Presented at: Fyssen Foundation Vision Workshop
Versailles France, March 1986, Organiser: M. Imbert
The proceedings were never published. A related paper entitled "On designing a visual system (Towards a Gibsonian computational model of vision)" was later published in 1989 in JETAI.
(Question about eyes added to caption of Fig. 5 on 15 Dec 2017)

What Are The Purposes Of Vision?

Aaron Sloman
School of Cognitive Sciences, University of Sussex Brighton,
Now at University of Birmingham
http://www.cs.bham.ac.uk/~axs

Contents

1 Introduction

The richness, variety and speed of human and many animal visual processes are a constant source of amazement to those who try to design artificial visual systems. By comparison, machine vision still limps along far more slowly and with significantly less functionality. This could be because we don't yet know much about human vision and therefore don't really know what we should be trying to simulate, or 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. It could be both. I suspect the former is the main reason, so that until we have a much clearer understanding of what is required, technology will not begin to catch up.

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. Does it feed information only to a central database, where other sub-systems can access it, or does it feed information direct

to a variety of sub-systems? What sorts of information does it feed – is it mostly a set of descriptions of spatial properties of the environment, or are there other sorts of descriptions, and other outputs besides descriptions? Is there a sharp boundary between vision and cognition? What sorts of input does the visual subsystem use?

I shall attempt to survey the uses of human vision, with the hope of deriving some design constraints and requirements both for theories about biological visual systems and for machine vision. I shall propose a very broad view of the functions of vision in human beings, and suggest some design principles for mechanisms able to fulfil this role, though many details remain unspecified.

The range of possible visual mechanisms to be found in the biological world and in present and future robotics laboratories is vast. Most of this paper will focus on human or human-like visual systems, but it should be remembered that in principle other systems might perform a different, but overlapping, set of tasks, and could use different mechanisms.

The discussion will revolve around the following key questions.

- Are descriptions of (possibly changing) spatial structure and location, the only descriptions produced by a visual system?
- If not, what other kinds of descriptions should a visual system produce? E.g. should descriptions of image features be output? Should descriptions of non-spatial properties be produced by the visual system, or are they inferred from the visual output, by separate modules?
- Is producing descriptions the only function of vision?
- If not, what other functions should a visual system have? E.g. should it also be able to trigger processing in other subsystems?
- 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?
- Is it possible to draw a boundary between visual processing and other kinds of processing, or is the brain best thought of as a very large richly interconnected system with, for example, increasingly multi-modal or amodal layers of processing as information moves from sensory transducers?

I shall contrast two extreme theories. The truth may be somewhere in between. On the "modular" theory, vision is a clearly bounded process in which optical stimuli trigger the production of descriptions of 3-D spatial structures, which are stored in a database where they can be accessed by other sub-systems. On this view all processes that make use of visual input have to go via this common database. This modular theory is defended at length in Fodor (1983), and is often taken for granted by workers in AI.

The alternative non-modular theory proposes that the visual system produces a wider variety of descriptions, that its outputs include more than just descriptions, that it makes use of a wider

variety of inputs, and that it can change its outputs as a result of training. On the non-modular theory there will still be a visual module[1], but its boundaries will be less clearly defined.

Discussion of these theories requires analysis of the uses of vision. Part of the 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 type of visual system would require a general-purpose trainable associative mechanism.

2 The 'modular' theory

A statement of the modular view is to be found on page 36 of David Marr's book (1983), 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 most vision work in Artificial Intelligence. For example, it pervades the wonderful book on seeing by John Frisby (1979), partly inspired by Marr, and the same "standard" view is expressed in the textbook on AI by Charniak and 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'.

This view of the purposes of vision is very attractive, as it holds out some hope for a *principled* design of visual mechanisms. For example, if the task of vision is to discover geometrical facts, such as facts about the shape and location of objects, then perhaps these facts can be inferred from the geometry of the optic array using principles of mathematics and physics, since the optic array impinging on the retina (or camera) is a richly structured array of information systematically derived from the shapes and locations of objects in the environment by a well understood projection process. A similar argument suggests that optical properties like colour and reflectance of visible surfaces may also be inferred from retinal stimulation in a principled fashion.

If the visual mechanism is a principled solution to very specific 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, as recommended by Marr and other workers in AI. However, I was struck by the fact that very few of the participants at this Fyssen workshop attempted to relate their work to a characterisation of the visually accessible

[1] **Note added 8 Oct 2012:** The module may be composed of a collection of sub-modules

properties of the environment. Could this be a fundamental methodological flaw? Or is it a reflection of a feature of the visual system, namely that it is not specifically geared to the physics of our environment, apart from the fact that the retina is sensitive to optical stimulation? Let's look more closely at this "modular" theory of vision before considering alternatives.

Although the modular view conceives of the visual system as having a well defined boundary it is not thought of as internally indivisible. It is assumed that there is a collection of different internal databases in which intermediate descriptions of various kind are stored, and used within the visual system in order to derive subsequent descriptions. (See Barrow and Tennenbaum, 1978, and Nishihara 1981). For example, among the intermediate databases, may be edge 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. Some of the 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. The only information available to other subsystems would be the output descriptions of objects and processes in the 3-D scene.

