WHY VISUAL SYSTEMS PROCESS SKETCHES

Aaron Sloman and David Owen [*1]
Cognitive Studies Programme,
School of Social Sciences,
University of Sussex,
Brighton, BN1 9QN, England

Abstract

Why do people interpret sketches, cartoons, etc. so easily? A theory is outlined which accounts for the relation between ordinary visual perception and picture interpretation. Animals and versatile robots need fast, generally reliable and "gracefully degrading" visual systems. This can be achieved by a highly-parallel organisation, in which different domains of structure are processed concurrently, and decisions made on the basis of incomplete analysis. Attendant risks are diminished in a "cognitively friendly world" (CFW). Since high levels of such a system process inherently impoverished and abstract representations, it is ideally suited to the interpretation of pictures.

1. Is the study of impoverished pictures relevant to 'real' vision?

AI vision work concerned with pictures, including digitised photographs, straight-line drawings and cartoons, etc. has recently been criticised as irrelevant to visual perception of objects in the environment, Clocksin [1978]. Related themes can be found in Horn [1978]. It can be argued that studying impoverished pictures with great local ambiguity leads to overemphasis on top-down, knowledge-guided visual processes, as in Shirai [1975] and Minsky [1975], and on complex control structures, as in POPEYE [*2]. Lack of detail in artificial images causes difficulties of interpretation which, it may appear, do not arise in ordinary perception, where disambiguating detail is provided by colour, stereopsis, optical flow, etc. Admittedly images interpreted by most A.I. programs lack many features available even in monocular perception of static scenes, from which useful information can be extracted with powerful algorithms and computational resources. [Marr 1976, Horn 1978 and papers cited therein]. Horn's claim: 'we may have closed our eyes to the raw image for too long', is reasonable, and supported by his own excellent work on images. But we mustn't now close our eyes to all else.

Extraction of low level image and scene features is but a sub-process of the visual mechanism. That a powerful subsystem is normally used does not imply that it is essential for vision. Stereopsis certainly occurs, and needs to be explained, but our ability to perform everyday tasks with just one eye also needs explanation. Similarly, we can often recognise things when detail is missing, or spurious information added, through poor visibility, eye defects, strong back-lighting, restricted view angle, or intervening shruberry. Normally, we use perceived detail to segment the scene into objects, but sometimes the grouping must go beyond consideration of image continuities and discontinuities because of occultation of some objects by others, camouflage, shadows or spurious juxtapositions. All this suggests considerable modularity: various sub-systems produce information, perhaps partly duplicated by other sub-systems, and less precise information may suffice if the ideal is not available. This modularity could allow a component which ideally should be driven by the data, to be driven instead by prior knowledge activated by other data. This might explain both our facility with impoverished pictures and the occurrence of misperceptions even in excellent conditions (well-known examples

are the hollow mask [Gregory 1970], and the triangle containing "PARIS IN THE THE SPRING").

How can we function so well when so much detail is lost? Recognition of sketchy drawings can be rapid and effortless [Hochberg 1978, page 193, citing Ryan and Schwartz]. Perhaps, when we look at pictures, intermediate results of the interpretation process are similar to some intermediate (sketchy) results of the processes of normal perception? Perhaps normal perception uses mechanisms with built-in characteristics designed to cope with abnormal, specially difficult, situations? Our central idea is that visual systems process many different domains of structure in parallel. So analysis of relatively "high-level", abstract, incomplete, representations, sometimes occurs in parallel with detailed analyses of visual data. [*3] Higher level processes would then be driven in part by prior knowledge of specific sorts of objects (e.g. generalised cylinders, humanoid figures), lower levels mainly by very general (implicit) knowledge about 3-D surfaces, lighting, motion, etc. Occasionally such high-level processes would reach conclusions which are overturned by more detailed analysis, e.g. the "double take". However the different processes would normally produce compatible results, making possible the modularity referred to above. How?

A basic assumption is that the visual system has evolved to work in a "Cognitively Friendly World", a CFW, (which may be very unfriendly in other respects). Here are examples of cognitive friendliness:

(A) The optic array is rich in useful information about the environment -- as noted above. This is due in part to the sorts of surfaces objects have, in part to a plentiful supply of short wave-length radiation and a transparent atmosphere. (N.B. the last two conditions are very variable.)

