

# 6

## Developing Sensibility – SURPRISE

### 6.1.1 Introduction

**T**HIS chapter addresses four questions concerning control of attention: how to direct attention to novel or ‘interesting’ events in the world, how to separate novel or interesting information from everything else, how to perform this segmentation of the world at multiple levels, and how to develop perceptual machinery capable of this. Each of the answers proves to be effective in both an engineering sense (having contributed to the control of WRAITH in natural surroundings) and a scientific sense, having sufficient biological plausibility to be a working hypothesis about natural visual systems.

The SURPRISE algorithm described in this chapter has been designed to accept any output produced by the JIGSAW ordering algorithm, or the DIEM sampling algorithm, described in previous chapters. It works equally well on standard linear orthogonal arrays and other space-variant mappings. Only with a motion pattern detection and reaction algorithm such as SURPRISE, can a robot begin to interact physically with the world. This is therefore an important component of WRAITH.

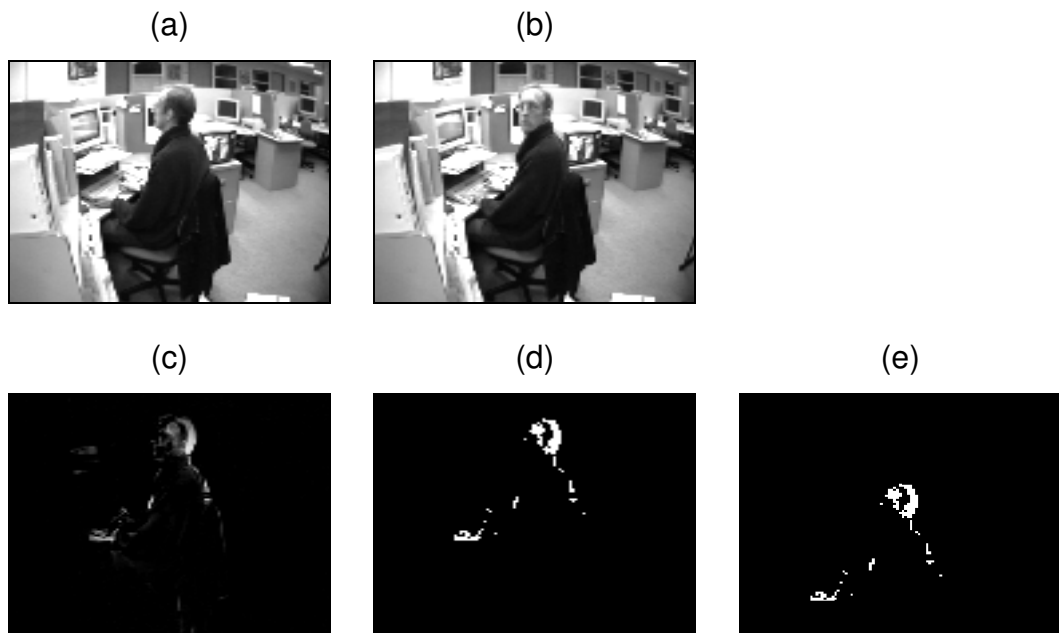
### 6.1.2 Highly reactive responses

In early WRAITH experiments the fundamental identification of motion was determined by simple pixel-level subtraction of one image from its predecessor. The resultant abso-

lute image difference was then thresholded, and actual values reduced to one or zero (see **Figure 6.1**). This produced a binary image whose dimensions were determined by the number of samples selected. The thresholding of differences eliminated most noise created by minor fluctuations in lighting, which caused pixels to record continuous spurious changes of one grey level, and sometimes more. This level of noise was unbounded, spread over the entire image regardless of the location of object edges in the scene, and was consequently highly detrimental to coherent motion analysis. The actual value of the difference at any particular pixel location is linked to several factors 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 simple difference values. However, it was found that, rather than deal with actual difference values, it is often sufficient to simply aim to divide the image into areas of movement versus non-movement, and hence use of the binary image.

For the purposes of guiding the camera so that it retains the highest level of real movement in the field of view, it was useful to develop a low-level (reactive) response which can function in the absence of all world knowledge or inferences based on previous images. In real time environments it is often preferable to respond quickly rather than accurately or with circumspection. Motion tracking is one such situation, because if a system delays camera repositioning until it has performed high level scene analysis, 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. This is reminiscent of the subsumption architecture (Brooks 1986) but differs in one significant respect: subsequent higher levels of response in WRAITH (corresponding to Brooks's subsuming layers) must be internally constructed via *adaptation* rather than externally imposed by the developer.

The low level tracking response in WRAITH is concerned solely with *movement* tracking, and not *object* tracking. To infer the presence of an object, *per se*, is a higher level process that depends on the extraction of a number of invariant pieces of information (Palhang and Sowmya 1999) from a sequence of images, and a probabilistic analysis of this data. By dealing with movement alone, regardless of object model inferences, it is possible to achieve high degrees of success in following rapidly moving objects, simply because a real



**Figure 6.1: Low-level motion tracking.**

(a) and (b) are closely spaced video images converted to data arrays by the frame grabber. (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. (d) is the thresholded binary version of (c). After calculating the centroid of the white areas in (d), which is somewhere in the subject's head, the system sends move commands to the motors controlling pan and tilt, resulting in a scene something like (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 (c). Fortunately in this instance the thresholding eliminated this in (d) before the centroid was calculated. Subsequent developments of the algorithm removed this, and related problems.

moving object creates coherent movement patterns through space and time, regardless of much of its object-level qualities.

