# WHAT ARE THE PURPOSES OF VISION?

Aaron Sloman

1987



Cognitive Studies Research Papers

Serial No. CSRP. 066

The University of Sussex
Cognitive Studies Programme,
School of Social Sciences,
Falmer, Brighton BN1 9QN

Presented at:
Fyssen Foundation Vision Workshop
Versailles. March 1986
Proceedings edited by M.Imbert

# WHAT ARE THE PURPOSES OF VISION?

Aaron Sloman
School of Cognitive Sciences
University of Sussex
Brighton

## CONTENTS



## 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 avalailable 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 subsystems can access it, or does it feed information direct to a variety of subsystems? 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, 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.

## 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-dimensinal 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 workshop attempted to relate their work to a characterisation of the visually accessible 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.

I shall offer an alternative model, 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 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

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.

## Previous false starts

I believe this modular theory of vision 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 seen before: merely attaching a known label does not do this. Of course recognition of substructures and of relationships is part of the process of producing a structural description.

- 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 is needed, as well as analysis.

- 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 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.

- 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.

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

It is important to distinguish three different sorts of question, empirical, normative and theoretical. The empirical question asks what actual biological visual systems are like and what they are used for. The normative question asks what sort of visual system would be desirable for particular classes of animal or robot. The theoretical question asks what range of possible mechanisms and purposes could exist in intelligent behaving systems, natural or artificial and how they might interact with other design options.

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

## 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.

## 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 be 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.