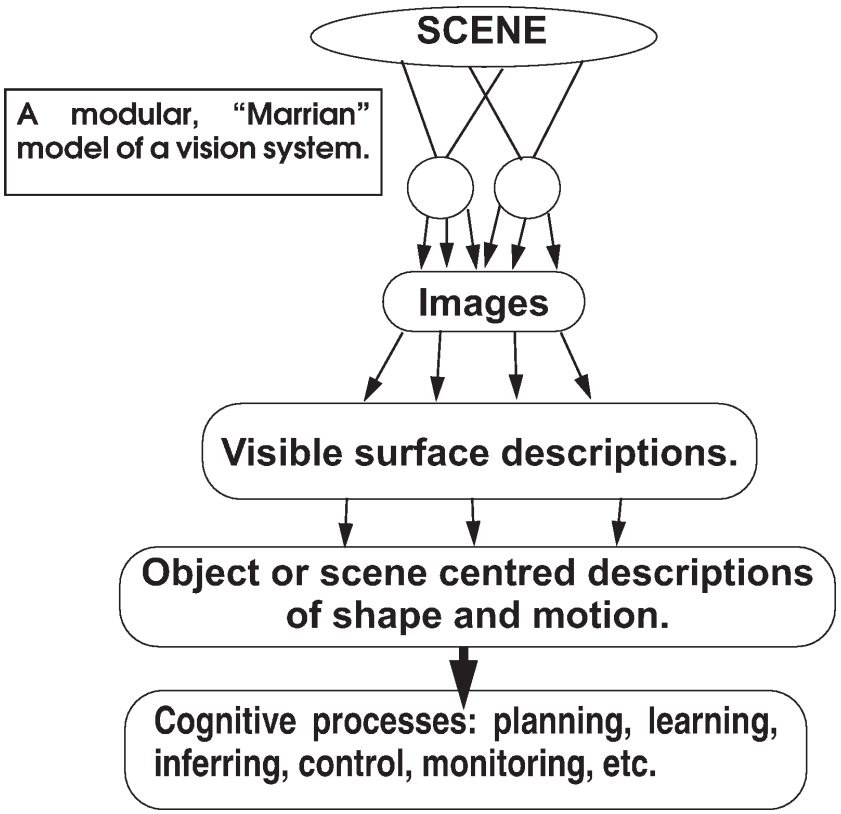

Figure 1: *The monolithic (Marrian) model of vision, in a simplified form. The internal structure of this model might be somewhat as represented here, with information propagated mostly uni-directionally from sense organs to central cognitive mechanisms*

The kind of visual system depicted in Figure 1 is as essentially a mechanism for transforming descriptions of optic arrays into descriptions of spatio-temporal structures, possibly including descriptions of optical features, such as colour, reflectance, or illumination. The inputs to the retina(s) are in the form of a (possibly changing) array of measures of local illumination and the outputs are in the form of a (possibly changing) collection of descriptions of spatio-temporal structures and relationships.[2]

Implicit in this sort of model is a view of vision as one among a number of sensory modules feeding information into some kind of central cognitive system, where each sensory module is relatively isolated, autonomous and unchanging.

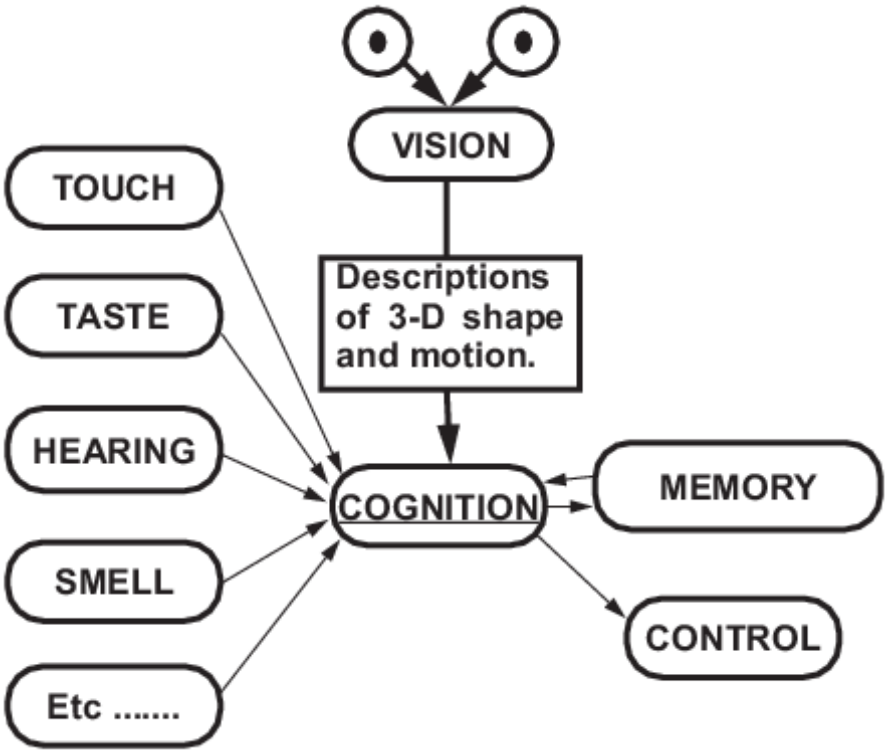

Figure 2: *The model of perceptual modules all interacting with some central cognitive system, for instance as suggested by Fodor, and implied by Marr. There could be more bi-directional flow of information and control than indicated.*

Both the models in Fig 1 and Fig 2 are too simple, as we'll see. I shall offer an alternative model[3], in which the visual system is not merely concerned with producing *descriptions*, and the descriptions it makes available to other sub-systems are not restricted to spatio-temporal and

[2]Added 8 Oct 2012: This view is largely influenced by the development of digital camera technology. It may seriously distort the nature of the problems solved in biological vision discussed in Craik (1943)

[3]Later described as requiring a "labyrinthine" architecture in a paper published in 1989: A. Sloman, On designing a visual system (Towards a Gibsonian computational model of vision), in *Journal of Experimental and Theoretical AI*, 1989, 1, 4, pp. 289–337, `http://tinyurl.com/BhamCog/81-95.html#7`

optical properties of 3-D objects in the environment. Further, I'll suggest that it can accept a variety of different types of input, can be linked to other sensory modalities, and can change its capabilities over time. This sort of system can perform a substantially wider range of functions than a monolithic rigidly restricted spatial description transducer.

