of scene features or fragments. This is an area in which much more progress has been made, and will be made, than used to be thought possible. However, there are some equally important problems not receiving so much attention:

(1) What are the functions of a visual system?

(2) What needs to be represented, and how?

(3) What sort of global architecture can enable a system to perform those functions?

In an attempt to elucidate the nature of these problems I shall show that common assumptions about the functions of vision are too restrictive, and that representations used in current models are not adequate, even for such apparently simple things as straight Lines. I shall offer some speculations about the architecture required by a human-like visual system and the way in which it should relate to the rest of the information processing mechanism. In particular, it should not be restricted to producing descriptions of the geometry and motion of the environment and it should have several subsystems which are themselves linked not only to each other but also to other mental mechanisms.

A.3. Methodological note

Being a philosopher and a programmer, my approach to an attempt to understand the space of possible computational systems is primarily top down: attempting to derive structure from functions and constraints. Analysis of function involves studying the purposes for which vision can be used and the circumstances in which it can achieve those purposes. Such things as required reaction times, error rates, type of degradation in various circumstances (e.g. occluded objects, poor lighting, mist, blizzards, loss of one eye, rapid motion, etc.) can be thought of as more detailed specifications of function. (Strangely, few text books on vision seem to discuss the functions of vision -- e.g. the relation between vision and action. HOCHBERG [20] mentions action on the last page!)

Although it is *very* important to understand human vision we can fruitfully aim for a higher level of generality, based only on assumptions common to natural and artificial systems: assumptions concerning the nature of the environment, the nature of the tasks of a visual sub-system of an intelligent system, and the need for rapid decisions relative to the available processing speed. We need to understand the space of possible visual systems and how they relate to the functions.

But for the crucial speed constraint, a visual system might systematically generate all possible 3-D scenes from all possible viewpoints, project them onto a representation of a retina, and compare with actual retinal stimulation. (Moving images would require representation of changing scenes.) Subject to a suitable finite quantisation of the search space and 'tolerant' matching, the selection of the best match could be done after an exhaustive search. The fact that problems and theories are formulated in such a way as (rightly) to rule out consideration of such absurd strategies shows that separating the theory of the domain from the design of algorithms and mechanisms may not always be as easy, or as useful, as MARR suggests (e.g. [25]). See also section C.8 below.

B.1. The functions of vision

A full survey of the functions of vision is not possible here, but it obviously provides information about the environment, which can be used for searching for things, controlling and monitoring actions, finding one's way, forming plans, reacting to opportunities and dangers, making predictions, understanding mechanisms, testing theories, interpreting communications, making pictures, building databases for future use, and even improving visual abilities. Vision can also trigger desires (e.g. sexual), emotions or reflex actions.

How can all this be done? It is often assumed that animals, and robots, need a visual subsystem whose function is to take in retinal stimulation and produce descriptions of (possibly changing) three-dimensional objects, their properties and relationships, thus:

```
      r     ____________     ________     __________
   -> e -> |Visual      |--> |3 or 4-D|--> |Non-visual|
   -> t -> |mechanisms  |--> |descrip |--> | sub      |
   -> i -> |            |--> | -tions |--> |systems   |
   -> n -> |____________|--> |________|--> |__________|
      a
```

I shall challenge the assumption that such a limited interface is compatible with the common uses of vision. I shall try to indicate the need for visual mechanisms to communicate more than just the geometry and motion of 3-D scenes.

B.2. Is there only one kind of output?

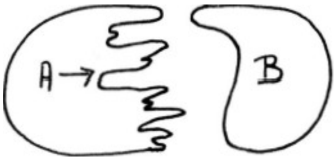
Information from lower levels is often useful, for instance when 2-D information, not 3-D information, is needed. Putting a finger on the edge of a table (usually) requires information about the three-dimensional location of the edge. However, being able to run one's eye along the edge (e.g. Looking for defects in the workmanship) requires only a two-dimensional representation of the location of the edge within the visual field. You can run your eye along a horizontal wire looking for kinks even if you cannot judge the distance of the wire. Similarly, moving a paint brush smoothly along an edge may require only the 2-D representation for monitoring the motion (once it has started) and providing adjustments if the hand is going too far up or down, in situations where the depth information comes from a non-visual source, such as touch, or is not needed because the geometry of the arm constrains the motion. Using disparity in a 2-D image to control movement *may* be far easier than using the 3-D disparity, for instance where there is insufficient time, or depth information, to compute the 3-D disparity. 2-D disparity will be specially useful when the only corrections to position of the arm are essentially 2-dimensional, e.g. merely raising or lowering.

2-D image structure is useful for deciding which way to move to see more of a partially hidden object. In some cases Local information near an occluding edge (especially when there's motion) suffices to determine the best direction, whereas in others more global computation is required. Compare the problems of how to move to see more of A or more of B in the figures below. Absolute depth information is not needed.

```
  __________                --------                    ---------
 |          |              |      \_______________/         |
 |          |______        |       \             /          |
 |          |      |       |        \     B     /           |
 |          |  A   |       |         \_________/            |
 |          |______|       |          \       /             |
 |__________|              ------------------    ------------------
```

2-D image structure could also be used by higher levels in: planning a route across a room, finding space for a new package on a cluttered floor, anticipating a trajectory, working out what someone is looking at, keeping to the centre of a path, sighting with a gun, moving out of sight of an enemy. A 2-D image map may help to constrain searching, both by providing rapid access to information about nearness, collinearity, inclusion etc., and by constraining proposals for solutions. How exactly such maps can be used remains a research problem. FUNT [11] deals with some simple cases. (See ch. 7 of [30] for a discussion of pros and cons of 'analogical' and 'applicative' representations.)

Some of these tasks merely extract information from the image representation, whereas others require an ability to manipulate the representation, as in FUNT's program which computes collision points of arbitrarily shaped objects by sliding and rotating images. This could use the mechanism MARR ([24],[25]) postulates in claiming that 'place tokens' representing structures inferred by higher level processes can be added to the primal sketch). An empirical question is what sorts of manipulation people can do. In the following figure it is much easier to see where A will first touch B on approaching it than to see where A' will first touch B':



In a computer it is simple to define array manipulations which will find both tasks equally easy. Is the human brain a poor array processor? Or is there some important generally useful design feature of the human visual system which is responsible for its limitations in this case? For many tasks array manipulations are not general enough, e.g. detecting the kind of approximate symmetry exhibited in many living organisms.

Different tasks require access to different layers of interpretation, **i.e.** the results of different sub-modules of the visual system. For some of the tasks 3-D structure could be used. For others, e.g. drawing a realistic picture, it is essential to use image relationships. Much more detailed information about the original image is needed for a realistic oil-painting than for a sketchy line-drawing. The latter and the guidance examples could use relatively high level intermediate results of visual processing: two-dimensional yet quite abstract.

B.3. Not the retinal image.