(B) The space of physically possible objects and processes is sparsely instantiated in the actual world (unlike science fiction), i.e. there is limited independent variation of features and relations: this makes images redundant. This is illustrated by planarity, continuity, rigidity, etc. (Marr [1979]) and the fact that no animal has the ear of a zebra and the body of a giraffe.

(C) Confusing coincidences (e.g accidental alignments and juxtapositions) are rare. This depends both on the kind of environment and on the low probability of such viewpoints for any given scene.

To make use of (A) a visual system needs good detectors for features of the optic array. Since these depend on laws of physics they don't vary much from one part of the world to another and can be usefully compiled into hardware. If we have evolved mechanisms to take advantage of (A), might we not also have evolved mechanisms to take advantage of (B) and (C)? Using (B) requires using knowledge of what sorts of objects actually occur, e.g. knowing about cylinders, about rigidity, and about zebras and their ears. Some of this (e.g. many objects are locally rigid) is useful in nearly all environments, and might be built into genetically determined mechanisms, whilst some (e.g. what sorts of plants, animals, or buildings, are common) will vary considerably and must be left to individual learning. Making use of (C) involves having good process organisation, to find the 'best' percepts [Hinton 1977].

A consequence of (B) and (C) is that usually any good interpretation of a visual image will be unique, and therefore the best one. (B) and (C) could also justify higher level processes jumping to knowledge-guided conclusions on the basis of partial results from lower levels. This could enable good decisions to be made in poor viewing conditions, and in good conditions would enable decisions to be made faster. (All of this is demonstrated in a very simple world, by the Popeye program [*2].) So, assumption (A) is of use in good viewing conditions where objects are unfamiliar, whilst (B) and (C) are of use where

conditions are bad but objects are familiar. A system designed with this flexibility might acquire a speed advantage where all of (A) (B) and (C) are satisfied, if different sub-systems work in parallel. It would still have to be basically data-driven (bottom-up) if serious mistakes are to be avoided, but it need not be pass-oriented, with each layer waiting for lower levels to "complete" their analysis (if completion has any meaning).

If higher level systems can operate thus on impoverished data available in adverse conditions, and on incomplete, partial, results of lower levels in good conditions, they should also be able to interpret some highly impoverished artificial data, such as we find in pictures. If so, the interpretation of pictures is not merely a culturally specific, learned, process. If ordinary perception of objects and relationships requires learning, then interpretation of pictures of the same objects will not normally require additional learning, on this view: toddlers we have observed respond naturally to cartoon drawings of familiar situations, without anything like the struggle which characterises learning to read. [Cf. Hochberg and Brooks 1962.] This is not the theory criticised by Gombrich [1960] and Goodman [1969] that realistic paintings and drawings produce the same visual stimulus as the things depicted.

2. Unarticulated, semi-articulated, and articulated representations.
We have claimed that vision requires far more than efficient detection of features of the optic array, and that several different domains of structure are processed. To explain why, we must ask: what is vision needed for? An animal, or robot, uses perception to make decisions in pursuit of its goals and to tell whether they have been achieved. It also needs to detect unexpected dangers and opportunities. All this requires construction of representations which articulate the environment into objects with properties and relationships of varying sizes and degrees of abstractness. Rarely will the detection of a particular feature in a particular location on the retina be very significant. Similarly, huge data-bases of unarticulated information, like depth-maps, surface colour or texture maps, surface orientation maps, primal sketches [Marr 1976], can be of little use without considerable further processing. They are effectively new, enhanced, images, even though they may contain 3-D information. Though important for further processing, these unarticulated databases are not directly useful for decision and action: only generalisations related to global image statistics can be learned or invoked e.g. 'lots of green', but not 'plum on tree', might be recognised.

To some extent groupings of fragments of information into larger wholes can be achieved by parallel "local" computations, e.g. relaxation techniques linking items subject to constraints [Hinton 1977, Radig 1978, Frisby and Mayhew this conference]. If the links exist without explicit description of the properties and relations of the linked groups, the database is semi-articulated. The process of growing such links may enable some useful global statistics to be collected, but represents objects only implicitly. Though providing a useful intermediate stage, a semi-articulated database does not explicitly represent one object as above, inside, between, or able to fit into, others. Such information is then not available for deciding, planning and learning. (Compare Marr's 'principle of explicit naming' [Marr 1976]. The same point was made in Minsky [1961].)

