ON DESIGNING A VISUAL SYSTEM
(Towards a Gibsonian computational model of vision)

Aaron Sloman

School of Computer Science,
The University of Birmingham, UK
http://www.cs.bham.ac.uk/~axs/

Abstract

This paper contrasts the standard (in AI) "modular" theory of the nature of vision with a more general (labyrinthine) theory of vision as involving multiple functions and multiple relationships with other sub-systems of an intelligent system. [1] The modular theory (e.g. as expounded by Marr) treats vision as entirely, and permanently, concerned with the production of a limited range of descriptions of visible surfaces, for a central database; while the "labyrinthine" design allows any output that a visual system can be trained to associate reliably with features of an optic array and allows forms of learning that set up new communication channels. The labyrinthine theory turns out to have much in common with J.J.Gibson's theory of affordances, while not eschewing information processing as he did. It also seems to fit better than the modular theory with neurophysiological evidence of rich interconnectivity within and between sub-systems in the brain. Some of the trade-offs between different designs are discussed in order to provide a unifying framework for future empirical investigations and engineering design studies. However, the paper is more about requirements than detailed designs.

---

[1]This was a sequel to some earlier papers on vision, and built on, but did not repeat all their contents, including:
(1) A.Sloman, Chapter 9 of *The Computer Revolution in Philosophy*
    http://www.cs.bham.ac.uk/research/projects/cogaff/crp/chap6.html
(2) Sloman, A., (1983), Image interpretation: The Way Ahead?, in Eds. O.J. Braddick and A.C. Sleigh, *Physical and Biological Processing of Images*, Berlin, Springer-Verlag.
The labyrinthine theory proposed that in addition to providing *factual* information about the environment (e.g. for use in reflective, deliberative and communicative processes) visual mechanisms could also provide *control* information, e.g. in visual servoing and posture control. Around that time, unknown to me, the theory became popular that there are two visual pathways (ventral and dorsal) associated with 'what' and 'where' processing. When I later learnt that these pathways were thought to separate out processing concerning objects and locations, I thought that was incoherent. Later I believe Goodale and Milner reached a similar conclusion and proposed a theory much closer to the one suggested here, explained in their paper summarising their book *The Visual Brain in Action* (1995), available here
    http://psyche.cs.monash.edu.au/v4/psyche-4-12-milner.html
The ideas presented here were extended in later papers and presentations, e.g.
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#gibson

Contents

1 Introduction

A squirrel, trying to get nuts in a bag hung up for birds, runs along the branch, hangs down, precariously supported by hind feet with tail curved round the branch, then, swaying upside down in the breeze, holds the bag in its forepaws while nibbling at the nuts protruding through the mesh. On seeing some fall to the ground, he swings up, runs further along the branch, leaps onto a railing on a nearby balcony and by a succession of runs and leaps descends to the nuts lying on the lawn below.

From my window I gaze out at this scene, both entranced by the performance, like a child watching a trapeze act, and also deeply puzzled at the nature of the mechanisms that make a squirrel possible.

The squirrel sees things, and puts what it sees to excellent use in selecting what to do next, controlling its actions, and picking up information that it will use next time about where to find nuts. I see things and enjoy and wonder at what I see, and try to think about the problems of designing a squirrel like THAT. How much is there in common between the squirrel's visual system and mine? How much is different? How much would a robot with visual capabilities have to share with either of us?

Why have we not yet been able to build machines with visual capabilities that come close to those of human beings, squirrels, cats, monkeys, or birds? It could simply be that the engineering tasks are very difficult, e.g. because we can't yet make cheap highly parallel computers available and we haven't solved enough of the mathematical or programming problems. Alternatively, it could be because we don't yet know much about human and animal vision and therefore don't really know what we should be trying to simulate. It could be both. I suspect the latter is the main reason - and that much improved hardware, better programming languages and design tools, faster mathematical algorithms, or whatever, would not in themselves bring us much closer to the goals of either explaining or replicating natural vision systems. We need a theory of what vision is for and how it relates to the other functions and sub-functions of intelligent systems. That is the main topic of this essay.

A good theory of human vision should describe the interface between visual processes and other kinds of processes, sensory, cognitive, affective, motor, or whatever. This requires some knowledge of the tasks performed by the visual subsystem and how they relate to the tasks and requirements of other subsystems. I shall attempt to analyse some uses of human vision, in the

3

hope of deriving some design constraints and requirements for visual systems for intelligent agents, whether natural or artificial - though I shall identify design requirements for which I do not have design solutions. More precisely I shall point to trade-offs between different sorts of designs rather than trying to prove that some are better than others in absolute terms. A popular "standard" theory implying that animals and robots should be designed in such a way that vision forms a well-defined module will be identified and criticised as too limited.

In principle this theoretical analysis should be tied in with detailed surveys of empirical facts about of human and animal visual systems (including their neurophysiology - cf. (Albus, 1981)), but my objective at present is not to establish a correct theory of human vision, so much as to provide a general theoretical framework within which empirical and design studies can be conducted.

Some time after writing an early version of this paper for a workshop in 1986, I began to read J.J. Gibson's book *The Ecological Approach to Visual Perception,* and, somewhat to my surprise, found considerable overlap, despite fundamental differences. I have therefore adopted some of his terminology, including the notion he defined in his chapter 5 of an "optic array", the array of information provided by light coming towards a viewing location. I shall use this to define the problem domain and formulate a set of key design questions.

2 What is vision?

In order to delimit the topic I assume, like Gibson, that vision is concerned with deriving information about the environment from (possibly changing) structure in one or more optic arrays. An optic array is structured in the sense that information coming to a location is different in different directions: different colours, intensities (and patterns of change in colours and intensities) are transmitted from different directions (mostly reflected but not always) towards any given viewing point. It is a two-dimensional array in the sense that the directions from which information comes vary in two dimensions, though if the array is a changing one, time adds a third dimension. As Gibson points out, a system does not need to have a retina onto which a 2-D optical image is projected in order to take in and make use of the 2-D structure of the optic array: compound eyes made of units aimed in different directions can also do this, as could scanning systems.

Defining vision as extraction of information about the environment from structure in optic arrays is not an attempt to legislate usage, or define arbitrary terminology, but merely to identify an important range of design issues addressed in this paper. For example, I am not concerned with how a plant might use measurements of daily incident light to determine when to bud or drop its leaves: this process does not (as far as I know) make use of the two dimensional structure of the optic-array to derive information about the structure and properties of the environment. It lacks other interesting features of vision, described below.

Despite this restriction, the concept of vision used here is very broad. It leaves open what information is derived from the optic array, how it is derived, what other information is used in the derivation, what the derived information is used for, and how many other kinds of subsystems there are in the total system of which vision is a part: enormous variation is possible

on all of these points, both in biological organisms or present and future artefacts. For now I shall assume that we are dealing with a total system that has many human-like and squirrel-like capabilities, including a range of different sensory and motor abilities, the ability to plan and control complex movements, to acquire and store information about the environment for later use, to pursue a variety of types of motives, and so on. This variety of capabilities will be left vaguely specified for now. It has architectural implications that will be mentioned later, as the discussion of design issues unfolds.

The aim of the paper is not to put forward empirical theories but to explore "architectural" design issues: that is questions about in what way, and at what level of scale, an intelligent system needs to be constructed from (or decomposable into) distinct components with different clearly defined functions and interfaces. The theory to be criticised takes a visual system to be a relatively large-scale module with restricted input and output channels: I shall contrast it with a theory postulating smaller components more richly connected with modules outside the visual system. The components to be discussed need not be physically separable: any more than the separate software modules in a computing system have to map onto usefully separable physical components. Some of the components in an information processing system may be "virtual" structures, like the linked lists, networks, trees, arrays and other structures created in a computer yet never to be found by applying physical measuring instruments to their innards - rather, such structures are abstract interpretations by programs (and by us) of the physical state of the machine.

I am not contrasting "virtual" with "real" as Gibson does: the contrast is with "physical". The virtual machine corresponding to a high level programming language running on a computer is as real as the physical machine running instructions in the low level machine language.[2] But it is a different machine with different states, properties and transitions, and different causal interactions. Similarly the components of a visual system will be described at a fairly high level of abstraction, leaving open the question whether the neurological basis of human vision adds further design constraints to those considered here.

Whether human vision is exactly as I say does not matter as much (for this paper) as whether systems *could* be designed thus, and what the costs and benefits would be of these and alternative designs. In other words, this essay is a partial exploration of the space of possible designs for visual systems. It does not aim to be a complete exploration: that is too big a task. Instead I shall focus on a relatively small subspace close to how human beings appear to me to be designed. Checking out whether humans or other animals really fit into this subspace and if so exactly where they are located, is a task I must leave to others with more detailed knowledge of human abilities and their neurophysiological underpinnings. I shall also leave to others the task of specifying mechanisms able to meet the requirements I'll outline.

2.1   Some key questions

The discussion will revolve around the following key questions.
- What kind of information can or should a visual system derive from the optic array?

- Should the information be expressed in descriptions in some kind of internal language?

- If not, in what way can the information stored, or used?

- Should the information be produced by a visual system in a single general purpose form, leaving it to other modules to transform it to suit their needs, or can a visual system directly produce forms of information suited to other specific modules?

- What other functions should a visual system have besides producing information about the environment? E.g. is part of its function to trigger or control processing in other subsystems? (This could be done either by sending descriptions of the environment or optic array, from which inferences about what to do would have to be drawn, or by deriving control signals as well as descriptions from the optic array and transmitting the control signals directly to the modules concerned.)

- Are descriptions of (possibly changing) 3-D spatial structure and location, the only descriptions that should be produced by a visual system?

- If not, what other kinds of descriptions should a visual system produce? E.g. should descriptions of 2-D, time varying, image features (or features of the optic array itself) be output? Should descriptions of non-spatial properties of objects (e.g. functional or causal properties) be produced by the visual system, or are they inferred from the visual output, by separate modules?

- What kinds of input should a visual system make use of? Is it purely, or mainly, optical data, or do other data play a significant role, e.g. data from other sensory subsystems, or data from higher level processes, or prior knowledge about objects in the environment?

- Is it possible to draw a sharp boundary between visual processing and other kinds of processing, or is it best to design intelligent agents around a richly interconnected processing system with increasingly multi-modal or amodal layers of processing as information moves from sensory transducers? In other words are there sharply distinguished modules for vision, touch, hearing, reasoning, memory, etc., or are the functional boundaries blurred and different subsystems closely integrated with one another?

2.2   Two opposing theories of vision

Although truth is rarely extreme, I shall contrast two extreme theories: the *modular* theory and the *labyrinthine* theory. The former treats a vision system as an integrated module with a clearly defined and very limited set of inputs and outputs, while the latter treats it as a collection of modules having a much wider, and more changeable, collection of input and output links to other sub-systems. I'll sometimes refer to the modular theory as the "standard" theory since something like it is currently a sort of orthodoxy among AI vision researchers and at least some cognitive scientists, though there may not be any adherent of its most extreme form. My arguments will mostly go against the modular theory, not on the basis of empirical evidence that it is wrong, but because it puts forward a poor design. Of course, that leaves open the possibility that we are poorly designed, and the modular theory is empirically correct.

I believe that Gibson reached conclusions about vision as multi-functional that are close to the labyrinthine theory, though he had a different standpoint: he used excessively narrow and old-fashioned concepts of "representation" and "processing" that led him wrongly to reject the idea of visual mechanisms that use representations or process information (as (Ullman, 1980) points out). He apparently thought of representations as essentially picture-like objects, isomorphic with things represented, and requiring something like perception for their use. He seems to have had no conception of an abstract structured internal store of information accessible via processes better described as computational than as perceptual, and indeed so different from perception that they are actually capable of playing a role in *explaining* perception.

He also seemed to think that theories of cognitive processing aim to show how retinal images get mapped onto images in the brain, which are operated on in various ways to produce states of consciousness (op.cit. page 252), and he seems to have thought (op.cit. p. 238) that all theories of cognitive processing relied on what he described as "old fashioned mental acts" such as recognition, interpretation and inference.

Gibson apparently had not learnt about work in Computer Science and Artificial Intelligence that postulates better understood processes invoked to explain mental acts. I am not claiming that current work in AI, including connectionist AI, has got good explanations as yet, merely that the space of possible explanatory designs acceptable for scientific or engineering purposes is richer than Gibson seems to have thought of.

So, although I shall not dispute Gibson's assertion that vision is best thought of as concerned with interpreting optic arrays rather than retinal images, unlike Gibson, I do not rule out a prior process of analysis and description of the input array as a 2-D structure: this pre-processing of the optic array to reduce the information content may be essential for reducing the combinatorics of mappings from the array to other things. Gibson's own formula (p.169) "equal amounts of texture for equal amounts of terrain" seems to depend on the assumption that the amount of texture in various regions of the optic array can be detected as a basis for inferring the sizes of the corresponding regions of terrain. Moreover, for some purposes, e.g. drawing and painting, it is clear that we need and use information about 2-D properties of the optic array. So we have an existence proof of the possibility of deriving and using such information, and this is consistent with (but does not prove) the claim that it plays a role in intermediate stages of vision, even if it is not part of the output of vision.

In fact, I shall suggest that one of the things that characterises a visual system, as opposed to other kinds of photo-sensitive mechanisms, is the use of internal representations that relate information to the two dimensional structure of the optic array. This allows indexing to be based on relationships in the array, such as direction and distance. This use of one or more 2-D maps as information stores may provide the only clear boundary between a visual system and the rest of the brain.

Consequently we need not stick with Gibson's mystifying and unanalysed notions of direct information "pickup" (chapter 14) and "resonance" (p. 249), though I shall sketch a design for visual systems that has distant echoes of these notions.

The notion of a representing structure can be generalised if we think of it as a sub-state of a system that can simultaneously have a collection of independently variable, causally interacting,

sub-states. Different kinds of variability and different kinds of causal interaction of substates will lead to interestingly different behavioural capabilities of the whole system. This notion subsumes both the kinds of data-structures used as representations in conventional AI and the kinds of information structures embedded in neural networks studied by connectionists. A detailed analysis of the way in which such substates can be interpreted as having semantics is beyond the scope of this paper. A beginning is made in (Sloman, 1987b), where I suggest that a mixture of causal relations and Tarskian model theory (generalised to cope with non-propositional representations) may suffice. The model theory delimits the mathematically possible models for a given representing structure and the causal relations select the portion of the world forming the actual model (or the set of models in the case of ambiguity or indeterminacy).

2.3   Sketch of the "modular" theory

Returning to the "modular" theory: it claims that vision is a clearly bounded process in which optical stimuli (usually, though not always, thought of as passively received via a retina) are interpreted in a principled fashion in order to produce descriptions or representations of (possibly time-varying) 3-D spatial structures. These descriptions are then stored in some kind of database (perhaps a short term memory) where they can be accessed by other sub-systems for a wide variety of purposes, such as planning, reasoning, controlling actions, answering questions, solving problems, selecting information for a long term memory store, and so on.

On this standard view all processes that make use of visual input have to go via this common database of descriptions of the spatial structure of visible surfaces, which therefore provides an interface between vision and everything else. It is possible for this database to be shared between vision and other sensory modules, all feeding in their own characteristic kind of information about the environment. There could also be a collection of output modules controlling muscles or motors, driven by plans or programs based on information from the central store.

So on the modular theory we can think of an intelligent agent as having an architecture something like a sunflower with radial florets attached to the edge of a central disc. Each floret represents an input or output module, or perhaps some other specific processing module such as a planning module, while the central core is a store of information fed in by sensory modules and expanded or interpreted by inference modules. An extreme version of the theory would require all information to go from one radial floret to another only via the central store. One of the florets would correspond to the visual system, others to smell, taste, touch, and various output modules, e.g. perhaps one for talking, one for walking, one for manipulation with hands, etc. Something like this modular theory is defended at length in (Fodor, 1983), and is often implicitly taken for granted by workers in AI.

