

# Aiming for More Realistic Vision Systems

Aaron Sloman and members of the CoSy project

## Abstract

An earlier position paper made a number of claims about requirements for vision systems that matched functionality of advanced biological vision systems, including human vision systems. It was suggested that most vision research ignored important requirements for vision to support action. Since then our understanding of those requirements has been enhanced by analysis done within an EC-funded robotics project. This paper extends points made in the earlier paper in the light of more recent investigations. Whereas previously it seemed that the function of vision was best thought of as providing information about spatial structures and affordances at different levels of abstraction, there are now reasons to regard it as primarily concerned with providing information about processes at different levels of abstraction, in partial registration with the optic array. This can require running multiple concurrent simulations. There are precursors to this idea, including Max Clowes' slogan: 'Vision is controlled hallucination'. In this context static scenes are merely a special case of processes, where nothing is happening, and the perception of affordances, construed as possibilities for and constraints on change, can be explained as using the ability to run simulations. In addition, reasons are given for regarding much of the information provided by vision as a-modal, and also as intimately bound up with causation, in a way that relates to the human ability to do mathematical reasoning visually. There are many hard problems regarding how to represent spatial information. This paper outlines some of the requirements rather than designs.

## 1 Introduction

This is a sequel to a paper written about five years ago that drew attention to the excessively narrow focus of much vision research and cast doubt on the relevance of much of it to the design of complete systems (e.g. robots) that make use of vision and other senses in learning about and acting in a 3-D physical environment, as humans and other animals do. That paper was written before Betty, the New Caledonian Crow amazed the world by demonstrating an untutored ability to make hooks from wire in order to lift a bucket containing food out of a vertical transparent tube [13].<sup>1</sup> As far as I know there is no artificial visual system that is coupled to a robot manipulator that comes near providing a creative artificial system with the information that might provoke the realisation that a straight piece of wire could be bent by sticking one end into a crack or hole in a rigid object or by sticking the end into an object made of a substance that will allow the end to penetrate while resisting subsequent sideways motion.

---

<sup>1</sup>See the videos on this web site: [http://users.ox.ac.uk/~kgroup/tools/tools\\_main.html](http://users.ox.ac.uk/~kgroup/tools/tools_main.html)

Most of the work on vision (much of it driven by benchmarks that allow ‘progress’ on a very narrow type of task to be measured precisely) is not concerned with perception of 3-D structure and motion of non-rigid objects, or the causal powers and limitations of objects in the environment determined by their shape, the materials of which they are made and their relationships to other things. Instead there seems to have been a vast amount of effort expended on systems that

- aim to recognise or classify objects or people in static images, without acquiring, using or reasoning about information about 3-D structure,
- aim to track moving objects treated as points or blobs or other simple shapes – often regarded as 2-D shapes,
- explore an environment building a 2-D map of walls, doorways, passages, etc., without necessarily emulating a human’s understanding of such maps,
- aim to control motion of a mobile robot that is essentially regarded as a moving point that can alter its location and direction of view, rather than as something that can manipulate things in the environment,
- aim to obtain some 3-D information about the environment but only sufficient to allow generation of new images, for instance images showing things from a different viewpoint, rather than, for instance, images showing possible future actions in which the objects are manipulated by being grasped, moved, bent, stretched, prodded, disassembled, inserted into gaps, used in constructing a new object, etc.

