

Active vision and adaptive learning

Mark Peters and Arcot Sowmya

Department of Artificial Intelligence, School of Computer Science and Engineering,
University of New South Wales, Sydney 2052, Australia.
markpeters@cse.unsw.edu.au, sowmya@cse.unsw.edu.au

ABSTRACT

Active vision is identified by a closed loop linking sensing with acting. Thus, an active vision system's behaviour is directly determined by what it senses. To date however, the responses produced by active vision systems have tended to be relatively low-level, generally designed to facilitate improved sensing, by enhancing the duration or speed of object tracking, for example, or optimising the focussed application of more intensive image processing. This is probably adequate if the active vision system is designed as a front end to other processes or to specialised application systems, or if it is a demonstration in support of a theoretical vision model.

However, this leaves unanswered the problems of i) how to select an appropriate action when many different alternatives are available, and ii) how best to modify the behavioural repertoire of the system. These problems are especially important in two situations: firstly, when an autonomous system faces a novel situation and must respond adaptively without the benefit of *a priori* knowledge, and secondly, when systems attempt higher levels of perception and response, and links between the absolute properties of the incoming image data and the actual objects of perception become increasingly attenuated.

This paper discusses methods for linking learning with active vision so that the behaviour of the system is optimised over time for the achievement of goals. We argue the necessity of system goals in learning vision systems, and discuss methods for propagating goals through all levels of loose hierarchies. In the last section we outline an architecture in which high and low level perception operate interactively and in parallel.

Keywords: active vision, adaptive learning, motion tracking, object tracking, human movement, foveating saccades

1 INTRODUCTION

This paper describes WRAITH, an active vision system (see **Figure 4**) currently under development in our vision lab. We give particular attention to those aspects of the work that indicate the need to introduce learning capabilities into the system. Section 2 outlines current research in relevant areas. Section 3 sets out some of the rationale behind our choice of domain and techniques. Section 4 details our work. In section 5 we draw hypotheses about some of the learning and adaptation questions raised by the work, and section 6 attempts to create a framework for linking various kinds of system learning within our domain.

2 CURRENT RESEARCH

The research areas relevant to this paper are hierarchical architectures of vision, and the convergence of machine learning and active vision. Recent work on hierarchical models includes work done on the analysis of dynamic traffic scenes, conducted by Dance and Caelli.^{1 2 3} In this work, images are preprocessed to find the positions and orientations of cars, and domain knowledge is provided to enable understanding of the scene. The architecture is described as a partial implementation of Minsky's society of mind.⁴ It uses a parallel, concurrent, object-oriented structure (Parlog++). The algorithm is a hybrid model-based and data-driven technique, with connectionist antecedents, in which multiple bottom-up data paths either refute or confirm downwards moving hypotheses about the scene, all operating in parallel. Such processing is also quite reminiscent of Dennett's multiple drafts theory of mind.⁵

Each object in the hierarchical network deals with a particular 'thing' in the image. Linkages between objects are made *a priori*. When a low level object is triggered it sends messages to related higher level objects, which are effectively being instructed by this message to seek independent corroboration of the particular state of affairs which the low level object flags. If satisfied that the condition of other low level objects is consistent with this state of affairs, these higher level objects will in turn trigger similar requests to still higher objects, and so on. Whatever hierarchical conclusions are built about the scene in question, all objects have access to the spatial database.

Dance and Caelli realised the need for some measure of certainty/uncertainty to be introduced into this system. They implemented the Dempster-Shafer formulations in preference to the Bayesian uncertainty calculus. Concurrent logic programming is a collection of horn clauses together with the parallel evaluation of goals with respect to these clauses. This allows goals to be regarded as a system of concurrent processes. There is a limit to the message passing within their architecture - it is not extended down into the image processing layers. These layers are actually handled by another, preprocessing system. Yet it is not inconceivable that a complete synthesis of all processing levels might be achieved, and that learning and system goals may be the method for adapting high and low level visual operations to each other. For an example of very low level learning in vision, Linsker⁶ demonstrates *unsupervised* learning techniques for development in a perceptual network of preferential sensitivity to features such as edges and orientations.

Tsotsos *et al*⁷ have developed a different hierarchical feedback architecture, with the express purpose of directing gaze or attention. This involves two passes through a pyramid of interpretive units. The first, feedforward, cycle computes the initial activation level of each unit. The second cycle, operating in the opposite, top-down, direction, detects the winner in the top layer, then detects the winner within that winner's receptive field in the next layer down, and so on. Units that are not within the receptive field of any layer's winner are pruned from the pyramid. All values in higher layers that are dependent on these units are then recalculated. Once a winner has been identified in each layer the focus of attention has been found.

While citing the obvious attractions of hierarchical architectures, Tsotsos *et al* identify four inherent problems of information flow:

- context effects: whether an interpretive unit correctly reacts to a particular event in an image is dependent on how well it is able to handle confusing information received from other parts of its receptive field.
- blurring: a single event may activate many interpretive units, all those whose receptive fields it occupies. Subsequent layers of interpretive units will find the event progressively harder to localise.
- crosstalk: two unrelated events taking place within a common sub-pyramid of interpretive units will activate all units whose function it is to detect a conjunction of such events.
- the boundary problem: the distribution of connections ensures that centrally appearing events will be reported more strongly than peripherally located ones, even in cases when the peripheral event is stronger.