The 2-D structure accessed from outside the visual system should not necessarily be a representation of the retinal image. That will constantly change as different parts of a scene are fixated while a problem is being solved. Something less transient is needed, such as a representation of the available 'optic array' at a viewpoint (DRAPER [9]), possibly built up over a period of time from different samples. Or it might be a map of what BARROW and TENENBAUM [3] call 'intrinsic scene features' - closely related to what MARR called the two and a half D sketch. Several such representations are needed, dealing with different levels of abstraction, all with array-like qualities and some sort of registration to enable information at different levels to be combined easily when appropriate (compare BALLARD [1]).

The fact that (especially with the aid of relative motion), we can easily perceive two scenes superimposed in a television screen, or one seen through a window and another reflected in it, suggests that different retinal flow elements may be separated into different arrays, and then two (or more?) sets of scene interpretations built on them in parallel. This may be related to the ability to see through mist or muddy water. Stereoscopic vision reverses the process, merging two arrays. (Recent experiments suggest that locations of different sorts of features even of the same object might be stored in different maps in human vision, requiring some time for inter-map relations to be computed TREISMAN [34]).

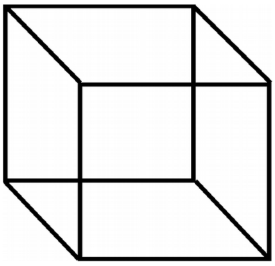
B.4. Is only geometry represented?

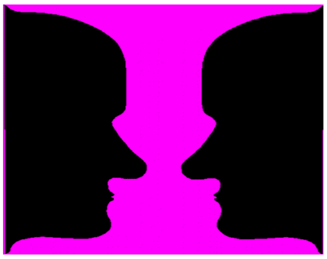
Vision does not produce only geometric information. Consider a view of a busy workshop. At any moment there will be complex changing retinal images, representing 3-D structures composed of walls, floor, furniture, machines, tools, materials with changing shapes, and human bodies in a variety of changing postures. Among the relationships we can see are many which are not purely geometrical. For instance, seeing a trestle as supporting a table involves seeing it as applying an upward force and as preventing downward motion. This interpretation can play an important role in assessing the consequences of moving the trestle, or predicting the effects of placing a very large object on the table. We can see many other relationships which are not purely geometrical, though they involve geometrical components (a fact which is relevant to their being seen as opposed to merely inferred or believed). Examples include: holding, pushing, pulling, cutting, turning, moulding, using, controlling, approaching, avoiding, looking at, catching, and so on. MARR [25] argues that this is not visual perception because the ability can be selectively impaired by brain damage. I'll try to show that he may have missed something.

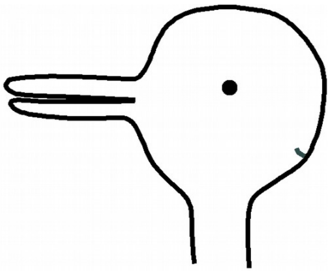
This is not merely a semantic quibble. The claim is that the kinds of representations which are used for geometric structures can also be used for non-geometric information which maps usefully onto geometric structure, and that the same processes can operate on them. In diagrams, we often find it useful, or even essential, to combine representations of non-geometrical properties or relationships with the representation of geometrical structure. Fields of force, lines of possible movement, causal relationships, can all be usefully represented this way, for problems in which spatial relationships are important. For similar reasons an intelligent system can benefit from integrating geometrical and non-geometrical information in combined representations.

There are many different detailed ways this can be done. For instance, in a computer it could be done by storing a variety of different sorts of information in 2-D arrays in registration with an array storing information about the structure of the current field of view. Alternatively a single array could have lists of pointers to different types of descriptions. Switching attention between classes of tasks or subtasks would be facilitated by having different 'maps' (array-like structures) storing different sorts of information, but with a mechanism for rapid access of corresponding locations in different maps, so that relations like contiguity and collinearity may be detected readily even among entries in different stores.

So non-geometrical concepts may generate geometrical problems, and it would be useful to design a visual system to include interpretations in terms of such non-geometrical concepts in order to facilitate practical problem solving.







B.5. Seeing mind in matter

Even mental states of other agents may usefully be represented in registration with spatial information. An animal may find it as important to see what another animal is looking at or aiming to pick up, as to see whether it is big, or coming closer. The computation of relations between another's field of view and objects he might react to can benefit from an integrated 'analogical' representation. Problems about the intersection of objects, or trajectories, with another's view field may make use of similar representations to those used for detecting overlap of two physical objects, or for controlling the motion of the hand, mentioned above.

When a Necker cube drawing 'flips', you see different sets of geometrical relations. But many visual ambiguities go beyond geometry. The vase/face figure includes both geometrical ambiguities concerning relative depth and also abstract ambiguities of grouping: what 'goes with' what. But when you look at the duck-rabbit picture, the flips feel just as visual, though the change is not geometrical. Instead you see different functional components (eyes, ears, etc.) and even more importantly the direction the animal is facing changes. Abstract functional descriptions like "front" and "back" are involved. The duck or rabbit is seen as Looking in one direction or another. Here "front", "back", "looking" are not just arbitrary labels, but imply links to elaborate systems of concepts relating to the representation of something as an agent, capable of moving, of taking in information, of making decisions, etc. For instance "front" indicates potential forms of motion and also the direction from which information is obtained about the environment. The attribution of 3-D slope and relative depth to parts of the Necker cube involves going beyond what is given. So does interpretation of natural images in terms of surface orientation or curvature, convexity of edges, occlusion, etc. Given the need to be able to interpret 2-D structures in terms of these quite different structures, why stop at mapping onto geometrical structures? We shall see that it can be useful for a visual system to map arbitrary image features onto arbitrary structures and processes, including direct triggering of actions.

B.6. Not only descriptions

Vision need not be restricted to the production of descriptions, whether of geometrical or non-geometrical, structures and processes. It may, for example, involve the direct invocation of action. In situations of potential danger or opportunities it would be useful for a disturbance in a peripheral part of the visual field to trigger an eye-movement to re-direct attention. A tendency to react quickly to a global pattern of optical flow produced by a large object moving rapidly towards the viewer would be very useful. A fencer or boxer who does not have time for processes of planning and deciding can also benefit from direct coupling between scene fragment detectors and the invocation of actions. Where speed of response is crucial, visual feature detectors should be able directly to trigger stored action routines without the mediation of a decision maker examining a description of what has been detected. (This seems to be the only way some animal visual systems work). The triggering could make use of general-purpose associative mechanisms which seem to be needed for other purposes. Visual learning would then include the creation of new detectors TREISMAN and GELADE [34] refer to a process of 'unitisation') and the creation of new links between such detectors and other parts of the system. Some of the links might provide new feed-back loops for fine control of actions. Some of the reflex responses may turn out to be misguided because the central decision making system is not given time to take context into account.

This use of special-purpose processors to provide 'direct coupling' could also provide the basis for many more abstract visual skills, such as fluent reading (including sight-reading music). It may also be very relevant to the fact that visual experiences can be rich in aesthetic or emotional content. It is argued in SLOMAN and CROUCHER [33] that in an intelligent system in a complex and partly unpredictable world, it is necessary for perceptual processes to be able to activate dormant motives, motive generators, and thereby generate processes which may be emotional. How to integrate the interrupt function of vision with its use in more relaxed planning and monitoring of actions is an important research issue.