Further study of visual articulated representations requires analysis of types of actions performed by different animals. (Some birds can learn to use a foot to depress one end of a lever, exposing food behind the other end. This probably involves articulating the lever into parts, e.g. ends, capable of different though causally linked motions.) It seems unlikely that a small number of mathematically simple structures (e.g. generalised cylinders) with a small

number of mathematically simple relationships (e.g. equations linking co-ordinates) will suffice for human perception. Besides crumpled newspapers (despaired of by Marr [1979]) we see fields and forests. Similarly, cluttered scenes made even of "clean" cylinders will have messy structure at larger scales, like large sets of axioms in a theorem-prover's database. To discern significant objects and relations in large and messy collections of image and scene features we need a much richer descriptive vocabulary than AI vision programs have hitherto incorporated. This is why multiple domains are important.

3. Multiple domains

Clowes [1970,1971] and Stanton [1970] stressed that visual perception and picture interpretation do not simply involve description of image structures. They described "mapping rules" linking different non-isomorphic domains. A domain is a class of structures defined by a "grammar" or set of axioms (e.g. 3-D Euclidean geometry). Scenes have quite different "grammars" from images. This needs to be generalised (as in Hearsay and Popeye) to allow many domains, with different though possibly overlapping grammars. Very briefly, this is because using many different domains allows: (a) 'structure sharing' between processes of recognising different sorts of objects, (b) intermediate results of processing to be relatively secure even if back-tracking is required at higher levels, (c) higher levels to recognise important scene features before lower level processing is complete (see below) (d) high level recognition despite poor low-level detail, (e) data derived from an image to be usefully structured (compare 'Scripts' and 'Frames'), (f) goal-directed activation or de-activation of large chunks of knowledge (e.g. 'mental set'), and (g) communication between different sensory modalities.

A.I. vision work has so far focussed on a small number of mathematically tractable special cases. A good survey of the different domains of structures useful in visual perception is still lacking. Likely relevant domains include 2-D arrays of changing colour and intensity, 2-D configurations of lines and regions and of texture, domains involving patterns of motion in both 2-D and 3-D, overlapping 2-D silhouette shapes [Paul 1976], curved and flat 3-D surfaces, both 2-D and 3-D stick figures [Palmer, 1975], various domains involving forces and a variety of cause-effect relations, intentional actions etc., properties like flexibility, rigidity, elasticity, hardness, etc. Besides plane surfaces, edges, vertices and generalised cylinders for representing shapes, we probably need generalised spheres, hemispheres, bags, tubes, strings, etc. In addition we need models for significant parts of such objects and their surfaces, like: hollows, grooves, holes, lumps, ridges, openings, rims, etc., and models for relating one to another (the groove runs across the hollow). Features and relations invariant under non-rigid transformations are particularly important in our world. We also need a large collection of schemas for types of motion and action: moving towards, moving away from, moving into, flattening, twisting, folding. Compare Hayes on 'naive physics' [1979]. Studies of pictures and cartoon movies can yield useful insights into the structures deployed in perception [Draper 1980]. Of course, it is hard to specify how such models may be represented, invoked, etc. in a working, system. Is all this "cognition" relevant to vision? A major feature of visual learning is linking new domains into the visual system - e.g. learning to see the muscular structure of human bodies, for artistic or medical purposes, learning to see when it is safe to cross the road. There is no sharp boundary between practically useful vision and cognition.

4. The domain of images

The structure of the 2-D image domain is important for both picture interpretation and normal vision: why? Goodman [1969 p.38], rejecting the idea that pictures and objects produce similar visual input, accounts for the "realism" of some pictures in terms of familiarity. But this fails to explain why even a two year old child can learn some pictorial styles, whilst others, though mathematically equally adequate, seem much harder. The human visual system does not work with arbitrary combinations of image elements, but, as the Gestalt psychologists noted, is largely constrained to use continuity, proximity, smoothness, concurrency, symmetry, containment, and other geometric and topological relationships, for linking low-level features into cues which invoke more abstract or global representations, which may themselves be similarly treated. A grasp of such relationships is required for interpreting pictures also. However, much richer image description languages are required than existing AI programs can handle: many can only describe the topology, and a few metrical properties, of networks of straight lines or picture regions. Others provide a simple semi-articulated description with no grasp of the implied structure [e.g. Radig 1978].