While the motors reposition the camera it continues to transmit images, which are asynchronously processed as just described. WRAITH does not toggle between 'seeing' and 'saccading' modes, it simply discards most images received while the camera is moving (smeared images) or while residual mechanical vibration is present in the camera (blurred images). Difficult smeared or blurred images are easy for WRAITH to distinguish as they usually report motion across the full width and height of the image, or at an unusually high density (above a threshold of, say, 25% of all 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 a bimodal (seeing versus saccading) method because it reacts well to unpredictable vibration effects, in effect delaying further motion tracking just long enough to get good data. One heuristic avoids processing any data from the central foveal region – to save time. Another heuristic cancels any camera movements below a certain angle – to avoid unnecessary jitter. 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, such as a tree in a breeze. The benefits are reduced use of the motors and clearer images.

The low-resolution binary report of motion makes it an easy task to define a bounding rectangle, ellipse, etc., and this facilitates tight efficient focussing of more discriminating, but resource-consuming, clustering or segmentation algorithms.

Different methods for calculating the centroid of binary motion make little apparent difference to the system's performance, particularly when camera movements are taking place roughly every 300-400 msec. 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 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 correctly points between the two edges, straight at the person, even though the area in the centre of the image has not reported any movement. However, if two very thin objects are moving side by side (such as a couple of saplings in a breeze) they create a very similar motion image (two roughly vertical coordinated motion lines against a background of non-motion). It would not always 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 (and which, by implication, is more interesting): it cannot usefully be indiscriminate at *all* levels of analysis. Clearly this points to the next step in system enhancement: discrimination of separate motion sets, and decisions about which of these to select for further observation. Ways of deciding which object is more interesting without knowing very much

about the object itself, on the basis of its movement patterns, are described later in this chapter.

Referring to **Figure 3.4**, we see that at this reactive stage in its development WRAITH implemented levels of analysis two and three (existential and ontological). In work described later in this chapter, WRAITH encroached on level five (the spatio-temporal level of analysis), in being able to detect where and how people moved, and having the ability to adjust direction of gaze and sampling regime accordingly.

### 6.1.3 Less reactive responses

Differentiation of motion types can 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 travel, spatially or temporally coordinated motion such as formal dance steps, sports, etc. Motion without travel 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 6.1(c)**) are reported as rectangular slabs of intense movement regardless of what they display. Despite all the apparent energetic movement, however, the screens stay in place. This might be classified as ‘motility without mobility’. WRAITH does not differentiate by categorisation, but by a more direct connection between sense and action. In this way it adopts the ecological approach to vision first promoted by Gibson (1979) who held that vision can only be properly understood within a framework that recognises the idea of an eye literally roving around in not just physical space but a visual environment having a specific relationship to the owner of that eye.

The most important aspect of this approach is that motion of both the subject and the object is the norm. Secondly, what the vision system is dealing with is not an image, but events in the world, events that are projected on to an imaginary spherical optic array centred on the vision system. Thirdly, within this array *transformation* is the primary characteristic; ‘What gets (sic) displayed are disturbances of structure in the array’ (Gibson p. 293).

Gibson’s approach differs strongly from the computational and much more explicit

theories of vision epitomised by the work of Marr (1982). Marr held to an immutable one-to-one mapping of internal and external states of affairs. Gibson did not. Gibson pointed out ‘the experimental psychologist should realise that he cannot truly control the perception of an observer, for the reason that it is not caused by stimuli [alone]’ (Gibson p. 305). In other words, the past cannot be separated from the present, and pre-experimental factors always affect the experiment. Similar factors apply to active vision; two algorithms cannot be compared on exactly the same data because, if they are *functionally* different, they actively change the data in different ways. The ecological approach has been given considerable support in the last few years, at a strategic level by the school of situated cognition (Clancey 1993, Norman 1993), and at a tactical level with the advent of active vision (Blake and Yuille 1992).

Gibson’s work is particularly relevant to active vision because of its emphasis on behaviour, embodiment, real world operation, real-time, and goals. Gibson pointed to some of the key problems, conveniently side-stepped by much disembodied vision research and image processing: problems such as self-calibration, separation of ego-motion from other motion, and so on. Of course, work developing solutions to these problems is now being done (e.g., Prokopowicz 1995).