B.7. What makes it VISION?

To sum up so far, instead of an output interface for only one type of result of visual processing, there is a need for different sorts of communication between visual sub-processes and other sub-systems. Sometimes descriptions of three or four¯dimensional structures may be useful, sometimes only two-dimensional And sometimes the descriptions needed are not purely geometrical, but include extra layers of interpretation, involving notions such as force, causation, prevention, function, or even the mental states of agents.

The suggestion that vision involves much more than the production of descriptions of three-dimensional structures, at least in the higher animals, conforms with the common-sense view that we can <u>see</u> a person looking happy, sad, puzzled, etc. Seeing into the mind of another, seeing an object as supporting another, seeing a box as a three¯dimensional structure with invisible far sides, may all therefore use the same powerful representations and inference mechanism as seeing a line as straight or curved, seeing one shape as containing another, etc.

Is there then no difference between vision and other forms of cognition, for instance reasoning about what is seen? A tentative answer is that the difference has to do with whether the representations constructed are closely related to 'analogical' representations of a field of view. This is a different boundary from the one postulated by MARR, between processes which are essentially data-driven and use only very general information about the physics and geometry of the image-forming processes, and processes which may be guided by prior knowledge, and which make inferences about non-geometric properties and relations.

Research issues arising out of this discussion include the following. What range of tasks can vision be used for? Is there a useful taxonomy of such functions? What range of pre-3-D structures is useful and how are they useful? What sorts of non-geometrical information can usefully be extracted from images and embedded in visual representations? What sorts of operations on these representations can play a useful role in reasoning, planning, monitoring actions, etc.?

B.8. Problems for a neural net implementation

How should the addressing be done in a neural net? It is relatively easy to use arrays on a conventional computer, with locations represented numerically and neighbourhood relations represented by simply incrementing or decrementing co-ordinates. This enables any subroutine to 'crawl' along the array examining its components, using the number series as an analogical representation of a line, following Descartes. On a multi-processor (e.g. neural net) representation, where location is represented by location in the network, problems of access may be totally different. Will each module accessing the array have to have physical links to all elements of the array? If so how will it represent those links and their properties and relations? Moreover, there is a risk of a 'combinatorial explosion' of connections. Would it be useful to have a single 'manager' process through which all others communicate with the array, using symbolic addresses? I suspect that even these questions are probably based on too limited a view of possible forms of computation. There is a need for further investigation of possible models [15].

Assuming that a temporary memory is needed in which information about the optic array can be accumulated, how should a neural net cope with changing scenes? E.g. if you swivel slowly round to the right does the stored structure really 'scroll' gradually to the left as new information is added at the right edge? This would happen automatically if the whole thing were constantly being re-computed from retinal stimulation -- but we need something more stable, built up from many fixations. Scrolling an array-like representation in a computer can be done easily without massive copying, merely by altering a table of pointers to the columns, provided that all access goes via symbolic addresses. When the representation is embodied in a network of active processors, simultaneous transmission across network links could achieve a similar effect, but the problems of how other sub-systems continue to be able to access information are considerable. Do neurones use symbolic addresses: using these rather than physical links might solve the problem. Some recent work explores the possibility that alternative mappings from one structure to another might be represented explicitly by different active 'units' which would mutually inhibit one another. (See HINTON [17,18,19], BALLARD [1]). This seems to require an enormous explosion of physical connections, to accommodate all possible translations, rotations, etc.

C.1. Problems of representation: what and how?

Even if there were some perfect mechanism which analysed retinal stimulation and constructed a detailed representation of all visible surface fragments, their orientation, curvature, texture, colour, depth, etc., this would not solve the problems of vision. This unarticulated database would itself have to be analysed and interpreted before it could be used for the main purposes of vision. And right now we don't know very much about what such an interpretation process would be like. In particular, what structures should be represented for different purposes, and how should they be represented? HINTON [16] demonstrates that different representations of so simple a structure as a cube can profoundly influence the tasks that can be performed. We know little about the ways of representing (for instance) a face to facilitate recognition from different angles or with different facial expressions, or to facilitate interpretation of subtle mood changes, or lip-reading.

I am not talking about the problems of detecting structures, (in the world or in images) but about how to represent them in a useful way.

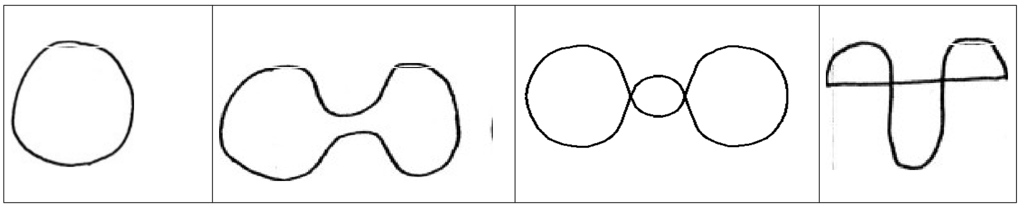
Even if we consider only two-dimensional geometric structures, such as the visible outlines of objects and the visible textures and markings on surfaces, we find a richness and variety that defeats existing representational schemes. Representations must not merely be mathematically adequate: they must also be epistemologically and heuristically adequate - i.e. including all the information required for a variety of tasks and facilitating computation in a reasonable time [23].

A representational scheme has two main components, the representation of primitives and the representation of composition. For instance, in many systems for representing images, primitives are local quantitative measures (e.g. of intensity gradients, or optical flow), and composition is merely the embedding of such measures in a two-dimensional array. Often a hierarchical mode of composition is more useful, e.g. using a relation "part of" to define a tree structure as used by linguists and extended by MINSKY [27a] to vision (elaborated, for example, by MARR and NISHIHARA [26]). However, a hierarchical tree structure is often not general enough to capture perceivable relationships in a useful way. For instance, the choice of a top-level node may be arbitrary, and computing relations between the 'tips' of the tree (e.g.. computing relations between a person's finger and the tip of his nose from a hierarchical body representation) may be difficult and time consuming. Yet we often see such relationships apparently effortlessly (hence letters to the press complaining about drivers who pick their noses whilst waiting at traffic lights). Moreover, many objects do not have a tree-like topology, for instance a wire-frame cube. So, instead, a network is often used, with links representing a variety of relationships, e.g. "above", "touches", "supports", "same-size", "three feet away from", etc. (See WINSTON [39], MINSKY [27]). This allows both global relations between major components, and useful information about arbitrary sub-components to be represented in a quickly accessible form. The network may still be hierarchical in the sense that its nodes may themselves be networks, possibly with cross-links to other sub-nets. One problem with such networks is their sensitivity to change. Relatively simple movements may drastically change the structure of the net. Sometimes parts of an object are not represented in terms of their mutual relationships, but in terms of their relationship to a frame of reference selected for the whole object. This can facilitate recognition of rigid objects using a generalised 'Hough transform' and cooperative networks of processors (BALLARD [1], HINTON [19]). It reduces the problem of representing changed states, since each part merely changes its relation to the frame of reference. However, relational networks seem to be more suited to non-rigid objects which preserve their topology rather than their metrical properties, like a snake or a sweater. (The Hough transform uses what BARLOW [2] calls 'non- topographical' representations, i.e. mapping objects into abstract spaces other than physical space.)