Further, articulated 3-D interpretations, required for planning actions, can be linked to image structures to facilitate processing. For instance, to answer the question "What is Y going to hit?", "Will I pass near A if I go straight towards B?" one can "traverse" the relevant part of the image to find the relevant bit of the 3-D interpretation. Moreover, our theory implies that in visual perception and in picture interpretation, descriptions of parts of a complex 3-D scene are built up in parallel. The linking of incomplete descriptions of different parts of the scene to form larger structures, will be facilitated if the 3-D structures are closely related to the network of descriptions of 2-D image structure - the latter providing indexing or addressing routes. [*4]. This applies to both real vision and interpretation of pictures. (More on this below.)

So, against Goodman we claim that "familiarity" of pictorial representations is not a matter of frequency, but depends in part on the way 2-D relationships are used in normal vision. Of course, mere similarity of domains does not suffice to explain facility with pictures. Maps also make use of 2-D structures and relationships, yet learning to use a map to find one's way around is harder than interpreting pictures. This is partly because our stored knowledge of objects is addressable by means of the kinds of articulated representations produced by both retinal images and artificial pictures, whereas our 'cognitive maps' of familiar surroundings are not normally addressable by the kinds of structures created when we look at maps. Things might be different if we could fly!

5. Reasons for using impoverished articulated representations

There are additional reasons why impoverished picture structures might be related to normal vision. We have already given a general reason why a visual system needs to be able to cope with impoverished representations: articulation of the scene implies reduction of information. Other reasons concern processing, the purposes of vision and the environment:

5.1. Some details may interfere. Much of the detail available to the eye arises from variable conditions, including lighting, atmosphere, viewpoint, non-rigid motion, and changing relations. The use of abstract schemas implies less memory space, faster matching, smaller searches among stored specifications and enables recognition of individuals or types (abstracting from individual details) in novel circumstances. It also provides the basis for forming generalisations.

5.2. Some details aren't needed in a "cognitively friendly" world. It may be possible to distinguish objects on the basis of only a few features. E.g. a colleague once remarked that he could recognize a zebra with just its ear visible. In a CFW where the space of possible structures is known to be sparsely instantiated ((B) above), inferences can be made from fragmentary evidence.

5.3. Details may be missing or spurious. As already noted, poor visibility, natural or artificial camouflage, eye defects, rapid motion, or the presence of visual obstacles, can produce degraded images. Injury can remove stereopsis. Optical flow is not always available. Stereopsis and optical flow don't help with distant stationary scenes. Extracting global features (e.g. silhouette descriptions) from such degraded data sometimes enables recognition of useful cues to overcome the difficulties. Once again, this depends on friendliness: e.g. important objects having distinctive outlines from most views. This requires assumptions (A) and (B).

5.4. Shared structure in memory entries. The system may share recognition processes between different objects by using a discrimination net. As partial specifications are built up, the set of remaining possibilities narrows. [Birch 1978 describes such an extension to Popeye.] Different recognition processes thus share significant sub-processes, minimising back-tracking or breadth-first searching. This uses incomplete descriptions, i.e. intermediate nodes in the discrimination net.

5.5. The need for speed. Even in a CFW, unfriendly circumstances may demand rapid decisions. The next section discusses the relevance of incomplete data.

6. Speed and the processing of incomplete representations

Complex articulated representations cannot be created instantaneously. Fast parallel processing at low levels depends on each processor being concerned with a relatively small well-defined portion of the data, and being able to work independently or co-operate with a relatively small set of neighbours. Thus, even data-flow channels can be 'hard-wired'. (Such mechanisms permit certain non-local interactions, via information propagated through the net.) But locality and independence do not characterise the process of articulating a mass of data into objects whose contributory regions change from one image to another. Portions of images relevant to a triangle or tiger vary in size and shape, and may be split into separate regions by intervening objects. Hence data-flow cannot be pre-determined, and organising data from particular images will therefore take a significant amount of time, compared with localised parallel computations. Though detectors for all possible edges may be 'hard wired' in advance, detectors for all possible triangle or tiger shapes could not be similarly pre-determined, partly because of the explosion of connections, partly because not all environments include them. The task of segmenting, aggregating, recognising, and building useful scene descriptions is therefore inherently much slower than low-level tasks. Thus there are limits to the speed-up available from hard-wired parallelism, and other mechanisms to speed things up could be useful: milliseconds may matter when life, or food, is at stake.