2.4   Proponents of the modular theory

Perhaps the clearest statement of the modular theory, at least as far as vision is concerned, is to be found on page 36 of David Marr's book (Marr, 1982), where he describes the 'quintessential

fact of human vision – that it tells about shape and space and spatial arrangement.' He admits that 'it also tells about the illumination and about the reflectances of the surfaces that make the shapes – their brightnesses and colours and visual textures – and about their motion.' But he regards these things as secondary '... they could be hung off a theory in which the main job of vision was to derive a representation of shape'. This echoes old philosophical theories distinguishing 'primary' and 'secondary' qualities.

Something like this view, perhaps without the distinction between shape as primary and other visual properties as secondary, underlies much vision work in Artificial Intelligence, including the work of some of the most brilliant researchers. It pervades John Frisby's excellent book on seeing (Frisby, 1979), partly inspired by Marr, and the same "standard" view is expressed in the textbook on AI by Charniak and McDermott (Charniak & McDermott, 1985), who write: 'Unlike many problems in AI, the vision problem may be stated with reasonable precision: Given a two-dimensional image, infer the objects that produced it, including their shapes, positions, colors and sizes'. If pressed, Charniak and McDermott would no doubt have included 'their motion'. A similar task definition is given by Longuet Higgins 'What the visual system ultimately has to do is to infer from a (2+1)-dimensional image – or two such images – the spatio-temporal structure of a (3+1)-dimensional scene' (Longuet-Higgins, 1987) pp 293–4. Although these authors apparently subscribe to something like what I am calling the "modular" theory of vision, they don't necessarily all embrace every aspect of the extreme version I've sketched.

So on this theory we have to think of the squirrel's visual system as extracting information from the optic array, processing it in various ways (outlined below) and storing descriptions of the (changing) 3-D structure of branches, leaves, railings, peanuts, or whatever in some central database, where it can be accessed by inference mechanisms and planning mechanisms e.g. to work out what actions to produce and by motor control mechanisms to ensure that the actions are performed correctly.

The standard theory leaves open precisely how motion and change are dealt with. Swaying branches and other continuous changes perceived in the environment might cause the descriptions in the central database to be continuously changed. Or persisting descriptions from a succession of snap-shots might be stored with some kind of time-stamp, until they are no longer needed (or they 'decay'). Alternatively instead of producing only time-stamped descriptions of spatial structure the visual system might produce descriptions of motion structure (e.g. "swaying"), in which case a fixed description might correspond to a changing scene, as unchanging differential equations can describe a changing physical system.

3   Must visual processing be principled?

One attractive feature of the modular theory is that it holds out some hope for a *principled* design of visual mechanisms. For example, if the task of vision is to discover facts about the shape, location, colour and texture of visible surfaces, then it is to be hoped that these facts can be inferred from the optic array using principles of mathematics and physics, since the optic array is a richly structured collection of information systematically derived, via a well

understood projection process, from the shapes, locations and reflective properties of objects in the environment, and the illumination falling on them. This notion of principled inference from optic array to scene structure is close to some of Gibson's ideas, though I shall end up using Gibsonian arguments against it.

Early proponents in AI of a principled derivation of scene structure from intensity discontinuities (from which a "line drawing" showing visible edges was assumed to be derivable) were (Huffman, 1971) and (Clowes, 1971). They showed that in a world of trihedral polyhedra, only certain interpretations of edge maps were consistent, work that was later extended by other researchers to more general scenes including shadows and curved objects. A different generalization came from Horn who argued that a lot more information about shape could be inferred from intensity changes (e.g. (Horn, 1977)). Marr (op.cit) also stressed the need for a principled theory as a basis for image interpretation, and inspired a considerable amount of relevant work. (Ballard & Brown, 1982) present an introductory survey of relevant mathematical techniques. Part IV of (Longuet-Higgins, 1987) is a collection of mathematical studies of what can and cannot be inferred from two images taken from different places or at different times.

If the visual mechanism is a principled solution to very specific mathematically statable and solvable problems intimately bound up with the geometry and optical properties of the environment, then a study of visual mechanisms should always be related to the nature of the environment. Yet it is interesting that many vision researchers are now investigating trainable neural networks rather than mechanisms designed from the start to work on principled algorithms that invert the supposed projection process. Is this new work fundamentally flawed? Or might it be justified because our visual system is not specifically geared only to the geometry and physics of our environment, but can process whatever it is trained to process? (I am not disputing that work on neural nets is based on a principled theory: but it need not be a specific theory about how to derive 3-D structure from 2-D optic array information.)

Might not a more general design of this kind be preferable in a world where, besides spatial properties and relations, objects have very many additional properties that are potentially relevant to a squirrel, including for instance edibility, graspability, support strength, and other causal properties not directly entailed by properties of the optic array?

Fluent reading (including musical sight-reading) illustrates the usefulness (at least for humans) of the ability of a visual system to be trained to associate non-spatial information with information in the optic array, where the association is arbitrary and unprincipled. But I'll argue that this general capability is useful for many other purposes that we share with other animals and could share with robots. So the aesthetic advantage of the modular theory, namely that it postulates a principled process of interpretation, may be outweighed by the biological or engineering advantage of something more general.

It must be said that even if the mechanisms in natural visual systems don't use principled algorithms for inferring geometrical structure but simply learn useful associations, the mathematical study of what can and what cannot be inferred in a principled way is still very worth while, since it may help to define what is learnable, and it may provide important algorithms for artificial vision.

3.1 Towards a "labyrinthine" theory

An alternative more labyrinthine theory can be based on the following ideas.

- A well designed visual system should produce not just descriptions of (changing) 3-D spatial structure but descriptions of a far wider variety of features of the environment - in fact anything that can be reliably detected and which is useful (compare Gibson's "affordances"). In particular, some of the output of vision might be *partial* results of analysis of the optic array, rather than information about the environment. (I'll give examples later.)

- The outputs of a visual system should not simply be descriptions of what has been detected or inferred, but might for example be motor control signals fed directly to motor sub-systems as part of a feedback loop.

- Closely related to the previous point, a visual system should not have a single output channel, but should be able to transmit descriptive or control information directly to any module that needs it.

- A visual system should not simply make use of the optic array but should be able to use a wider variety of inputs, including low or intermediate-level information from other sensory transducers and high level conceptual information, as well as control information about actions that change the information coming from the optic array (e.g. information about eye or neck movements, or bodily motion).

- A visual system should not have rigidly fixed channels of input and and output, nor fixed limits on the kind of information that it can produce, but instead should be capable of changing all of these as a result of training. In particular, it should be possible to extend the descriptive capabilities, and to set up a new information channel from any intermediate stage of visual processing to some other sub-system that can make good use of the intermediate information.

- The interpretation processes employed by a visual system need not be mathematically derivable from principles of physical optics and projective geometry but may make use of any cues that are empirically found to be useful: i.e. the process of extracting information from the visual array need not be *principled,* even if it is the result of a principled learning process.

This labyrinthine theory admits that there is such a thing as a visual module specially geared to processing optic arrays, but it does not insist on fixed and sharp boundaries as the standard modular theory does, like the single attachment point for each radial floret on a sunflower. In particular, it does not assume a fixed type of output restricted to descriptions of spatial structure and changes.

Discussion of such design options requires analysis of the uses of vision. Part of my argument is that in order to do what the modular view proposes, the visual system needs a type of mechanism that would in fact enable it to do more than just produce spatial descriptions: for even the more restricted modular type of visual system would require a general-purpose

associative mechanism. This is because it requires vision to be conceptually creative, as we'll see.

4 The innards of the "standard" visual module

Let's look more closely at the "modular" theory of vision. Although it conceives of the visual system as having a well defined boundary it is not thought of as internally indivisible. Modular theorists often postulate a collection of different internal sub-modules and databases in which intermediate descriptions of various kind are stored, and used *within* the visual system in order to derive subsequent descriptions. (See (Barrow & Tenenbaum, 1978), and (Nishihara, 1981)). For example, the intermediate databases postulated include edge (or surface discontinuity) maps, binocular disparity maps, depth maps, velocity flow maps, surface orientation maps, histograms giving the distribution of various kinds of features, descriptions of edges, junctions, regions and so on. I use the word "map" here to suggest that the information is stored more or less in registration (some at fine and some at coarse resolution) with the optic array. (NOT with the retina: retinal images change too rapidly during saccadic motion.) Some of these databases may contain viewer-centered, others object-centred or scene-centred, descriptions of objects, or fragments of objects, in the environment.

On the modular view, these internal data-bases are purely for use within the visual subsystem. They contain information that is of use only as an intermediate stage in computing information about 3-D spatial structure and change. The only information available to other subsystems would be the descriptions of objects and processes in the 3-D scene that are fed from the visual module to the central store where other modules can make use of them.

But why should a visual system be restricted to such a narrow function? If the intermediate databases contain useful information why should it not be made directly available to other non-visual modules, such as motor-control mechanisms? It could be useful to have a system that can perform a substantially wider range of functions than this monolithic, rigidly restricted, spatial description transducer. In particular, since the optic array is inherently ambiguous in several respects (e.g. as regards hidden parts of objects, or non-geometrical properties such as the strength of an object), it would be useful if a visual system could at least sometimes make use of information from other sources to help disambiguate the purely visual information. If this requires the use of learnt associations that cannot be inferred from general principles, then it is necessary to have a system that can be trained, and will therefore change over time.

If it is possible to build a visual system that can extract useful non-geometrical information about objects in the environment, e.g. information about causal relations like support or rigidity, or information about edibility, or information about the intentions of other agents, then it would be worth giving those tasks to the visual system *provided that* the information is derivable more easily, more quickly, or with greater precision, or in a more useful form, from the optic array (or intermediate stages of processing of optic arrays) than from descriptions of 3-D structure and motion. In that case the visual system might as well produce different sorts of information in parallel, rather than requiring one to be derived from the other by a separate module.

Notice that I am not claiming that visual systems don't produce geometrical information

about the environment. Obviously they do. Moreover geometrical descriptions produced in a principled way may be part of the process of learning a less principled way: if non-rigidity is first detected on the basis of changing shape, it may later, as a result of learnt associations, be detected on the basis of the type of material and its texture or colour.

The richer conception of vision as having many different purposes rather than simply producing descriptions of the structure and motion of visible surfaces, has implications both for the architecture of a visual system and for the types of representations that it uses internally. The architectural implication is that instead of a single input channel from the retinas and a single output channel to the central store of 3-D information, a visual system may require far more connections, so as to receive inputs from more sources and so as to be able to output different kinds of information to other modules that need it. The neat sunflower model would then have to be replaced by a tangled network of interconnected modules.

5 Previous false starts

The modular theory of vision provides useful insights into some of the sub-tasks of a visual system, but tells only a small part of the story, like all the other 'fashions' that have characterised AI work on image analysis since the 1960s. The history of attempts to make machines with visual capabilities includes several enthusiastic dashes down what proved to be blind alleys, or at best led to small steps forward. Here I'll list some of the over-simplifying assumptions that have proved tempting, as an introduction to a more detailed analysis of the purposes of vision.

- Vision is essentially a process of image enhancement: if only you can make a computer produce a new image showing clearly where the edges of objects are, or how portions of the image should be grouped into regions, then you have solved the main problems of vision. However, the production of images cannot be enough - for something would then have to see what was in these new images. (This seems to be the most common trap for engineers who start working on vision.)

- Vision is pattern recognition: if only we could make machines recognise patterns in images (or optic arrays), all the problems would be solved. This ignores the need to perceive and negotiate complex structures and situations not seen before: merely attaching a known label naming a recognized pattern does not do this, although recognition of known substructures and of relationships is *part* of the process of perceiving new structures.

- Since optic arrays and retinal images are two dimensional, vision is a process of analysing 2-D structures, for instance finding edges, grouping the array into regions, describing relationships within the image. Clearly this cannot be the whole story, even if it is a part of the correct story, for the whole story must include perception of 3-D structures. There may be some primitive organisms that need only 2-D information. But the squirrel's actions have to be intricately related to 3-D structure, distances, shapes, and so on. In short, *interpretation* is needed, as well as *analysis,* where interpretation includes mapping given structures to quite different structures (e.g. mapping 2-D structures onto 3-D structures).

13

- Vision is essentially a process of segmentation: if only images (or optic arrays) could be segmented into parts belonging to different objects, the rest would be easy. This is a tempting strategy if the 2-D segmentation can somehow be made to correspond to boundaries between objects in the environment. However, even if image segmentation may be part of the story, it does not meet the need to describe 3-D relationships between objects and parts of objects, and it doesn't account for perception of smoothly varying shapes that have no clear segmentation into parts, e.g. a human torso. (Though what such perception amounts to remains untold.)

- Vision is syntactic analysis - finding the hierarchical structure in images, just as a parser finds structure in sentences. (This idea was inspired by work in theoretical linguistics in the 1960s, and is expounded at length in (Fu, 1977) and (Fu, 1982).) However, this is simply a more sophisticated variant of the previous erroneous views: it is not enough to find and describe structures in images or optic arrays. In order to work out a path from its branch to the nuts on the grass the squirrel needs to grasp the structure in the environment, not in viewpoint-centred 2-D patterns.

- Vision is heterarchic (non-hierarchic) processing, mixing top-down and bottom-up analysis: if only the right control structure is used, and enough prior knowledge available about possible objects in the environment, stored hypotheses about likely objects can be triggered by cues in the input in order to control analysis and disambiguate evidence. This view was partly inspired by Winograd's work (Winograd, 1972) on heterarchy in language understanding and is supported by many examples of human abilities to see things in inherently ambiguous pictures and views. However, it says nothing about the perception of shape, about the ability to see quite unfamiliar structures (where top-down guidance is therefore unavailable), and about the way in which vision relates to other processes. Moreover, the claim that high level hypotheses can influence low level analysis risks being defeated by the combinatorics, except in special cases mentioned below: there are far too many ways of mapping the hypothesis that an elephant is in front of you into detailed hypotheses about edges, optical flow, intensity gradations, etc.

- Vision is essentially a matter of getting 3-D information about the environment: if only we could find a way of deriving from retinal images or optic arrays a 3-D depth map of distances to the nearest surface in various directions, the rest would be easy. However, a 3-D depth map is just another unarticulated database, and, as will be shown later, it would still require considerable processing in order to provide useful descriptions of what is in the scene. In particular, it has the unfortunate problem of being dependent on viewpoint, so that it captures no viewpoint-independent facts about the scene, such as that there is a table in the middle of the room with edges parallel to the walls.

- Vision is highly parallel - if only we had powerful enough parallel computing engines everything would be easy. This ignores the question of what should be computed. For instance it would leave us with the problem of how to represent spatial structure and how to derive it from optic arrays. How to make use of massive computing power in vision remains a problem that cannot be addressed properly until the purposes of vision have been clarified.

14

- Vision requires connectionist machines capable of doing parallel distributed processing, as defined, for instance in (McClelland, Rumelhart, & *et al*, 1986). Mechanisms of this sort seem to be good for learning associations and then generalising by interpolation, and for rapid detection of low level features like intensity discontinuities and optical flow. It is not yet clear whether they can cope with tasks that involve hierarchical structure description in unfamiliar situations (seeing a new whole made of parts which are made of parts etc.) Moreover, merely describing a general type of processing leaves unanswered a host of specific questions about how vision works including the question about what kind of information has to be extracted from optic arrays or how it is to be used. In particular I see no reason (so far) to believe that connectionist mechanisms will help us with the hitherto intractable problem of representing arbitrary shapes in a useful way.