C.2. On perceiving a line

Despite the richness and complexity of such representational schemes, many percepts do not seem to be captured adequately by them. For instance, a circle might be represented approximately as made of a number of straight or curved line segments, or by parameters for an equation. But neither representation does justice to the richness of the structure we perceive, which has the visible potential for decomposition in indefinitely many ways, into semi-circles or smaller arcs, or myriad points, etc. The decomposition perceived may change as the surrounding figure changes or as the task changes. Yet, there is also a unchanging percept: we see a persistent continuous structure. (Phenomenal continuity is an interesting topic requiring further analysis. We certainly do not see what quantum physicists tell us is really there.)

This perception of continuity through change can also occur when an object changes its shape. If a circle is gradually deformed, by adding dents and bumps, the mathematical representation in terms of its equation suddenly becomes grossly inadequate, but we can see a continuous change. We can see the identity of a continuously changing object. A 'chain' encoding in terms of length and orientation of many small segments may be less sensitive to change, but will not capture much of the global structure that is perceived. It also fails to capture the perception of the space surrounding the line which is also seen as continuous. For instance it makes it hard to discover the closeness of two diametrically opposite points after the circle has been squashed to a dumbbell shape.



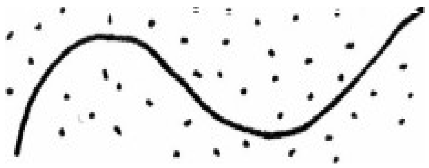
The line may gradually be deformed into a very irregular shape with many changes of curvature, sharp corners, lines crossing, etc. The algebraically representable mathematical properties will change drastically and discontinuously, and so will any network representing the more obvious decomposition based on discontinuities and inflection points. Yet we easily see the continuing identity. We don't have to switch into totally different modes of perception as the figure changes (though there may be sudden recognition of a new shape or relationship, emerging from the basic perception of the layout of the line in space). We need some account both of the relatively unchanging perception of the line in a surface as well as the awareness of higher level patterns and relationships which come and go.

C.3. The special case trap

Many computer programs make use of representations of straight lines, circles, and other mathematically tractable shapes. One obvious moral is that even when we can represent certain special cases in a computer model, we may have totally failed to explain how they are actually represented in human or animal vision. Living systems may deal with the simple cases using resources which are sufficiently powerful to cope with far more complex cases. Seeing a straight line is probably just a special case of seeing an arbitrary curve. How to represent arbitrary curves in a general purpose visual system remains an unsolved problem.

Similarly, theorems about perception of smooth surfaces may tell us little about a system which can see a porcupine and treats smooth surfaces as a special case. Many existing programs represent static configurations of blocks, but cannot cope with moving scenes. Perhaps a human-like system needs to treat such static scenes as special cases of scenes with arbitrary patterns of motion? This would imply that much work so far is of limited value, insofar as it employs representations which could not cope with motion. Of course, this does not prove that the simple models are totally irrelevant: continuous non-rigid motion may perhaps be seen in terms of locally rigid motion, which in turn may be represented in terms of a succession of static structures, just as arbitrary curves may be approximated by local straight lines. However, it needs to be shown that such a representation is useful for tasks like unfolding a garment in order to put it on, or trying to catch a rabbit which has escaped from its hutch. It may be possible for a visual system to use many special modules which deal only with special cases, when they are applicable. If so, our account of the global organisation of a visual system needs to allow for this.

C.4. Criteria for a satisfactory representation



How can we tell when we have found a satisfactory 'general purpose' representation for lines in 2-D or 3-D space? Any answer will ultimately have to be justified in relation to the tasks for which it is used. But our own experience provides an initial guide. The representation should not change in a totally discontinuous fashion as the shape of the line changes, for we sometimes need to see the continuity through change. We also need to be able to examine the neighbourhood of the line, for instance looking for blemishes near the edge of a table, so we need to be able to represent the locations in the space surrounding the line as well as the locations on the line: the 'emptiness' of other locations may be significant for many tasks. The representation should allow arbitrary locations on the line to become the focus of attention, e.g. watching an ant crawling along the line.

The representation should not change totally discontinuously if the line gradually thickens, to become some sort of elongated blob, with an interior and boundary. The representation should allow the potential for many different ways of articulating an arbitrary curve, and the spaces it bounds, depending on current tasks. It should be usable for representing forms of motion. Potential for change should be representable even when there is no actual change -- for instance seeing the possibility of rotation of a lever about a pivot. Perception of the possibility is not just an abstract inference: we can see which parts will move in which directions. [NOTE 1].

C.5. An 'active' multi-processor representation?

The discussion so far suggests that it would be useful to represent arbitrary structures both by projecting them onto 'analogical' representations in registration with a representation of the optic array, and also by more abstract symbolic networks of relationships, some of them object-centred, some scene-centred. There is also a need for a large number of independent processors all simultaneously attempting to analyse these structures in a variety of different ways and offering their results to other sub-systems. As a perceived structure changes, so will the pattern of activity of all the processors accessing the arrays. At lower levels the changes will be approximately as continuous as the geometrical changes. At higher levels, new processes may suddenly become active, or die away, as shapes and relationships are recognised or disappear. Hence the impression of both continuity and discontinuity through change.

Here we have a representation not in terms of some static database or network of descriptions, but in terms of a pattern of processing in a large number of different sorts of processors, including for instance some reporting on the "emptiness" around a perceived line. It is not going to be easy to build working models.

C.6. The horrors of the real world

For limited purposes, such as recognition or guiding a simple automatic assembly system, some simple cases, e.g. rigid, plane-sided objects, with simple forms of motion (e.g. rectilinear or circular) can be represented using conventional mathematical techniques.

But when it comes to the peeling of a banana, the movements of a dancer, the ever changing visible patterns on the surface of a swiftly flowing river, we have to look for new ideas. Even for static scenes, conventional AI representations tend to make explicit only singularities in space -- edges, vertices, surfaces of objects -- and not the visible continuum (or apparent continuum) of locations in which they are embedded, a problem already mentioned in connection with the simpler 2-D case.

The use of integrative array-like analogical representations described above may not be feasible for three and four dimensional structures, owing to prohibitive storage and connectivity requirements. (This is not obvious: it depends both on brain capacity and on what needs to be represented.) Perhaps the desired integration, for some purposes, can be achieved by projecting three and four dimensional representations into 'place tokens' [24] in one or more changing two-dimensional analogical representations in registration with the optic array. Some arrays might represent surface structure more closely than retinal structure, for instance, if a square table top seen from the side is represented by a square array, not a trapezium. Array representations would help to solve some problems about detecting relationships between arbitrary components of a scene but would still Leave the necessity for articulation, and a description in terms of recognised objects their properties and relationships, in a manner which is independent of the current viewpoint.