## 2 What Freddy did in 1973

A robot [1] that was actually built in Edinburgh and demonstrated in 1973, is shown in Figure 1.<sup>2</sup> Because the hand was so heavy and most of the desired movements would have been very difficult on a mobile arm, the horizontal movements were done by moving the table in two directions. Because it took over 20 minutes to find regions and region boundaries in an image in those days, and because the hand got in the way of the camera, it was impossible for Freddy to perceive what it was doing while acting: it had to look, think, then act blindfold – which is still how some people think AI systems have to work, using a ‘sense–decide–act’ cycle. A very different view is presented below (some aspects of which were described long ago in [12]). Another serious limitation of Freddy was that it was not able to perceive 3-D shape. Instead, objects were represented in terms of known types of 2-D patterns, and prior learnt knowledge about relationships between those patterns and actions required to grasp and move things was used, albeit in a flexible manner that coped with a wide variety of configurations of objects on the table. If presented with a new object, such as a tea-cup, Freddy would not have been able to perceive its 3-D shape: unlike humans and many other animals it could not see objects it had not learnt to recognise in a training process. Despite these limitations, Freddy was a very impressive achievement for its time.

---

<sup>2</sup>The main programs ran on an Elliot 4130 computer. Imagine using a computer with 128KB RAM for a robot now, with a CPU frequency of about 0.1Mhz! The operating system and the AI systems were all implemented in the Edinburgh AI language POP2. An auxiliary Honeywell computer provided interfaces to camera and arm. There is more information on Freddy here <http://www.ipab.informatics.ed.ac.uk/IAS.html>

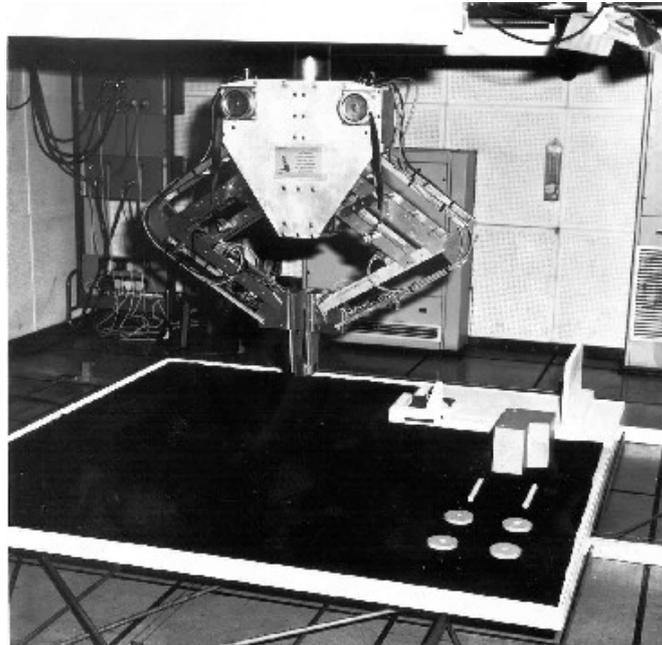


Figure 1: Over 30 years ago, Freddy, the Edinburgh Robot was able, in 1973, to assemble a toy car from the components (body, two axles, two wheels) shown. They did not need to be laid out neatly as in the picture. However, Freddy had many limitations arising out of the technology of the time. E.g. Freddy could not simultaneously see and act: partly because visual processing was extremely slow, and perception of the car parts required them to be separated (using a ‘heap-smasher’ routine) if jumbled, and recognition depended on the fact that each part had only a few stable configurations. (Picture © University of Edinburgh )

The limitations of Freddy were recognised of course, and during that decade there were some attempts to move beyond 2-D pattern recognition towards understanding 3-D structure. One of the people who had worked on Freddy was co-author of the (temporarily) very influential 1978 paper [2] discussing use of 2-D monocular cues to support perception of local 3-D spatial structure (represented in ‘intrinsic images’). Other work around that time began to explore structure from motion, shape from intensity variation, and the use of stereo.

However, the work was seriously hampered by limitations of computer power and the difficulty of selecting good representations for the task, and it seems that the investigation of shape perception was mostly abandoned in favour of work on recognition, tracking, and localisation, with little or no work on understanding of 3-D spatial structure.

There has been a great deal of progress on specialised image processing tasks, driven by collections of benchmarks which have little or nothing to do with the ability to act in the environment, for instance benchmarks concerned with recognition of hand-written characters, faces, or types of animals or vehicles, but without perceiving any 3-D spatial structure or providing any information relevant to action.