Each layer in their hierarchy contains biases which are applied to influence the relative strength of signals produced by different feature interpretive units. Biases may be set to discriminate for or against particular areas of the image or particular kinds of features. Thus biases take the form of guides to both the target to be found and particular search strategies to be used. Such biases can be used to select for motion as easily as any other image feature.

Tsotsos *et al* used their selective tuning model to segment images into regions of consistent optical flow. As with most motion algorithms, there was some degree of uncertainty and error at the margins of regions, but the overall architecture provided a useful framework for effective winner take all (WTA) selection of key areas. They also convolved a two-image sequence with multiple difference of gaussian filters to create onset and offset detectors within a WTA network. Within each spatial or temporal scale a WTA selected the strongest response, then an across-scale winner was selected via another WTA. In their particular example the temporal scale was effectively constrained to one change. The WTA strategy is an effective way of selecting focus of attention without knowledge of the content of the image, though there is doubt that in practice it is any better than alternatives. For the recognition of articulation of motion, a layer of more sophisticated interpretive units, capable of detecting conjunctions of motion features, would be required.

Tsotsos *et al* present the control of an active vision system as a decision process: when to move and when to stay stationary. The first practical problem to overcome is the boundary problem, that of difference between the way an event is reported depending on whether it is foveated or peripherally located. This must be accomplished by compensating weights attached to peripherally originating signals. They compute the overall most salient event in the image, and then compute the independent annulus salient event. If these points differ, the annulus event is compensated and compared to the most salient event. If the annulus event is more salient, it is foveated. In changing the direction of gaze however, new events are brought into the visual field. There needs to be some kind of damping so that these newcomers do not immediately cause another saccade, and possible endless oscillation between targets. In response, Tsotsos *et al* created a crude spatial map with a temporally decaying inhibitory value attached to attended locations. So, once a location has been attended to, its power to distract is temporarily weakened, and the system is capable of further visual exploration.

Many other similar models of vision are discussed by Tsotsos *et al*. Hierarchical and pyramidal architectures are commonplace and well-established. The same cannot be said of active vision and learning however. Whilst there have been several projects initiated to link learning with vision (see, for example, the PAMI special section on learning in computer vision⁸), *active* vision has not figured largely in this work. In two recent reviews of the active vision field,^{9 10} the importance of learning is acknowledged, but little is said of *how* active vision systems might be made to teach themselves to see better. One exception is work in Japan where learning has been applied to active vision at least once.¹¹

3 AIMS

Our vision work has the following primary aims:

- **To develop an experimental active vision test-bed.**

Our long term goal is to develop systems which are capable of understanding something about the activities of humans taking place in real scenes. We feel it is important that this work be done in real time and in real situations because the limited generalisability of systems developed in artificial domains suggests that such crippling generalisation problems are *inherent* to all toy domains. It may be argued that problems solved by reducing the number of variables and limiting parameter space do eventually contribute to work done in real situations but our choice is to avoid this strategy. One of the endemic problems of simulated or simplified domains is that, rather than just reduce problems to their essential components, they also introduce and depend upon an artificial level of *certainty*, one which cannot be relied on in real domains. The need to avoid rash extrapolation of hypotheses from simulations underlines the necessity of a real active vision test-bed.

Level of analysis	Questions to be answered
geographical	where is the person?
physical	what is, and what is not, included in the person's body?
spatio-temporal	where and how does the person's body move?
functional	what effect is the person having on their surroundings?
behavioural	what factors are driving the person?
procedural	what plan is the person following?
intentional	what does the person mean by their behaviour?
socio-cultural	how will third parties interpret the person's behaviour?

Table 1: Levels of analysis for understanding human movement.

Understanding human movement (specifically, a person's behaviour) implies being able to answer different questions about such behaviour. This, in turn, implies the ability to analyse behaviour at multiple levels. This hierarchy of analysis owes something to Dennett.¹²

On the other hand, there is clearly no practical limit to the potential levels of understanding that may be brought to the domain of human activity. So any solution to the problems posed by this domain is likely to be always incomplete. Human movement is, by choice, a domain in which all complexities and multiple-level semantics (geographical to socio-cultural, see **Table 1**) exist. In this domain we cannot expect to solve all problems to the degree that humans do (at least not in the foreseeable future), but our goal is nevertheless to develop *real* solutions to *real* problems, however partial.

- **To implement a motion recognition system.**

It is evident from a cursory inspection of biological systems that detecting motion is important: many sensory systems have evolved for this purpose. Things that move tend to be significant to living things, in special ways, and should be taken, all other things being equal, more seriously than things that don't. So, whilst things (objects) may be significant, so too is change (movement). It is also likely that movement is a vision primitive for many species, possibly including ourselves; neurologists report that patients in deep comatose states have been known to retain the ability to track motion with movements of their eyes, despite loss of all higher functions.¹³ This primitive status of motion sensitivity suggests that objects might be inferred from motion as well as, or even instead of, vice versa. In our research we therefore do not presuppose the existence of objects per se. The concept of an object is one we consider to be rather inferential, dependent on probabilistic analysis of lower level change and correlation in image information.