So, for now, it seems sensible simply to regard connectionist (PDP) mechanisms as part of the stock of design options, along with other forms of computation, that need to be considered in designing intelligent systems. We can then try to find out which mechanisms are best suited to which subtasks. I shall identify a number of subtasks involving mapping information from one domain into another, for which connectionist mechanisms may be well suited.

6 Interpretation vs analysis

All this shows that there are several key ideas that are easily forgotten. One is that visual perception involves more than one domain of structures. This is acknowledged by those who claim that vision involves going from 2-D structures to 3-D structures, which is why *analysis* is not enough. Besides analysing image structures, or the structures of optic arrays, a visual system has to *interpret* them by mapping them into quite different structures. One strength of the standard modular view is that it acknowledges this. Gibson's talk of "affordances" also implicitly acknowledges this: the affordances that he claims the organism "picks up" from the optic array are not properties of the optic array but of the environment. The squirrel is interested in nuts, not features of its optic array. I shall later describe some quite abstract affordances that can be seen.

Of course, analysing and describing the 2-D structure of the optic array could be an important *part* of the complete system, and might be an essential part of the interpretation process. It cannot be the *whole* process, since that analysis does not produce all the required information.

Another key idea that has played an important role in AI work, especially the standard modular theory, is that vision involves the production of *descriptions* or representations, in some kind of internal formalism. For instance, in saying that image structures are mapped into 3-D structures it is often assumed that the mapping involves producing descriptions of the 3-D structures. Nobody knows exactly what sorts of descriptions are needed, but at least it seems that vision produces at least hierarchical descriptions of 3-D structures such as vertices, edges, surfaces, objects bounded by surfaces, objects composed of other objects, and spatial properties and relationships such as touching, above, nearer than, inside, etc. So any system that merely

produces data-bases of measurements (e.g. a depth map), or that merely labels recognised objects with their names, cannot be a complete visual system.

However, it can hardly be said that AI work or even work on computer-aided design has produced anything like a satisfactory language for describing shapes. Mathematical descriptions suffice for simple objects composed of planes, cylinders, cones, and the like, but not for the many complex, partly regular and partly irregular, structures found in the natural world, such as oak trees, sea slugs, human torsos, clouds, etc. Moreover, there are deep philosophical problems about what it means for a mechanism to produce structures that it interprets as referring to something else, though I shall not discuss them here, for my main point is that even if all these gaps can be filled, what has been said so far is not enough. Interpretation of the optic array need not involve only the production of descriptions, and it need not be restricted to extraction of information about 3-D spatial structures.

Not enough attention has been given to the fact that vision is part of a larger system, and the results of visual processing have to be useful for the purposes of the total system. It is therefore necessary to understand what those purposes are, and to design explanatory theories in the light of that analysis. The rest of this essay addresses this issue. I'll try to show below that besides the domains of 2-D and 3-D spatial structures, a truly versatile visual system should be able to cope with yet more domains, interpreting 2-D optic arrays in terms of abstract domains involving functional or causal relationships, and perhaps even meanings of symbols and perceived mental states of other agents. I'll outline some architectural principles for achieving this, but will have little to say about the detailed sub-processes.

7   What is, what should be, and what could be

A methodological digression is necessary, in order to prevent misunderstandings about this exercise. It is important to distinguish three different sorts of question, empirical, normative and theoretical. The empirical question asks what actual biological visual systems are like and what they are used for. The normative question asks what sort of visual system would be desirable for particular classes of animal or robot (given certain objectives and constraints). The theoretical question asks what range of possible mechanisms and purposes could exist in intelligent behaving systems, natural or artificial and how they might interact with other design options.

It is possible for these questions to have different answers. What actually exists may be a subset of what is theoretically possible. It may also be different from what might be shown to be optimal (relative to some global design objectives).

I shall probably confuse my audience by mixing up all three sorts of questions in the discussion that follows. This is because in discussing design possibilities and trade-offs, my real concern in this paper, I am occasionally tempted to express some empirical conjectures about biological visual systems, including human ones, e.g. the conjecture that they have a broader range of functions than the modular theory admits. However, establishing this is not my main aim. I am concerned only to make the weaker claim that alternative designs with interesting trade-offs are possible and worth exploring. That this is relatively weak does not

make it trivially true or unimportant: it provides a framework for formulating and exploring stronger theories.

Even if my empirical biological conjectures are false, the normative claim about what designs would be best (in relation to certain biological needs) might be correct: biological visual systems might be sub-optimal.

Moreover, even if the empirical claim is false, and the normative arguments about optimality are flawed, the theoretical claim that these alternative designs are possible might be true and interesting. For example, by analysing the reasons why a labyrinthine design is not optimal we increase our understanding of the optimal design. Moreover, by studying the biological factors that ruled out the alternative design we may learn something interesting about evolution and about design trade-offs.

My own interest is mainly in the theoretical design questions. This is part of a long term investigation into the space of possible designs for behaving systems with some of the attributes of intelligent systems, including thermostats, micro-organisms, plants, insects, apes, human beings, animals that might have evolved but didn't, and machines of the future. Surveying a broad range of possibilities, studying the implications of the many design discontinuities in the space, and attempting to understand the similarities and differences between different subspaces, and especially the design trade-offs, is a necessary pre-condition for a full understanding of any one subspace, including, for instance, the subspace of human-like designs.

8 Problems with the modular model

A well known problem with the view that 3-D scene descriptions are derived from image data in a principled manner by a specialised visual module is that the system can quickly reach definite interpretations even when the information available at the retina from the optic array is inherently ambiguous. A principled system would have to produce a list of possible interpretations, or perhaps fail completely.

In particular, in many monocular static images it is easy to show, e.g. using the Ames Room and other demonstrations described in (Gregory, 1970), (Frisby, 1979) and even Gibson (op.cit. p.167), that human visual systems rapidly construct a unique (possibly erroneous) 3-D interpretation even though the particular optic array is mathematically derivable from a range of actual 3-D configurations, and hence there is no unique inverse to the process that projects scenes into images. Johansson's films with moving points of light that we reconstruct as moving people, provide another example. The ambiguity can be present even when the images are rich in information about intensity, colour and texture, as shown by the Ames room. More precisely, 3-D information about structure or motion is often lost by being projected into 2-D, but that does not prevent human visual systems rapidly and confidently coming up with 3-D interpretations.

Notice that I am not drawing the fallacious conclusion criticised by Gibson (op.cit. p 168) namely that normal visual perception has to rely on information as ambiguous as the illusory contexts. My point is only that the human visual system has the *ability* to form percepts that are not mathematically or geometrically justified by the available information: and indeed are even

mistaken sometimes. If it has that capability, then perhaps the capability can be put to a wider range of uses.

A similar problem besets optical characteristics of visible surfaces other than shape and motion. Information about illumination, properties of the atmosphere, surface properties and surface structure gets compounded into simple measures of image properties, which cannot generally be decomposed uniquely into the contributory factors. For example there are well-known pictures which can be seen either as convex studs illuminated from above or hollows illuminated from below. A rooftop weather-vane seen silhouetted against the sky can also be ambiguous as to its orientation. Yet the human visual system has no difficulty in rapidly constructing unique interpretations for many such inherently ambiguous images – often the wrong interpretation! So it must, in such cases, be using some method other than reliance on a principled correct computation of the inverse of the image-formation process.

This is not to dispute that in some situations, or even most normally occurring situations, a great deal of the scene structure may be uniquely inferrable, e.g. from binocular disparity or especially from changing structure in the optic array – a point stressed by Gibson. The argument is simply that visual mechanisms seem to be able to deliver clear and unambiguous interpretations even in *some* situations where they have no right to. So it follows that they are able to use mechanisms *other* than principled inverse inferences from (changing) optic arrays to scene structures. Moreover, from the point of view of a designer, having these more general mechanisms is potentially more useful than being restricted to geometrical transformations.

Theoreticians faced with uncomfortable evidence grasp at straws as readily as squirrels grasp branches. A standard response to the problem of explaining how unambiguous percepts come from ambiguous data is to postulate certain general assumptions underlying the visual interpretation process and constraining the otherwise unmanageable inference from image to scene. Examples are:

- the "general viewpoint" assumption, (e.g. assume there are no coincidences of alignment of vertices, edges, surfaces, etc. with viewpoint),

- the assumption that objects are locally rigid,

- assumptions about surfaces such as that they are locally planar, mostly continuous, mostly smooth, not too steeply oriented to the viewer, mostly lambertian, uniformly textured, etc. (Gibson's own rule relating "equal amounts of texture" to "equal amounts of terrain" is based on such an assumption of uniformity),

- assumptions about the source of illumination, for instance that it comes from a remote point, or that it is diffuse, etc.

On the basis of such assumptions it is sometimes possible to make inferences that would otherwise not be justified.

These assumptions may well be useful in certain situations, but all are commonly violated, and a visual system needs to be able to cope with such violations. Instead of rigidly making such assumptions, a visual system has to find out the best way to make sense of currently available information, and this may involve violating one or more of these assumptions. For instance, if the size of texture elements on a surface varies across the surface then Gibson's rule has to

18

be violated. ((Scott, 1988) criticises assumption-based approaches to solving the problem of inferring structure from image correspondences.)

Another response is to postulate mutual disambiguation by context, subject to some global optimising principle. Constraint violations are dealt with by using designs in which different constraints are computed in parallel, and violations of some of them are tolerated if this enables *most* of the image to be interpreted in a convincing manner. (E.g. see (Hinton, 1976) and (Barrow & Tenenbaum, 1978)).

This requires the visual system to be designed as an optimiser (or minimiser): interpretations are selected that optimise some global property of the interpretation. Connectionist approaches to vision extend this idea (e.g. see (Hinton, 1981)). The measure to be optimised does not always seem to have any very clear semantics, as it depends on the relative weights assigned to different sorts of constraint violations and there does not seem to be any obviously rational way to compare different violations - though perhaps some kind of learning could account for this.

These "co-operative" network-based mechanisms may be part of the story and may even hold out some promise of explaining how high level hints (e.g. "look for the Dalmation" - see Frisby (1979) page 20) can help to direct low level processing in situations where image information is so radically ambiguous that there is no natural segmentation or grouping. A suitably structured network could allow some high level information to alter low level constraints or thresholds in such a way as to trickle down through the net and change the stable patterns that emerge from lower level processing.

The Ames demonstrations ((Gregory, 1970) (Frisby, 1979)), in which a distinctly non-rectangular room viewed through a small opening is perceived as rectangular, and a collection of spatially unrelated objects is perceived as assembled into a chair, suggest that in some situations what counts as globally optimal for the human visual system is either what fits in with prior knowledge about what is common or uncommon in the environment or what satisfies what might be regarded as high level aesthetic criteria, such as a preference for symmetry or connectedness. Note that a preference is not the same as an assumption: preferences can be violated, and therefore require more complex processing.

At any rate it begins to look as if vision, instead of using principled inferences from the structure of the optic array, may be highly opportunistic and relatively unprincipled. And why shouldn't it be, if that works well? Moreover, there may be higher level principles at work.

8.1 Higher level principles

A co-operative optimisation strategy may well be partly principled, in that the competing hypotheses are generated mathematically from the data, even if the selection between conflicting hypotheses is less principled.

The process may also be principled at a different level, for instance if the selection among rival interpretations of an ambiguous image is determined in part by previous experience of the environment, using a principled learning strategy, such as keeping records of previously observed structures and preferring interpretations that involve recognised objects.

Another kind of principled design would be the use, in some circumstances, of a mechanism

that favoured *rapid* decisions, even at the cost of increased error rates. This would be advantageous in situations where very rapid responses are required for survival. The satisfaction of getting things right is not much compensation for being eaten because you took too long to decide what was rushing towards you.

Mechanisms that favour speed against reliability would also be useful in situations where there is so much redundancy in images that quick and unprincipled processes generally produce the right answer. If most things that look approximately like branches are safely able to support the squirrel on its travels it does not need to go through detailed processes of analysis and interpretation that might be necessary to distinguish safe from unsafe branches in a less friendly environment. So for some purposes it may suffice to jump rapidly to conclusions (and therefore to branches) from branch-like characteristics.

Moreover, in such a cognitively friendly environment where unprincipled cues are statistically reliable guides, the processes controlling actions following a decision (i.e. running along the branch) may be able to make use of very rapid and partial visual analyses that guide motor processes in a tight feedback loop, even though in a less friendly environment they would be too unsafe. If many of the branches were dead and fragile, slower and more cautious mechanisms that carefully analyse the visual information would be needed, to reduce the occurrence of dead and rotting squirrels.

Evidence that human vision takes rapid decisions in the basis of partial analysis of the optic array takes many forms including undetected misprints in reading, cases of false recognition that are spontaneously corrected after the stranger is out of sight, and a host of accidents on the road, in the home and in factories.

Another meta-level principle is that effects of inadequate algorithms or data should be minimised. What this means is that the system should be designed so that even if it can't always get things exactly right, it should at least minimise the frequency of error, or be able to increase the chances of getting the right result by collecting more data, or performing more complex inferences. This is sometimes referred to as "graceful degradation" – not often found in computing systems.

It is far from obvious that these different design objectives are all mutually compatible. Further investigation of the trade-offs is required.

8.2 Unprincipled inference mechanisms

Even if it is true that working visual mechanisms do not use mathematically principled methods for inferring scene structure from optic array structure, this does not imply that mathematical analysis of the problems by Horn, Longuet-Higgins, etc. is a waste of time: on the contrary it is very important insofar as it helps to clarify the nature of the design task and the strengths and weaknesses of possible design solutions. This scientific role for rigorous analysis is distinct from its role in working vision systems.

If a totally deterministic and principled mathematical derivation from images to scene descriptions is not always possible, then the visual system needs mechanisms able to make use of the less principled methods, which may nevertheless satisfy the higher order principled

requirements sketched above. The most obvious alternative would be to use a general-purpose associative mechanism that could be trained to associate image features, possibly supplemented by contextual information, with descriptions of scene fragments. This sort of mechanism could work at different stages in processing: some of the associations might require preliminary grouping of fragments. It could also work at different levels of abstraction. The design of suitable mechanisms for such tasks is the focus of much current research on neural computation, and will not be pursued here.

If general associative mechanism are available at all in the visual system, they could be put to far more extensive use than indicated so far. For example, *the very same* visual mechanisms might be used to make inferences that go well beyond spatial structures derivable from the physical properties of the optic array. And indeed human vision seems to do that.

It is significant that the first example of vision mentioned in the textbook on psychology (Lindsay & Norman, 1977) is 'the conversion from the visual symbols on the page to meaningful phrases in the mind'. Here the detection of shape, colour and location of marks on paper is at most an intermediate phase in the process: the important goal is finding the meaning.

On the modular theory (in its extreme form), finding meanings (or meaning representations, in the sense defined in section 2.2) would be done only *after* the visual system has done its general purpose interpretation of the optic array and stored 3-D descriptions in some central database. Even if this is what happens in a novice reader, it appears that in a fluent reader the visual system itself has been trained to do new tasks, so that it no longer merely stores the same spatial descriptions in the same database. If a general associative mechanism can be trained to set up direct associations between visual structures and abstract meanings, why should it have to go through an indirect process, that would presumably be more complex and slower?

There is plenty of evidence from common experience that visual phenomena have a very wide range of effects besides providing new information about 3-D structures. The effects include being physically startled, reflexes such as saccades, or blinking, being aesthetically or sexually moved, and subtle influences on motor control and posture. On the modular theory, these effects would all be produced by non-visual systems reacting to a central store of 3-D descriptions produced by vision. The labyrinthine alternative offers a potentially more efficient design by allowing a visual system to have a broader role than producing descriptions of 3-D structures.

