

WRAITH: Ringing the changes in a changing world

Mark W Peters and Arcot Sowmya

University of New South Wales

Sydney 2052, Australia.

markpeters@cse.unsw.edu.au, sowmya@cse.unsw.edu.au

...novelty itself will always rivet one's attention. There is that unique moment when one confronts something new and astonishment begins. Whatever it is, it looms brightly, its edges sharp, its details ravishing, in a hard clear light; just beholding it is a form of revelation, a new sensory litany. But the second time one sees it, the mind says, Oh, that again... Ackerman.¹

Keywords

active vision, adaptive memory, perceptual hierarchy, segmentation, human movement

Abstract

This paper documents techniques we have used to control the attention of a visual robot assigned the task of following (in the sense of both tracking and understanding) human movement. We believe that even simple tasks like intelligently controlling the direction of gaze must involve a complex synthesis of all levels of perception, from those of 'early' vision to sophisticated socio-cultural levels. In order to produce real-time real world intelligent behaviour, two key questions must be answered: how are such levels of perception generated in the first place, and how do they combine to create coherent, useful behaviour. We explain how we have approached these problems and present some results of current work in the WRAITH project. The paper also discusses our hierarchical conceptual framework of motion perception, which has guided the work. We suggest that, in addition to having the essential faculty of pattern recognition, intelligent systems must detect and react to pattern change, often before they have had time to identify the patterns, and that it is just as important to ring the changes in the world as it is to notice its regularity.

1 Introduction

1.1 Defining the problems

Understanding physical actions is an essential part of our cognitive repertoire, to the extent that people without this ability excite considerable neuropsychological curiosity.² This understanding, though usually taken for granted, probably involves a complex synthesis of many levels of perception, from simple motion detection up to an apprehension of the socio-cultural import of an action. These levels of perception are interactive. Lower levels seem to provide much of the information used by higher levels, operating in some sense as filters of unnecessary or irrelevant information. Simultaneously, higher levels, though dependent on lower level input, have been shown³ to influence the content of lower level perception.

Standard conceptualisations of different perceptual levels tend to be rather heterogeneous: mathematical when it comes to low levels,⁴ and by turn logic-based,⁵ linguistic,⁶ and informal⁷ at higher levels. This seems somewhat at variance with the largely homogeneous neural activity of the brain, but the heterogeneity of common models poses more serious problems when we attempt to vertically integrate several levels in a learning system. It is rather difficult to propagate a reinforcement function through a heterogeneous system. It is also difficult to autonomously and automatically generate a higher heterogeneous level of

perception, with its own reinforcement function, from the activities of existing lower levels, or even to know when to do so. Homogeneity therefore has attractions.

The direction of attention, a seemingly simple response, needs to be reactive to all levels of perception. So, when we attempt to build robots that exercise intelligence (albeit in a way often described as reactive⁸) in their choice of where to look, we must also decide at what levels they will perceive the world.

1.2 Our approach

We have attempted to establish a consistent way of thinking about levels of perception, in which we have started to sketch some of the inter-level dependencies, and we have drafted a hierarchy of loosely differentiated levels.⁹ This provided the guiding framework for our active vision robotics experiments, and has also clarified certain information prerequisites for action perception (see **Table 1**).

Organisation of perceptual system	Maximum level of response this affords
non-directional motion detection	activation of agent (the 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 in motion at a rudimentary abstract level	prediction of future states, planning, cooperation
integration of motion information with memory, inference, other modalities and sub-modalities	human-level socio-cultural modelling of the world and other agents

Table 1: Motion sensitivity and corresponding afforded responses.

The left column suggests how perceptual systems might organise incoming motion data. The right column describes responses that are only possible once the corresponding perceptual organisation is in place.