The richer multi-purpose conception of vision has implications both for the architecture of a visual system and for the types of representations that it uses internally.[4]

3 Previous false starts

I believe the modular theory of vision proposed by Marr and others provides very useful insights, but misses out some important aspects of vision. It seems to me to be just the latest in a series of '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. Examples of previous errors include the following:

- Since retinal images are two dimensional, vision is a process of analysing 2-D structures. We have already seen that this cannot be the whole story, even if it is a part of the correct story.
- 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 images.
- Vision is essentially a process of segmentation: if only images could be segmented into parts belonging to different objects, the rest would be easy. This may be part of the story, but it ignores the need to describe 3-D relationships between objects and parts of objects.
- Vision is pattern recognition: if only we could make machines recognise patterns in images, all the problems would be solved. This ignores the need to describe complex structures and relationships not previously encountered: merely attaching a known label does not do this. Of course recognition of familiar substructures and well known relationships is part of the process of producing a novel structural description. (Compare parsing, in language understanding mechanisms.)
- Vision is syntactic analysis - finding the 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 Fu 1977 and Fu 1982.) However, it is not enough to find structures in images: many of the structures we need to see are structures in the environment, not in retinal images. Interpretation and inference are needed, as well as analysis and parsing.
- Vision is heterarchic processing, mixing top-down and bottom-up analysis: if only the right control structure is used, with enough prior knowledge about possible objects in the

[4] Added 8 Oct 2012: There are also implications for biological evolution and individual development.

environment, everything will be easy. This view, partly inspired by Winograd's work on heterarchy in language understanding, ignores unsolved problems about how to represent scene structures and does not account for cases where we see complex structures not known previously.

- Vision is essentially a matter of getting 3-D information about the environment: if only we could find a way of deriving a 3-D depth map from retinal images, the rest would be easy. However, a 3-D depth map is just another unarticulated database, and, as will be shown later, would not be able to serve the main purposes of vision: it would still require considerable processing in order to provide useful descriptions of what is in the scene, for instance descriptions of what Gibson (1979) called "affordances".
- Vision is highly parallel - if only we had powerful enough parallel computing engines everything would be easy. This ignores the question whether there is something special about the requirements for vision, for instance the need to be able to represent spatial structures.
- Vision requires connectionist machines. See the previous comment.

There are several key ideas that are easily forgotten when people enthuse over the latest approach to vision. 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, the visual system has to *interpret* them by mapping them into quite different structures. This is acknowledged by the modular view described above. I shall argue later that besides the domains of 2-D and 3-D spatial structures, yet more domains may be involved, e.g. abstract domains involving functional or causal relationships, and perhaps even meanings and perceived mental states of other agents. I am not denying that a process that describes or labels 2-D image structures can play a role in vision. This may be one of many important sub-processes in a complete visual system.

Another key idea that has played an important role in AI work is that vision involves the production of descriptions. Nobody knows exactly what sorts of descriptions, 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 merely labels recognised objects with their names, cannot be a complete visual system. However, it can hardly be said that AI work 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.

Besides the key ideas already mentioned, I think there is a very important idea that has not been given sufficient attention, namely 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.[5]

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

It is important to distinguish three different sorts of question, empirical, normative and theoretical. Empirical questions ask what actual biological visual systems are like and what they are used for. The normative questions asks what sort of visual system would be desirable for particular classes of animal or robot. Theoretical questions ask 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 I have an empirical conjecture that some biological visual systems, including human ones, have a broader range of uses than the modular theory permits. I also have a normative proposal that a broader design would be preferable, given certain constraints such as the unpredictability, the variability, and the speed of changes in the environment. Finally I make the relatively weak claim that alternative designs are possible and worth exploring.

Even if my empirical conjecture is false, the normative claim might be correct. In that case biological visual systems would be non-optimal.

Moreover, even if the empirical claim is false, and the normative claim can be shown to be flawed, the theoretical claim that these alternative designs are possible might be true and interesting. For example, by analysing the reasons why an alternative 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.

Anyhow, in what follows I'll simplify exposition by using a mode of expression that suggests that I am making empirical claims. I hope readers will appreciate that the normative and theoretical conjectures may have some worth even if the empirical claim turns out to be false.

My own interest is mainly in the theoretical question. I regard this as part of a long term investigation into the space of possible behaving 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, and attempting to understand the similarities and differences between different sub-spaces, and especially the design trade-offs, seems to me to be a necessary pre-condition for a full understanding of any one sub-space, including, for instance, the sub-space of human beings. This is analogous to comparing existing inverse

[5] Added 8 Oct 2012: In some papers published after this I used the label "scaling out" to describe the requirement for a mechanism to be embedded in an architecture with other mechanisms with which it has to interact. This includes being able to provide control information as well as factual information, as Gibson noted.

square laws with alternative possible action-at-a-distance laws in physics, in order to discover exactly what the inverse square law rules out.

5 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 module is that the information available at the retina is inherently ambiguous.

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) and Frisby (1979), that a particular optic array is usually derivable from a range of actual 3-D configurations, and hence there is no unique inverse to the process that projects scenes into images, even when the images are rich in information about intensity, colour, texture, etc. More precisely, 3-D information about structure or motion is lost by being projected into 2-D. A similar problem besets optical characteristics of the environment. 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.

Yet the human visual system has no difficulty in rapidly constructing unique interpretations for many such inherently ambiguous scenes – 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 structure is uniquely inferrable, e.g. from binocular disparity. The argument is simply that vision cannot always depend on that, and therefore more general mechanisms must be available. (I am also not disputing the importance of theoretical analysis of what can and what cannot be inferred from different kinds of retinal evidence.)

A standard response to the problem of ambiguity is to postulate certain general assumptions underlying the interpretation process. These can be used to constrain the 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, etc.
- 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. (Scott 1986 criticises assumption-based approaches to solving the problem of inferring structure from image correspondences.)

An alternative 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, Barrow and Tenenbaum 1978).