However, many aspects of Gibson’s work raise questions that remain unanswered. One that has particular relevance to the work discussed here is his dichotomy of so-called ‘perceptual adaptation’ and ‘awareness of transformations’. On the one hand, he wrote that he had given up theorising about perceptual adaptation (Gibson p. 248), but on the other, draws attention to the ‘false dichotomy between past and present experience’ (Gibson p. 253), and says his ‘pickup theory can assume an awareness of transformation’ (Gibson p. 247). It is not clear that the processes that Gibson believed in are really in opposition to what he did not believe, since, from the point of view of a truly *functional* vision system there can be no distinction between changes in the system (e.g., perceptual adaptation) and changes outside the system (i.e., awareness of transformation). Any modelling system can only apprehend one class of change (internal), and can at best merely act *as if* this bears some sensible relationship to another kind (external and unknown). As has already been stressed, it is never possible for a system to verify its perceptions of the world by stepping over them and comparing them to the real thing. Gibson’s idea of direct pickup of affordances, when all is said and done, must somehow be executed by mechanisms or programs. The perceptual adaptation of a vision system has to be thought

of as ‘awareness of transformation’ in the optic array. In the case in point, perceptual adaptation, as typified by the self-modulating SURPRISE algorithm of WRAITH, which is described below, does *exactly* what Gibson sought: it creates both a synthesis of past and present *and* a direct flagging of change.

### 6.1.4 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 (Campbell, Landis and Regard 1986, Hess, Baker and Zihl 1989). Although usually taken for granted, such understanding probably involves a complex synthesis of many levels of perception, from simple reactive motion detection up to an apprehension of the full socio-cultural import of an action (see **Table 3.1**). These levels of analysis are almost certainly interactive. Lower levels seem to provide much of the information used by higher levels, probably operating in some sense as filters of unnecessary or irrelevant information. Simultaneously, higher levels, though dependent on lower level input, have been shown (Ernst 1986) to influence the content of lower level perception.

As mentioned in Chapter 3, *Constructing a Framework - WRAITH*, standard conceptualisations of specific perceptual levels tend to be rather heterogeneous: mathematical when it comes to low levels (Marr 1982), and by turn logic-based (Waltz 1972), linguistic (Chella, Frixione and Gaglio 1997), and informal (Gibson 1979) at higher levels. How we learn to synthesise information at all levels is hard to say, since it is rather difficult to propagate a reinforcement function through a set of heterogeneous systems, though taking an economic approach may be a possible answer (Simon 1969). More significantly, 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. We commonly write programs that are able to instantiate other programs to do a task, but this task is never, in turn, to create another program-creating program, and so on. It is likely that if we are ever to solve the hierarchical integration problem, we will do so by discovering suitable recursive techniques. The output program of a parent program will probably resemble that parent in most basic respects, but its different position in the architecture will mean that its input will be different, and *therefore* its role will also be different. Uniform programs with non-uniform roles might then be proliferated to a degree commensurate with the complexity of the goals or input.

Homogeneity, not heterogeneity, across the complete architecture therefore has attractions. Crick (1994) writes:

The secret of the neo-cortex, if it has one, is *probably its ability to evolve additional layers to its hierarchies of processing*, especially at the upper levels of those hierarchies. Such extra layers of processing are probably what distinguishes (sic) higher mammals, like man, from lower mammals, such as the hedgehog. I suspect that the neo-cortex uses special learning algorithms that permit each cortical area to extract new categories from experience, even though each area is embodied in a complex processing hierarchy.

It is suspected that these ‘special learning algorithms’ are adaptively homeostatic, each causing an additional layer to evolve only when it fails to maintain equilibrium on its own. The development of a new adaptive layer might then even up the score, allowing the organism to successfully adapt to higher level patterns in the world not previously recognisable. This then becomes *de facto perception of that higher level of pattern*.

## 6.1.5 Designing the solutions

In building machines that operate on such hierarchical principles we are forced to decompose the problem. This work adopts a bottom-up approach. This chapter first asks what does attention really mean *in practice*, and what basic decision-making processes are required for directing attention. Then, answers are progressively pursued back up the causal chain, and at each stage the assumptions that enabled all previous questions are themselves questioned. The four stages are as follows:

### 6.1.5.1 Directing attention to novel or interesting events in the world

Assuming that we have already been able to make a map of our surroundings that indicates our level of interest in every direction, we must now find a way to select a single point to which to turn and pay attention. It may be appropriate to clarify what we mean by ‘paying attention’. It must mean more than simply remembering, or re-orientating towards something, since it implies being able to receive more information, or higher resolution information, from the object of attention. This entails the *concentration* of sensors upon the object of attention. One way to achieve such a concentration is by employing a space variant mapping, so that once a particular orientation towards the object is made, there are more pixels devoted to that object than had been the case beforehand. DIEM performs this function; it enables standard reorientation of attention. The method of selecting the point of interest is not dependent on any inferences about agents or activities, it is highly simplified: calculation of a centroid of activity, as described in the following

section. This is very much like the execution of reflexive eye movements triggered by a population of superior colliculus neurons described by Sparks, Lee and Rohrer (1990).