We assume there are always time constraints on processing. This implies that changing data must be monitored more closely than static data. That change in the world is important is uncontentious, but there is less agreement on exactly what constitutes change. As far as any system is concerned, change is how it measures it. We measure change as a conjunction of the outputs of several layers of ‘surprise generators’, each building upon its predecessor and effectively measuring a derivative of change, or the way change itself, in a previous layer, has been changing. Each layer tunes itself to recognise a pattern (the current pattern) in its input and reacts only when this pattern is interrupted. Its reaction is passed to the next (higher) layer, forming another input, with a pattern of its own. The sum of the parts is thus a pattern reaction system in which only the intervals between patterns are reacted to (ringing the changes). Pattern intervals are complex multi-level phenomena. Thibadeau¹⁰ noted the importance of identifying the moments when particular actions start and finish. We attempt

to make an active vision robot, WRAITH, react to these special moments both adaptively, and *without a priori knowledge of any form*, so that future higher functions, whose task might be to identify specific actions, receive a representation which has already carved the world at its joints. This then, is a bottom-up approach to the segmentation problem.

Initially, we have conducted experiments in which WRAITH has been set the task of finding what is most interesting in a typical laboratory scene, and keeping its eye on it. ‘Interesting’ is the quality attributed to areas of the optic array that cause the greatest total stimulation, after appropriate weighting of higher and lower surprise values in the system. Our results show that WRAITH acts rather as we do, paying progressively less attention to repetitive movements, or movements that do not travel through space (eg, hands continuously waving, or flickering computer screens), in favour of novel movements and movements in, or to, new places.

2 Methods

WRAITH is fitted with a small video camera delivering 768×576 pixels at 25 frames per second with a greyscale of 256 values. We start by treating each pixel location in the visual field as a miniature processing unit, complete with a proto-memory consisting of a single value.ⁱ The value of the memory is updated continually, using the equation:

$$memory_t = signal + (memory_{(t-1)} \times retention)$$

where *signal* is the brightness of the pixel, and *retention* is a constant between zero (meaning no previous *signal* value is retained in *memory*) and one (all previous values of *signal* are retained at full strength, summed and stored in *memory*). Varying *retention* adjusts the weight given to more recent values of *signal*, effectively determining whether *memory* can be thought of as long- or short-term.

If *retention* is near one then *memory* will be large in relation to *signal*. So, as we want to compare *signal* to *memory* to derive a *surprise* value, we first need to renormalise *memory*, and arrive at a recent average of *signal*. To do this we use another constant:

$$persistence = 1 / (1 - retention)$$

hence:

$$prediction = memory / persistence$$

and so:

$$surprise = difference(signal, prediction)$$

The value of *surprise* represents the difference between the *signal* just detected and that which was expected, the *prediction* (a weighted average of previous values of *signal*). Thus *surprise* corresponds to the ‘figural’ content set against the ‘ground’ of *prediction*.¹¹ Each pixel’s *surprise* is combined with its coordinates to localise its call to the machine’s attention. The strength and location of the call are combined with those of all other pixels to produce a

ⁱ Such a simple memory minimises computer requirements and does not strain biological plausibility.

single point which is the centroid of surprise. This is the point to which the motors then turn the camera.

To show quantifiably how WRAITH behaves we set up a scene containing two 5 cm discs with alternating black and white quadrants, attached to small motors set 12 cm apart (see **Figure 1**). When the discs rotated they created local changes in pixel levels, but as their motion was uniform and fixed in location WRAITH, after initially paying attention to them, almost immediately started returning to its neutral central position. Only when each disc suddenly ceased rotating did WRAITH's attention momentarily snap back to it.

3 Discussion

WRAITH achieves these results by passing information through three layers, each applying its own *surprise* function. This is not sufficient to go beyond level three in **Table 1**, and therefore the best that we can hope for with the current arrangement is the development of a range of responses to be used for control purposes. While aware of the vast gulf that lies between current performance and full understanding of human movement, we feel our approach is sound and will remain a foundation stone of further work.

The *surprise* function is central to this work. It provides sophisticated noise-reduced differencing of video frames, and multiple independent temporal resolution of data. It adapts to local (both spatial and temporal) prevailing conditions, obviates the need for thresholding, overcomes fixation, signals change, has space variant versatility, and can be used recursively.