C.7. What sort of "construction-kit" would help?

We have suggested a two-tier representation of spatial structures, using both projection into a family of array-like structures and networks representing structural and functional relationships between parts. There seems to be a very large "vocabulary" of recognisable visual forms which can be combined in many ways. The attempt to reduce them all to a very small set of 'primitives' such as generalised cylinders MARR [26] does not do justice to the variety of structures we can see. We seem to need a vocabulary of many sorts of scene-fragments including: surface patches -- concave and convex, corners of various sorts, surface edges, lamina edges, tangent edges (relative to a viewpoint), furrows, dents, bumps, ridges, rods, cones, spheroids, lamina's, strings, holes, tubes, rims, gaps between objects, etc. Besides such shape fragments, there may be a large vocabulary of process fragments - folding, twisting, moving together, coming apart, entering,flowing, splashing, etc. [NOTE 1]. Compare the "Naive Physics Project" of HAYES [13]. Larger structures might then be represented in terms of a network of relationships between these "primitives". (Christopher Longuet-Higgins, in a discussion, suggested the analogy of a construction kit.) Do we also need primitives for entities with fuzzy boundaries and indefinite shapes like wisps of smoke or a bushy head of hair?

Primitives are not enough: we also need to represent their composition into larger wholes, and once again there are gaps in existing techniques. How is the relation between a winding furrow in a field and the field itself represented? Conventional network representations allow a relatively small number of 'attachment' points to be represented. But we see the furrow embedded along its whole length. Is this adequately catered for by combining a conventional network-like description with a projection back into a shared array-like representation? Even for a block structure Like an arch, the conventional network representation (e.g. [39]) of one block as "above" or "supported by" another does not adequately represent the perceived line of contact, which can be seen as a continuum of possibilities for inserting a wedge to separate the blocks. How is a happy expression represented as embedded in a face?

At a lower level the primitives themselves would need a representation which permitted them to be recognised and their relationships characterised in detail. If this were based in part on analogical representations both of the image forms and of the 3-D structures, then this might provide a means for linking percepts together by embedding them in a larger map-like representation, e.g. Linking 3-D edge descriptions to 2-D image features detected by edge-image recognisers. Could some generalisation of this help to explain the perception of a composite object as a continuous whole, unlike the relational network representation? (I am not saying that the representation is continuous, like a drawn map, only that it may be sufficiently dense to represent continuity, especially if the resolution can be changed as required.) At a very general level both forms of representation are equivalent to collections of propositions: the network is equivalent to propositions about object parts and their relationships, the map is equivalent to propositions about locations and their occupants. But they have different heuristic power.

C.8. Why the system will be messy

The use of a large number of different visual primitives for describing scenes is from a mathematical point of view redundant: for instance the geometrical structure of any scene made of objects with nonfuzzy boundaries can be represented with as much precision as required in terms of suitably small plane surface fragments. But that representation will not be useful for many tasks, such as recognition of a non-rigid object. The use of a large number of mathematically inessential primitives is analogous to the use of a BECKER's 'phrasal Lexicon' [4], instead of a non-redundant grammar in a language understanding system. It is also analogous to a mathematician's learning many lemmas and rules which are redundant in the sense that they can all be derived from a more basic set of axioms. The use of the redundant system can constrain searching in such a way as to save considerable amounts of time, since searching for the right derivation from a non-redundant set of general axioms can founder on the combinatorial explosion of possible inference steps. Imagine trying to find a derivation of the theorem that there is no largest prime number from Peano's axioms.

The process of compiling the strictly redundant rules would to a considerable extent be influenced by experience of dealing with successively more complex cases, and storing the results for future use. In that case the total system at any one time, instead of having a neat mathematically analysable mode of operation, might be a large and messy collection of rules. The rules could even be partly inconsistent, if mistakes have been made and not all the rules have been thoroughly tested. The theory of how such a system works could not satisfy a craving for mathematical elegance and clarity. In MARR's terms this is a 'Type 1' theory explaining why no 'Type 1' theory can account for fluent visual processing. (I believe this is a very general point, which applies to many forms of intelligence, including language understanding and problem solving.)

C.9. Further problems.

There are many open research issues associated with the discussion so far. Which sorts of representation primitives and modes of composition are useful for seeing different sorts of environments and for performing different tasks? I've made some tentative proposals above, but they need further study.

Ethologists may be able to provide some clues as to which animals make use of which primitives. This may help in the design of less ambitious computer models and help us understand our evolutionary history.

A visual system which is not allowed to take millions of years of trial and error before it becomes useful as a result of 'self-organising processes' needs some primitives to be built in from the start, as the philosopher Immanuel Kant argued a long time ago. What needs to be built in depends in part on how much time is available for the system to develop before it has to be relied on in matters of life and death. We still don't know what would need to be built in to a general purpose robot, nor how it could synthesise new primitives and new modes of composition.

Investigation of these issues needs to be guided by a number of constraints. The representations must be usable for the purposes of human, animal, or robot vision, such as controlling actions, making plans, making predictions, forming generalisations, etc. They need not be effective for all possible goals, such as recognising the number of stars visible on a clear night, or matching two arbitrarily complex network structures. Is there some general way of characterising the boundary between feasible and over-ambitious goals?

The representations must be useful for coping with known sorts of environment, but they need not be applicable to all physically possible environments. In particular, it may be the case that the design of a visual system as powerful as ours must use assumptions about what I've called the 'cognitive friendliness' of the environment [32]. (More on this below.)

The representations must be processable in a reasonable time. This may prove to be a very powerful constraint on theories of animal vision, given the slow speeds of neuronal processing.

The representations must be capable at least in part, of being developed through some sort of learning. This allows adaptation to new environments with different properties.

They must be capable of representing partial information, for instance when objects are partially obscured, the light is bad, vision is blurred, etc., or when lower-level processing is incomplete but decisions have to be taken. (Partial information is not to be confused with uncertainty.)

More specific constraints need to be considered in relation to specific tasks. When the task is recognition, the representation should be insensitive to non-rigid deformations of the object, changes of view-point, etc. When the task is manipulation of a fragile object, the reverse, great sensitivity, is required. (Compare [26]).

D.1. The architecture of a visual system

The space of possible computational architectures is enormous, and only a tiny corner has been explored so far. It is hard to make sensible choices from a dimly visible range of options. Far more work is needed, on the properties of different forms of computation, especially multi-processor computations and their relations to the functions of vision. I shall try to show how design issues can usefully be related to different forms of 'cognitive friendliness' which may be present or absent in the environment.

Architectures, like representational schemes, may be discussed in terms of primitives and modes of composition, at different levels. Given basic limitations in processor speed, low-level visual processing needs a highly parallel organisation, in order to deal with massive amounts of information fast enough for decision-making in a rapidly changing environment. This is the biological solution, and, for the lower levels, may be the only physically possible solution given constraints on size and portability of the system. Of course, mere parallelism achieves nothing. Quite specific algorithms related to the physics and geometry of scenes and image production are required. There is plenty of work being done on this, and I'll say no more about it.