Cues invoking previously computed information can speed things up. This old idea [e.g. Roberts 1965] is now associated with the 'frames' theory [Minsky 1975]. Compare the idea of a 'phrasal lexicon' [Becker 1975]. But the theory leaves many questions unanswered: on encountering a new scene where should one start looking for cues in the image? At which level of analysis (in which domain) will the most useful cues be found? How can cues be recognised rapidly? The last question is very difficult, and will not be answered here. Our answer

to the first two is that as far as possible analysis should proceed simultaneously in many locations and at many levels, since the location or domain of the most useful cues cannot be predicted. This should be concurrent with general purpose image processing. Analysis of higher-level domains cannot begin until after some flow of data from lower-levels, but it need not wait for completion. The structure of such a network of processes will vary from image to image, so time and resources may be saved if its growth can be constrained, eliminating or suppressing portions which are not required, and giving priority to those yielding useful results -- e.g. activating and deactivating whole domains. This can be achieved if high-level structures (where the networks are relatively small) can be recognised whilst lower level networks are still incomplete. Thus construction of the network of communicating sub-processes which interpret the image, may itself be controlled by partial interpretations.

If, at any level, there is a lot of partially processed information, things may be speeded up by treating the partial results as a new image, in which gross features provide useful higher-level cues: using redundancy in a CFW [Sloman 1978, ch 9]. A specific purpose (e.g. finding a tool) might be achieved using this gross structure, without waiting for details [*5]. So, in some CFW environments, allowing many domains of structure to be analysed in parallel, could speed up actions. Even marginal advantages may influence biological evolution when resources are scarce, or predators plentiful. There is a kind of recursion in our argument, and possibly also in biological evolution. Where speed is important, the pressure towards further decomposition into parallel sub-systems is great, provided images have sufficient redundancy, i.e. provided it is a CFW.

We have not claimed that higher level processes can influence lower levels, except perhaps by aborting, or re-directing them. But it may be useful for partial results to affect some thresholds or even the invocation of specific forms of analysis, at low levels. Alternatively, cognitive processes may simply control the direction of attention, without modifying the nature of the processing. Even if animal physiology permits no direct downward influence on the processes which generate, say, a primal sketch, there might still be good reasons for designing artefacts differently. It would be no different in principle from making high levels influence direction of gaze, dilation of pupils, convergence of two eyes, etc. all of which affect the low-level image.

7. Some implications

In a CFW, multi-layered processing can improve flexibility, graceful degradation and speed. This applies to any kind of activity requiring intelligent analysis and interpretation of a large amount of data, based on expertise in the field, e.g. solving a complex mathematical problem, debugging a program, etc. One consequence is that demands on sub-systems are relaxed. For instance, if processing of level P has to be completed before processing at level Q can be begun, then it is important that P terminate. However, if Q can get started early, then it does not matter if P refines its analysis indefinitely! In vision, input is continuous, so lower levels cannot "finish" their analysis. Thus higher levels must in any case operate in parallel with them.

Moreover, in a CFW, mistakes at lower levels can be tolerated without disaster. The system must be conservative about transmitting items to higher-level domains, i.e. only sending well-supported reports. Then occasional mistaken reports will not combine usefully with other reports received at that level: (compare the role of 'impossible fragments' in Birch 1978). If a relatively large object is recognised on the basis of several different fragments reaching a high level, then the chances of it being a mistake will be small, assuming limited independent variation of object features. So the system need not

guarantee finding the best interpretation of any image, as in Woods [1977], since any good one will normally be unique, as we noted above. So it will often pay to accept a high level decision, abandon lower level analysis, and re-direct attention to the next task [*5].

All this depends on knowledge enabling fragmentary evidence to invoke specific larger structures, i.e. the principle of limited independent variation. General-purpose knowledge about 3-D structures and the principles by which they map into 2-D images does not constrain the space of possible scene structures so as to permit the inference that any good interpretation of an image is probably the best one. E.g. it does not rule out the existence of animals combining features in bizarre ways. Without specific knowledge of the world, detection of a zebra's ear would not rule out an animal with a trunk, six legs and two tails. The world would then not be a CFW. (This is like employing frequently useful theorems as well as axioms, to control search for proofs in a theorem-prover.) Our arguments are not relevant to the design of a machine whose visual system will never need to act quickly, which will always have perfect viewing conditions and which will often be transferred to a totally new environment where only the most general and primitive knowledge of 3-D structure, lighting, etc. will be of use to it.