In biological systems individual neurons, and even semi-discrete neural systems, over time exhibit reduction in response to unvarying stimuli.¹² The processes are called habituation, adaptation, or depletion, depending on the context. Such neural units effectively report onset and offset, not absolute values. This neural model becomes interesting when we look at sequential sets of units, each feeding its output to a successor. To get the maximal response from the final unit in a sequence, we must present the initial unit with a pattern of stimuli that propagates excitement through the entire sequence. What kind of signal would produce this response, and how might this apply to our work?

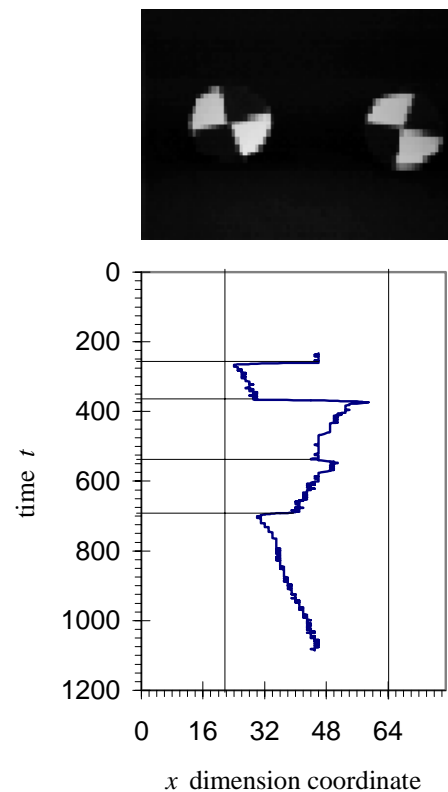


Figure 1: Accommodation to uniform motion.

The image (top) shows the scene containing the two spinning discs. The chart (above) shows the x coordinate of WRAITH's direction of gaze plotted against time (the trace records approximately 10 secs). At $t = 261$ the left disc starts spinning, it grabs attention for a moment, then slowly starts to lose it again. At $t = 367$ the right disc starts spinning, pulling attention to the right, then also losing it. At $t = 537$ the right disc **stops**. WRAITH responds by turning right **towards** the sudden cessation of motion, despite the continuing activity of the left disc. At $t = 690$ the left disc also stops, now causing a sudden swing of attention to the left, followed by a neutral centring of attention due to low ambient noise (c. 2%) across the scene.

The temporally varying signal that produces the greatest level of excitement is one which defies pattern wholeheartedly, at all levels. There are two ways in which we may think about how such a signal has the property of compelling attention: firstly, it has the property of driving our sequential sensing system to maximal response, the system cannot accommodate to the signal, cannot predict it, and is compelled to report a high level of surprise; secondly, from our objective third party perspective, we can see that such a signal, far from being mere noise, would contain multi-scale surprises and discontinuities in pattern that indicate something out of the ordinary.ⁱⁱ Things that are extraordinary defy our tacit predictions, and therefore grab our attention. It is highly suggestive that a change-reporting sequence of units reacts maximally to signals generated by activities which would also maximally attract our attention.ⁱⁱⁱ

Activities which we call ordinary can be accommodated (note the ambiguity of this term) in a finite number of neural layers. The number of layers ranges upwards from one in the case of inanimate, stationary objects, to an arbitrary number. Watching people working on a production line may fail to excite more than, say, six layers in our kind of neural system. Extraordinary patterns, on the other hand, are defined as those that exceed some arbitrary number of activation levels. What is extraordinary at one level of analysis may, of course, be quite unremarkable at the next level. A system built of several layers may therefore be able, based on the conjunction of reactions at different levels, to develop responses that represent activities at many different levels in the world, separating the extraordinary from the ordinary.