- **To learn motion concepts.**

Domains of any significant complexity are laden with unforeseeable characteristics. The question of how a system deals with this is important. There are several approaches: gamble that the unforeseen won't happen, develop recovery responses so that temporary system disorientation in the face of the unforeseen is quickly overcome, or have the system learn the domain such that the unforeseen is appropriately responded to (eventually, most of the time). Our intuitions are that learning is the correct approach, that recovery responses are also required, and that the domain must always be assumed to be open and infinitely variable. Our preliminary work seems to confirm these ideas.

The kind of learning we expect to be required is that which attempts to derive a higher level motion concept from a disjunction of lower level motion concepts. An example of this could possibly be observing alternating swinging areas which we might infer to be limbs. At a higher level of analysis, when this is observed in close proximity to, and below, another pair of alternating swinging areas this could lead to further inferences that the lower pair are legs, the upper pair arms, and that a body is walking. This suggests a hierarchical structure if only because it is clearly possible for low level concepts to exist independently of higher level concepts. Details of the kind of conceptual hierarchy we propose are in section 6. We hope that the system, finding that such swinging motions are nearly always found in concert, is able to make reasonable interpretations of motion even if not all the expected components of the higher level motion (walking) are to be found.

4 METHODS

Our experimental test-bed consists of a small video camera mounted on a pan-tilt motorised gimbal. The camera has fixed focus, automatic correction for brightness, and delivers 30 monochrome interlaced images (512 x 512 pixels) per second to a frame grabber. The data received by the frame grabber is sampled using a space-varying matrix which can independently concentrate sampling in both horizontal and vertical dimensions, either towards the periphery or the image centre (see **Figure 1**). Sampling has been varied from (16 x 16) to (128 x 128), and non-square sample matrices have also been used. We have found that (64 x 64) has produced the least problematic trade-off of speed and accuracy.

An elegant method for deriving a space-variant mapping is the optical approach adopted by Kuniyoshi *et al.*,¹⁴ though this gives a fixed mapping bias. It turns out that the mapping bias strongly affects the behaviour of the system. Sampling more densely at the periphery causes the system to be relatively more distractable yet sensitive to objects moving into its visual field, while concentrated central sampling reduces the number and size of movements made by the motors, and therefore facilitates slow tracking.

Clearly there are situations in which either of these sampling biases might be advantageous over the other. In situations where it is important that movement is responded to as soon as it appears in the field of view, a peripheral bias is desirable. When the system has found a concentrated area of movement and needs to track it, then a central bias should be chosen. It is just as obvious that an active vision system may have to employ both these modes, and possibly others. It is therefore important to overall efficiency that sampling bias be under system control, and that bias changes become part of the system response to the scene before it. The simplest way of controlling bias is to devise a single coefficient of bias (γ) for each dimension. For any sample number s of a total of S samples in a horizontal line across an image of width w , the image pixel to be sampled i_p is given by:

$$i_p = \frac{(w-1)}{2} \left(\frac{2s}{(S-1)} \right)^\gamma \quad (1)$$

This exponential curve covers half the image, it is then trivial to invert it and map the other half of the image with the symmetrical inversion, thus producing a sigmoid. Sampling density and bias are our first items for consideration by learning (see also **Figure 2**).

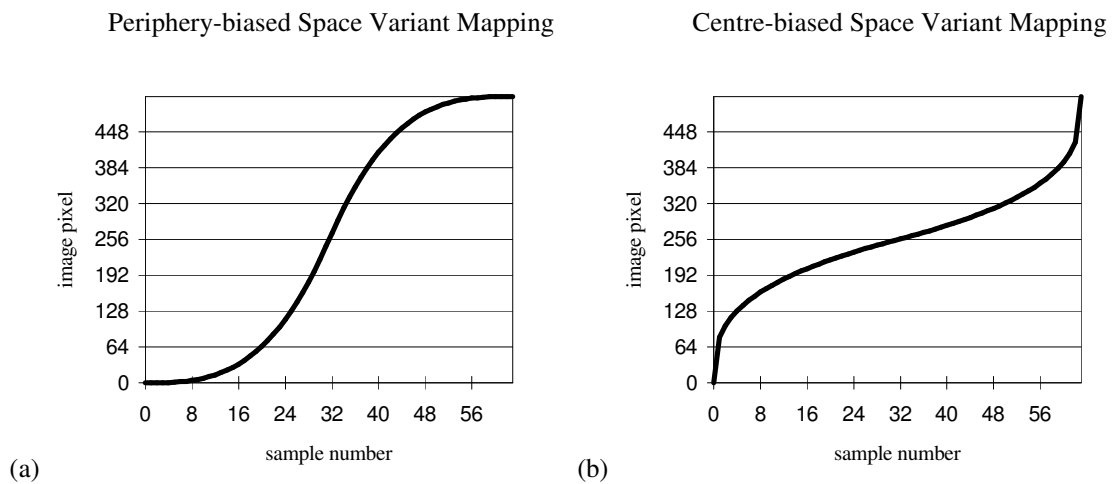


Figure 1: Comparison of space variant mappings.