Figure 2: *Most people see the 3-D scenes depicted here fairly clearly, though not with great precision, as the images are noisy and have low resolution (160x120 pixels – taken using a webcam in a hotel room with poor lighting). A normal adult looking at the pictures can easily visualise a sequence of actions that would transform the configuration on the left to the configuration shown on the right, and vice versa, though not with perfect precision. Despite what humans can do with such images, it is still the case that seeing the 3-D spatial structure and the affordances in such images is beyond the capabilities of current vision systems. If humans can see a lot of structure in such static, monocular, monochrome images, that demonstrates that the problems do not require use of higher quality cameras, colour vision, motion or stereo. Something deeper than better technology is required.*

There has been a considerable amount of work on visual acquisition of information about surface structure, e.g. using laser range finders. The capabilities of such systems are often demonstrated by projection of images from new viewpoints. But those vision systems show little or no *understanding* of 3-D structure and its relevance to *action* (e.g. through understanding affordances [7]). For example such systems don't understand how features like curvature, orientation, hardness, smoothness, graspability, vary from one part of the surface of an object to others. (Apologies to anyone whose work is an exception – it has been hard to find!)

### 3 What current systems cannot do

Consider what a human can see in Figure 2(a) (which has low resolution, poor lighting and much noise). People tested by the author looking at that picture can select different parts of the objects that could be grasped using finger and thumb or a tool bringing two surfaces together. Moreover, they can see roughly the orientation required for the grasping surfaces for different grasping locations. For instance the orientation required varies smoothly all around the rim of the cup, along the edge of the saucer, on the handle, etc. Understanding how to align grasping surfaces with grasped surfaces can be an important part of shape perception, since, for example, if the cup is grasped without the surfaces first being aligned then the result will be enforced rotation of the cup if the fingers are firmly compressed and the wrist rigidly held in a fixed orientation. Of course, use of a compliant wrist along with force feedback can reduce the need for precision in the pose

of the grasping fingers.

Nevertheless there is a requirement for visually controlled grasping to use some structural information. For instance if the fingers approach the rim from entirely the wrong direction they will merely hit a vertical surface and not be able to grasp the rim. Moreover, different approach angles would be required if the cup were lying on its side on the saucer, in which case the opening could face in any direction.

Perceiving the relationships between shape and grasping is far more important when the grasped object is too big to fit as a whole between the grasping surfaces at their maximum separation. However even if the maximum diameter is small enough it may still be necessary to understand the affordances involved in the shape, for instance if there are dangerous sharp points sticking out of some parts of the surface, or if some parts of the object are more delicate and fragile than others. In all cases the perception of affordances goes beyond seeing what actually exists in the scene. It requires using what exists as a basis for understanding things that could happen, and constraints on things that could happen. In other words it requires a grasp of counterfactual conditionals involving *processes* that are not actually occurring in the scene.

Apart from the (positive and negative) affordances for action visible in the shape of the surface of the cup, there are many more affordances concerned with lifting the cup, pouring something into or out of it, sipping from it, throwing it, etc. All of these involve future possible processes: process that would or would not occur if such and such an action were performed. So once again, insofar as vision provides information about affordances, it provides information about what could exist, and not just what does exist in the environment.

However, as far as we know, there are no AI vision systems that can perceive surface structure in such a way as to produce an understanding of affordances that typical human children of about 2 years have. Of course, they do not understand all the positive and negative affordances provided in 3-D objects which they can manipulate, as is evident if a 2-year old is given plastic meccano or other toys requiring insight into shape and motion constraints. For example, a child whose ontology does not include the shape of a boundary or the possibility of two non-symmetric boundaries being aligned or mis-aligned, will not see the requirement for putting a jigsaw puzzle piece into its recess. A simpler ontology suffices for seeing the affordances in stacking circular mugs, because their symmetry does not allow certain kinds of mis-alignment to occur. So such stacking competence comes earlier. Over several years, a child learns to see and use increasingly complex affordances based on visible surface structure, materials of objects, etc.