One of the main differences between different computational architectures concerns the way in which local ambiguity is handled. Some early programs used simple searching for consistent combinations of interpretations, but were defeated by the combinatorial explosion, e.g. [8], as well as a rather limited grasp of scene or image structures. Speed is crucial to an animal or robot in an environment like ours and rules out systems based on combinatorial searching. Co-operative models, such as 'Waltz filtering' [36] and relaxation use a parallel organisation to speed up the search enormously, when local features, together with their remaining interpretations, can reduce the ambiguity of their neighbours, in a feedback process. However, Waltz filtering is subject to what HINTON [14] called the 'gangrene' problem: eliminating local hypotheses which violate constraints can ultimately cause everything to be eliminated, in cases where the only good interpretation includes some local violations. Relaxation methods get round this, but it is not easy to apply them to networks of modules which need to be able to create new hypotheses on the basis of partial results of other modules. (MARR [25] argues that co-operative methods are too slow, if implemented in human neurones -- I don't know whether this argument is sound. It depends on how neurones encode information.)

New forms of co-operative computation based on massive parallelism (BALLARD [1], HINTON [18,19]) seem to be potentially very important for visual processing. BALLARD calls them 'unit/value' computers, HINTON 'mosaic' computers. They replace combinatorial search in time with massive spatial connectivity, and an open question is whether the combinatorial explosion can be controlled for realistic problems by a suitable choice of representations. The combinatorial possibilities are not so great at the lower levels, so perhaps they are restricted to the detection of relatively simple, relatively local, image features.

It is not so clear whether higher levels need a parallel organisation, and to what extent the processes are essentially serial and unable to benefit
fit from parallelism. Our discussion of functions and representations suggests that it would be useful to have a number of sub-processes dealing with different aspects of analysis and interpretation. I shall show that allowing different processes to be relatively independent, so that they can operate in parallel, makes it possible to take advantage of certain forms of cognitive friendliness of the environment, in order to compensate for unfriendliness in other dimensions.

My discussion bears a superficial resemblance to claims which used to be made about the need for 'heterarchic' control. The heterarchy/hierarchy distinction played an important role in the early 1970's in relation to the problems of designing systems to run on a single essentially sequential processor. (See WINSTON [37] SHIRAI [2] BRADY [7].) In that context, 'heterarchy' was hard to distinguish from 'anarchy'. The restriction to a serial processor generated problems of control which no longer seem relevant. If many modules can operate in parallel we need not argue about how they transfer control to one another, and there is no worry that one of the modules may assume control and never relinquish it. Instead the important questions are: what information flows where, and when, and how it is represented and processed. We attempted to address such questions in the POPEYE project [31,32]

D.2. The relevance of the environment to architecture

Cited work by Horn, Marr, Barrow and Tenebaum, and others, has shown how prior assumptions about the general nature of the environment can reduce search spaces by providing local disambiguation: for instance assuming that surfaces are rigid, or continuous, and have clear boundaries, or that illumination is diffuse. These are examples of 'cognitive friendliness' of the environment. Another example is the assumption that there is adequate short wavelength illumination and a clear atmosphere. The availability of space to move, so that parallax and optical flow can be used for disambiguation is another. The relative infrequency of confusing coincidences, such as edges appearing collinear or parallel in an image when they are not in the scene is another (often described as the 'general viewpoint' assumption). The reliability of certain cues for directly triggering predatory or evasive action (section B.6.) is another form of cognitive friendliness, in an environment which may be unfriendly in other respects. An oft-noted form of friendliness is Limited independent variation of object features, implying that the space of possible scenes is only sparsely instantiated in the actual world, so that scenes and therefore images have redundant structures, which can be useful for disambiguation. (E.g. BARLOW [7]. (The assumptions of rigidity and continuity of objects are special cases of this.) The 'phrasal lexicon' strategy sketched above in C.8. presupposes this sort of friendliness - limited variation implies re-usability of the results of computations.

Some of the general assumptions can be 'hard wired' into some of the processing modules, such as edge detectors, detectors of shape from shading or shape from optical flow. More specific assumptions, e.g. concerning which object features tend to co-occur in a particular geographical region, would be learnt and represented symbolically, like knowledge of common plant forms.

But the degree of friendliness can vary. Implicit or explicit assumptions about constraints can prove wrong, and if HINTON's 'gangrene' is to be avoided the organisation used needs to allow for local violations, if that provides a good global interpretation of an image, e.g. in perceiving camouflaged objects, or coming across a new sort of plant.

A common form of temporary unfriendliness involves poor viewing conditions (bad light, mist, snow storms, intervening shrubbery, damaged lenses, etc.) which can undermine the performance of modules which work well in good conditions. A traditional way of dealing with this is to allow incomplete or unreliable data to be combined with previously stored information to generate interpretations. This implicitly assumes that not all forms of cognitive friendliness deteriorate at the same time: creatures with novel shapes don't suddenly come into view when the light fades. (Feeding children false information about this can influence what they see in dim light.)

The idea that intelligent systems need to degrade gracefully as conditions deteriorate is old. However, it is often implicitly assumed that the main or only form of cognitive unfriendliness is noise or poor resolution in images. There are several dimensions of cognitive friendliness which need to be studied, and we need to understand how visual systems can exploit the friendliness and combat the unfriendliness. Human vision seems to achieve this by great modularity: many independent modules co-operate when they can, yet manage on their own when they have to. Binocular stereo vision is certainly useful, but normally there is no drastic change if one eye is covered -- driving a car, or even playing table tennis, remain possible, though with some reduction in skill. Similarly loss of colour information makes little difference to the perception of most scene structure, though various specialised skills may be degraded. Motion parallax and optical flow patterns are powerful disambiguators, yet a static scene can be perceived through a peep hole. We can see quite unfamiliar structures very well when the light is good, but in dim light or mist when much disambiguating information is lost, we can 'still often cope with relatively familiar objects.

D.3. Speed and graceful degradation

Previously, it was argued that non-visual subsystems need to obtain information from different visual subsystems. It can also be useful to have information flowing into visual data-bases not only from other parts of the visual system, but also from other sub-mechanisms, including Long-term memory stores. For instance, if some data-bases have to deal with incomplete information or possibly even incorrect information, because of some form of cognitive unfriendliness in the environment, then it will be useful to allow prior knowledge to be invoked to suggest the correct information. A more complex use of prior knowledge is to interact with partial results to generate useful constraints on subsequent processing. PAUL [28] showed how the layout of dimly perceived limb-like structures could interact with knowledge of the form of a puppet to specify which are arms and which legs, indicating roughly where the head is, and even suggesting approximate 3-D orientation. This sort of process seems to be inconsistent with the first of BARLOW's two 'quantitative laws of perception', which states that information is only lost, never gained, on passing from physical stimuli to perceptual representations [2].