The vertical axes refer to the coordinates (0 to 511, in this case) of pixels in either horizontal or vertical lines through the image. The horizontal axes refer to the samples (numbered 0 to 63, in this case) taken from a single such line. In (a) the densest sampling takes place at the ends of the line ($\gamma = 3$). In (b) the sampling is concentrated in the central part of the line ($\gamma = 1/3$). With $\gamma = 1$ the mapping is uniform across the whole image.

Interestingly, it has been shown that human vision also has biased sensitivity to motion. The foveal part of the retina is unusual in that it has no connections to the superior colliculus, a neural body involved in eye orientation. More significantly however, we also exhibit selective sensitivity to the *direction* as well as location of movement. Centripetal motion is more easily detected, and motion anisotropy is larger in the temporal than the nasal hemifield. Raymond¹⁵ conjectures that this biased sensitivity may exist because our normal forward locomotion habituates us to centrifugal motion in the visual field, that is, it is an adaptive response to our prevailing motion environment.

In WRAITH the fundamental identification of motion is determined by simple pixel-level subtraction of one image from its predecessor. The resultant absolute image difference is then thresholded, and actual values are reduced to 1 or 0. We

thus produce a binary image whose dimensions are determined by the number of samples selected. The thresholding of differences eliminates noise created by continuous minor fluctuations in lighting which cause some pixels to record spurious changes of one grey level, and possibly more. This level of noise is spread over the entire image, regardless of the location of edges in the scene, and is consequently highly uncondusive to coherent motion analysis. Secondly, the actual value of the difference at any particular pixel location is linked to several factors which are irrelevant to motion: reflectance of the moving object, reflectance of the background, variation in surface reflectances, variation in light levels. This makes it difficult to perform meaningful real-time motion calculations on difference values. However, we have found that, rather than deal with actual difference values, it is sufficient to simply divide the image into areas of movement versus non-movement, and hence we derive the binary image.

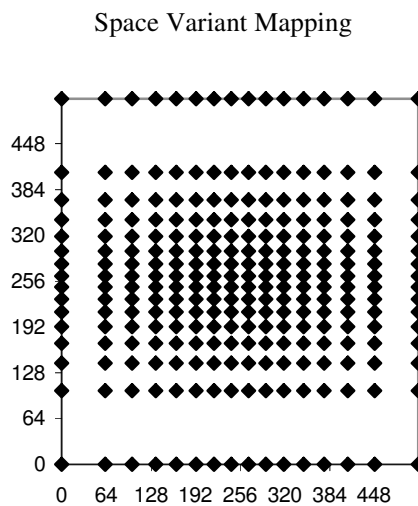


Figure 2: Demonstration of space variant mapping.

Although sampling independently across vertical and horizontal dimensions introduces non-uniform radial effects, it is practical in real time operation as each video scan line mapped to memory can produce a full-width set of samples (16, in this case). Additionally, in the case of human movement, very little happens in the top and bottom bands of the image, so it is useful to be able to reduce sampling in these regions without compromising sampling in the extreme left and right areas of the image. In this example 16 sample lines are taken from an image of 512 x 512 pixels, and 16 sample points are taken from each line. $\gamma = 0.7$ in the horizontal dimension, and $\gamma = 0.45$ in the vertical dimension.

For the purposes of guiding the camera so that it retains the highest level of real movement in the field of view, we have developed a low-level response which can function in the absence of any world knowledge or inferences based on previous images. This is therefore a reactive response. We observe that in real time environments it is often preferable to respond quickly rather than accurately or circumspectly. Motion tracking is one such situation because, if the system delays camera repositioning until it has performed high level analysis on the scene, there is a danger that the subject may actually move out of the visual field altogether before the system is ready to make its move. The continued presence of a default low-level tracking response is insurance against such an eventuality. Operating in parallel, this response may interrupt higher level analysis to reposition the camera, and intermediate responses may also interrupt higher level responses. This is reminiscent of the subsumption architecture¹⁶ but differs in one significant respect: subsequent higher levels of response in our system (corresponding to Brooks's subsuming layers) must be internally constructed via *learning* rather than externally imposed by the developer. The mediation of control is obviously another consideration for learning.

The low level tracking response in our system is concerned solely with movement tracking and not *object* tracking. We consider the inferring of the presence of an object, per se, as a higher level process which would depend on the extraction

of a number of invariant pieces of information¹⁷ from a sequence of images, and a probabilistic analysis of this data. By dealing with movement alone, regardless of object model inferences, we are able to achieve high degrees of success in following rapidly moving objects, simply because a real moving object creates coherent movement patterns through space, regardless of much of its object-level qualities.

Our tracking response simply calculates the centroid of all pixels reporting motion, and uses the difference between the centroid coordinates and the centre of the image to derive the pan and tilt motions necessary to centre the centroid in a single movement (see **Figure 3**). For n pixels reporting motion, the centroid coordinates (c_x , c_y) are calculated as:

$$c_x = \frac{\sum_{i=1}^n i_x}{n}, \quad c_y = \frac{\sum_{i=1}^n i_y}{n} \quad (2)$$

Given an image of width w and height h , the relative pan and tilt angles (p , t) are thus:

$$p = k_x \left(c_x - \frac{w}{2} \right), \quad t = k_y \left(c_y - \frac{h}{2} \right) \quad (3)$$

where k_x and k_y are calibration constants. If changing cameras or lenses the calibration constants may need to be varied. This is a further possible item for consideration by learning.