## 4 Stacking non-blocks

After a while that competence can be used in planning and executing quite complex sequences of actions, including novel sequences (as Betty did when making hooks from wire in several different ways).

Many people (e.g. [4] – though in fairness that author has recently changed his views) criticised early work in AI because of its concern with allegedly trivial tasks in toy worlds, e.g. stacking blocks, and the assumption that symbolic reasoning, such as explicit planning, is relevant to how robots need to act. One might have expected such critics to go on and demonstrate how their proposed methods could work in more demanding

situations, such as transforming the configuration of objects in Figure 2(a) into the stack shown in Figure 2(b) and vice versa. However extensive internet searches and letters written to leading vision researchers have indicated that this is still far beyond the state of the art.

We suggest that the reason for disappointing progress and unfulfilled promises in early AI work was wrongly diagnosed: it had nothing to do with choice of toy domains or use of specific representations and modes of reasoning. Rather the problem was over-ambitious promises based on inadequate analysis of *requirements* for a vision system that is capable of doing what young children and many animals can do. The requirements may *seem* to be much simpler than they really are if the tasks are not analysed in sufficient depth. If they are not analysed properly no implementation techniques will solve the problem. A failure to recognize the complexity of the requirements leads to over-optimistic predictions, as happens repeatedly in AI.

To illustrate the unobvious complexity: one of the features of human competence in relation to the stacking task (whether rectangular blocks are used or objects with more complex shapes) is that that different aspects of the task require different sorts of representations, i.e. different information structures encoding information at different levels of abstraction.<sup>3</sup>

For example, if vision is used to control the action of grasping a mug in a particular place, then relatively precise information is needed about whether the hand is moving in the right direction, rotating to the correct angle of approach, and moving the fingers in an appropriate trajectory to the selected grasping point. In many cases this does not require millimetre precision since getting within a few millimetres and then closing the grasp will suffice. However, higher precision is required when manipulating tweezers to select a hair.

Much lower precision may be required for seeing that something is not graspable, e.g. because it is too far away, too big, has no suitable protrusions, is moving too fast, etc. Intermediate levels of precision may suffice for thinking about how an action could be done, what the consequences would be, and what might go wrong if a mistake is made or if some potential obstacle is moved during the action.

For some tasks it may be enough to discretise possibilities, for instance when planning future actions, explaining past events, or giving someone else verbal instructions ‘Put the cup on the table then put the saucer on the cup, and then the spoon on the saucer’.

All of this implies that a visual system that is capable of representing shape and surface structure only in one way will not be cognitively adequate, even if it has great mathematical precision, for instance using formulae from differential geometry with precise parameters.

It seems that we need generalisations of work in qualitative representation and reasoning to be applicable to 3-D features and relationships. A more detailed analysis requires further research, but it is expected that that different kinds of learning process will provide both topological and metrical ontologies, allowing different ways of chunking continuous

---

<sup>3</sup>Like most of the scientific community outside ‘Nouveau AI’, we use the word ‘representation’ to cover a very wide variety of means of encoding information in a usable form, whether transient or enduring, whether implicit in process activations or explicit in manipulable structures, whether expressed in a simple fixed-complexity syntax such as vector forms, or in variable complexity grammatical forms, whether continuous or discrete, whether stored internally or stored externally. However in this paper we use ‘representation’ to refer to *internal* information structures unless the qualifier ‘external’ is used. Of course, internal structures may refer to external, states, processes, objects, relations, etc.

spaces, reducing resolution, and developing measurements of varying precision relative to features of the perceiver such as proportions of the visual field, proportions of various kinds of stretch the agent is capable of, proportions of lengths of body parts of the agent, and also the use size and angle representations relative to other things in the scene. (Some of these are discussed in connection with infant development in [9]. Related points are made by [3].)