Of course, our parallel, schema driven, system will sometimes make mistakes: but people make mistakes and sometimes learn from them. How? Decomposition into sub-systems processing different classes of structures provides opportunities for learning about new rules for linking the different domains, and for inhibiting the invocation of schemas, as well as defining new types of structures in terms of previously known substructures.

8. Problems of incompleteness

This theory raises many unanswered questions. Frank O'Gorman has pointed out in an unpublished manuscript that in a pass-oriented system, where each level of analysis is completed before the next begins, incompleteness of information at a certain location and level has a definite meaning: i.e. it represents the absence of something in the image. We have found it important to distinguish two sorts of incompleteness. It is not too difficult to cope with a gap in a known structure, for instance a hypothesised letter "E", for which the lower "ell" junction has not yet emerged from lower levels. We call this explicit incompleteness: a filler is missing for a slot in a frame. Here there are only two candidate letters "E" or "F", and the word-recogniser can decide which is correct on the basis of other letters which have emerged - even if they too are ambiguous. This depends on limited independent variation of letters in the domain of possible words. Implicit incompleteness occurs when trying to link features together to form cues to drive recognition -- for instance two previously unattached strokes to form a stroke-junction. Whether such features should be linked often depends on which other features are present nearby. From the absence of neighbours it cannot be decided whether this is because there is no evidence at lower levels, or because processing in that region has not yet finished.

In early versions, every level of Popeye[*2] simply used whatever information had already emerged, and then relied on context, or later bottom-up processing, to correct mistakes. Errors were reduced by delaying processing of any one level until a certain amount of information had been received at that level, using thresholds determined by image statistitcs. But even this left the garbage-collection problem of undoing mistakes and their consequences. So higher levels confronted with this incompleteness were allowed to ensure that everything up to that level, within a restricted region of the image, had been processed, making use of image-related addressing routes. This caused the focus of attention to

jump about, centering on important image features such as junctions between "bars". A better, more psychologically realistic solution, might be to let each level constantly recompute its hypotheses on the basis of the most recent information from other processes. This would be a generalisation of mechanisms using local co-operative processes, like relaxation. It could be very expensive on current computers, and hard to control.

9. Testing the theory experimentally

The fact that very young children learn to interpret cartoons and other 'impoverished' pictures so easily seems to support this theory. More detailed studies of what they find easy might be helpful. There is some additional evidence for our claim that higher level processing begins before lower level analysis is complete. People often think they've recognised a person or object, then spontaneously realise that a mistake has been made, even after the object has passed from view. Informal experiments with messy pictures of overlapping capital letters forming a word suggest that people often see the word before seeing all the letters. More detailed studies could provide clues as to domains and analyses being processed in parallel, in ordinary vision. Studies of brain damage might indicate which domains of structure (section 3) can be selectively disabled. Useful evidence should come from a study of visual errors. Our theory predicts that even in good visibility, humans and other animals moving rapidly will make more mistakes in an environment containing unfamiliar sorts of objects. (Testing this could be difficult, expensive and dangerous!) Experiments could test whether increasing familiarity improves performance (of survivors!). Different mixtures of familiar and unfamiliar features could be used, to find out if more obvious familiar features lead to errors concerning the other features. Additional experiments would vary lighting, foggy atmosphere, etc. as well. In poor viewing conditions, our theory would predict that visual judgements (especially at speed) would be more accurate when the environment contains familiar objects. Comparative studies might show that only some animals with visual systems possess the ability to process a variety of different domains in parallel.

FOOTNOTES

[*1] Acknowledgements:
This work is supported by the U.K. Science Research Council. We have benefitted from discussions with: Geoffrey Hinton, Frank O'Gorman, Steve Draper, Margaret Boden, Max Clowes, Monica Croucher, Steve Hardy, Christopher Longuet-Higgins, David Hogg, Larry Paul, Phil Pettitt, John Rickwood, Robin Stanton and Sylvia Weir, among others. Mike Brady and an anonymous referee made useful comments on a previous version. Judith Dennison helped with production.