While the motors reposition the camera it is still transmitting images, and these images are asynchronously analysed as just described. WRAITH does not toggle between ‘seeing’ and ‘saccading’ modes, it simply discards images received while the camera is moving (smeared images) or while residual mechanical vibration is present in the camera (blurred images). In some ways this is avoiding the problem of egomotion rather than solving it. Difficult smeared or blurred images are easy to distinguish as they usually report motion across the full width and height of the image, or at an unusually high density (say more than one in four pixels) across the area of the image. Using heuristic methods to discard these images, the system simply waits until it receives images that meet the heuristic criteria, and only then calculates a new camera position. This method has proved superior to alternatives such as the bimodal method alluded to above because it reacts well to unpredictable vibration effects, in effect delaying further motion tracking just long enough to get good data. The heuristics may be learned too. One in particular is a heuristic which cancels any camera movements below a certain angle. This prevents the camera from embarking on a series of rapid, tiny adjustments when observing an object that has some surface motion, but is essentially stationary. The benefits of this are reduced use of the motors and clearer images. Exactly how large a planned movement must be before it overcomes the threshold is only answerable in a particular motion context, the optimum setting cannot be foreseen.

Different methods for calculating the centroid make little apparent difference to the system’s performance, particularly when camera movements are taking place roughly every 250 msecs. The effect of any discrepancy seems to be limited to a single camera movement, and is usually compensated for by the next movement. Indeed, the significant factors are not camera repositioning algorithms (these can obviously be quite crude), but the characteristics of objects observed in the scene. A person wearing non-reflective black clothing and moving close to the camera may be reported simply as two lines; their left and right sides, where their clothing contrasts with the (lighter) background. When a single motion centroid is calculated using both left and right parts of the person’s outline the camera is correctly repositioned pointing straight at the person, even though the area in the centre of the image has not reported any movement (we only know it is there because we are sophisticated humans with the appropriate world knowledge). However, if two very thin objects are moving side by side (such as a couple of saplings blowing in the wind) they would create a very similar motion image (two roughly vertical coordinated motion lines against a background of non-motion). It would not normally be desirable, in such situations, to point the camera into the gap between the two areas of movement. It is desirable that *at some level of processing* the system makes a choice about which active object it is watching, it cannot be usefully indiscriminate *at*

all levels of analysis. It may well be that the level of selection is very low. If this is so then the current response is inadequate. Clearly this points to the next step in system enhancement: discrimination of separate objects, and decisions about which of such separate objects to select for further observation.

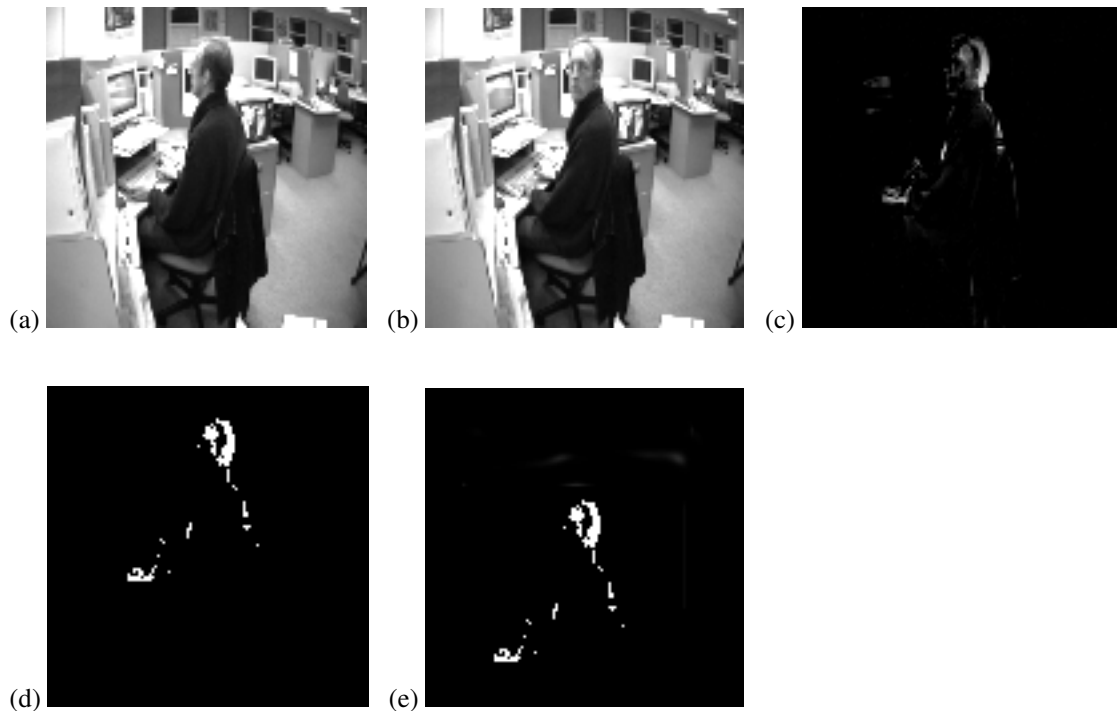


Figure 3: Low-level motion tracking.