## 5 Perceiving processes

All of this has implications for perception of processes – the most general and pervasive form of perception since very few animals are located in static environments: normally they, conspecifics, prey, predators and other objects are constantly on the move.

It might be thought that the only implication is that a visual system needs to build optical flow maps. However the previous discussion implies that even though such maps could be useful in various ways (some discussed by Gibson in [7]) much more is needed.

All spatial objects in addition to having spatial relations to other objects also have many parts that have spatial relations to each other and to other objects. For example a cube has faces, edges and vertices that are related to one another and if there is another cube in the vicinity the faces, edges and vertices of the two cubes will all be related. Moreover, not only is this a geometrical fact about the cubes it is also the case that many such relations can be perceived, even if not all of them will be perceived in any given situation. Likewise different parts of a cup will be related to different parts of a spoon or a saucer or a grasping hand. We can express this by saying that perception can involve multi-strand relationships requiring much richer forms of representation than a logical form like  $R(a, b)$ .

Moreover, when things move and processes occur, many such relationships can change in parallel. Thus perception of changing scenes can involve representation of several concurrent processes. Our previous discussion of the different levels of abstraction that need to be represented for different purposes carries over to processes. So in general perceiving a process, such as a hand grasping a cup will involve acquiring information about multiple concurrent sub-processes involving changing relations between different parts of the hand and different parts of the cup, at different levels of abstraction, some using high metrical precision, some low precision but still representing metrical information, and some using topological and other relationship that change only discretely.

It seems to follow from all of this that a human-like robot will need a visual system that is capable of simultaneously representing or simulating multiple concurrent processes of different sorts and different kinds of abstraction in partial registration with retinal images, or to be more precise with optic arrays, since retinal images keep changing with saccades.<sup>4</sup>

It may be objected that the need to represent all these concurrent processes at different levels of abstraction is not a requirement for vision, but for a central cognitive system, which will use information from vision and other senses. That may be partly correct, but a sharp separation between cognitive and visual processes is not possible given the requirement for the process representations to have complex structures that are in partial registration with the contents of the optic array. However the mapping will need to be constantly updated because of saccades, head and eye movements, and changes in

---

<sup>4</sup>If the paper is accepted references to our more detailed recent papers will be included.

amounts of detail as objects or parts of objects go in and out of sight, either because they are occluded or because they move out of the field of view without being forgotten. (You probably know about many things in your immediate environment that you looked at recently but cannot see now, e.g. if talking at a dinner table and turning to face different people at different times.)

## **6 Beyond sensorimotor understanding**

It is often thought that since the environment can be sensed and acted on through different modalities and different body parts, understanding the environment involves making use of correlations between contents of different sensory and motor modalities.

This may be adequate for simple organisms (e.g. insects), but for humans and many other species, the fact that (a) some body parts can move and be seen to move independently of the motion of the eyes and (b) similar actions (like grasping, pushing, pulling, twisting) can be performed by different body parts (left hand, right hand, mouth, etc.) implies that an enormous reduction in complexity can be achieved if there are modality neutral, objective representations of objects, relations, and processes in the environment that can be used for reasoning, planning, learning correlations etc., and separate mappings learnt between those objective states and particular modalities.

One consequence of the availability of amodal representations is that they can be used to think about past, future, or distant entities and processes without those thoughts having to specify sensory details. For instance you can think about whether to go to a conference near Paris next year without knowing what the location looks or smells like.

Another consequence is that the same modes of representation can be used in perceiving or reasoning about affordances for others, e.g. predators, prey or young animals needing help. These can be called ‘vicarious affordances’. If this is right then visual learning needs to be linked not only to sensorimotor contingencies but to ‘objective’, amodal, condition consequence contingencies which are facts about the environment, not about one’s experience of the environment.

Since the variety of types of surfaces and 3-D orientations of graspable, touchable, pushable, pullable surface fragments is astronomical any attempt to learn about such affordances by storing sensorimotor correlations will founder on a combinatorial explosion.