This requires the visual system to be designed as an optimiser: interpretations are selected that optimise some global property of the interpretation. Recent work on connectionist approaches to vision extends this idea. (See Hinton 1981, and connectionist papers in this volume.) Unfortunately, the measure to be optimised does not generally 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.

The Ames demonstrations, 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, suggests 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 aesthetic criteria, such as a preference for symmetry or connectedness. We are no longer dealing with a principled derivation of scene structure from image structure.

6 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 decision making, even at the cost of increased error. 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.

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.

If a totally deterministic and principled mathematical derivation from images to scene descriptions is not 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. 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. There seems to be plenty of evidence of general associative mechanisms in animal nervous systems, even though the details of how they work are still unknown.

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 far beyond spatial structures derivable from the physical properties of the optic array.

It is significant that the first example of vision mentioned in the textbook on psychology by Lindsay and 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. The kind of question I am raising is whether finding meanings is 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, as the modular theory would assume. Even if this is what happens in a novice reader, it is possible 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.

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. On the modular theory, these effects would all be produced by non-visual systems reacting to a central store of descriptions produced by vision. The alternative theory is that the visual system has a broader role than producing descriptions of 3-D structures.

7 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 rival 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 is a single-purpose visual module whose results are then accessed by a variety of secondary processes, with the conjecture that the visual module itself, i.e. the sub-system that produces 3-D spatial descriptions, can also 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 would turn out to be just a special case of this broader function. This conjecture can be supported by examining more closely what is involved in deriving 3-D descriptions from retinal stimulation.

8 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. This requires the visual system to have what we might call "conceptual creativity".

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 which are applicable to retinal images themselves.

Conceptual creativity is characteristic of all perception, since the function of perception is rarely simply to characterise sensory input. It includes at least the *interpretation* of that input as arising from something else. Hence descriptors suitable for that something else, namely the environment, are needed, and in general these go beyond the 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. At most they extend beliefs, not concepts.

Since everyone agrees that vision is not merely concerned with happenings on or close to the retina, but at least includes the function of describing 3-D spatial structure and motion, it follows that visual mechanisms require not just inference capabilities but conceptual creativity: the output language is richer than the input language, in that it uses concepts not definable in terms of the input language.

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. (I have discussed this general question elsewhere, e.g. Sloman 1986.)

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. 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.

A special purpose visual mapping system might somehow have evolved into a more general associative mechanism. What is more likely in biological terms is that a general associative mechanism became specialised for vision.

9 The biological need for conceptual creativity

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.

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?

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 only higher level processes make use of the semantics. Similarly, a 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.

We should expect biological evolution 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 might involve the ability to re-direct specialised output to other sub-systems, as required, instead of always going through 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.

10 The uses of a visual system

What is vision actually used for and what can it be used for? 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 controlling action. 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 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 basic. 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 are 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.[6]

The distinction between active and passive uses is orthogonal to the distinction between theoretical, practical and aesthetic uses.

If vision is capable of being used both actively and passively this imposes global design requirements on the architecture of the system.[7] Most AI work seems to treat vision as passive – i.e. mainly data-driven. There are many unsolved problems about the kind of architecture, representations and algorithms that could facilitate active vision.

It is not obvious how a visual system can function in active top-down mode. In most cases there is no simple translation from a high level question or hypothesis to low level questions to be posed by feature detectors, segmentation detectors, and the like. Perhaps the most that can be done in most cases is to direct visual attention to an appropriate part of the scene, 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.

The human visual system seems to be capable of more direct and powerful top down influence, since 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 which 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 high-level expectations, questions, goals or preferences which simply provide additional inputs to the network. 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

[6] Often both active and passive processes may miss an opportunity. Something before your eyes that is relevant to an important task may go unnoticed. It may even be the very item you are looking for. Proof reading is notoriously difficult, because of this. Sometimes what is not at first noticed will be seen spontaneously a moment later, or after someone else has mentioned seeing it, suggesting that visual processing is not entirely a data-driven process even in cases where the data are perfectly capable of driving the whole process.

[7] E.g. data-driven input to an array of databases where demons inserted top down can be triggered?

now I'll simply elaborate on some of the less obvious points.

11 Sub-tasks 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:[8]

- 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? 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 distance between a pair of approaching surfaces will 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 is to make a plan involving getting successive parts of the chair through the doorway, using perceived 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. A particularly common case is visual tracking by the eye: here the result of the action controls the action.
- Ordinary life teems with examples of visual control and monitoring, 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, 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 monkey failing to catch a branch as it leaps through tree tops. It is therefore particularly important to take advantage

[8] Compare Chapter 6 of Sloman (1978).

of any opportunity to compute only what is 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.

- Noticing unexpected relevant information
  During the course of executing a plan, new dangers, problems, opportunities, may arise that need to be detected even though there is no specific provision for them. 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.

12 Perceiving functions and potential for change

It is often useful to perceive not only structure but also function, and causal relationships. For example, seeing something as a window catch, or seeing a plank as holding a shelf up, is potentially useful in planning or guiding actions: the catch must be moved if the window is to be opened, and the plank may be moved if the shelf is to be brought down. Brady (1985) uses the design of some familiar tools to present examples of 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 relations between possibilities. For instance, in order to understand the window catch fully one must see that whether movement of the window can occur depends on whether the catch can rotate.

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 perceive possible changes inherent in the structure of the situation and the relationships between different possibilities is not merely an adult human capability. A dog can apparently see that putting its paw on a bone is a way of preventing the movement that defeats its attempts to tear meat off it. The process of assembling some of the more intricately constructed bird's nests must involve at least local planning on the basis of perception of possibilities for change. 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 had seen the potential for change in a complex structure, presumably helped by analogy with a different but more familiar situation.

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. 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 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.

Whether this 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. Ultimately it will have to be settled by discovering in detail how the brain works from anatomical and physiological studies. From a design point of view the main advantage of the non-modular 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.