Images (a) and (b) are closely-spaced video images received as raw data by the frame grabber. Image (c) is the absolute result of pixel-by-pixel subtraction of (a) from (b). Note that the movement of the subject's head and hand are clearly separated from the background even at this early stage. Image (d) is the thresholded binary version of (c). After calculating the centroid of the white areas in image (d), which is somewhere in the subject's head, the system sends move commands to the motors controlling pan and tilt, resulting in something like image (e), with the motion centroid correctly centred (or foveated). Note: the computer screen (top left) has been reported as a very faint area of motion in image (c). Fortunately in this instance the thresholding eliminated this in image (d) before the centroid was calculated.

If we refer to **Table 2**, we will see that WRAITH has implemented levels 1 and 2 (non-directional and directional motion detection) already, and must implement level 3 (differentiation of motion types, and the corresponding development of a range of responses) in order to improve. Differentiation of motion types could mean a number of things, depending on the environment. It is not simply limited to identifying single object motion versus group motion or uncoordinated mass motions. Motion may consist of rigid or non-rigid body motion, repetitive periodic patterns, motion without travelⁱ, spatially or temporally coordinated motion such as formal dance steps, sports, etc. We expect to direct such differentiation via learning.

ⁱ Indeed, this has proved to be a phenomenon that must be dealt with. Due to the higher temporal discrimination of video cameras compared to the human eye, visual display screens with scanning rates such as those used on computers (see **Figure 3**) are reported as rectangular slabs of intense movement regardless of what they display. Despite all the apparent energetic movement taking place, however, the screens stay in place. This might be classified as 'motility without mobility'.

It may also be worthwhile to note that using the low-resolution binary report of motion makes it an easy task to define the bounding rectangle, ellipse, etc., and that this facilitates tight efficient focussing of more discriminating, but resource-consuming, clustering or segmentation algorithms.

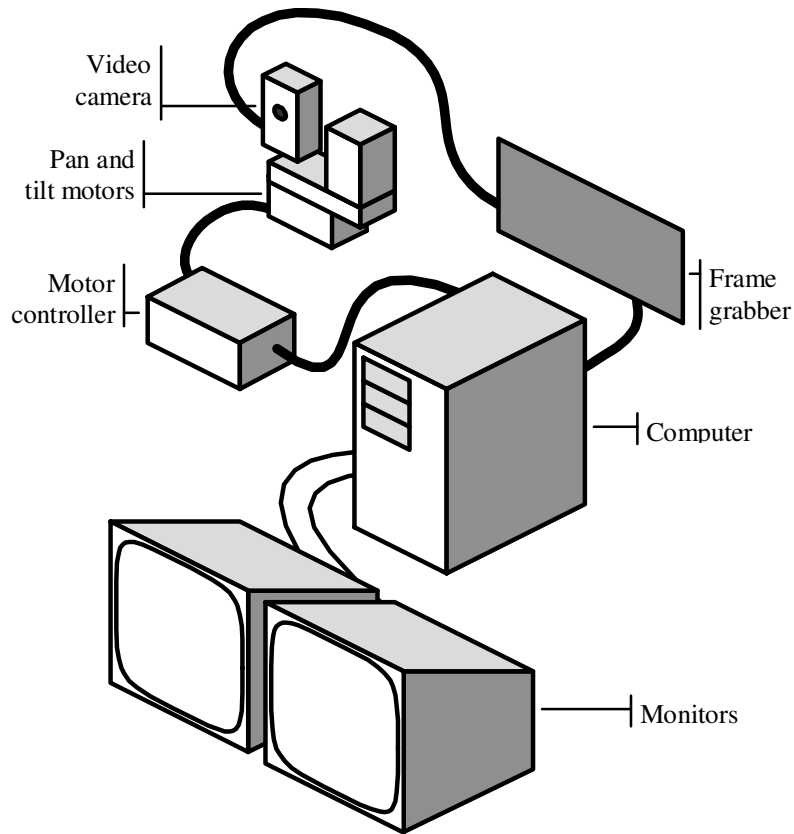


Figure 4: WRAITH hardware.

The closed action-vision loop is shown by the heavy line linking components.

5 LEARNING

Simple vision systems may be able to achieve their goals in the absence of learning capabilities, but this does not hold true for more demanding goals. In-system learning is a reasonable method for dealing with the unforeseeable, more reasonable than blind faith that the unforeseen will not take place, or the placing of limitations on the domain to artificially ensure that that will be the case. Furthermore, at a practical level the complex mechanics and optics of active vision systems present difficult calibration problems that encourage the use of learning in favour of hard-wired solutions. So far we have identified the following areas for consideration by learning: sampling density and bias, threshold levels, interrupts between processes, construction of subsuming layers of response, calibration, image discard heuristics, motion discard heuristics, and motion type differentiation. It is obvious that this list may be extended even further.

The concept of learning introduces the challenge of how to organise selection of response. It is a prerequisite of a learning system that it has at least two alternative responses available to some particular stimulus. The learning consists of selecting the response which, according to some criterion, is more appropriate than its alternative(s). The appropriateness needs to therefore be quantified in some way, and this quantification must represent the proximity of the

consequent outcome to a particular goal. This, briefly, is the justification for dealing with hierarchy, feedback, learning and goals in vision systems.