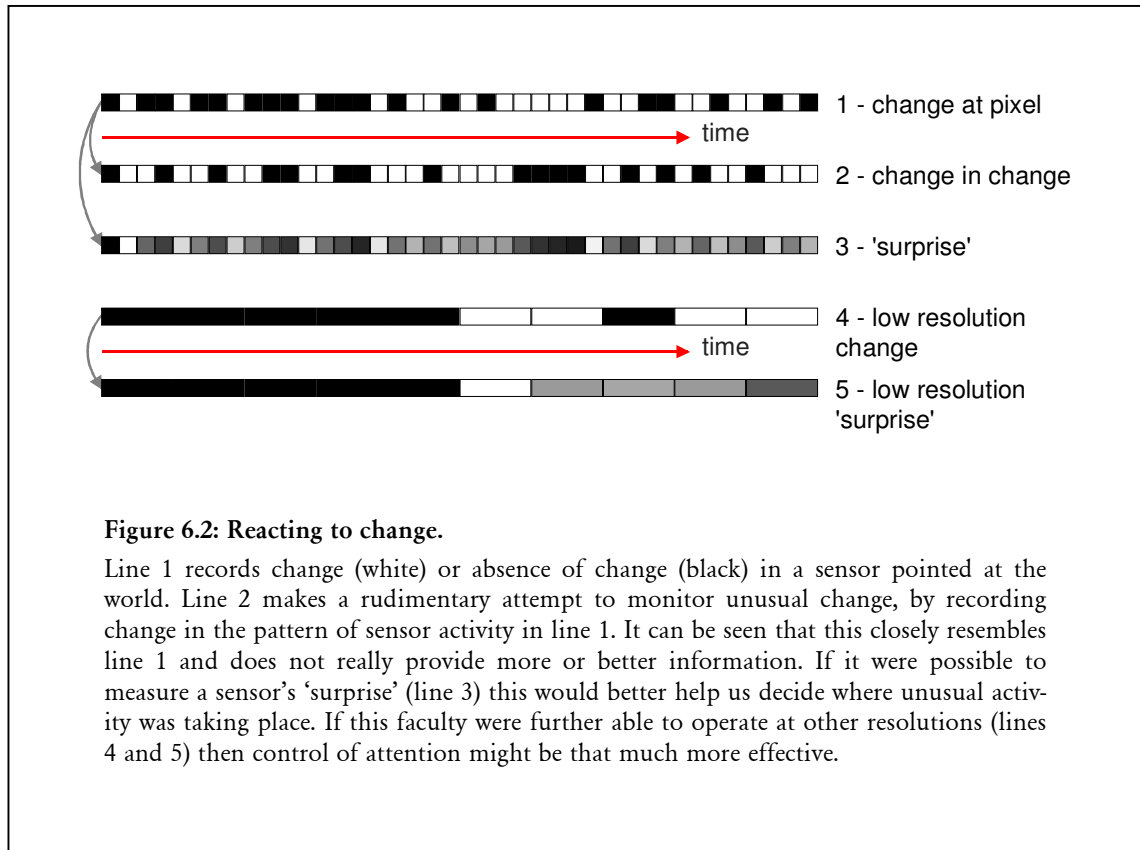
### 6.1.5.2 Separating novel or interesting information from everything else

For all biological systems some states are more conducive to life than others, and the basic biological function is to preserve homeostasis by moving from less conducive states to more conducive ones. This applies to states both internal and external. Once homeostasis is achieved, a system need not do anything different until there is some change of state. In other words, it does not even need to keep telling itself that things are still the same. It is not surprising then, that over time individual neurons and even semi-discrete neural systems exhibit reduction in response to unvarying stimuli (Day 1972). Such changes in response are called habituation, adaptation, or depletion, depending on the context. Importantly, such neural units effectively report onset and offset, not absolute values.

However, it grossly oversimplifies to say that perception is ‘on’ in the presence of change and ‘off’ otherwise, because perception is actually ‘off’ much more frequently than suggested. We habituate not only to no change, but also to consistency of change. Tuning out constant background phenomena such as a ticking clock is an example. That this happens should not be surprising, as it is a predictable effect of certain multi-layered neural systems described later, in which the output of one layer is the input of another. A constant signal will cause an anterior layer to habituate, and thus cease to activate the change-measuring function in a posterior layer.

The step from change to surprise is simple, though the concept of change can be a little elusive. What is experienced as change at one level of resolution may not be at another, and simply compiling records of change does not really provide any advantage either (see **Figure 6.2**). To decompose the phenomenon of surprise let us agree that there are two primary components: an *expectation* and a *departure*, without either of which no surprise can be experienced. The expectation is based on previous experience, and might loosely be thought of as a form of pattern recognition. The departure is the discrepancy between what is expected to happen and what actually does happen.

To measure the discrepancy we need memory to compare what is happening now with what has just happened. We start by providing each pixel location in the visual field with a miniature memory unit. This possesses a proto-memory consisting of a single value  $M$ .