One of the unanswered questions is how possibilities for change should be represented. At this stage I'll merely point out that if the visual system is able to represent actual optical flow, as many assume it can, then a similar symbolism or notation might be used for representing possible movements. Representing possible relative motions is a little harder. Representing IMpossibilities harder still.

David Young has pointed out in conversation that Gibson's notion of perceiving 'affordances' in objects seems closely related to my notion of perceiving possibilities for change (and some of the other more abstract percepts discussed in this paper). As I am not familiar with Gibson's work, I cannot comment on this, except to remark that unlike Gibson I am assuming that all such percepts are embodied in internal computational processes in which representations are constructed and manipulated.[9]

13 Figure and ground

It is often noticed that perception involves a separation of figure from ground. 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. For example, it includes 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. A full analysis would require investigation of the concept of an "object", another concept 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 familiar vase/faces figure, Figure 3, 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: anyone who tries can see the picture as depicting two faces with a vase wedged between them. The lines in the picture then depict cracks between adjacent

[9]Note added 8 Oct 2012: See also `http://tinyurl.com/BhamCog/talks/#gibson`

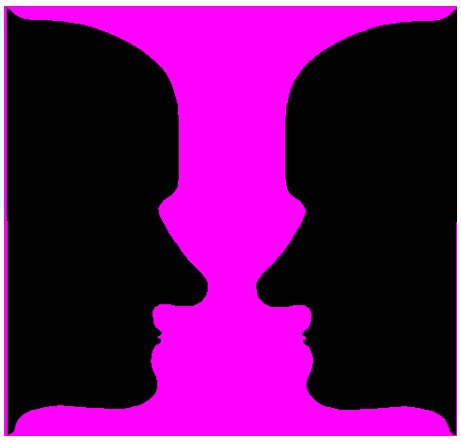

Figure 3: *An example of figure/ground ambiguity*

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 than 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. In other words, an object is an entity to which labels describing potential for change can be attached. This is just a special case of a more general role for segmented objects, namely that they can enter into relationships and have properties ascribed to them. In other words they can occur in articulated representations, described below. (Further discussion of this notion would relate to the "Naive Physics" project of Pat Hayes (1979).)

14 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. 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. (For my lecture at the workshop Geoffrey Hinton provided me with a sheet of paper folded to form a bird that flapped its wings when the tail was pulled. Close visual examination explains why, though the perceptual process is quite sophisticated.)

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". An example of a high-level cognitive task for which a visual representation is very useful is an argument to show 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. These relationships between spatial and abstract structures are often used by programmers.

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.

15 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 an 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 years. Yet for adults the relationship is blindingly obvious: what has changed?

It may be useful if complex abstract descriptions of potentiality for motion, and constraints on motion, can be collapsed into single functional labels, like "hole", "furrow", "exit", "opening", etc. Perhaps practical need trains the visual system to apply such labels on the basis of low level cues, leaving other subsystems to interpret them. These are not simply geometrical descriptors

but provide pointers to functional or causal information about what can happen or be done. From these it is a short step to functional descriptions like “lever” “pivot” “support” “wall”, “container, “lid”, etc. which point to a combination of possibilities and constraints on motion.

Much work remains to be done, classifying different sorts of compact functional descriptions and showing (a) how they can be derived from images and (b) how they can be used for planning and the control of actions. Let’s now look at yet more abstract visual descriptions.

16 Seeing mental states

Compare the Necker cube, with the duck-rabbit picture in figure 4. 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 the cube, 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 would expect on the modular view.

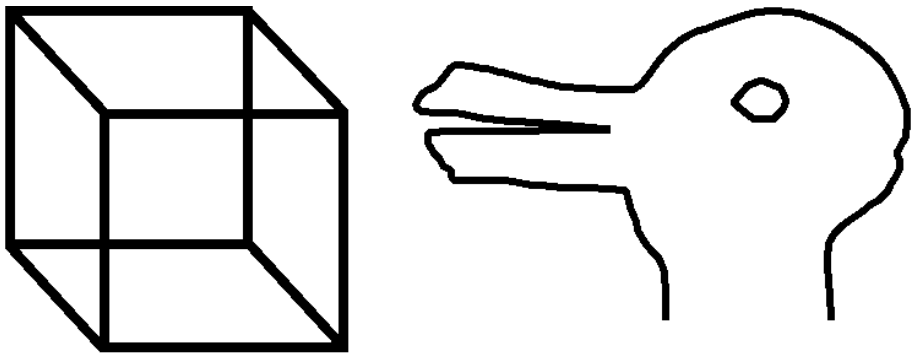

Figure 4: *Both figures are ambiguous and can be seen to switch between two views. The changes when the Necker cube “flips” are all geometric, involving relative distances and orientations of lines and planes. The changes in the duck-rabbit are not geometric and involve both identifications of biological parts and functions and also a change of category.*

Descriptions of the ‘flip’ experienced with the duck-rabbitt are very different. There is no significantly different perceived spatial structure. 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 harder to explain what this means. I offer the following conjecture.

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 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 of other agents and descriptions of what is visible to other agents.

On this theory the "flip" between duck and rabbit percepts would involve 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. 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*. I believe it is no accident 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 puzzle, as it would have to be on the modular theory.

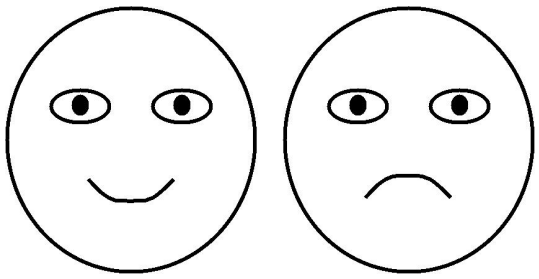

Figure 5: *Is the perception of mood visual, or post-visual? Or a mixture? Perhaps there is feedback from higher level processes to representations in registration with the "optic array". Do the eyes in the left face look the same as the eyes in the right face? If not, why not?)*

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 the two in Figure 5 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. Apparently 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.