Of course, the number of responses available, given some stimulus, is invariably more than two. And the most appropriate response might not consist of an externally observable action, but instead may be the creation of new behavioural descriptions, which are stored and modified for later use. Indeed, we see the development of the behavioural repertoire over time as a significant goal of our system.

To think in terms of a behavioural repertoire is useful. It places the emphasis on appropriate selection and development of responses. We may think of the behavioural repertoire developing along lines that reflect the active vision system's perceived structure of the world. In other words, as the hierarchy of the world is discovered through association and correlation of data, so must the hierarchy of responses develop to match this. Though WRAITH has only one kind of physical response (repositioning its camera), what is important is the process it uses to do that. In the sense that human expectations are built upon all our knowledge of the world, the system's camera positioning is similarly the result of all the levels of analysis it is able to perform in the given time. Its goal is to correctly anticipate movement. This has a single externally discernible response, but a complex hierarchy of responses interacting internally to produce the simple external response. To be successful in this simple act, it needs to develop a sophisticated understanding of the human world.

6 ARCHITECTURES

Hierarchies seem to describe visual phenomena fairly well. This may be because simple visual concepts naturally aggregate to form more complex visual concepts. Despite this, it seems that bottom-up (data-driven) hierarchies of this type are inadequate as an explanation of vision as an inferential (theory driven) process capable of operating successfully with underdetermined data. For this some elements of top-down activity are also required. This top down activation may perform a compensatory role in human vision where, for example, reasonable expectation or good prediction of the world may explain the clear superiority of our perceptions over the relatively poor optics and discrimination of the eye alone.

The coexistence and interdependence of bottom-up and top-down activation, crucial features of any learning system, also raise questions about the very nature of the hierarchical model. We can defend the hierarchy on the basis of afferent signals alone, and call it a representation model, or point to the efferent signals and call it a control model. Either of these models is relatively straightforward. However, in a system containing both afferent and efferent signals the activation from any one point can propagate in all directions and may skip certain levels altogether. There is no longer any real sense in which a point may be said to be above, below or at the same level as any other point in the system. Multiple internal feedback loops are difficult to accommodate hierarchically, regardless of whether we develop from the representation model or the control model. Hierarchical models, though useful, only describe some aspects of the vision process. To deal with this we have formulated the idea of different levels of stimulus-response pairings. We feel this is a tractable way to give structure to vision systems. This idea is elaborated below.

Since Marr¹⁸ it is commonly agreed that differences between absolute image properties can be collectively interpreted as features, that features can be assembled into objects, and objects into scenes. It is less obvious that, for example, absolute pixel value differences between images in a sequence can be collectively interpreted as motion, that 2D motion can be interpreted as 3D movement, and that configurations of movement can be assembled into meaningful actions, or activities. Tsotsos,¹⁹ noting the importance of lateral inhibition in vision, applied to temporal as well as spatial relations detected through *change* in intensity or colour, suggests that temporal information is probably necessary for the construction of a primal sketch but remarks that Marr did not consider this. Tsotsos divides vision into four main kinds of sensation: form, motion, colour, depth, implying that these sensations are somewhat like the primary colours or tastes in that they are not derivable from each other. This gives considerable significance to motion.

Multi-level motion interpretation as just described implies qualitatively different interpretive mechanismsⁱⁱ operating in concert. Low level mechanisms may simply compare absolute values in an image, whilst high level mechanisms may depend on many contextual clues and assumptions before making a semantic interpretation of a scene. Such high level interpretations and abstractions therefore make additional demands on the sensory input, for not only must low level patterns be present in the data, but so must some indication of high level structure which may consist of, for example, long sequences of similar movements, or spatially separated, yet temporally coordinated movements performed by separate individuals. High level responses can therefore be said to be *dependent* on high level stimuli, where these stimuli coexist with, and (in fact) consist of, low level stimuli.

Articulation of perceptual system	Maximum level of response this affords
non-directional motion detection	activation of agent (startle response)
directional motion detection	orientation of agent, modified observation
differentiation of motion types	development of a range of responses
identification of simple spatial or temporal patterns of motion	appropriate prior selection from a range of responses
separation of invariant aspects in motion from other data	sustained responses to intermittent stimuli
recognition of patterns of motion at an rudimentary abstracted level	prediction of future states, planning, cooperation
integration of motion information with memory, inference, other modalities and sub-modalities	human-level spatio-temporal modelling of the world and other agents

Table 2: Motion sensitivity and related response affordances.

The left column describes the progressively more sophisticated levels of articulation of incoming data performed by perceptual systems. It is assumed, of course, that the data is conducive to such articulation. The right column describes correspondingly sophisticated levels of response which are only possible in the presence of such articulation.