8.3   Is this a trivial verbal question?

It may appear that this is just a semantic issue concerning the definition of the term 'vision'. Defenders of the modular theory might argue that the broader processes include two or more distinct sub-processes, one being visual perception and the others including some kind of inference, or emotional or physical reaction. In other words, the labyrinthine theory is simply making the trivial recommendation that the words 'vision' 'visual' 'see' should be used to cover two or more stages of processing, and not just the first stage.

This, however, misses the point. I am not recommending that we extend the word "visual" to include a later stage of processing. Rather, I am countering the conjecture that there has to

be a single-purpose visual module whose results are then accessed by a variety of secondary processes, with the design proposal that the visual module itself, i.e. the sub-system that produces 3-D spatial descriptions, should also be able to produce a variety of non-spatial outputs, required for different purposes. This is not a question about how to define words.

If the very mechanisms that perform the alleged 'quintessential' task of vision are capable of doing more, are used for more in humans and other animals, and would be usefully designed to do more in machines, then far from being quintessential, the production of 3-D descriptions could turn out to be just a special case of a broader function. The design proposal and the empirical conjecture can be supported by examining more closely what is involved in deriving 3-D descriptions from optic arrays.

8.4   Interpretation involves "conceptual creativity"

It is not often noticed that on the modular model, the description of scenes requires a much richer vocabulary than the description of images or the optic array. This requires the visual system to have what we might call "conceptual creativity": a richer set of concepts is required for its output descriptions than for its input. This extra richness can include both the mathematical property of admitting more syntactic variability (as 3-D structures admit more variation than 2-D structures), and also the semantic property of describing or referring to a wider range of things.

A retinal image, or the optic array, can be described in terms of 2-D spatial properties and relations, 2-D motion descriptors, and a range of optical properties and relations concerned with colour or intensity and their changes over space or time. Describing a scene, however, requires entirely new concepts, such as distance from the viewer, occlusion, invisible surface, curving towards or away from the viewer, reflectance and illumination. None of these concepts is applicable to a retinal image or the optic array itself. I.e. visual perception, even on the standard theory, involves moving from one domain to another.

Conceptual creativity is characteristic of all perception, since the function of perception is rarely simply to characterise sensory input. It often includes the *interpretation* of that input as arising from something else. Hence descriptors suitable for that something else, namely features of the environment, are needed, and in general these go beyond any input description language.

This would not be the case if all that was required was classification or recognition of sensory stimuli, or prediction of new sensory stimuli from old. For classification and prediction, unlike interpretation and explanation, are not processes requiring conceptual extrapolation beyond the input description language. (I am here talking about classification of features or structures in a retinal image or optic array, not classification of objects depicted. The latter includes interpretation.)

This touches on a very old philosophical problem, concerning the origins of concepts not directly abstracted from experience. How can visual mechanisms go from a set of image descriptors to a significantly enlarged set of descriptors? Closely related is the question how scientists can go from observations to theories about totally unobservable phenomena. More generally how can anything relate manipulable structures to remote objects - i.e. assign a semantics to symbols or representations? (I have discussed these general questions elsewhere,

e.g. (Sloman, 1987b).)

Production of 3-D descriptions on the basis of 2-D features requires a mechanism with the following powers. When presented with stimuli which it can analyse and describe in a particular formalism, it should somehow associate them with a quite different set of descriptions, with different semantic variability. We have already had reason to believe that this association is not always a principled inference. It might, for example, be based in part on training using an associative memory.

Evolutionary history would determine the precise mechanisms actually used. For instance, a special-purpose visual mapping system might somehow have evolved into a more general associative mechanism, or a general associative mechanism might have became specialised for vision, or there might have been parallel developments for a time after which the two were combined.

8.5   The biological need for conceptual creativity

If a visual inference mechanism can make the conceptual leap from 2-D image descriptions to 3-D scene descriptions, is there any reason why *the very same mechanism* should not be capable of producing an additional set of biologically important descriptors?

From a biological point of view it would be very surprising if a perceptual mechanism of considerable potential were actually restricted to producing purely geometrical descriptions of shapes and spatial arrangements of objects and surfaces, perhaps enhanced by descriptions of optical properties. For, although these properties are of importance to organisms, so also are many other properties and relationships, such as hardness, softness, chewability, edibility, supporting something, preventing something moving, being graspable, movable, etc.

A powerful language for representing and reasoning about spatial relations might be an applicative language, with explicit names for spatial properties and relationships. The mechanisms for manipulating such a language would work just as well if the symbols named non-spatial properties and relationships. In fact, a representing notation is neutral except in relation to an interpreter. What makes certain symbols describe 3-D structures is the way in which they are interpreted and used. A visual sub-system that produces such symbols may know nothing of their interpretation if the semantics play a role only in higher level processes. Similarly, a suitably trained (associative) visual system could produce non-spatial descriptions which *it* couldn't interpret, but which 'made sense' to the part of the brain that received them, in virtue of the uses to which they were put there.

Biological evolution can be expected to search for general and flexible visual processing mechanisms. One breakthrough would be a mechanism which did not simply transform an input array of measurements to another array of measurements (eg. a depth map, orientation map, or flow field) but instead produced databases of descriptions of various sorts. Another breakthrough, if the modular account ever was true, might involve the ability to re-direct specialised output to other sub-systems, as required, instead of always going through a single channel to a central database. We'll return to these issues later. In order to provide a context for the discussion, let's now look at ways of classifying purposes of vision, in order to see what different outputs might be used for.

23

9   The uses of a visual system

There are several different ways of classifying the purposes of vision. For example, we can distinguish theoretical, practical and aesthetic uses. We can also distinguish active and passive uses.

- Theoretical uses
  Acquiring new information about the environment, forming new beliefs, or modifying old ones, checking hypotheses, answering questions, removing puzzles, generating new puzzles, correcting false beliefs, explaining observations, suggesting generalisations, producing new concepts. The beliefs affected by vision may be high-level conscious beliefs or low level details about the world that are used unconsciously in producing actions. Sometimes visual input gives an entirely new belief such as that there is a person in the doorway. At other times it merely modifies or amplifies a belief that was there already, for instance by providing more detailed information about the object in question, such as its precise shape, the speed at which it is moving, whether it is accelerating, etc.

- Practical uses
  Using visual input in relation to actions, e.g. making plans or choosing between options, monitoring and controlling execution, triggering new actions (reflexes), generating new motives (e.g. the desire to help someone or to eat a new visible tempting morsel), learning new skills from perceived examples, communicating with other agents, controlling other agents, e.g. by threatening them or visually indicating what is to be done. There appear to be several practical applications of vision that we are not conscious of, for instance using visual information to control posture and balance, and using it to control eye-movements. In many cases the practical use of vision requires not merely the perception of structure but also the perception of functional relationships and *potential for change*, as explained below.

- Aesthetic uses
  This is a very ill-understood function of vision, yet it seems to be very important in human life and culture. It is not so evident whether or to what extent this applies to other animals, since there is no unambiguous behavioural manifestation of aesthetic appreciation. Although aesthetic appreciation of objects is normally thought of as peripheral to vision, Guy Scott has suggested in personal communications that it may in fact be a basic function underlying other visual processes. At any rate it is found in all known human cultures, suggesting that it has some deep biological role.

Another way of classifying uses of vision is to distinguish active and passive uses.

- Active uses of vision
  These are cases where a goal is being pursued and the visual system is in some way controlled or directed by processes involved in achieving the goal. This includes searching for an object, attempting to answer a question, checking whether a goal has been achieved, using vision for fine control of actions, using vision to predict what will happen (e.g. extending a visible trajectory of a moving object), comparing two items to

see whether or how they differ, attempting to understand or interpret something, copying something, for example imitating a movement or making a sketch, learning how to do something.

- Passive uses of vision
  In these cases, events occur under control of incoming data rather than because they were brought about by a pre-existing goal or intention. This includes both *noticing* an object or event, and a range of phenomena in which a visual experience *triggers* a new process, for instance saccadic reflexes, a startled reaction, the occurrence of a thought or reminder, the production of a new motive, the detection of a violated expectation, and many aesthetic experiences, sexual reactions, reactions of disgust, and the like.

The distinction between active and passive uses is orthogonal to the distinction between theoretical, practical and aesthetic uses. For example an active practical use of vision would be the purposeful visual monitoring of an action in order to obtain fine control, whereas a passive practical use would be reacting to a totally unexpected and unlooked-for event by rapidly moving out of danger.

If vision is capable of being used both actively and passively this imposes global design requirements on the architecture of the system. Most current AI work seems to treat vision as passive, though work on movable cameras in robotics is an exception.

It is not always obvious how a visual system can function in active top-down mode, though it may be straightforward in special cases, such as checking how motion of an object under observation continues, since the observed location and previous motion of the object constrains the search for its "next" location (as in (Hogg, 1983)). In most cases, however, there is no simple translation from a high level hypothesis or question (such as "Where is the telephone?") to low level questions for feature detectors, segmentation detectors, and the like. Perhaps the most that can usually be done is to direct visual attention to an appropriate part of the scene or optic array, then operate in bottom-up mode, letting low level detectors, re-tuned if appropriate, find what they can and feed it to intermediate level processes: this is simply top-down selection of input for bottom up processes. It may also be possible top-down to switch certain general kinds of processes on or off, or change their thresholds, such as increasing sensitivity to horizontal edges.

The human visual system seems to be capable of more direct and powerful top down influences than this re-direction of passive processing: very high level information can sometimes affect the way details are seen or how segmentation is done. For instance, there are well known difficult pictures that begin to make sense only after a verbal hint has been given, and many joke pictures are like this. The mechanisms for such abstract top down influence are still unknown. Some cases might be handled by connectionist designs in which all processing is the result of co-operative interactions, including both visual input and also high-level expectations, questions, goals or preferences which provide additional inputs. How this works in detail, though, remains to be explained, especially as it presupposes a mapping from purposes, expectations, etc. to patterns of neuronal stimulation suitable as input to a neural net.

The different sorts of uses I've listed are not mutually exclusive. The practical purpose of controlling actions may be served in parallel with the theoretical purpose of acquiring

information about the environment in order to answer questions. A detective may enjoy watching the person he is shadowing. Whilst performing a complex and delicate task one can simultaneously control one's actions and be on the lookout for interesting new phenomena.

A full analysis of all the different uses and their requirements would need a lengthy tome. For now I'll simply elaborate on some of the less obvious points.

9.1   Subtasks for vision in executing plans

There are several different ways in which new information can be relevant to an intelligent system carrying out some plan. At least the following tasks can be distinguished:

- Checking achievement of goals and preconditions for actions.
  Often it is important at the end of executing a plan, or sub-plan, to check whether the effect has been achieved, or before starting a new action to check whether its pre-conditions are satisfied. This means that the visual system is given a particular question to answer: is the nail head flat against the surface? Are the two parts lined up so that the next step can be executed? Has the hand reached out far enough for the grasping action to begin? Is the car far enough into the garage for the door to be shut? Is the road clear enough to be safe to cross? Has the squirrel reached the point on the branch above the bag of nuts? I've already commented on the difficulty of accounting for such top-down processing.

- Providing information about discrepancies.
  If a goal has not been achieved, or a precondition is not satisfied, then, instead of producing a full description of the situation, it may suffice for the visual system to describe the nature of the discrepancy. For example, in which direction should an object be moved, or how far should motion continue? In some cases a 2-D projection of the discrepancy is enough. This sort of restricted information may be much simpler to compute than a complete description of the shapes of all the objects involved and their spatial relationships. For example, checking the visual distance between the edges of a pair of approaching surfaces may be simpler than describing their shapes, their orientations in space, and so on. Whilst trying to get a chair through a narrow doorway by a combination of movements and rotations, it could be quite difficult to represent the total 3-D situation and plan appropriate motion. An easier task might be to make a plan involving getting successive parts of the chair through the doorway, using perceived 2-D discrepancies to control the action.

- Continuous monitoring and control.
  A generalisation of static checking of goals, preconditions and discrepancies is the use of vision to supply continuous feedback in a motor control loop. Continuous feedback can lead to finer control and robust execution of plans. A particularly common case is visual tracking by the eye: here the result of the action controls its trajectory. The squirrel running along the branch probably has to be continually making fine adjustments to its acceleration and velocity. It is not at all obvious what information is required for doing this, nor how it is used. It might, for instance use the rate of change of some 2-D aspect of the optic array rather than 3-D spatio-temporal changes.

Ordinary life teems with examples of visual control and monitoring, even for those of us who don't leap through tree tops, for instance walking or running on a narrow pathway, parking a car, pouring a liquid from one container to another, running to catch or intercept a moving object, controlling the motion of a pen, or a paint brush, aiming a hosepipe or paint-spray, and so on.

If information comes too slowly in a feedback loop the result can be "hunting", or even complete disaster, such as the car crashing into the wall or a squirrel failing to catch a branch as it leaps through tree tops. It is therefore particularly important to take advantage of any opportunity to compute the minimum required, if that can improve the speed of feedback. This speed requirement has important implications for the design of the system. For example, speed may be traded for accuracy and reliability in some situations: and when this works we can say that the environment is 'cognitively friendly', in the sense that it allows partial processing to suffice. (There are several other dimensions of cognitive friendliness.)

- Noticing unexpected relevant information.
  During the course of executing a plan, new dangers, problems, and opportunities may arise that need to be detected even though there is no specific provision for them in the plan. Since by definition these are not things that can be specifically predicted or looked for this is a passive use of vision. Yet it may include setting up specific monitors or "demons" operating on lower level descriptions instead of just waiting for 3-D outputs.[3] The extent to which this is done can vary, and ordinary language indicates this by describing actions as involving more or less care, attention or caution.

In some cases, simply lowering thresholds for lower level processes (e.g. for 'change detectors') might suffice for achieving greater receptiveness to new information that might imply a need to change the current plan or action. However people appear to be capable of being trained to detect specific signs of danger, and this could involve the creation of subroutines that can be turned on or off, rather than always being active once learnt (Ullman, 1984). In that case being more cautious might involve turning on specific detectors relevant to the current situation and current task, which then react passively to incoming information.

9.2   Perceiving functions and potential for change

What kinds of information can be obtained from the optic array to serve all these different purposes? I have previously discussed the need for conceptual creativity, i.e. the ability to map structures in a 2-D image or optic array onto objects or relationships in some totally different domain, such as a domain of 3-D structures. In this section I shall discuss more abstract domains of interpretation required for perception of physical objects, and in a later section move on to even more conceptually creative forms of perception, namely those required for dealing with

---

[3]This was suggested in chapter 6 of (Sloman, 1978), available online here
http://www.cs.bham.ac.uk/research/projects/cogaff/crp/chap6.html

other intelligent agents. These all seem to be closely related to Gibson's notion of perceivable affordances.

Although perception of 3-D structure is important, it is often equally important to perceive potential for change and causal relationships, including the kind of potential for change and causal relationships that we describe as something having a certain function: for example seeing the cutting capability of a pair of scissors requires seeing the potential for relative motion of the two blades and the potential effect on objects between them. Seeing A as supporting B involves seeing A as blocking the potential for downward motion of B. By analogy with modal logic, I call these facts modal facts about physical objects, and descriptions of them modal descriptions.[4]

Not all the theoretical possibilities are usually perceived. For example, every surface has, in principle, the mathematical potential for all kinds of deformations, including developing straight, curved or jagged cracks, becoming wrinkled or furrowed, folding, shrinking, stretching, etc. However only a subset of these logical or mathematical possibilities will be relevant to a particular perceiver in a particular situation, and different subsets may require different kinds of descriptive apparatus, some of it expressed in terms of changes that can occur in the objects and some of expressed in terms of opportunities for action or constraints on action by the perceiver.