The signal entering each unit is represented by a single value  $G$ . In the initial case,  $G$  is just the brightness of the pixel at the memory unit's location. The value of  $M$  is updated from  $G$  continually, using the equation:

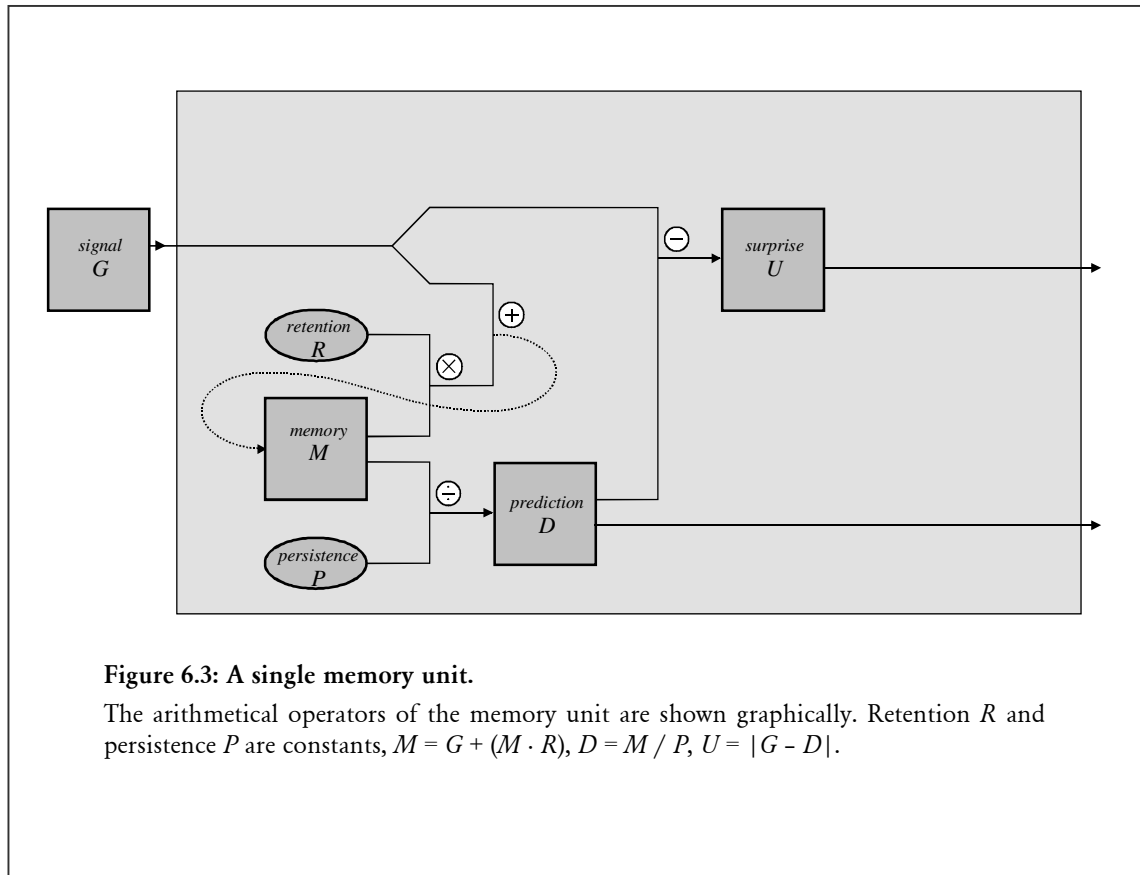
$$M_t = G + (M_{(t-1)} \times R), \text{ where } 0 < R < 1 \quad (6.1)$$

Varying retention  $R$  adjusts the relative weight given to less recent values of  $G$ , effectively determining whether memory can be thought of as long-term or short-term, and thereby having significant influence over the behaviour of the system.

If  $1 - R$  is small, then the value of  $M$  will rapidly grow large in relation to that of  $G$ , due to its accumulative effect. So, as we wish to compare  $G$  to  $M$  to derive a surprise value  $U$ , we first need to renormalise  $M$ . To do this we use another constant,  $P$ , which is derived from  $R$ :

$$P = 1/(1 - R) \quad (6.2)$$

$P$  measures the persistence of memory. If the system has strong retention (i.e.,  $R$  is high), this is reflected in a correspondingly high value in  $P$ .  $P$  is specifically calculated to provide a measure of expectation or prediction when applied to  $M$ . Now:



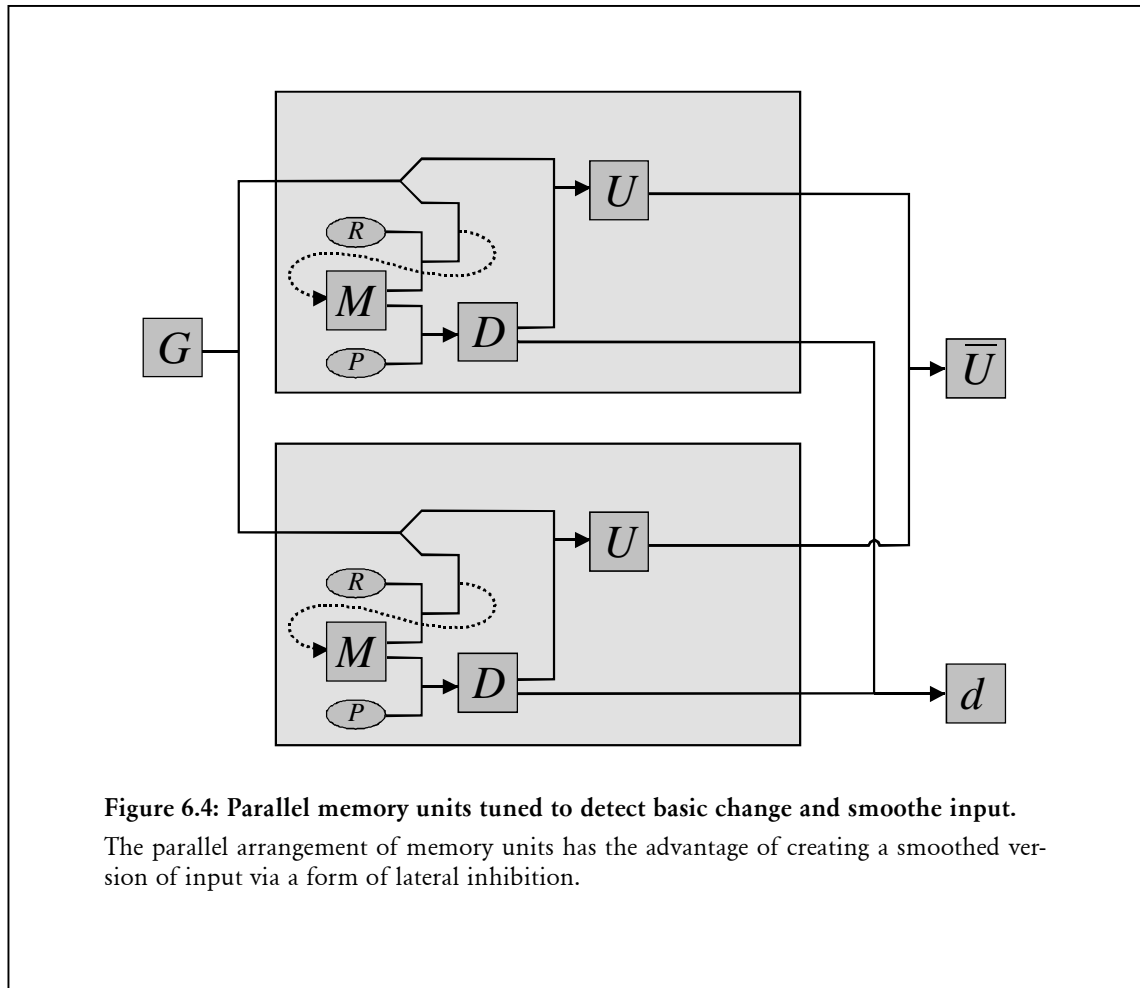
$$D = M/P \quad (6.3)$$

where prediction  $D$  is an exponentially decaying moving average of  $G$ . Then:

$$U = |G - D| \quad (6.4)$$

Thus surprise  $U$ , the discrepancy at any one moment between the signal  $G$  and the prediction  $D$ , corresponds to the ‘figural’ content set against the ‘ground’ of  $P$  (see Pribram 1991). The preceding concepts are represented diagrammatically in **Figure 6.3**.

$U$  is an output of the memory unit, one that responds to change in single time intervals because its value depends directly on the input value of  $G$ . It is therefore subject to any noise carried by  $G$  and does not, in itself, provide any temporal smoothing. Unfortunately, it transpires that noise levels are quite high in video data. The effect of noise added to an unchanging stream of uniform input values is to produce a spurious and continuous form of reported change where sometimes only stasis should be reported, so some degree of smoothing is desirable. Fortunately, the memory unit described here already has a form of internal smoothing in  $D$ . Consequently, the first two memory units (called *channel1* and *channel2*) in our visual system are implemented in parallel, and the  $D$



**Figure 6.4: Parallel memory units tuned to detect basic change and smooth input.**  
The parallel arrangement of memory units has the advantage of creating a smoothed version of input via a form of lateral inhibition.

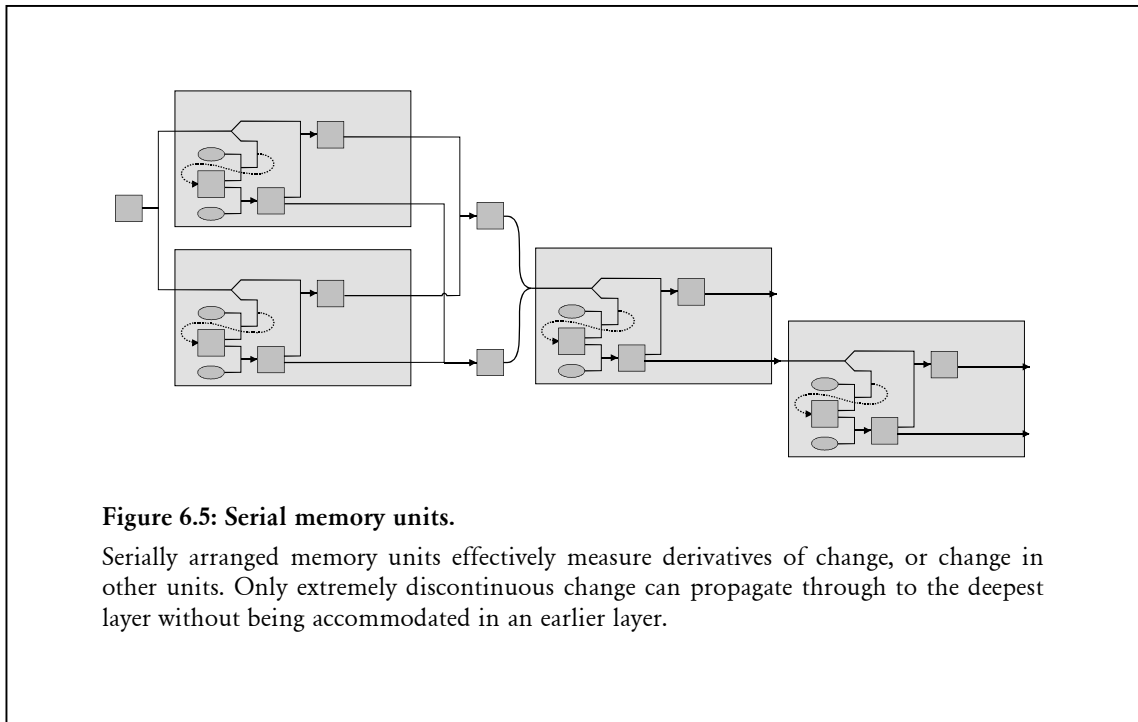
outputs used to produce a difference  $d$ , that is