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, how they are detected, what sorts of representations are used, 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, how exactly other systems use the information, and so on. All these are questions for further investigation.

17 Practical uses of 2-D image information

So far I have been arguing that in addition to spatial information a visual system can 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 image structures.

For example someone sighting a gun uses co-incidence in the retinal image (or optic array) rather than a full 3-D description. 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". A 2-D discrepancy measure may be easier and quicker to compute for the purpose of controlling action than the full 3-D discrepancy. 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 tell that the corners in the image are not. 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: the visual system can output information about the 2-D structure of the images, 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, 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. It appears that muscular control of balance depends on global patterns of optical flow which provide information about one's own forward or backward motion. Experiments by Lee (1975, 1977) 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 seems likely 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.

In any case, 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. 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, then using 2-D structure will make more sense than if there is a very large collection of different views all generated from an invariant 3-D structure. The latter is more likely to be the case with a non-rigid object, such as a sweater, which has an unchanging topology even though the 3-D details change as it is 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 probably has to learn to react to a variety of 2-D cues in order to be able to take very rapid evasive action.

18 Varieties of descriptive databases

We can distinguish different types of databases that may be produced from visual images.

- Descriptive
  Structures of arbitrary complexity, either in the image or in the scene are given explicit labels and described using explicit labels for their properties, their parts, and relationships to other labelled structures, the relationships also having labels. Logical languages and semantic nets are examples of formalisms for constructing articulated databases. Descriptive databases have several advantages, such as: reduced amount of information to process; invariance over viewpoint, distance, orientation, etc.; independence of sensory modality, so that information from different senses can be combined; applicability of very general inference methods, modularity.
- Articulated
  Structures are linked together, and have links to other structures including links to their parts, but there are no explicit labels, and there are no labels for relationships. Rather relationships are implicit in the ways things are linked together. An unlabelled parse tree for a sentence would be a simple example. If the components are themselves made of linked structures, the database is hierarchical-articulated, otherwise flat-articulated.
- Semi-articulated

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 the whole structure by starting from any of its parts and following links to their neighbours. The structuring is all at one level. For example, all edge points in an image may be linked to their neighbouring edge points but there is no link from one set of edges (a line) to another, since this would presuppose some explicit representation of the higher level structures.

- Pre-articulated
  Elements of the image or scene description are not linked together and there are no names for larger structures. But elements may have been labelled in some way to indicate which ones belong together. From each element it is possible to discover what its label is (or what its labels are) but not possible to go directly from labels back to the elements or from elements to others with the same label. In the original image, intensity measures are labels of this kind. After some processing pixels may be labelled according to whether they are at edges of a certain orientation, which region they belong to, or the direction and magnitude of optical flow. Co-ordinates and things computable from them are implicit labels. E.g. a set of collinear points have the same implicit label.

- Non topographic transforms
  There are many kinds of transforms from an image to a database where spatial location is lost. The most common example is a histogram recording numbers of picture points with a particular colour or intensity, or falling within a range of values. Closely related are Hough transforms (See Ballard and Brown 1982), in which each element of the original is mapped into a set of functions of properties of the element. The histogram provides a means of accumulating spatially disparate evidence in support of conflicting interpretations.

- If the histogram contains only measures of how many elements map onto each possible value then it gives no information about which parts of the image contributed. In more complex cases, each "bucket" may contain descriptive, articulated, semi-articulated or pre-articulated information about contributing portions of the image. It then turns into a separate mini-database. 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 features with that orientation, this constitutes a database of information about (roughly) aligned image fragments.

- Feedback and indexes
  If labels are created for the 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. This may also be done by creating new "pseudo-images" in registration with the original images. This sort of (frequently re-invented) strategy seems to be what Marr referred to as the use of 'place-tokens', and what Barrow and Tennenbaum described as 'intrinsic images'. The advantage is that by locating abstract labels within or in registration with the image structure we obtain a useful index for finding things during active visual processing. For example, if you wish to find out what a moving object is likely to hit first, you can project its trajectory into the

image then scan it for labels pointing to objects of various kinds. There are many tasks for which this kind of spatial addressing is useful. I suggest that this is the main feature that distinguishes vision and visual processing from other kinds of perception.

All of the above types of representations may contain information about 2-D structures, 3-D structures, or more abstract structures. 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, for example a framework defined by the walls of the room. All these different cases 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. So we see a need for many different kinds of descriptions, for different purposes.

Offset against different uses are different demands made by various representations. 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.

Some types of representations release a visual system from slavery to spatial structures closely related to retinal structures, since arbitrary labels may be used and the groupings need not be determined solely on the basis of spatial and optical properties. Mechanisms enabling this flexibility could at the same time release vision from slavery to spatial structures altogether, since the same symbolisms and inference mechanisms would be applicable to descriptions of non-spatial structures.

A particular representational task that requires the use of fairly abstract descriptive labels is the representation of potential for change, discussed above. This is of profound importance for intelligent planning and control of actions, yet I know of no detailed investigation of the kinds of representational structures that will support this, or algorithms for deriving them from visual information. (Perhaps I am just ignorant of relevant work.)

A naive approach might be to try to represent all the different possible situations that could or could not arises 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, already mentioned in connection with the description of empty spaces, 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 an 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 map in which each visible scene fragment has a label summarising its possible movements.

19 Kinds of visual learning

Any mechanism that supports the use of an abstract description language sufficiently general to contain explicit symbols for parts, properties, relationships, possible changes in the scene and functions of objects, would also, in principle, provide the basis for a powerful visual learning capability, since the syntax of the language would enable yet more descriptors to be introduced. The problem is what sort of mechanism would enable this to occur as a result of training, or experience.

There is plenty of evidence 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 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 vehicle adequately. (This may, however, depend on the frequency with which they have to decide.)