These "functional" or causal aspects of physical structures are not directly represented by the kinds of geometrical descriptions that are typically used to represent shapes in a computer, for instance in terms of coefficients in equations and topological relations between vertices, edges and surfaces. It may be possible to *derive* the information about possibilities from the geometrical descriptions, but the derivation is likely to be a complex process, and if a visual system can be designed or trained directly to associate such information with aspects of the 2-D input array, just as it appears to be able to associate 3-D structure, then the direct association may be more suitable for rapid processing than a two stage procedure in which the 3-D structures are first described and then the more abstract properties and relationships computed.

This view has something in common with Gibson's notion that perception of affordances is direct, though our accounts are subtly different. Gibson means that vision is "a one-stage process for the perception of surface layout instead of a two-stage process of first perceiving flat forms and then interpreting the cues for depth" (op.cit. p 150). My use of the word "direct", by contrast is intended to imply only that aspects of the 2-D input array (not necessarily a flat image on a surface) can be directly associated with abstract descriptions, instead of *always* depending on a prior process of production of 3-D descriptions. But this does not rule out a prior stage of analysis of the 2-D structure of the optic array. So I am simply saying that (some) non-spatial descriptions can (sometimes) be computed as directly as 3-D spatial descriptions. I am not saying that that process is as direct as Gibson suggests.

If it is true that our perception of causal and functional relations does not have to depend on prior creation of 3-D descriptions, then this might account for our natural tendency to say things like "I *see* that the plank is propping up the shelf (i.e. preventing realisation of its potential for downward motion)", rather than "I *infer from what I see* that the plank is propping up the

---

[4]Footnote added in 2006: there is another use of 'modal' mean linked to a particular sensory modality, and 'amodal' meaning not linked to any particular sensory modality.

shelf". Gibson (op.cit. page 138) quotes Koffka and Lewin as making similar remarks about the directness of many forms of perception, though he criticises them for treating the perceived 'valences' as phenomenal or subjective. The potentialities and relations between potentialities that I have been discussing are not subjective.

Exactly what kind of language or representational formalism is suitable for expressing these modal facts about spatial relationships, or, put another way, what internal substates in an animal or robot can store the information in a useful form, is a hard problem, and is likely to depend both on the needs and purposes of the agent and also what it is able to discriminate in the environment. But for now I shall simply assume that some suitable language or representation or set of addressable substates is available. The claim then is that it would be useful for a visual system to be able to include such descriptions or representations of modal facts in its outputs. This is just a special case of what Gibson apparently meant by perceiving "affordances".

Seeing something as a window catch, or seeing a plank as holding a shelf up, is potentially useful in selecting, synthesising or guiding actions: the catch must be moved if the window is to be opened, and the plank must be moved (or broken, etc) if the shelf is to be brought down. (Brady, 1985) uses the design of some familiar tools to illustrate our ability to perceive the relationship between shape and function.

So in general it is not enough to perceive what is the case. We also need the ability to perceive what changes in the situation are or are not *possible* and also *relations between possibilities*. For instance, in order to understand the window catch fully one must see that whether movement of the window is possible depends on whether rotation of the catch is possible. So perception of function sometimes depends on perception of second order potentialities.

Both the examples involve seeing *potential for change* in the situation. This includes seeing the constraints on motion, the possibilities left open by those constraints, and dependencies between the possibilities. The shelf cannot move down but it would be able to if the plank were not there. The plank would cease to be there if it were slid sideways, which is possible. The catch can rotate, removing a restriction on motion of the window.

This ability to detect and act on possible changes inherent in the structure of the situation and the relationships between different possibilities is not merely an adult human capability. However, it is not always clear to what extent the perceived possibility is explicitly represented, and to what extent combinations of goals and perceived structures are mapped directly onto actions by stored associations without going via explicit representation of modal facts.

Does a dog perceive the possibility of using a paw to restrict the possibility of motion of the bone off which it is tearing meat, and does the squirrel perceive the possibility of the branch supporting it upside down as it attacks the bag of nuts, or do they simply 'respond' to the combination of current goal and detected 3-D structure in the situation, using stored associations? Perception of possibilities seems to be needed for planning action sequences in advance (as well as for other tasks like explaining how something works). But it may be that for "reflex" or trained actions the possibilities themselves are not explicitly represented, and instead the result of visual processing is direct control signals to motor-control systems.

A lot depends on task analysis: until we know in detail how certain tasks could or could not

be performed it is hard to speculate about other animals. However, the process of assembling an intricately constructed bird's nest looks as if it must involve at least local planning on the basis of perception of possibilities for change. Similarly I've watched a very young child accustomed to levering the lid off a large can with the handle of a spoon, baffled one day by the lack of a spoon, eventually see the potential in a flat rigid disk and use that as a lever by inserting its edge under the lid. He saw the potential for change in a complex structure and then knew exactly what to do. Perhaps only a subset of animals can do that. Kohler's apes could not do it in all his test situations.

People also see causal relations in changing situations: the billiard cue is seen to cause the ball to move, the cushion is seen to cause the ball to change direction. Michotte's studies of human responses to displays of two squares moving in one dimension indicate that relatively impoverished information about relative motion in the optic array can determine a variety of different causal percepts, such as colliding, launching, triggering, and passing through, with the interpretation sometimes influenced by non-verbal context or by the visual fixation point (Michotte, 1963).

All these examples of abstract perceptual capabilities raise the question whether we are talking about a two stage process, one visual one not. On the modular theory, vision would yield a description of spatial structure, then some higher level cognitive process would make inferences about possibilities and causal relations. Of course, this sometimes happens: we perceive an unfamiliar structure and explicitly reason about its possible movements. The alternative is that the visual system *itself* is designed or can be trained to produce 'directly' not only 3-D structural descriptions, but also descriptions of possibilities and causal relationships, so that the two sorts of interpretations are constructed in parallel, in at least some cases. (I am not claiming that all such affordances are detected infallibly.)

Whether this direct perception of modal facts ever occurs is an empirical question. It is not easy to see how it could be settled using behavioural evidence, though reaction times might give some indication, if combined with detailed analysis of the task requirements for different kinds of observed behavioural abilities. Anatomical and physiological studies of how the brain works may also help by showing some of the routes by which information flows. From a design point of view the main advantage of the labyrinthine mechanism would be speed and economy. It may be possible to avoid computing unnecessary detailed descriptions of spatial structure in situations where all that is required is information about potential for change inferrable directly from fairly low level image data, perhaps with the aid of prior knowledge and current goals.

One of the unanswered questions is how possibilities for change and other abstractions should be represented. If the visual system is able to represent actual velocity flow in a 2-D map of the optic array, as many researchers assume it can, then a similar symbolism or notation might be used for representing the spatial distribution of possible movements.

Although the representation of potential for change, and other modal information, appears to be of profound importance for intelligent planning and control of actions, I know of no detailed investigation of the kinds of representational structures that will support this, or algorithms for deriving them from visual information.

A naive approach might be to try to represent all the different possible situations that could or could not arise from small changes in the perceived situation. How small should the changes be?

The larger the allowed time, the more vast the space of possibilities. In any moderately complex scene explicit representation of all possible developments will be defeated by a combinatorial explosion, since there are so many different components that can move in different ways.

One strategy for avoiding the explosion is to compute only possibilities and constraints that are relevant to current purposes. This requires some "active" top-down control of the interpretation process. Another strategy, also relevant to the description of empty spaces (see below), is to use summary representations in which the different local possibilities are represented by abstract labels, which can be combined as needed for purposes of planning or prediction. For example, describing an object as "pivoted at this edge" implies that it can rotate about the edge in a plane perpendicular to that edge. Given this summary description, it may not be necessary to represent explicitly all the different amounts and speeds of rotation. It might be useful to build a 2-D map in which each visible scene fragment has a label summarising its possible movements. (Topographic maps of the optic array are discussed below.)

Representing possible *relative* motions is harder. Longuet-Higgins has suggested (op.cit. p306) that the human visual system may possess channels tuned to four basic types of relative motion. Activation of units associated with such channels when the motion is absent might be one way of representing its possibility. Representing impossibilities, like the impossibility of a shelf falling while a plank is propping it up, is more complex: it requires the representation of a possibility and something to indicate its unachievability.
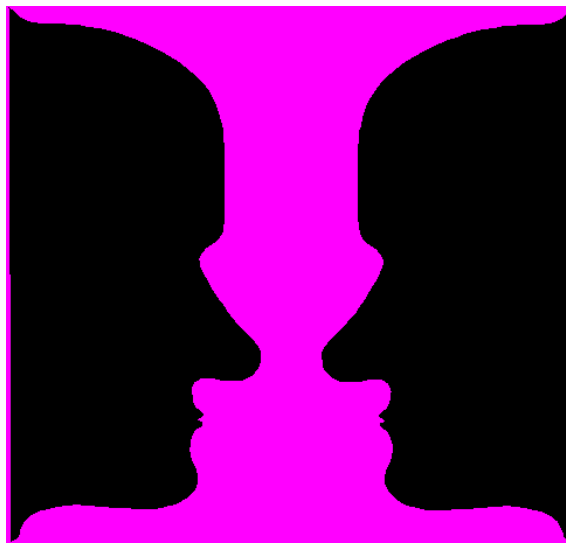

Figure 1: *Caption: This can be seen as two faces, or as a vase, or as a vase wedged between two faces.*

9.3   Figure and ground

It is often noticed that perception involves a separation of figure from ground, as illustrated in figure 1. Exactly what this *means* is not easy to explain. It is more than just the perception of 2-

D or 3-D structure. My suspicion is that it involves quite abstract relationships analogous to the modal relations just discussed, including the notion that the image elements forming the figure in some sense belong together. The concept of being part of the same object is a deep concept often used without analysis in designing segmentation algorithms. For example part of the concept seems to involve the possibility of common motions and restrictions on possibilities of independent motions. A full study would require detailed analysis of the concept of an "object", a concept that is generally taken for granted, yet fundamental to intelligent thought and perception.

Evidence for the general lack of understanding of the concept of figure ground separation is the often repeated claim that in the vase/faces figure it is possible to see either the vase as figure and the rest as ground, or the two faces as figure and the rest as ground, but never both at once. This is just untrue: people who try can easily see the picture as depicting two faces with a vase wedged between them. The lines in the picture then depict cracks between adjacent figures, rather than occluding edges. This, incidentally is an example of the way top-down suggestions can make a difference to how things are seen.

The notion of figure, therefore, is not inseparably tied to the notion of a "background" to the figure. If it were then the alleged impossibility would exist, since it is impossible for A to be nearer to B at the same time as B is nearer than A. How does the concept work then? Part of the answer is that figure ground separation is related to the concept of an enduring object. The "figure" is conceived of as an object composed of portions capable of moving as a whole, without the rest of the scene. One implementation for this might be treating an object as an entity to which labels describing potential for change can be attached, with related labels attached to the different parts, indicating the mutual constraints on possibility of movement.

So it may be that even perception of the environment as composed of distinct objects sometimes requires the production not only of descriptions of spatial structure and motion, but also of far more abstract relationships between possibilities and impossibilities in parts of the scene. The full semantics of such descriptions will be determined by the limitations on how they are used by the agent, e.g. how they affect planning, reasoning, predictions and motor control.

I am not claiming that the idea of common possibilities for motion suffices to define the concept of an "object" or "figure". This is just a special case of a more general role for segmented objects, namely that they can enter as wholes into relationships and have properties ascribed to them. In other words they can occur in articulated representations, described below. What counts as a "whole", or how segmentation is to be done, will depend on internal and external context. Whether a portion of water is seen as a whole can depend on whether it forms a puddle in the road or an undifferentiated part of a lake, or whether it is the intended target for a diver poised on a diving board.

9.4   Seeing why

Closely related to perception of function, constraints, and potential for change is the use of vision to provide *explanations*. Very often one knows some fact, such as that an object is immobile, or that when one thing moves another does, but does not know *why* this is so.

Knowing why can be important for a whole range of tasks, including fixing things that have stopped working, or changing the behaviour of something so that it works differently. Vision is often a powerful source of explanatory insight.

A verbal description of the mechanism of a clock would be quite hard to follow, whereas seeing the cogs, levers, weights, chains, etc. can make the causal connections very much clearer, and can give insight relevant to controlling and predicting behaviour. Similarly, it is possible to fold a sheet of paper into the form of a bird with the entertaining property that it flaps its wings when the tail is pulled. Close visual examination explains why, whereas describing the structure and relationships in words is quite hard. There is something about the visual presentation of information, including not just geometrical information, but also causal and functional information, that seems to make use of powerful cognitive mechanisms for spatial reasoning in humans, a fact that is increasingly being used in human-computer interfaces. Graphs, charts, trees, diagrams, maps etc. have long been preferred to tables of numbers, equations or lists of facts, for some purposes.

A possible way of thinking about this is to note that all reasoning, whether logical or visual, requires symbolic structures to be built, compared, manipulated. It may be the case that mechanisms have evolved for manipulating the spatial representations created at various stages in visual processing and that some of these manipulations are useful both for the interpretation of images (which requires inference) and for other tasks, generally thought of as more cognitive, or more central, such as predicting the behaviour of others, or understanding how things work. If this (often re-invented) idea is correct then instead of being a self-contained module separate from cognitive processes, the visual system must be inextricably linked with higher forms of cognition.

One indirect piece of evidence often cited for this is the prevalence of spatial metaphors for talking about difficult non-spatial topics. For example, programmers often use flow charts to represent algorithms. Another commonplace example is talk about a "search space" and its structure. We can also think about different search algorithms in spatial terms, and use diagrams and other spatial representations for them, for example when we talk about depth-first and breadth first searching. Similarly physicists talk about "phase spaces". Computer programmers often use relationships between spatial and abstract structures, for instance the fact that depth first search corresponds to a last-in/first-out STACK of options, whereas breadth first search corresponds to a first-in/first-out QUEUE of options. Another example is the relationship between two nested "for" loops and a path through a 2-D array. (The generalisation to higher dimensions is harder for people to visualise.)

Alas, the increasing use of microelectronics means that we can make less and less use of our biological endowments to understand the machines around us, and we have to depend increasingly on abstract logical and mathematical explanations.

9.5   Seeing spaces

Another aspect of the practical role of vision involves the perception not of objects but of empty yet structured spaces. A simple example is perception of a hole or doorway capable of being used as a way in to an object or room. A more complex case is perception of a possible route

across a cluttered room, where the route is constructed from a succession of spaces through which it is possible to walk or clamber. Seeing gaps, holes, spaces and routes is closely bound up with seeing the potential for change in a situation. There are toys that help children learn to see such relationships – seeing the relationship between the shape of an opening and the action required to insert a tight-fitting object is not innate in humans and apparently does not develop for several months. Yet for adults the relationship is blindingly obvious: what has changed? Perhaps this uses the same mechanisms as learning to read, after which the meanings of written words cannot be ignored when we see them.

It might be useful if complex abstract descriptions of potentiality for motion, and constraints on motion, could be collapsed into single functional labels, something like "hole", "furrow", "exit", "opening", etc. Perhaps practical need trains the visual system to create and apply such labels on the basis of low level cues, leaving other subsystems to interpret them. But how? These are not simply geometrical descriptors but provide pointers to functional or causal information about what can happen or be done. From general labels relating possible changes and causal relationships it is a short step to functional descriptions like "lever" "pivot" "support" "wall", "container, "lid", etc. which summarise a combination of possibilities and constraints on motion.