$$d = |\text{channel1}.D - \text{channel2}.D| \quad (6.5)$$

Thus we compare two smoothed versions of the input, one having a retention value of 0.2 and the other 0.4, say. This is effectively a smoother version of each channel's input  $G$ , and the difference between the two provides a smoother counterpart to  $U$ . This is a form of lateral inhibition, shown in **Figure 6.4**.

We assume the system's initial focus is directly ahead, at the centre of its visual field. This position is incrementally modified as we examine the output of each memory unit. After a pass (which does not necessarily have to be exhaustive) through the memory units the new location of the focus of attention  $(x, y)$ , the centroid of interest, is given by:

$$x = \frac{\sum_{u=1}^n u_x u_d}{\sum_{u=1}^n u_d}, \quad y = \frac{\sum_{u=1}^n u_y u_d}{\sum_{u=1}^n u_d} \quad (6.6)$$

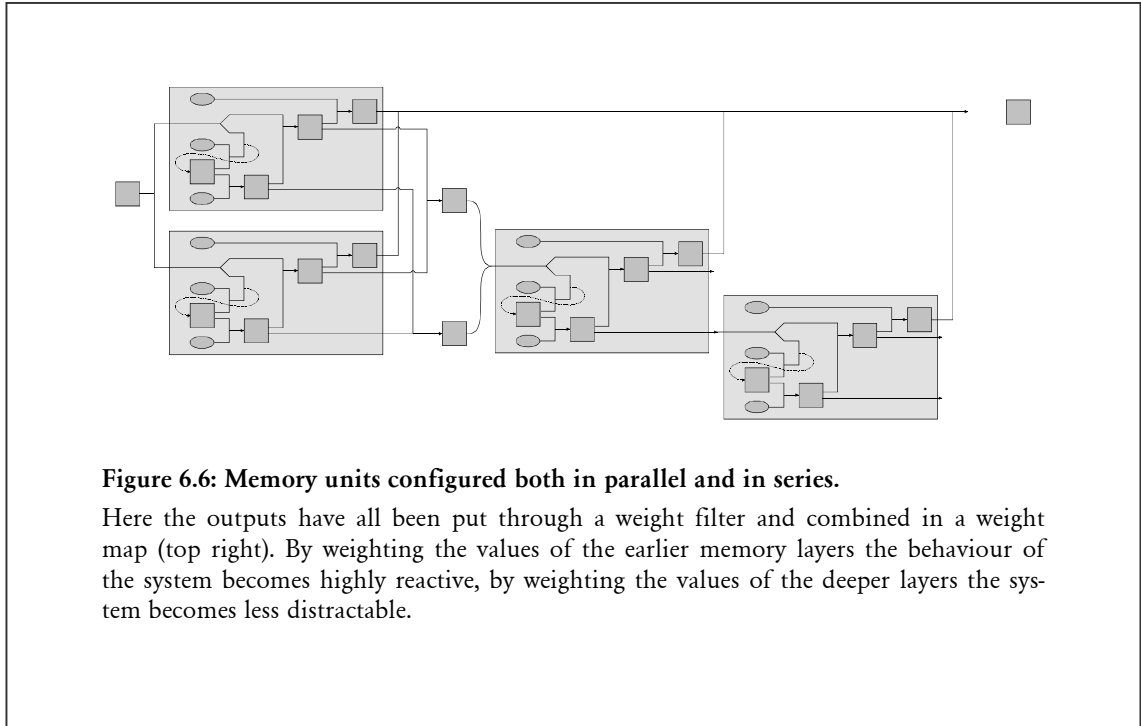


where  $n$  is the number of memory units examined, each of which ( $u$ ), has an  $x$  co-ordinate ( $u_x$ ), a  $y$  co-ordinate ( $u_y$ ), and a difference value ( $u_d$ ). This emergence of attention can be likened to a pandemonium model in which every memory unit location is constantly calling attention to itself though it is only effective if other units comply by being relatively quiet for a moment.