Identical twins provide a very interesting example of visual learning. 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. It is as if the frequent need to tell the difference somehow causes the visual system to enrich its descriptive output so that it includes features required for the particular discrimination task.

Many sporting activities also seem to involve the training of new discriminative abilities. A boxer has to learn to detect incipient movements that indicate which way the next punch is coming. Batsmen in cricket, like tennis players, have to learn to see features of the opponent's movements that enable appropriate actions to be initiated at a very early stage.

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". They go further if they require the creation of new concepts, not just new combinations of old concepts.

Some of these forms of learning are 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.

There are many kinds of high-level learning, such as learning new faces or the names of new kinds of objects. This may or may not involve consciously associating a name with the object. Recognition is often thought of as involving the production of a name. But this is just one kind of response to recognition. Physical reflex responses without the intervention of explicit recognition or description also appear to be learnable.

The need for speed in dangerous situations suggests a design in which the triggering of a response is done directly, 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. The sporting activities mentioned above may involve both the training of new kinds of discriminative capabilities and the development of new routes for the output of visual processing.

Fine control of physical movements is another use where it might be advantageous in some cases to have a direct link from intermediate stages of the visual system to whichever part of the brain is executing the action, instead of going through a central database.

I conjecture that learning to sight-read music makes 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, 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.

20 What changes during visual learning?

The discussion so far suggests that there are two very different sorts of changes in visual learning. (a) The output of the visual system for a particular type of input may change, and (b) *where it goes* may also change.

During conventional processes of learning to read, 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 seems to involve direct stimulation of semantic knowledge, not only by-passing phonetic representations, but almost as if the recognition and checking of printed characters or words had been by-passed.

This suggests that instead of substructures always being recognised and associated with names that can enter into descriptions of more global spatial structures, combinations of low-level features may, as a result of training, be directly associated with lower-level units in a non-visual sub-module in the brain. Direct stimulation of such modules could invoke non-visual processes, such as the construction of sentence interpretations.

Learning to read fluently seems to illustrate both sorts of learning, namely making new visual discriminations and categorisations and sending the output direct to new sub-systems in the brain instead of going through a central database of 3-D scene descriptions, as suggested by the

modular theory. If full 3-D structural descriptions 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.

If there are these two kinds of learning then we should expect to find at least two ways in which the process of learning to read can go wrong, and two different kinds of effects of brain-damage. The same argument would apply to other sorts of learning with both facets. For example the boxer has to learn both to discriminate different kinds of incipient movements and to route the visual information to appropriate motor sub-systems. So 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.

Resolution of the empirical questions about what actually happens in humans may have to await substantial advances in our understanding of the functional organisation of the brain. However, from a theoretical design point of view we can see that the sorts of mechanisms sketched here would enable an intermediate stage to be bypassed, which could be important where speed mattered.

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 and MacDonald 1976).

21 Triggering mental processes

Besides triggering physical responses visual stimulation can trigger new mental processes. The most obvious case is silent fluent reading. Another common example 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 produce sexual responses. 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 1987).

None of this is 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 the general purpose database, it is at least possible that similar mechanisms could directly trigger other mental processes, in some cases after appropriate training. Which happens in humans is an empirical question.[10]

[10] 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,

22 The enhanced model

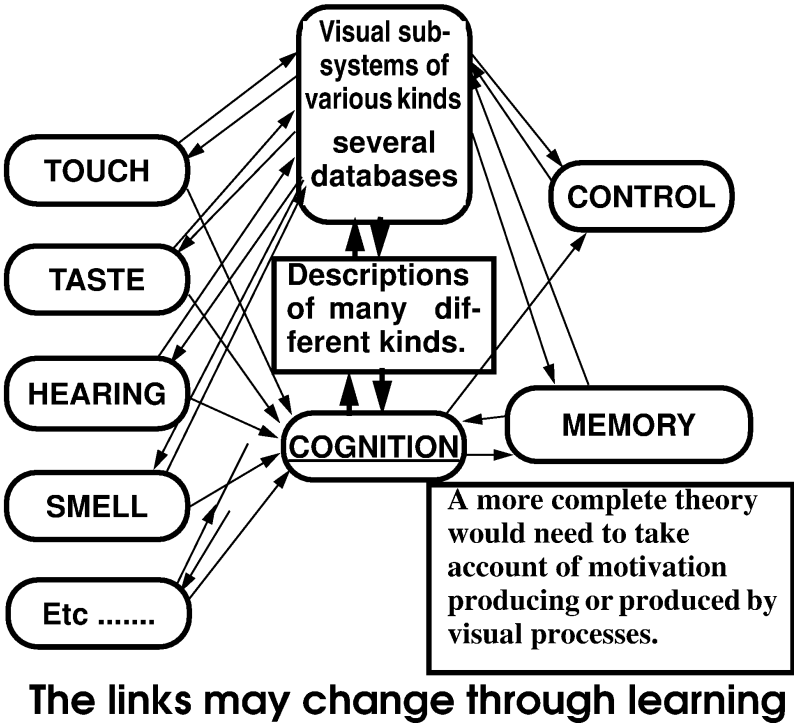

Figure 6: *A more "integrated" model of vision. A crude and incomplete depiction of an architecture in which perceptual subsystems are more closely integrated with one another and with the rest of the architecture. Action subsystems are not shown.*

I have contrasted the modular theory of vision with a theory postulating a wider variety of functions performed by the visual sub-system. On our revised model the visual system is not a narrowly constrained sensory module taking in only image data and transforming it to 3-D descriptions of the environment stored in a central database where they can be accessed by other sub-systems. Rather, the inputs to vision may include information from other sensors, and hints, questions, or tasks specified by planners and other cognitive mechanisms. Further the outputs may include a range of descriptions including 2-D image structure, functional and causal descriptions, descriptions of mental states of agents, and perhaps the meanings of printed text. In addition to descriptions, the output can include stimulation of other modules that may need to react quickly to produce either physical responses or new mental processes. Moreover, the range of descriptive outputs and the range of connections to other sub-systems can be modified by training, rather than being rigidly fixed. The visual system and its relationships

such as a violin.