So for real progress in this field to occur it will be necessary for substantially new kinds of AI vision systems to be developed using new ontologies including means of representing processes that involve different concurrent sub-processes, all at different levels of abstraction that can change according to task demands.

## **7 Understanding causation**

It seems that at least humans, though probably not new born infants, can not only perceive processes of the sorts described above in which many things change concurrently – they can also visualise or imagine them when they are not happening. This is an important part of the ability to make plans or to solve problems involving spatial configurations (as discussed long ago in [11]). Insofar as this includes the ability to propagate constraints in changing representations and to understand why certain changes necessarily produce others it provides the basis for a kind of causal understanding that is structure-based and deterministic (Kantian), rather than the purely correlational (Humean) type of causation that has been most studied recently in AI and philosophy.

This also seems to be the basis for much human mathematical competence especially in learning about geometry and topology. The full implications of this will need to be discussed on another occasion.

## 8 Work to be done

There is much work still to be done on the forms of representation, mechanisms, architectures and varieties of learning that can occur in animals or robots of the sort discussed here. It seems likely that currently understood learning mechanisms that depend on large numbers of repetitions of examples will be inadequate for the tasks in which humans and some other animals seem to be able creatively to solve new problems as reported, for example, in [8], [9], [13], [5], [10], and many others. There are some overlaps between the requirement for intelligent playful exploration of a richly structured environment and creative learning about mathematics, guided in part by a notion of ‘interestingness’, as described in [6].

All of these ideas need to be made more precise. But it is likely that considerable benefits can be gained by working backwards from very detailed, even painstaking, requirements analysis for very long term goals, instead of always only looking at marginal improvements that can be made to existing systems.

## 9 Acknowledgements

This work arises from collaboration with a number of others who will be listed in the final version if the paper is accepted.

## References

- [1] A. P. Ambler, H. G. Barrow, C. M. Brown, R. M. Burstall, and R. J. Popplestone. A Versatile Computer-Controlled Assembly System. In *Proc. Third Int. Joint Conf. on AI*, pages 298–307, Stanford, California, 1973.
- [2] H.G. Barrow and J.M. Tenenbaum. Recovering intrinsic scene characteristics from images. In A. Hanson and E. Riseman, editors, *Computer Vision Systems*. Academic Press, New York, 1978.
- [3] Alain Berthoz. *The Brain’s sense of movement*. Perspectives in Cognitive Science. Harvard University Press, London, UK, 2000.
- [4] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.
- [5] J Chappell and A Kacelnik. New Caledonian crows manufacture tools with a suitable diameter for a novel task. *Animal Cognition*, 7:121–127, 2004.
- [6] Simon Colton, Alan Bundy, and Toby Walsh. On the notion of interestingness in automated mathematical discovery. *Int. Journ. of Human-Computer Studies*, 53(3):351–375, 2000.
- [7] J.J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986. (originally published in 1979).

- [8] W. Kohler. *The Mentality Of Apes*. Routledge & Kegan Paul, London, 1927. 2nd edition.
- [9] Philippe Rochat. *The Infant's World*. Harvard University Press, Cambridge, MA, 2001.
- [10] L. R. Santos, N. Mahajan, and J. L. Barnes. How Prosimian Primates Represent Tools: Experiments With Two Lemur Species (*Eulemur fulvus* and *Lemur catta*). *Journal of Comparative Psychology*, 119:394 – 403, 2005. 4.
- [11] A. Sloman. Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, London, 1971. Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971.
- [12] A. Sloman. *The Computer Revolution in Philosophy*. Harvester Press (and Humanities Press), Hassocks, Sussex, 1978. Online at <http://www.cs.bham.ac.uk/research/cogaff/crp>.
- [13] A A S Weir, J Chappell, and A Kacelnik. Shaping of hooks in New Caledonian crows. *Science*, 297(9 August 2002):981, 2002.