In good viewing conditions this sort of mechanism is not necessary, and a modular design can allow what is found in the data to dominate the interpretation (though it doesn't always in humans, for instance in proof-reading). When very rapid decisions are needed, higher levels may start processing more quickly, and if lower levels have not completed their analysis, decisions may have to be based on data which are as bad as when viewing conditions are bad. The experience of the 'double take', thinking you've seen a friend then realising that it was someone else, could be explained in this way.
So both speedy decision making, and graceful degradation, can be facilitated in related ways. If modules have to be able to cope with incomplete information by using prior knowledge of the environment, then sometimes a high-Level decision can be taken before all lower level analysis has been completed, either because part of the field of view has been processed fully, revealing an unambiguous detail, or because coarse-grained global analysis has quickly provided information about a large scale structure, e.g. the outlines of a person seen before fine details have been analysed. (The presence of visual modules which process sketchy, incomplete information, indexed by location relative to the optic array, may account for the ease with which children learn to interpret very sketchy drawings.)

Decisions based on partial information are, of course, liable to error. In an environment where different forms of cognitive friendliness do not all degrade simultaneously, errors will be comparatively rare. This liability to error coupled with tremendous power and speed, is indeed one of the facts about human vision which requires explanation. It points to the importance of designs which may not be totally general and may not always find optimal solutions, but which achieve speed and robustness in most circumstances.

An animal visual system need not be guaranteed to be error-free, or even to find the best interpretation, so long as it works well most of the time. The 'good is best' principle states that in an environment with limited independent variation of features any good interpretation is usually the only good interpretation, and therefore the best one. So designs guaranteeing optimal interpretations (e.g. [40]) may not be relevant to explaining human perception. An open question is whether task constraints will often require a guarantee of optimality to be sacrificed for speed, even for robots. This could have implications for how robots are to be used and controlled. A lot depends on how friendly the non-cognitive aspects of the environment are, i.e. what the consequences of errors are.

Besides factual information, it may sometimes be useful for information about goals to flow into visual sub-modules. What sorts of interactions are desirable? An omnivore's goal of finding food could not interact directly with edge-detectors -- but what about the goal of looking for cracks in a vase? Higher level goals could not normally be fed directly into visual modules. Nor can they normally be translated into specific collections of lower level goals, except in special cases (e.g., finding vertical cracks requires finding vertical edges). However, goals may be able to constrain processing by directing fixations, and possibly by influencing which stores of prior information are used by certain modules. An example might be the use of different discrimination nets linked into some object-recognising module depending on the type of object searched for. If you are searching for a particular numeral on a page, it may be useful to use a different discrimination net from one relevant to searching for a particular letter, even though, at a lower level, the same set of feature detectors is used. In that case, at higher Levels, the process of searching for the digit '0' will be different from the process of searching for the letter '0' despite the physical identity of the two objects.

D.4. Recapitulation

I have tried to illustrate the way in which task analysis can precede global design or theorising about mechanisms. I've suggested that a visual system should not fit into, the mental economy like a black box computing some function from 2-D image structures to 3-D scene structures. Instead, we have a sketch of the visual system as a network of processes feeding many sub-databases which may be linked to different non-visual subsystems. (The general form of this sketch is not new, and not restricted to vision: e.g. [10].) Among the visual databases will be a subset which have an array like structure (with location representing location, and relationships implicit). For indexing, and problem-solving purposes, these should be mapped onto each other and a representation of the field of view, possibly built up over several fixations. The contents of higher level data bases, with more explicit representation of relationships, can be projected back into these array-like structures. Some of the databases should be able to trigger actions which bypass the central decision making process. Some may include amodal abstract, representations shared with other sensory subsystems. (This might allow some essentially visual modules to be used for spatial reasoning by the blind, even if they get no information via the eyes.) The central decision-making process needs to have access to a large number of the visual databases, though it will not be able simultaneously to process everything. (If the information available is not used it may not be stored in long term memory. So inability to recall does not prove that something has not been processed visually.)

The enormous redundancy in such a system makes empirical investigation a difficult and chancy process. For without being able to switch modules on and off independently it will be very hard to observe their individual capacities and limitations. Perhaps the effects of brain damage, combined with performance in very (cognitively) unfriendly situations, will provide important clues. Perhaps it won't.

It is likely that the design goals can be achieved in more than one way. However, there may be an interesting class of constraints, including the nature of the environment, the tasks of the system, and the maximum speeds of individual processors, which determine unique solutions (apart from the sorts of individual variations we already find between humans).

D.5. Further research

We need to explore in more detail the different dimensions of cognitive friendliness / unfriendliness of the environment, and how exactly they affect design requirements. Which sorts of friendliness can only be exploited by hard-wired design features and which can be adapted to through learning processes?

Given the nature of the environment, and the needs of an animal or the purposes of a robot, what kinds of data-bases are likely to be useful in a visual system, and what should the topology of their interconnections be? Can we get some clues from comparative studies of animals? I've made tentative suggestions about some of the sorts of data-bases which could play a role in human vision, and how they are interconnected. Could experimental investigations shed more light on this?

A problem not faced by most computer models is that in real life there is not a single image to be processed, nor even a succession of images, but a continual stream of information [7]. The problem of representing motion was mentioned in C.7. How constantly changing information is to be processed raises other problems. Once again we don't have a good grasp of the possible alternatives. As remarked in section C.3, it may be that only a system which is good at coping with changes will be really good at interpreting the special case of static images. The lowest levels of the system will probably be physical transducers which react asynchronously to the stream of incoming information. Is there a subset of data-bases which makes use of the "succession of snapshots" strategy? What are the trade-offs? Should higher level modules be synchronised in some way, or are they best left to work at their own speeds? If the environment is cognitively friendly in that most changes are continuous, and most objects endure with minimal change of structure, this provides enormous redundancy in the stream of information. The architecture of the system, and the representations used, could exploit this, avoiding much re-computation.

Much current research is aimed at finding out how much can be achieved by totally data-driven processing. We have seen that integration of prior knowledge with incoming data can provide speed and graceful degradation. We need to find out exactly which kinds of long-term memory need to be able to interact with which temporary visual databases.

My discussion has stressed the modularity and redundancy of a visual system. We need to explore in more detail the ways in which different sorts of failure of individual modules or connections between modules would affect total performance. There may be failures due to lack of relevant information or internal failures due to physical malfunction, or programming errors.

Our discussion has implications concerning the relationship between vision and consciousness. As usual, many questions remain unanswered. In particular, what exactly determines which databases should be accessible to consciousness?

D.6. Conclusion

I said at the beginning that I would present more questions than answers. I have outlined an approach to studying the space of possible visual mechanisms, by relating them to functions and properties of the environment. The study of the functions of possible mechanisms can have many levels. I have mostly stuck to a level at which it is indifferent whether the modules are embedded in brains or computers. As many AI researchers have pointed out, it's the logical nut the physical nature of the representations and manipulations thereon that we need to understand initially. However, we cannot try to build realistic models of the type sketched here until we know a lot more about what should go into the various databases. This requires finding out more about what needs to be represented and how it can be represented usefully.

This top-down research strategy is only one among several: we can learn from many disciplines and approaches. But analysis of function can provide a useful framework for assessing relevance. However, we must always bear in mind that our attempts to derive structure from function are inherently limited by our current knowledge of possible forms of representation and computation. The way ahead includes increasing this knowledge.