[*2] Preliminary reports on POPEYE can be found in Sloman and Hardy [1976], Sloman et. al. [1978], Birch [1978], and chapter 9 of Sloman [1978]. See also Owen [1980]. Popeye analyses artificially generated dot pictures representing words made of overlapping cut-out capital letters. It can recognise words whilst much of the lower level processing is incomplete. Details will be reported elsewhere. The 1978 conference paper discusses differences between Popeye and the Hearsay system [Erman and Lesser, Hayes-Roth and Lesser], which have much in common. In particular, both process different domains of structure in parallel, though Popeye eschews the 'blackboard' concept. A similar philosophy has been used in the 'Visions' system (IJCAI-5, pp 642-647).

[*3] Marr makes a similar but different claim in justifying his theory of the 'primal sketch', [e.g. Marr 1979]. He postulates a progression, from image to primal sketch to 2.5D sketch to 3D model, whereas we propose many more domains, processed in parallel. In Popeye, the domains mainly form a hierarchy, but there are two main routes from image data to letter hypotheses and both feed the

word recogniser. We suspect that real visual systems require a far more elaborate network of routes through domains.

[*4] In Popeye the need for this arises often, e.g. when two parts of a letter are separated because of occlusion. The two parts can sometimes only be related by using a combination of (a) geometrical relationships and (b) partial recognition of the letter, since there are no image cues for linking, like 'back-to-back' tee junctions. So having recognised what may be, say, an E or an F, the program works out roughly where in the image evidence of a missing bottom stroke might be found, and this constrains searching.

[*5] In Popeye, processing can be aborted when the highest level decides it has recognised the depicted word; lower level analysis will often be incomplete.

TRUNCATED BIBLIOGRAPHY
(Compressed owing to page limit)

Becker, J.D. 'The Phrasal Lexicon' T.I.N.L.P. Eds. R.C. Schank and B.L. Nash-Webber. June 1975.
Birch F. in Sleeman 1978.
Clocksin, W.F., in Sleeman [ed] 1978.
Clocksin, W.F. 'A.I. theories of vision', AISB Quarterly 1978
Clowes, M.B. 'Picture syntax' in Kaneff, 1970.
Clowes, M.B. 'On seeing things', in A.I. Journal vol 2, no. 1 1971.
Draper S.W. 'A reply to Clocksin', AISB Quarterly, 1979.
Draper, S.W. 'Psychological relevance..' AISB Quarterly, 1980
Erman L.D. and V.R. Lesser in IJCAI-4, M.I.T 1975.
Gombrich E.H. Art and Illusion Phaidon Press, 1962.
Goodman N. Languages of Art, Oxford University Press, 1969
Hayes, P.J., 'The naive physics manifesto', in D.Michie (ed), Expert Systems in the Microelectronic Age, Edinburgh Univ. Press, 1979.
Hinton G.E. Relaxation and its role in Vision, Ph.D. thesis, 1977.
Hochberg, J and V Brooks, 'Pictorial recognition as an unlearned ability' Am Jour Psych, 1962.
Hochberg, J Perception (2nd Ed) Prentice Hall, 1978.
Horn, B. Overview lecture on vision, in Sleeman [ed] 1978.
Kaneff S. Picture Language Machines Academic Press, 1970.
Marr, D. 'Early processing of visual information', in Royal Society 1976.
Marr, D, Proceedings IJCAI 1979.
Minsky, M.L. 'Steps towards artificial intelligence' 1961
Minsky, M.L. 'A framework for representing knowledge', in Winston [1975]
Norman D.A. and D.E. Rumelhart Explorations in Cognition, W.H. Freeman 1975
Owen, D.B. 'Intermediate representations in POPEYE', this volume 1980.
Palmer S.E., in Norman and Rumelhart 1975.
Paul, J.L, 'Seeing puppets quickly' Proc AISB Conference, 1976.
Radig, B., in Sleeman [ed] 1978.
Roberts, L.G. in Tippett et al, Electro-optical Information Processing, 1965
Shirai, Y., in Winston [1975]
Sleeman D (ed) Proc. AISB/GI Conference., Hamburg 1978
Sloman, A. and S. Hardy, Proc AISB Conference, 1976.
Sloman A, The Computer Revolution in Philosophy, Harvester Press, 1978.
Sloman A, and D. Owen, G. Hinton, F. Birch, in Sleeman 1978.
Stanton, R.B. 'Plane regions...', in Kaneff.
Winston, P.H. (ed), The Psychology of Computer Vision, McGraw-Hill 1975.
Woods, W.H. in Proc. IJCAI-5, M.I.T. 1977.