It is appropriate, when discussing the stimulus and response pairings of active vision systems, to establish as clearly as possible a dependency between a particular level of response, and the extent and content of prerequisite stimuli. This helps us avoid attempting to learn too sophisticated a response from inadequate stimuli. It may not always be easy to determine whether a stimulus is adequate for a particular response, but some cases are obvious. For example, the most primitive response to motion in the proximity of a proto-vision organism endowed with only non directional sensors might be a ‘startle’ response, which simply indicates that motion (change above some given threshold) has taken place, but fails to indicate just where the motion took place, what kind of motion it was, and so on (see **Table 2** for elaboration). The startle response may be terminally simple, a binary value corresponding to yes/no, motion did or did not take place. Such a response is adequate to alert internal functions to the existence of external motion but inadequate to orient the system towards the motion. A directional sensory articulation of the environment is required for this, and such an articulation depends on being able to make distinctions between the arbitrary sectors of the optic array that contain motion, and those that do not. Additionally, it must entail some low-level analysis of such motion information to make a decision about, for example, where next to direct its sensors. At the risk of belabouring the obvious, any attempt to produce a directional response from a system which has only non-directional articulation of incoming data is futile. And similarly, any higher response must have certain sensory dependencies: presence of the prerequisite characteristics in the data, and the sensitivity and selectivity of the sensory apparatus to actually articulate these characteristics within the system.

ⁱⁱ Such mechanisms are not necessarily different in principle. It is useful to ask what different forms of learning have in common and what it is that makes them all *learning* systems.

Of course, much of this discussion is dependent on the level of world knowledge possessed by the system. If, from experience, through close association of directional tactile sensation (feelers) and non-directional visual sensation (eye spots), our proto-vision organism has learnt that the sudden darkening of its world foretells contact at, say, its front end, then it has grounds for treating the non-directional visual stimulus as *directional* information, and responding accordingly. This serves to illustrate the importance of learning even in low level vision. Indeed, even in image processing it is extremely useful to first have some idea of what you are looking for.

In summary, our early work has indicated, emphatically we believe, that vision must be a learned capability, and this is the direction we intend to pursue. Our architecture shares characteristics with standard pyramidal hierarchies (eg. Tsotsos), object-oriented semantic nets (eg. Dance and Caelli), and subsumption architectures (Brooks), and is informed by ideas concerning the development of levels of articulation in perceptual systems that correspond to the many levels of meaning in the domain of human movement.

7 REFERENCES

-
- 1 S Dance and T Caelli, "On the symbolic interpretation of traffic scenes", *ACCV93: Proceedings of the Asian Conference on Computer Vision*, pp 798-801, Osaka, 1993.
 - 2 S Dance and T Caelli, "A symbolic object-oriented picture interpretation network: SOO-PIN", *Advances in Structural and Syntactic Pattern Recognition, Proceedings of the International Workshop*, Horst Bunke, pp 530-541, World Scientific Publishing Co., Bern, 1993.
 - 3 S Dance, T Caelli, and Z-Q Liu, "An architecture for a traffic scene interpretation system", Technical Report No. 94/12, pp 1-43, Department of Computer Science, University of Melbourne, 1994.
 - 4 M Minsky, *The society of mind*, Simon & Schuster, New York, 1985.
 - 5 D C Dennett, *Consciousness explained*, Penguin, London, 1991.
 - 6 R Linsker, "Self organisation in a perceptual network", *Computer*, 1988(3), pp 105-117, 1988.
 - 7 J K Tsotsos, S M Culhane, W Y K Wai, Y Lai, N Davis, and F Nuflo, "Modeling visual attention via selective tuning", *Artificial Intelligence*, 78, 1-2, pp 507-545, 1995.
 - 8 B Bhanu and T Poggio, "Introduction to the special section on learning in computer vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 9, pp 865-868, 1994.
 - 9 A L Abbott *et al.*, "Promising directions in active vision", *International Journal of Computer Vision*, 11, 2, pp 109-126, 1993.
 - 10 C Fermüller and Y Aloimonos, "Vision and action", *Image and Vision Computing*, 13, 10, pp 725-744, 1995.
 - 11 L Berthouze, S Rougeaux and Y Kuniyoshi, "A learning stereo-head control system", *World Automation Congress/International Symposium on Robotics and Manufacturing*, France, 1996.
 - 12 D C Dennett, *The intentional stance*, MIT Press, Cambridge, Massachusetts, 1987.
 - 13 R M Restak, *The brain has a mind of its own: insights from a practicing neurologist*, Crown, New York, 1991.
 - 14 Y Kuniyoshi, N Kita, K Sugimoto, S Nakamura and T Suehiro, "A foveated wide angle lens for active vision", *Proceedings of the IEEE International Conference on Robotics and Automation*, 1995.
 - 15 J E Raymond, "Directional anisotropy of motion sensitivity across the visual field", *Vision Research*, 34, 8, pp 1029-1037, 1994.
 - 16 R A Brooks, "A robust layered control system for a mobile robot", *IEEE Journal of Robotics and Automation* RA-2(1), pp 14-23, 1986.
 - 17 M Palhang and A Sowmya, "Learning object models by inductive logic programming", *First International Conference on Visual Information Systems, Melbourne*, Australia, pp 335-344, 1996.
 - 18 D Marr, *Vision: a computational investigation into the human representation and processing of visual information*, W H Freeman, New York, 1982.
 - 19 J K Tsotsos, "Knowledge of the visual process: content, form and use", *Proceedings of the Conference on Pattern Recognition and Image Processing*, pp 654-669, 1982.