Acknowledgements

I have Learnt from talking to Frank Birch, Margaret Boden, Steve Draper, John Frisby, Bob Hawley, Gerald Gazdar, Geoffrey Hinton, David Hogg, Christopher Longuet-Higgins, John Mayhew, Julie Rutkowska, Mike Scaife, Frank O'Gorman, David Owen, John Rickwood, and the Late Max Clowes who set me thinking about these problems. Prof. Barlow made useful comments on an early draft of this paper. Between drafts I read most of and learnt from Marr's book on vision. Alison Sloman drew the pictures. Alison Mudd and Judith Dennison helped with production of the paper. The work has been supported in part by SERC grant GR-A-80679.

[NOTE 1]
The motion primitives referred to in C.7 may be used to link static scene descriptions, E.g. the description of shut scissors may be linked via a description of relative rotation to a description of open scissors. A description of a ball may be linked via descriptions of a squashing process to descriptions of disks and cylinders. Such Linking of static and non-static concepts may both facilitate prediction and account in part for the experienced continuity as scenes change, referred to in C.4. MINSKY makes similar suggestions in [27]. If such links are accessible while static scenes are perceived, this could account for the perception of 'potential for change' referred to in C.4, which seems to play an important role in planning, understanding perceived mechanisms, and solving problems.

BIBLIOGRAPHY

1 Ballard, D.H. 'Parameter networks: towards a theory of low-level vision' in Proceedings 7th IJCAI, VOL II, Vancouver, 1981.

2 Barlow, H.B. 'Perception: what quantitative laws govern the acquisition of knowledge from the senses?' to appear in C. Coen (ed.) Functions of the Brain, Oxford University Press, 1982.

3 Barrow, H.G. and Tenenbaum J.M. 'Recovering intrinsic scene characteristics from images', in [12] 1978

4 Becker, J.D. 'The Phrasal Lexicon' Theoretical Issues in National Language Processing. Eds. R.C. Schank and B.L. Nash-Webber. Proc. Workshop of A.C.L., M.I.T. June 1975. Arlington, VA.: Association for Computational Linguistics.

5 Brady, J.M. 'Reading the writing on the wall', in [12] 1978

6 Brady, J.M. (ed.) Special Volume on Computer Vision Artificial Intelligence 17,1, 1981. North Holland.

7 Clocksin, W.F.. 'A.I. theories of vision: a personal view', AISB - Quarterly 31 1978

8 Clowes, M.B. 'On seeing things', in Journal of Artificial Intelligence, vol. 2, no. 1 1971.

9 Draper S.W. 'Optical flow, the constructivist approach to visual perception and picture perception: a reply to Clocksin', A.I.S.B. Quarterly, 33, 1979.

10 Erman L.D. and V.R. Lesser 'A multi-level organization for problem solving using many diverse cooperating sources of knowledge' in IJCAI-4, M.I.T 1975.

11 Funt, Brian V. WHISPER: A Computer Implementation using Analogues in Reasoning. Technical Report 76-09, Dept. of Computer Science University of British Columbia, Vancouver, 1976. (Summarised in IJCAI-77.)

12 Hanson, A. and Riseman E. (Eds.) Computer Vision Systems Academic Press, New York, 1978.

13 Hayes, P.J., 'The naive physics manifesto' in D.Michie (ed.) Expert Systems in the Microelectronic Age, Edinburgh University Press, 1979.

14 Hinton G.E. 'Using relaxation to find a puppet', in Proceedings A.I.S.B. Summer Conference, Edinburgh 1976.

15 Hinton G.E. and Anderson J.A. (eds.) Parallel models of associative memory Hillsdale, NJ, Erlbaum, 1981

16 Hinton, G.E. 'Some demonstrations of the effects of structural descriptions in mental imagery', Cognitive Science, 3, 231-250, 1979.

17 Hinton, G.E. 'The role of spatial working memory in shape perception', in Third Annual Conference of the Cognitive Science Society, Berkeley 1981.

18 Hinton G.E. 'A parallel computation that assigns canonical object-based frames of reference', Proceedings 7th IJCAI, VOL II, Vancouver, 1981.

19 Hinton G.E. 'Shape representation in parallel systems' Proceedings 7th IJCAI Vol II, 1981

20 Hochberg, J.E., Perception, (Second edition) Prentice Hall, 1978.

21 Hogg D, 'Model based vision: a program to see a walking person', in preparation, University of Sussex, 1982.

22 Horn B.K.P, 'Obtaining shape from shading information', in [38] 1975.

23 McCarthy, J and Hayes P, 'Some philosophical problems from the standpoint of Artificial Intelligence' in Machine Intelligence 4, eds. B. Meltzer and D. Michie, Edinburgh University Press, 1969.

24 Marr, D. 'Early processing of visual information', in Philosophical transactions of the Royal Society of London, pp. 483-519 1976.

25 Marr, D. Vision, Freeman, 1982

26 Marr, D. and Nishihara, H.K., 'Representation and recognition of the spatial organisation of three-dimensional shapes.' Proc. Royal Society of London, B. 200, 1978

27a Minsky, M.L. 'Steps towards artificial intelligence' reprinted in Feigenbaum and Feldman (eds.) Computers and Thought 1961

27 Minsky, M.L. 'A framework for representing knowledge', in [38] 1975.

28 Paul, J.L, 'Seeing puppets quickly' in Proceedings A.I.S.B. Summer Conference, Edinburgh 1976.

29 Shirai, Y. Analysing intensity arrays using knowledge about Scenes, in [38] 1975.

30 Sloman A. The Computer Revolution in Philosophy: Philosophy Science and Models of Mind, Harvester Press and Humanities Press, 1978.

31 Sloman A, and D. Owen, G. Hinton, F. Birch, 'Representation and control in vision' in Proceedings AISB/GI Conference, Hamburg, 1978.

32 Sloman Aaron, and David Owen, 'Why visual systems process sketches', in Proc. AISB Conference, Amsterdam, 1980.

33 Sloman A. and Croucher M, 'Why robots will have emotions', in Proceedings 7th IJCAI, Vancouver 1981.

34 Treisman, A.M. and Gelade, G., 'A feature-integration theory of attention' Cognitive Psychology, 12, 97-136, 1980.

35 Treisman, A.M. and Schmidt H., 'Illusory Conjunctions in the Perception of Objects' Cognitive Psychology, 107-140,

36 Waltz, D. 'Understanding line drawings of scenes with shadows', in [38] 1975

37 Winston, P.H. 'The M.I.T. Robot' in Machine Intelligence Vol. 7 ed. D. Michie and B. MeLtzer, Edinburgh University Press, 1972.

38 Winston, P.H. (ed.) The Psychology of Computer Vision, McGraw-Hill 1975.

39 Winston, P.H. 'Learning structural descriptions from examples', in [38] 1975.

40 Woods, W.H. 'Theory formation and control in a speech understanding system with extrapolations towards vision', in [12] 1978.