This ability to learn by watching an expert may be connected with the involuntary physical movements that sometimes accompany watching sporting events. It is as if our visual systems are directly connected to motor-control mechanisms. This would obviously be of biological value as a way of passing on skills from adults to the young.

to other sub-systems may be crudely characterised in Figure 6.[11]

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?

I have previously hinted that the answer to this lies in the nature of the primary input, namely the optic array, and in the way information is organised. Very roughly, in a visual system, input data and intermediate partial results of the interpretation process, are all indexed according to spatial location in a two dimensional field corresponding to the 2-D structure of the optic array sampled by the retina. In other words, information is indexed by means of location in a network of 2-D spatial relationships, an example of what I have previously called 'analogical' representations (e.g. Sloman 1978, Sloman 1985).

We may call these 'optically-registered' databases. They will not be closely tied to the retina itself, since rapid eye movements can constantly change which portions of the optic array are sampled by which portions of the retina. More likely the databases are related to the optic array itself. Moreover, it may be useful to construct a collection of separate 2-D databases for different perceived surfaces. For example the floor of the room would 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.

This indexing by location in 2-D structures is illustrated by Marr's use of 'place-tokens' (Marr 1982, p.51) and the collection of 'intrinsic images in registration' proposed by Barrow and Tennenbaum (1978) as well as some of the comments of Barlow (1983) about topographic maps in the brain. However, the information stored in this spatially-indexed way need not itself be spatial, as we have seen. In some cases it will not even be descriptive information, but may be procedural, for instance if steps in a plan are mapped into it. (If these 2-D databases 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.)

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, my claim is that it is the use of databases whose structure is closely related to the structure of the incoming optic array that makes a process visual as opposed to just a cognitive process. Even if some of the databases are not structured in this way, it is the fact that their contents point into the image-registered databases and are pointed to by such databases that makes them part of the visual system.

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. However, a complete theory of vision will require a general account

[11] Added 8 Oct 2012: These ideas were further developed in the CogAff schema and H-Cogaff architectures later developed at Birmingham university, as explained in `http://tinyurl.com/BhamCog/`.

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.

23 Conclusion: a three-pronged objective

This paper has three main targets. Confusing them can easily lead to arguments at cross-purposes. 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 analytical discussion says nothing about how any actual visual system works or how any practical robot should be designed. Second, and far more tentatively, I have produced some fragments of evidence suggesting that human perceptual systems can be construed as using the non-modular 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. Finally, the discussion suggests that in at least some cases, the multi-connection multi-function design may actually be useful for practical engineering purposes. This will turn out false if in practice it cannot be implemented at reasonable cost. The third claim says nothing about how any actual biological visual system works.

The three prongs, the theoretical, the empirical and the normative theses can be mutually reinforcing, but in fact they are distinct. So anyone wishing to kill this hydra will have to chop off three heads.

24 Acknowledgement

The work reported here was supported by a fellowship from the GEC Research Laboratories and a grant from the Renaissance Trust. Much of this paper expands ideas put forward in Sloman 1978 and Sloman 1982. I am grateful to Chris Darwin and David Young for references to empirical research results. The ideas reported here have been influenced by discussions over many years with colleagues at Sussex University, especially Geoffrey Hinton, David Hogg, Christopher Longuet-Higgins, Guy Scott and David Young.

25 References

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

Ballard, D.H. and C.M. Brown, *Computer Vision,* Prentice Hall 1982.

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

Barrow, H.G. and Tenenbaum J.M. 'Recovering intrinsic scene characteristics from images', in Hansen and Riseman (eds), 1978

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

Brady, J.M. 'Artificial Intelligence and Robotics', *Artificial Intelligence,* 26,1, 1985, North Holland.

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

Charniak, E and D. McDermott, *Introduction to Artificial Intelligence* Addison Wesley, 1985.

Craik, Kenneth, *The Nature of Explanation*, CUP, 1943.

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

Fodor, J. *The Modularity of Mind,* MIT Press 1983

Frisby, J.P., *Seeing: Illusion, Brain and Mind,* Oxford University Press, 1979

Fu, K.S. (ed), *Syntactic Pattern Recognition Applications,* Springer-Verlag 1977.

Fu, K.S., *Syntactic Pattern Recognition and Applications,* Prentice-Hall, 1982.

Gregory, R.L. *The Intelligent Eye,* Weidenfeld and Nicolson, 1970.

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

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

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

Hinton G.E. 'Shape representation in parallel systems' *Proceedings 7th IJCAI, VOL II,* Vancouver, 1981.

Lindsay, P.H. and D.A.Norman, *Human Information Processing: An Introduction to Psychology,* 2nd edition, Academic Press, 1977

McGurk H and J MacDonald, 'Hearing lips and seeing voices', *Nature,* 264, p.746-748, 1976.

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

Marr, D. *Vision,* Freeman, 1982

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

Nishihara, H.K., 'Intensity, Visible-Surface, and Volumetric Representations' in Brady 1981.

Scott, Guy L. *Local and Global Interpretation of Moving Images,* Cognitive Sciences, University

of Sussex, D.Phil thesis 1986.

Sloman A. *The Computer Revolution in Philosophy: Philosophy, science and models of mind*, Harvester Press and Humanities Press, 1978

Sloman A. 'Image interpretation: The Way Ahead?', in O.J.Braddick and A.C.Sleigh (Eds) *Physical and Biological Processing of Images* Springer-Verlag 1983.

Sloman A., 'Reference without causal links' in L. Steels, B. du Boulay, D. Hogg, editors, *Proc 7th European Conference on AI,* Brighton, 1986, North-Holland.