The accommodation capability would probably be enhanced if, in parallel, different levels of spatial resolution were being propagated through the neural layers. This is biologically valid, as the sensory systems do indeed contain mixes of both small and large receptive fields. When the presence of an unvarying response at layer i produces a non-response at layer $i+1$, no higher level patterns can be detected *unless* there is a parallel transmission of data using either a different resolution or a different accommodation function.

4 Future work

All the work discussed has used a retinotopic mapping of the world, so when the camera is moved the system must either remap its memory values to new locations based on the motion vector, or discard them completely. In parallel to this work we are experimenting with methods of storing memory values in a set of scene coordinates.

As with all mechanical implementations, there are difficult problems not encountered in simulations. Residual vibration and egomotion are present for much of the time WRAITH operates. By choosing to work in real time we have chosen to face these problems, and are currently looking at several ways of suppressing self-generated motion information.

ⁱⁱ 'Ordinary' things tend to remain still (if they are inanimate), move in periodic patterns (if they are machines, waves, wind-driven objects), or perform routines (if they are animals, labouring people, traffic).

ⁱⁱⁱ This is not the whole story, of course. Pattern recognition, and pattern reinforcement operate in parallel and counter to the phenomena presently discussed. But what is pattern to WRAITH? Pattern is recognised as incomplete penetration of the sensory layers. Any signal which does not propagate throughout the system must have been accommodated, and was therefore invariant, at some level of perception.

Multiple spatial resolutions may also provide a number of advantages. The *surprise* function uses an input of one pixel. For multi-resolution operations it will be necessary to provide the *surprise* function with smoothed and resampled versions of the highest resolution image.

5 Summary

It is interesting that, if systems like WRAITH can produce hierarchical responses to represent specific kinds and levels of activity in the world, then very little learning, in any conventional sense, is required to do so. All such a system has to do to actively seek out higher levels of pattern is maximise the *surprise* values of its innermost response layers, which thus take on the learning reinforcement function. Using absolute stimulation levels in this role may seem too simple, yet by observation of biological systems, we can see that maximising stimulation values is crucial to all kinds of processes, including the development of both sensory and control systems. A question we seek to answer, then, is how far can undifferentiated stimulation at posterior layers be used as a reinforcement function?

Our goal has been to make what we do at the current perceptual level fully amenable to modulation by subsequent subsuming levels. Foremost, we hope to use the motion representations of WRAITH as a means of inferring physical continuity and invariance in the world, and thereby construct more meaningful spatio-temporal models of human activities surrounding the robot. We see this work as a foundation for more application-specific motion recognition or behaviour generation research.

¹ Ackerman D *A natural history of the senses* Phoenix, London, 1990, p 305.

² Campbell R, Landis T, and Regard M *Face recognition and lipreading: a neurological dissociation* Brain 109, 1986, pp 487-499.

³ Ernst B *The eye beguiled* Taschen, Germany, 1986.

⁴ Marr D *Vision* W H Freeman, New York, 1982.

⁵ Dance S, Caelli T, and Liu Z-Q *An architecture for a traffic scene interpretation system* Technical report 94/12, Department of Computer Science, University of Melbourne, 1994.

⁶ Chella A, Frixione M, and Gaglio S *A cognitive architecture for artificial vision* Artificial Intelligence 89, 1997, pp 73-111.

⁷ Gibson J J *The ecological approach to visual perception* Houghton Mifflin, Boston, 1979.

⁸ Brooks R A *A robust layered control system for a mobile robot* IEEE Journal of Robotics and Automation RA-2(1), 1986, pp 14-23.

⁹ Peters M W and Sowmya A *Active vision and adaptive learning* Proceedings of Intelligent Robots and Computer Vision XV, SPIE Vol 2904, Boston, 1996, pp 413-424.

¹⁰ Thibadeau R *Artificial perception of actions* Cognitive Science 10, 1986, pp 117-149.

¹¹ Pribram K H *Brain and perception* Lawrence Erlbaum Associates, New Jersey, 1991, p 221.

¹² Day R H *Human perception* John Wiley, Sydney, 1972, pp 139-161.