These are still all very sketchy design conjectures and much work remains to be done, classifying different sorts of compact functional and modal descriptions and showing (a) how the need for them might be learnt, (b) how they can be derived from images and (c) how they can be used for planning and the control of actions. Let's now look at yet more abstract visual descriptions.
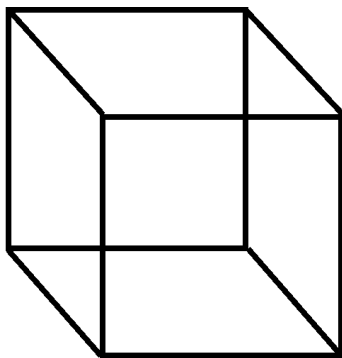
9.6   Seeing mental states

Figure 2: *The "flip" in this figure is describable in purely geometric terms (e.g. "nearer", "further", "sloping up", etc.)*
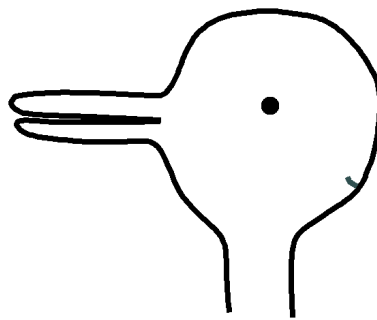
Figure 3: *The "flip" in this figure is not a purely geometric one: it faces in different directions, and parts change their functions.*

I shall try to show that we can use well known kinds of visual ambiguities as pointers to wide variations in the kinds of information handled by visual systems, though as always the

arguments are suggestive rather than conclusive.

Compare figure 2, the Necker cube, with figure 3, the duck-rabbit picture. Both are standard examples of visual ambiguity. In both cases the picture can 'flip' between two interpretations, where each interpretation corresponds to a distinct visual experience. If people are asked to describe what is different about the two views of the same figure, then in the case of figure NECKER, the answer supports the standard modular view of vision, for the two experiences differ in terms of how the lines are mapped into three dimensional spatial structures and relations. Before the flip one square face appears nearer the viewer, and after the flip it is further. Similarly the 3-D orientations of lines flip between sloping up and sloping down. These changes in perceived 3-D structure are what one might expect on the modular view of vision as concerned with the production of descriptions of spatial structure.

The visual 'flips' that people experience with figure 3 are very different. There is no significantly different perceived spatial structure in the two views. Instead, parts are given different *functional* descriptions in the two views: ears flip to become the duck's bill. A mark flips from being meaningless to being the rabbit's mouth. It is as if the labelling of parts as having a function is somehow 'painted' into the image: 'bill' or 'ears'. More subtly, the front and back of the animal flip over. The rabbit faces one way, the duck the other way. It is hard to explain what this means, but I think it can be expressed in terms of perceived possibilities for action and perception in another agent.

The notions of "front" and "back" are linked both to the direction of likely motion and also to what the creature can see. For intelligent perceivers both of these characterisations of a perceived agent could be very important. It is often useful to know which way prey or enemies are likely to move and what they can see. If the visual system, by virtue of its ability to store arbitrary useful associations, is capable of producing abstract descriptions of the possibilities for change in purely mechanical systems, then perhaps the same mechanisms could be made to produce descriptions of potential movements and potential percepts of other agents.

Of course, I am not suggesting that the information is encoded as we might describe it in English, any more than information about shape, or possibilities for motion are necessarily encoded in words or any other propositional form. All that is required is that information-rich sub-states be created that are accessible by other processes that need the information. The theoretical design of suitable forms of encoding of all this information, and empirical investigation to see which are used by people and animals are still difficult tasks that lie ahead. My conjecture is that in visual processing information is stored in a form that makes it accessible via some kind of map or index based on the 2-D structure of the optic array. This is what makes us say the two views *look* different, rather than simply saying that the image reminds us of different things, or that we can infer different things from it.

On this theory the "flip" between duck and rabbit percepts might involve something like different "visible by X" labels being planted into the scene map just as orientation labels, or depth labels are planted in the case of the Necker cube, and labels describing functions or modal facts in the cases of perceived causal relations discussed earlier.

If this is correct, the processing would occur within the visual system, since it would require access to the intermediate visual databases. This use of vision, like labelling directions of potential movement, would be useful for planning actions or predicting what a perceived agent

will do next. For example if you are attempting to collaborate with someone it may be important to know where you should put something so that he can see it, and if you wish to catch prey it will be useful to know where to move in order not to be seen.

By contrast, on the modular view, high level inference mechanisms would need to reason from 3-D scene descriptions plus prior knowledge that the duck can see certain things rather than others. This sort of reasoning, like a detective's deductions, would not produce the characteristic "feel" of a change in how a picture is *seen.* It would probably take longer too. So it is neither accident nor error that so many text books on vision include both the cube and the duck-rabbit as examples of the same kind of thing: a *visual* flip, rather than treating one as a visual ambiguity and the other as an intellectual non-visual puzzle, as it would have to be on the standard modular theory.
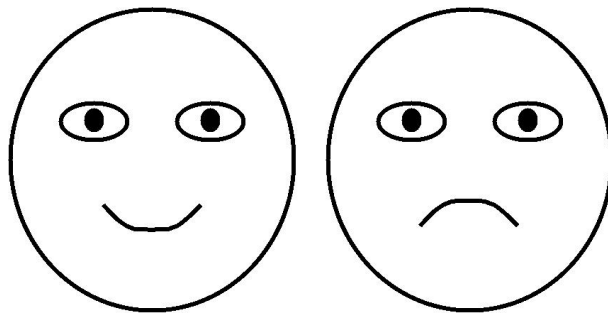
9.7   Seeing through faces

Figure 4: *Is the perception of happiness or sadness in a face visual, or is it a post-visual inference? Do the eyes in the two faces look the same? If not, why not?*

This ability to see which way another agent is looking could be just one among a large variety of ways in which vision is used to provide information about mental states of other agents, just as it provides information about unobserved physical states like rigidity and causal relations. Visual perception of other agents also illustrates another theme of this paper, namely that besides producing descriptions visual processes may produce control information that is somehow fed directly to other sub-systems.

Visual experiences are capable of being very moving. A delightful and disturbing fact of human existence is the richness of emotional interaction produced in face-to-face situations. Sometimes it is almost as if we see through the spatial aspects of physiognomy to some of the underlying mental states. The two appearances of the duck-rabbit as looking left or right are special cases of this more general ability to see more than physical structure. This is apparently a deep-rooted feature of human vision. For example, it is difficult to see images like those in figure 4 as merely *spatial* structures.

It is as if we see the happiness or sadness in a face as directly as we see the concavity in a surface or the fact that two dots are inside a circle. So perhaps descriptions of at least some

mental states are part of the output language of the visual system, rather than an additional inference from perceived shape. This is very similar to the experience of fluent reading. These abstract visual capabilities are puzzling only if you forget that being able to output information about 3-D structure on the basis of information in one or more changing 2-D optic arrays is no less puzzling. Both require conceptual creativity.

Moreover, in perceiving faces, we not only get factual information about the state of the other agent, we also seem to have a large collection of automatic and largely unconscious responses (including eye movements and facial expressions), that play an important and very subtle role in the management of interpersonal relationships. The powerful effect of an infant's smile on doting parents is just the beginning of a complex feed-back loop that develops over the years, sometimes disastrously.

We sometimes see mental states and processes even in the absence of human or even animal faces and bodies. The experiments of (Heider & Simmel, 1944) using moving geometrical patterns show that many people spontaneously interpret patterns of movement of triangles, circles and squares, in terms of intentions and even emotional states of agents. This kind of thing is used to good effect in some of the more abstract cartoon films.

Of course, I am not able to say *how* these processes work - what precisely the features of the optic array are which can have these effects, nor how they are detected, how the information is encoded, what kind of associative mechanism relates the geometrical features to the mental descriptions, at what stage in the processing the information flows from the visual system to other systems, which processes are innate and which learnt, how exactly other systems use the information, and so on. All these are questions for further investigation.

9.8   Practical uses of 2-D image information

So far I have been arguing that in addition to spatial information a well designed visual system should be able to produce descriptions of non-spatial facts. It is also worth pointing out that for some purposes it is not 3-D scene structure that the visual system should produce but rather descriptions of 2-D structure in the optic array. So not all geometric output of vision has to be concerned with 3-D scene structure.

For example someone sighting a gun uses co-incidence in the retinal image or optic array rather than information about the 3-D relationship between gun and target. For many sorts of continuous control, it may be far simpler and quicker to use 2-D relationships, such as keeping the line of motion central relative to edges of a road or path-way, or moving towards a target by keeping in line with two "sighting posts" (an example suggested to me by Christopher Longuet-Higgins). A 2-D discrepancy measure may be easier and quicker to compute for the purpose of controlling action than the full 3-D discrepancy.

Perhaps this effective use of 2-D relationships is part of the squirrel's secret: for instance the task of remaining upright while moving at speed along a thin branch might use the direction of optical flow at the visible contours of the branch. If there is a component of flow to the right at the left and right edges of the branch, then the squirrel is falling to the right and should compensate by tilting to the left. (For crooked branches a more complex story is required.) For animals that mostly leap from branch to branch, like some monkeys and apes, or fly between

them, like nest-building birds, different aspects of the visual field may figure in the control of motion. A gibbon (or Tarzan) in mid air, arm outstretched towards the fast approaching branch, may do best to use the 2-D projection along the line of sight, of the relationship between hand and upper edge of the branch.

I am not talking about introspectively accessible 2-D information: in fact most of the kinds of information produced by a visual system do not need to be accessible to consciousness, since what we need to be able to reflect on and talk about, for instance in analysing failures or making plans, may be very different from what is required for normal ongoing interaction with the environment. Often people cannot consciously access 2-D image structure without special training. People see the corners of a table as rectangular and may find it very hard to attend to the acute and obtuse 2-D angles. Painters need access to such 2-D structure in the visual field in order to produce a convincing depiction, and they often have to learn to attend to the required information. But the important thing is that it *can* be done: so the visual system *can,* at least sometimes, output information about the 2-D structure in the projection of a scene to a viewpoint, when it is useful.

I am not disputing that full 3-D descriptions are useful for many purposes. If, however, intermediate, 2-D information is also useful output, that suggests that the visual system should not be construed as an inaccessible black box, whose output always takes a certain form. Instead it may be possible for a range of different processes to access intermediate data-stores. In fact it seems likely that some reflex responses do just that, for example the blinking response to an object rapidly approaching the eye, or the posture-controlling reflexes that seem to react to optical flow patterns. Muscular control of balance could depend on global patterns of optical flow which provide information about one's own forward or backward motion. Experiments reported in (Lee & Lishman, 1975) suggest that even when people are unconscious of experimentally manipulated global flow changes they react with muscular changes, and can even be made to lose their balance without knowing why.

Although further investigation is required, it is possible that (a) this process makes use of 2-D flow patterns and (b) the information goes direct to posture control mechanisms rather than having to go through a central general purpose database recording a change in distance to the wall ahead. The latter design would require extra stages of processing and might therefore provide slower feedback to posture control mechanisms, a serious problem for inherently unstable upright two-legged animals or fast-moving squirrels.

Moreover, it is far from obvious that the most effective design for the purposes of recognising 3-D objects is always to use general methods to infer 3-D structure (describable in an "object-centred" frame), *and then* attempt recognition, rather using recognition of 2-D structure as a cue into information specific to the object. The latter requires that a range of viewpoint-dependent 2-D views of the object should be stored, and is therefore costly in storage, but has the advantage that 2-D structure matching is inherently less complex than 3-D structure matching. So we have a space-time trade-off here that could favour 2-D structures when speed is important, though neither strategy should be adopted exclusively.

Which is better will depend on task-relative trade-offs. For example, if an object has a relatively small number of distinct views that have a common structure adequate for discriminating it from other objects in the environment (like a fairly flat star-fish), or has

features that project into distinctive 2-D patterns (like a Zebra?) then using 2-D structure will be useful, unlike the case where the only way to use 2-D information reliably for recognition would be to use a very large collection of different views all generated from an invariant 3-D structure. I suspect that inferring 3-D structure and topology prior to matching is likely to be the best strategy with non-rigid objects, like sweaters, which can generate a huge variety of 2-D projections when crumpled, folded, worn, etc.

The usefulness of using stored 2-D views will also depend on how often the objects have to be perceived, how quickly they have to be recognised or discriminated, and what the costs of delay are. We probably learn to recognise a range of 2-D views of people we are close to, just as we learn to recognise their footsteps and all manner of indications of their presence or actions. Similarly a boxer may have to learn to react to a variety of 2-D cues in order to be able to take very rapid evasive action, though in this case it is not just descriptions that are required from the visual processing, but direct control signals to produce the necessary response.

9.9 Triggering and controlling mental processes

Besides triggering physical responses visual stimulation can trigger new mental processes. During conventional processes of learning to read text there is a first stage of learning to discriminate and recognise written marks (e.g. letters or letter clusters) and associating them with sounds (either portions of words or whole words, depending on the teaching strategy). The sounds, or combinations of sounds, being previously understood, are then used to make the links to meanings. By contrast, fluent reading, as remarked previously, seems to involve direct stimulation of complex processes that manipulate semantic information about whatever is represented in the text, by-passing phonetic representations. The process also seems to by-pass recognition and checking of printed characters or words.

This suggests that combinations of low-level features may be directly associated with lower-level units in non-visual non-motor modules in the brain. Direct stimulation of such modules could invoke non-visual processes, such as the construction of sentence interpretations, and many other mental processes.

There are several other examples from ordinary experience. One is being reminded of something: seeing one thing makes you think of another related thing. Often what is triggered is a new motive, for example a desire: seeing food, or a picture of food, can make you want to eat, seeing someone in distress can make you want to help. In many animals perceived displays apparently produce sexual desires. Visual stimuli can also have powerful aesthetic effects. Some visual reflexes seem to be part of the machinery involved in human and animal emotions (Sloman, 1987a).

In addition to initiating or triggering a new mental process, the visual system seems to be capable of ongoing control of enduring mental processes, as for example during the reading of a story: this can even take on some aspects of experiencing the events related, including joy, and sorrow sufficient for tears. A different case is the use of an external structure to store information about and control reasoning about some abstract problem. The use of diagrams in geometrical reasoning has something of this character, as does visual examination of an object, or a picture of an object, or a working model of the object, in order to gain an understanding of

how or why it behaves as it does.

The existence of these phenomena is not controversial. What is at issue is whether all these responses go via a central database of scene descriptions as the modular theory would imply, or whether some of them are produced more directly. If there are mechanisms for direct triggering of physical reflexes, without going through a general purpose database of descriptions, it is at least possible that similar mechanisms could directly trigger or control other mental processes, in some cases after appropriate training (discussed below). Exactly which kinds of human mental processes are directly driven by special-purpose output from the visual system is an empirical question.

10   Varieties of visual databases

I have argued that there is no reason to restrict the output of a visual system to be descriptions of spatial structure and change, and have suggested that (after suitable training if necessary) information of arbitrarily abstract kinds may be produced along with concrete geometrical information. However, there do seem to be some kinds of processing that are characteristic of vision, and have to do with the fact that the bulk of the information, and certainly most of the fine detail, comes via 2-D optic arrays. This is the basis of the idea put forward by (Barrow & Tenenbaum, 1978) that visual systems produce a collection of different databases of information in registration with input images. Others have referred to these as 'topographic maps', e.g. (Barlow, 1983).

This does not necessarily mean that the databases are arranged as regular rectangular arrays as commonly happens in computer models: for they might be hexagonal, or concentric rings (Young 1989), or simply irregular. As Vaclav Hlavac has pointed out to me, a visual mechanism might learn to make use of an irregular system for sampling the optic array. The precise form of connectivity and addressing modes within visual databases can vary as long as useful relationships like relative closeness and (approximate) direction are preserved. This indexing by 2-D spatial relationships allows questions like these to be answered relatively efficiently:

Is there an X near here?

If I start here and scan in this direction will I find an X?

Doing this in relation to a 2-D index is a good heuristic for cutting down the search space for the corresponding 3-D questions.

It may be useful now to sketch out some of the typical kinds of intermediate information that appear to be useful during visual processing, including some that are not indexed by location in the optic array, but in other 'spaces', e.g. histograms associating features with numbers of locations that have the feature.

I'll start my list with the most highly processed structures, of kinds that might be output on the modular theory, and continue through some less obvious kinds of intermediate databases. On the standard modular theory these would be used only *within* the visual system, as part of the process of producing descriptions of 3-D shape and motion. On the labyrinthine theory, their contents might be available to other modules that can make good use of the information.

- Descriptive databases

  In these, structures of arbitrary complexity, either in the image or in the scene are given explicit labels and are explicitly related to their properties, their parts, and their relationships to other labelled structures, the parts, properties and relationships also having explicit labels. A parse tree is a typical example of such a structure, though, for vision, networks generally seem more useful than trees. Logical languages and semantic nets are examples of formalisms for constructing such databases. Descriptive databases can serve a variety of purposes including reducing the amount of information to be processed during recognition, planning, or control; providing a viewpoint-independent representation of the scene; allowing generalisations to be made by abstracting from individual components; making general purpose inference mechanisms applicable for combining new specific information to general information, and so on.

  2-D maps of optic array information could include pointers to nodes in these high level descriptions, and the descriptions could include pointers back to the maps. However, when the viewpoint changes, the whole structure would have to be re-built, which could be very inefficient since the environment does not change. Further the links to the maps would need to be updated rapidly, a non-trivial processing task. The complexity of the task would be reduced if there are good strategies for systematically transforming the maps and their links on the basis of knowledge about the viewer's trajectory through space, instead of continually re-building the maps and derived structures from scratch. This would be an example of the way in which information about the agent's own motion and previous percepts could be important inputs for visual processing.

- Articulated but implicit descriptions

  In this kind of database, structures are linked together, and new nodes formed to represent linked wholes, and these have links to their parts and to other related structures. But there are no labels categorising the nodes. Instead, all the information about what the structures are is implicit in the ways things are linked together. For example three points and three lines suitably related would constitute a triangle, and a triple consisting of a point and two lines ending at that point would constitute a vertex of that triangle.

  An unlabelled parse tree for a sentence would be another example of an articulated implicit description.

  Construction of the network of links, i.e. the articulation of the information derived from the optic array, would normally be an important step towards the recognition and labelling of larger scale structures and their relationships, although in some ambiguous images the higher level recognition might be required in order to set up the low level links.

  If the components that are linked are themselves made of linked structures, the database is hierarchical-articulated, otherwise flat-articulated.

- Semi-articulated databases

  Structures are formed by linking things together if they belong to the same larger whole, but there is not necessarily any label or pointer to a whole that is accessible outside the linked structure. It may be possible to traverse a set of linked elements by starting from any of its parts and following links to their neighbours. But as there is nothing

representing the whole linked structure, there is no way of relating it to another such complete linked structure, so the structuring is all at one level.

For example, if edge points in an image are linked to neighbouring edge points with a similar orientation (and linked to at most two neighbours as a result of a 'competitive process'), then clusters of linked edges would form line segments. But in an unarticulated database there could be no link from one *set* of edges (a line) to another, since this would presuppose some explicit representation of the higher level structures.

The production of unarticulated databases is useful if local information and relationships provide evidence for linking things. Region growing and line growing algorithms can work like this, but will tend to get out of control and produce very messy results in complex images, if there is no feedback from higher level structures: one of the motivations for so-called 'heterarchic' processing.

- Pre-articulated databases
  In these, elements of the image or scene description have been labelled in some way to indicate implicitly which ones belong together, but they are not yet linked together, and there are no names for larger structures. For instance, if points of discontinuity in the optic array (edge points) have known locations and orientation discontinuities, then this represents a potential for linking edge points into lines, though not necessarily unambiguously. Similarly, if local elements of the optic array are labelled according to properties like colour, intensity, texture density, optic flow, etc. then this represents a potential for linking them into regions, again not necessarily unambiguously.

  In a pre-articulated database, from each element it is possible to discover what its features are but not possible to go directly from features back to the elements or from elements to others with the same or related features. Indexing elements by location, as in 2-D image maps, is one way of constraining the search for relevant elements to link, in order to build up a semi-articulated database.

- Non topographic transforms
  There are many kinds of transforms from an image to a database where spatial location is lost. Examples would be histograms recording numbers of optic array locations with a particular colour, intensity, intensity gradient, texture, optical flow, etc., or recording numbers of points falling within a range of values. Closely related are Hough transforms (explained in (Ballard & Brown, 1982)), in which each element of the original is mapped into a set of functions of properties of the element.

  Histograms provide a means of accumulating spatially disparate evidence in support of conflicting interpretations.

  If a histogram contains only measures of how many elements map onto each possible value then it gives no information about which parts of the image have contributed. If each 'bucket' contains descriptive, articulated, semi-articulated or pre-articulated information about the contributing portions of the image, it then turns into a separate mini-database linking items which are similar in certain respects.

  For example, it may be useful to map detected image features into an orientation

histogram. If, instead of simply counting contributions, each orientation record keeps a list of edge features with that orientation, this constitutes a database of information about (roughly) parallel image fragments. The Hough transform is often used to make a finer discrimination that stores information about collinear fragments. (I am ignoring problems about quantisation of orientations and other measures.)

Storing such pointers in histograms makes it possible to search for neighbours in a variety of abstract spaces while interpreting visual data: one process by which pre-articulated databases may be created.

- Feedback and spatial indexes
  If labels are created for the more abstract objects and relationships found during the interpretation process then it is possible for those labels to be "planted" back into the lower level representations such as pre-articulated databases or topographic maps. This may also be done by creating new "pseudo-images" in registration with the original array. This sort of (frequently re-invented) strategy seems to be what Marr referred to as the use of 'place-tokens' ((Marr, 1982), p.51), and what Barrow and Tenenbaum described as a collection of 'intrinsic images in registration'. The advantages of doing this were discussed above, e.g. it provides a useful spatial index for finding things during active visual processing, for example, working out what a moving object is likely to hit first, by projecting its trajectory into such a map.

- Object specific indexing structures
  In addition to linking information in topographic maps in register with the structure of the optic array and grouping items in more abstract histograms and databases, it may, for certain purposes, be necessary also to use specialised maps tailored to the perception of known types of objects. For example, when perceiving well known type of object that is not rigid it may be useful to have a topological map of its structure, into which is projected some of the detailed information about the particular individual: such as what the parts are like and what they are doing. Then subsequent searches for related information may not be bogged down by the problems of coping with the ever-changing 2-D projection of a non-rigid body.

  This object-related indexing of information is more or less what is currently known as "model-based" vision ((Ballard & Brown, 1982) p.217ff, (Hogg, Sullivan, Baker, & Mott, 1984) (Hogg, 1988)). If the maps have a simple enough structure, they can be manipulated (e.g. searched) using mechanisms similar to those that work on topographic maps. However, more general operations on topological models, for instance looking to see whether one network is a sub-network of another, are potentially combinatorially explosive, and this restricts their usefulness.

- Topographic maps of visible surfaces.
  It may also be useful to construct a collection of separate 2-D databases for different perceived surfaces. For example the floor of the room will often provide a useful spatial indexing function. If most of the floor is visible it would map systematically into a part of the optic array - so this sort of structure can be closely related to the 2-D image structure. Other surfaces, for instance table tops, walls, or landscapes may also be treated this way.

One benefit of building maps tied to scene surfaces rather than simply optic array maps (or retinal maps) is that some of these maps can be preserved while the optic array changes, because the viewer rotates or moves to a new location. If the motion is controlled by the agent and has known properties, then the relationship between the object-based maps and the optic-array-based maps can be continually updated, giving the perception of an unchanging environment that endures through changing experiences of it.

Another use of object-based maps would be to provide a useful way of preserving information about a moving object instead of constantly having to re-compute it from new locations in the optic-array.

All of the above types of representations may contain information about 2-D structures, 3-D structures, or more abstract objects, properties or relations, such as causal relations or potential for change. The descriptions may be either relative to the viewer (e.g. depth, visibility), or relative to frameworks defined by individual objects (which may, for instance, have a major axis), or relative to some global framework in the environment, such as the walls of the room.

If the 2-D maps and mechanisms that operate on them are accessible by higher-level cognitive processes, this might account for the pervasive use of spatial reasoning in human thought: even the congenitally blind might use this kind of visual processing.

Different information stores are useful for different purposes. Viewer-centred descriptions are specially useful for fine control of actions. Object-centred descriptions are useful for recognising objects seen from different viewpoints. Descriptions based on more global frameworks are useful for large scale planning, especially plans involving several objects or agents. Moreover, different scales of resolution will also be relevant to different tasks.

Offset against different merits are different demands made by various information stores. For example, they vary according to how long they take to derive from image data, how much space they require, how sophisticated the interpretative algorithms need to be, how sensitive they are to noise or slight changes in the scene, whether they engender combinatorial searches and so on.

11   Kinds of visual learning

The labyrinthine theory permits far more possibilities for visual learning and training than does the modular theory. This is because it allows:

- more kinds of output (descriptions of more kinds of things, along with control information, including control of mental processes),

- more output routes (i.e. descriptive or control information may be sent to wherever it is needed),

- more kinds of input (information from other sensory subsystems, or from higher level information stores),

- more ways of deriving output from input: the output does not have to be derived by means of general principles of optics and geometry, but can use arbitrary but useful learned associations.

If learning is the production of long term change in knowledge and abilities, then many kinds of learning are possible: new particular facts, new generalisations and associations, new concepts for expressing information, new skills. There are also forms of learning that don't change qualitative capabilities but simply increase speed or reliability.

Even the modular theory presupposes that vision uses descriptive languages sufficiently general to allow the conceptual creativity required for going from optic array features to characterisations of 3-D shape and motion, with explicit information about parts, properties and relationships at different levels of abstraction. I suggested above that mechanisms providing this kind of representational capability could also support the representation of information not included in the modular theory. The syntax of representing structures with this kind of power would enable yet more descriptors to be introduced, adding to the conceptual creativity of the system: a powerful visual learning capability. Exactly what sort of mechanism would enable this to occur as a result of training or experience remains a topic for further theoretical and empirical investigation.

Common experience demonstrates that, at least in humans, several varieties of visual learning can occur e.g. learning to read text or music, learning to discriminate the colours named in one's culture, learning to discriminate plants or animals, learning to see tracks in forests, learning to tell good from bad meat in a butcher's shop, learning to judge when it is safe to cross the road despite oncoming traffic etc. (My informal observations suggest that it is not until after the age of eight or nine years that children learn to discriminate the combinations of speed, distance and size of approaching vehicles adequately.)

The task of distinguishing identical twins provides an interesting example. Many people have had the experience of meeting twins and being unable to distinguish them at first, then finding several months later that they look so different that it is hard to imagine anyone confusing them. The same thing happens sometimes whilst getting to know people from another ethnic group. It is as if the frequent need to distinguish certain classes of individuals somehow causes the visual system to enrich its analysis capabilities and descriptive output so that it includes new features helpful for the particular discrimination task. Exactly how this is done requires further investigation: it may be that there is some modification of *general* shape description processes to extract more detailed information from the optic array, or it may be that a *specialised* face recognition module is trained to make new uses of *previously* available low level shape descriptors.

There are many kinds of high-level learning, such as learning new faces or the names of new kinds of objects, situations, or processes (e.g. dance movements). This may or may not involve consciously associating a name with the object. (What is meant by "conscious" or "consciously", or even whether these are coherent concepts, is another large question not addressed here.) Recognition is often thought of as involving the production of a name. But this is just one kind of response to recognition. Reflex physical responses tailored to the fine structure of the situation, but without the intervention of explicit recognition or description are another kind.

45

The need for speed in dangerous fast-changing situations suggests a design in which the triggering of a response is done as directly as possible, that is without the intermediate formation of an explicit description of what is happening, which then interacts with inference mechanisms to form a new motive or plan or set of motor-control signals. Using a faster, more direct, process may require new connections between sub-systems to be set up, through learning.

Many sporting activities seem to involve both development of new discriminative abilities and linking them to new control routes. A boxer has to learn to detect and react to incipient movements that indicate which way the next punch is coming. Cricket batsmen and tennis players have to learn to see features of the opponent's movements that enable appropriate actions to be initiated at an even earlier stage. I do not know how much of the squirrel's ability to judge where to put its feet next is learnt and how much innate. In all these cases, detection is not enough: rapid initiation of appropriate action is also required, which could be facilitated by developing new control connections from lower levels of visual processing, if those levels can be trained to make the relevant discriminations and store appropriate output mappings.

These forms of conceptual learning go beyond the kind of rule-guessing processes studied by psychologists and AI workers under the title of "concept formation". New combinations of old concepts (e.g. an X is something that is A or B but not C) will not always suffice: it may be necessary in visual learning as in the development of science to create new concepts not definable in terms of old ones, or new descriptive capabilities, a more general and powerful form of learning. Connectionist processing models may be able to account for this, but for now precisely how it is done is not my concern. A harder question is how new undefined symbolic structures get their semantics: part of the problem mentioned above and answered sketchily in terms of generalised Tarskian models plus causal embedding.[5]

Some of these forms of learning seem to be slow, gradual and painful. Others can happen as a result of a sudden re-organisation of one's experience, perhaps influenced by external prompts, like seeing a pattern or structure in an obscure picture with external verbal help, after which one sees it without help. Whether the former learning is inherently slow because of the nature of the task, or whether we just don't have very good learning mechanisms is a topic for further investigation.

Learning would be *inherently* slow if it involved setting up new associations between relatively low level viewpoint-sensitive 2-D optic array descriptors and appropriate actions. For any 3-D scene there will be indefinitely many significantly different 2-D views, so that far more descriptions would have to be analysed to find commonalities and set up associations than if viewpoint-independent descriptions of objects and events were used. The actual number will depend on what differences are significant, or how the continuous variation is quantised.

The trade-off is that as the level of abstraction goes up, the simpler the descriptions and the smaller their number, but the less detailed information is preserved. So part of the visual learning task is to find the highest level of abstraction that preserves sufficient information to make discriminations required for the needs and purposes driving the learning: optimising the space/information trade-off.

However, what satisfies this criterion may not be fast enough for some dangerous situations.

---

[5]Note added in 2006: this paper rejected symbol-grounding theory before it became popular.

So discrimination may have to happen at a lower level of processing, therefore requiring more different associations to be learnt, and new information routes from lower levels of the visual system to be set up. Thus a longer learning or training period would be required to improve speed within a fixed level of discriminatory performance. It is possible that good sports coaches have some kind of intuitive grasp of this and select training situations that help this process.

A similar trade-off applies to action planning and control mechanisms, which need to select the appropriate level of action description to generate a response: a high level plan may be more generally applicable, but it requires a complex interpreter to generate appropriate motor control signals in the light of the current situation. If the control signals associated with particular situations are lower level, they will be more complex and detailed, as required, but a larger number of different combinations will have to be learnt and stored, and it will therefore take longer to learn them. Moreover, if part of the learning process is finding the right level of abstraction to meet both requirements of specificity of description and speed of processing, then the search space to be explored can be very large and learning will be inherently slow.

Frisby's book includes some random dot stereograms that are quite hard to fuse into 3-D percepts (because they represent continuously varying 3-D surfaces, not sharp edges). But after exposure to some of them people seem to get better at those particular ones. This may be because something has been learnt about the vergence angles they require, or for more subtle reasons to do with storing higher level information that controls the detection of binocular disparity. However random dot stereograms have so little in common with ordinary optic arrays that the slow processes they require for binocular fusion and depth perception may have little to do with normal vision.

Fine control of physical movements (like painting a picture or catching moving insects) is another kind of use where it might be advantageous in some cases to have a direct link from lower or intermediate stages of the visual system to whichever part of the brain is executing the action, instead of going through a central database of geometrical descriptions. There are at least three possible reasons for this: (a) the lower-level descriptions may contain more information of the kind required for fine control (b) it may be easier to compute corrections on the basis of perceived 2-D discrepancies than on the basis of relations in 3-D (c) the extra time required for going via the higher level descriptions may introduce feed-back delays that produce clumsy and irregular movements. (This would be relevant to the effects of some kinds of brain damage.)

Learning to sight-read music could make use of the same mechanisms. The experience of an expert sight-reader suggests that the visual stimulus very rapidly triggers movements of hands, diaphragm, or whatever else is needed (e.g. feet for an organist), by-passing the cognitive system that might otherwise interpret the musical score and plan appropriate movements to correspond to it. It is as if the visual system can be trained to react to certain patterns by interpreting them not in terms of 3-D spatial structures but in terms of instructions for action transmitted directly to some portion of the brain concerned with rapid performance. This does not imply that the patterns themselves are recognised as unstructured wholes: there must be some parsing (structural analysis), for otherwise a pattern never seen before could not have any sensible effect, whereas the whole point about sight-reading is that the music has not been seen before, except at the very lowest level of structure.

Learning to read fluently seems to illustrate both making new visual discriminations and

categorisations and also sending the output direct to new sub-systems in the brain. If full 3-D structural descriptions of the printed page contain information that is not particularly suited to the purposes of fluent reading, then it may be more efficient to "tap" the visual information before the stage at which descriptions of 3-D spatio-temporal structures are constructed.

There is also some evidence that visual information can be used in early stages of processing of other sensory sub-systems. A striking illustration is the fact that what we hear can be strongly influenced by what we see. In particular, how people hear a particular acoustic signal can be strongly influenced by perceived motions of a face on a video screen (McGurk & MacDonald, 1976).

Another very interesting process capable of being driven by vision is the learning of skills by example. Often a complex skill cannot be imparted by describing it, or even by physically moving the learner's limbs in the fashion of a trainable robot, yet can be conveyed by an expert demonstration, though not necessarily instantaneously. This is often used in teaching dancing or the playing of a musical instrument requiring rather subtle physical co-ordination, such as a violin.

The process of learning by watching an expert may be connected with the involuntary physical movements that sometimes accompany watching sporting events, as if our visual systems are directly connected to motor-control mechanisms. This ability to learn by seeing would obviously be of biological value as a way of passing on skills from adults to the young. However, it requires a different kind of processing from any described above, because the motion of another agent, initially represented from a different viewpoint, would have to be transformed into motion from the perceiver's own viewpoint, and then mapped on to motor control information by the perceiver.

Whether this mapping (during learning) has to go via a viewpoint-independent 3-D structural description is an interesting question. It may be that, as mentioned above in listing intermediate databases, we have a specialised representing structure related to the topology of the human form, because of its potential general usefulness in vision (as in model-based computer vision systems). In that case the use of this specialised map to store detailed motion information about perceived agents could facilitate transfer of the relevant information to a map of the perceiver's own body, and from there to relevant motor control units.

If specialised maps are useful for indexing during visual processing, then another kind of visual learning may be the discovery of useful maps. As ever there will be tradeoffs: the more abstract mapping structures will be more generally applicable and will require less storage space and perhaps faster searching and matching, whereas the more detailed ones will have more useful information, but will require larger numbers to be stored as well as being slower to search or match. For high level recognition and planning tasks the more abstract structures will be more useful. For more detailed perception, planning and control, the lower level ones may be more useful. (Whether matching high level structures is faster or slower than low level ones depends on the kind of matching. Parsing a sentence, i.e. matching against a grammar, can be much slower than comparing two sentences word for word.)

Resolution of empirical questions about the extent to which human vision conforms to the labyrinthine design may have to await substantial advances in our understanding of the functional organisation of the brain. However, from a theoretical point of view we can see

that this design allows processing advantages and permits more generally applicable learning possibilities.

If the different kinds of learning sketched above really do exist in humans, then we should expect to find different ways in which learning can go wrong as a result of brain damage or other problems. For instance, the discussion implies that reading may go wrong because the ability to access the relevant 2-D descriptions is lost so that reading has to go via the 3-D descriptions rather than using only the faster lower-level visual processes, or because the specialised links between this visual processing and abstract semantic representations are lost. In the latter case other capabilities relying on the intermediate 2-D information would be preserved. Similarly, because a boxer has to learn both to discriminate different kinds of incipient movements and to route the visual information to appropriate motor sub-systems either type of learning might be impaired, though the second cannot work without the first, and either type of skill might be damaged after it has been acquired.

Another empirical question is how much variability there is in routes available for linking two sub-systems. If there is only one route available then if that gets damaged after it has developed, then re-training will not produce a cure. Whether alternative routes are available depends on empirical facts about the underlying physical mechanisms, which are not the topic of this paper.

(Selfe, 1977) describes an autistic child with amazing drawing abilities between the ages of 3 and 7 years, for instance capturing horse and rider superbly foreshortened in motion towards the viewer. The ability appeared to be considerably reduced after she began to learn to talk in later years. Selfe conjectured that brain damage prevented the development of normal higher-level processing of a kind required for language and this somehow facilitated compensatory development of other capabilities. These other capabilities, according to the theory sketched here, might have been concerned with analysis of relatively low level structure in the optic array, and associating such structure with relatively low level control of actions required for drawing.

For normal children the requirement to draw well is of far less significance than the ability to form and relate higher level perceptual and action schemata: a child can get on well without being able to draw, but being unable to communicate or make plans is a more serious disability. So, in normal children, the pressure to optimise the space/information/speed trade-offs discussed above would lead to construction of more general-purpose links at higher levels of abstraction. Perhaps the learning processes that drive this construction compete for resources with those that create the lower level links? Or perhaps the higher level links, once created, somehow mask the lower level ones so that they can no longer be used? This would be more consistent with Nadia's reduced drawing ability after beginning to learn to talk. However, the change might have been motivational, rather than a change in her abilities. Only when we have far more detailed theories about possible mechanisms will we be able to make progress interpreting such evidence.

12   Conclusion

I have contrasted the modular theory of vision (as one floret on a sunflower) with a possible "labyrinthine" design in which a wider, and extendable, variety of functions is performed by a visual sub-system composed of smaller modules using a wider variety of input and output links to other systems. On the labyrinthine model the inputs to vision may include information from other sensors and from long-term information stores, in conjunction with hints, questions, or tasks specified by planners and other higher level cognitive mechanisms. The outputs may comprise both descriptions (including 2-D image structure, modal, functional and causal descriptions, descriptions of mental states of agents, the meanings of printed text and other abstract interpretations) and also control signals and stimulation of other modules that may need to react quickly to produce either physical responses or new mental processes. Moreover, the range of descriptive and control outputs and the range of connections to other sub-systems can be modified by training, rather than being rigidly fixed.

An obvious objection can be posed in the form of a rhetorical question: What makes this a *visual* system, as opposed to yet another general computing system that takes in a range of information, computes with it, and produces some outputs, possibly after communicating with other systems?

The answer to this lies partly in the nature of the primary input, namely the optic array with its changing 2-D structure, and partly in the way information is organised during processing. Very roughly, in a visual system, input data and intermediate partial results of the interpretation process, are all indexed according to location in a two dimensional field corresponding to the 2-D structure of the optic array. In other words, information of various kinds derived from the optic array is indexed (in part) by means of location in a network of 2-D topographic maps, an example of what I have elsewhere called 'analogical' representations (e.g. (Sloman, 1975), (Sloman, 1978)). This does not rule out simultaneous use of other non-topographic maps and more abstract databases of information.

The 'optically-registered' databases are not necessarily tied closely to the retina, since rapid eye movements can constantly change which portions of the optic array are sampled by which portions of the retina. It seems more useful to have the databases in registration with the optic array itself, as this is less changeable.

Not all the information created or used by the visual system need be stored in optically registered databases. Various abstract 'non-topographic' databases, such as histograms and Hough transforms, may also be useful, including the abstract non-topographic mappings postulated by (Barlow, 1983) and (Treisman, 1983). Nevertheless, the central use of databases whose structure is closely related to the structure of the incoming optic array is, I suggest, what makes a process visual as opposed to just a cognitive process. Even if some of the databases are not structured in this way, if their contents point into the image-registered databases and are pointed to by such databases then they can be considered part of the visual system. Of course, this characterisation does not define a sharp boundary between visual and non-visual mechanisms: nor is there any reason why nature should have sharp divisions corresponding to the labels we use. (Where, precisely, are the boundaries of a valley?)

There is still much that is vague about the model sketched here. It will have to be fleshed out

by describing in detail, and building computer models of, some of the important components, especially the kind of trainable associative mechanism that can map image features to the required descriptions. Moreover, a complete design for a visual mechanism will require a general account of how spatial structure and motion can be represented in a manner that is adequate to all the uses of vision. We are still a long way from knowing how to do that, though we share with squirrels and other animals a rich intuitive grasp of spatial structure and motion.

This paper has two main objectives. First I have compared two abstract hypothetical design-schemas pointing out that if they can both be implemented then one of them may have some advantages over the other. This abstract analytical discussion says nothing definite about how any biological visual system works or how any practical robot should be designed, for there may be additional design constraints arising from the underlying physical mechanisms used.

Second, and far more tentatively, I have produced some fragments of evidence suggesting that human perceptual systems can be construed as using the labyrinthine design. I do not claim to have established this empirical thesis. At the very most some questions have been raised which may perhaps lead to further empirical investigations of how both human and (other) animal visual systems work. This is a task for specialists with more detailed knowledge than I have. My concern is primarily with the design suggestion that, in at least some cases, the multi-connection multi-function labyrinthine design will actually be useful for practical engineering purposes. This could turn out false in practice. However, at least some neurophysiologists interpret available evidence as suggesting that different sensory and motor sub-systems are linked in a manner that involves much richer interconnectivity than assumed by the modular theory, with "overlapping hierarchies that become increasingly interrelated and interconnected with each other at the higher levels" ((Albus, 1981) – see also his figures 7.1 and 7.2). The neat sunflower gives way to a messy spiders web.

As for the squirrel, I think its versatility and speed will far outclass anything we know how to design and build, for many years.

Acknowledgements

References

Albus, J. (1981). *Brains, behaviour and robotics*. Peterborough, N.H.: Byte Books, McGraw Hill.

Ballard, D., & Brown, C. B. (1982). *Computer Vision*. Englewood-Cliffs: Prentice Hall.

Barlow, H. (1983). Understanding natural vision. In O. Braddick & A. Sleigh (Eds.), *Physical and Biological Processing of Images.* Berlin: Springer-Verlag.

Barrow, H., & Tenenbaum, J. (1978). Recovering intrinsic scene characteristics from images. In A. Hanson & E. Riseman (Eds.), *Computer vision systems* (pp. 3–26). New York: Academic Press.

Brachman, R., & Levesque, H. (Eds.). (1985). *Readings in knowledge representation*. Los Altos, California: Morgan Kaufmann.

Brady, J. (1985). Artificial intelligence and robotics. *Artificial Intelligence*, *26*(1), 79–120.

Brady (editor), J. (1981). Special volume on computer vision. *Artificial Intelligence*, *17*(1), 1–508.

Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Reading, Mass: Addison Wesley.

Clowes, M. B. (1971). On seeing things. *Artificial Intelligence*, *2*(1), 79–116.

Fodor, J. (1983). *The modularity of mind*. Cambridge MA: MIT Press.

Frisby, J. P. (1979). *Seeing: Illusion, brain and mind*. Oxford: Oxford University Press.

Fu, K. (1977). *Syntactic Pattern Recognition Applications,*. Berlin: Springer-Verlag.

Fu, K. (1982). *Syntactic Pattern Recognition and Applications*. Englewood-Cliffs: Prentice-Hall.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Gregory, R. (1970). *The Intelligent Eye*. London,: Weidenfeld and Nicolson.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behaviour. *American Journal of Psychology,*, *57*, 243–259.

Hinton, G. (1976). Using relaxation to find a puppet. In *Proceedings AISB Conference.* Edinburgh.

Hinton, G. (1981). Shape representation in parallel systems. In *Proceedings 7th IJCAI, VOL II.* Vancouver.

Hogg, D. (1983). Model-based vision: A Program to see a walking person. *Image and Vision Computing*, *1*(1), 5–20.

Hogg, D. (1988). Finding a Known Object Using a Generate and Test Strategy. In I. Page (Ed.), *Parallel Architectures and Computer Vision.* Oxford University Press.

Hogg, D., Sullivan, G., Baker, K., & Mott, D. (1984). Recognition of vehicles in traffic using geometric models. In *Road Traffic Data Collection.* IEE Conference Publication 242.

Horn, B. (1977). Understanding image intensities. *Artificial Intelligence*, *8*(2), 201-231.

Huffman, D. (1971). Impossible objects as nonsense sentences. In D. Michie & B. Meltzer (Eds.), *Machine Intelligence 6.* Edinburgh University Press.

Lee, D., & Lishman, J. (1975). Visual proprioceptive control of stance. *Journal of Human Movement Studies*, *1*, 87–95.

Lindsay, P., & Norman, D. (1977). *Human Information Processing: An Introduction to Psychology, 2nd edition*. New York: Academic Press.

Longuet-Higgins, H. (1987). *Mental Processes: Studies in Cognitive Science*. Cambridge Mass,: Bradford Books, MIT Press.

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

McClelland, J. L., Rumelhart, D., & *et al* (Eds.). (1986). *Parallel Distributed Processing, Vols 1 and 2*. Cambridge Mass: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Michotte, A. (1963). *The Perception of Causality*. Methuen.

Nishihara, H. (1981). Intensity, Visible-Surface, and Volumetric Representations. In *Brady (1981)*.

Scott, G. L. (1988). *Local and Global Interpretation of Moving Images*. Pitman, London & Morgan Kaufmann, Los Altos.

Selfe, L. (1977). *Nadia: a case of extraordinary drawing ability in an autistic child*. London: Academic Press.

Sloman, A. (1975). Afterthoughts on analogical representation. In R. Schank & B. Nash-Webber (Eds.), *Theoretical issues in natural language processing (tinlap)* (pp. 431–439). MIT. (Reprinted in (Brachman & Levesque, 1985))

Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). Available from `http://www.cs.bham.ac.uk/research/cogaff/crp`

Sloman, A. (1982). Image interpretation: The way ahead? In O. Braddick & A. Sleigh. (Eds.), *Physical and Biological Processing of Images (Proceedings of an international symposium organised by The Rank Prize Funds, London, 1982.)* (pp. 380–401). Berlin: Springer-Verlag. (http://www.cs.bham.ac.uk/research/projects/cogaff/06.html#0604)

Sloman, A. (1985). Why we need many knowledge representation formalisms. In M. Bramer (Ed.), *Research and development in expert systems* (pp. 163–183). Cambridge University Press. (http://www.cs.bham.ac.uk/research/projects/cogaff/04.html#200406)

Sloman, A. (1987a). Motives mechanisms and emotions. *Cognition and Emotion*, *1*(3), 217–234. (Reprinted in M.A. Boden (ed), *The Philosophy of Artificial Intelligence*, 'Oxford Readings in Philosophy' Series, Oxford University Press, 231–247, 1990)

Sloman, A. (1987b). Reference without causal links. In J. du Boulay, D.Hogg, & L.Steels (Eds.), *Advances in artificial intelligence - ii* (pp. 369–381). Dordrecht: North Holland. (http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#5)

Treisman, A. (1983). The role of attention in Object Perception. In O. Braddick & A. Sleigh (Eds.), *Physical and Biological Processing of Images*. Berlin: Springer-Verlag.

Ullman, S. (1980). Against direct perception. *The Behavioural and Brain Sciences*, *3*, 373–381. (3)

Ullman, S. (1984). Visual routines. *Cognition*, *18*, 97–159.

Winograd, T. (1972). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. *Cognitive Psychology*, *3*(1).