### 6.1.5.3 Performing segmentation of the world at multiple levels

Distinctions between ordinary and extraordinary events must be made at multiple levels. Indeed, a single event may be simultaneously novel at one level and quite unremarkable at the next. For example, one morning a person who has been lying down suddenly rises and walks about but is known to do this every day. Within the context of the day their activity is novel, within the context of the week, month, year, lifetime, it is not. If this is any ordinary morning rising it may get a second or two of our attention, but if this is the first rising in many months it will be paid far more attention. To produce the same range of reactions in machines we need now to develop mechanisms that respond to an event *according to all its temporal contexts*.

WRAITH computes change as a conjunction of the outputs of several layers of surprise generators (memory units), each building upon a previous layer and effectively measuring



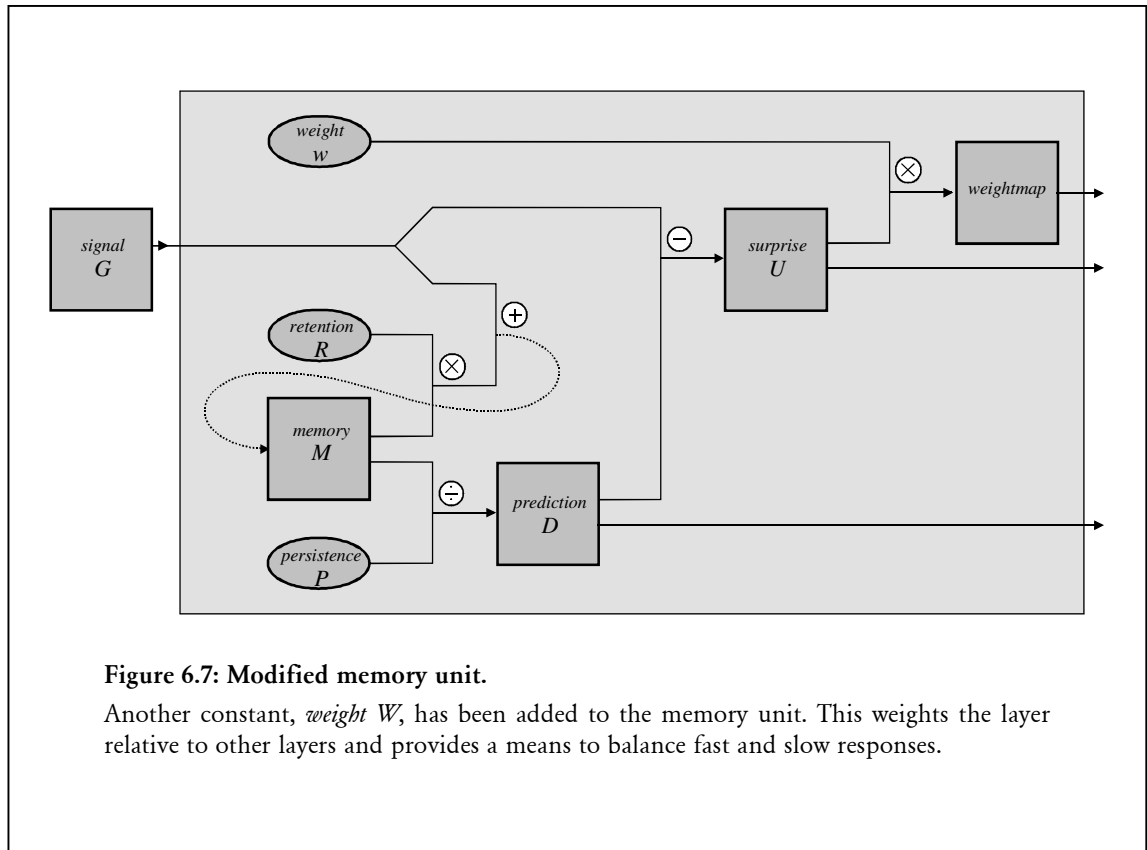
the way change itself, as a perception in its predecessor, has been changing (see **Figure 6.5**). Each layer tunes itself to recognise a pattern (the current pattern) in its input and reacts only when this pattern is interrupted. Its reaction is passed to the next (higher) layer, forming another input, with a pattern of its own.

The system's behaviour is considerably improved by the addition of more memory layers. These new layers become responsible for keeping memory of lower frequency change in the world. They therefore need to be connected, not to data such as  $G$ , but to a data stream that has already had higher frequencies removed, and  $D$  fits this role perfectly. Note that once memory units are arranged serially, there must be some relative weighting between the layers, even if this weighting is uniform and gives precedence to no layer in particular. Effectively, equations (6.6) now become:

$$x = \sum_{l=1}^m \frac{l_w \sum_{u=1}^n u_x u_{dl}}{\sum_{u=1}^n u_{dl}}, \quad y = \sum_{l=1}^m \frac{l_w \sum_{u=1}^n u_y u_{dl}}{\sum_{u=1}^n u_{dl}} \quad (6.7)$$

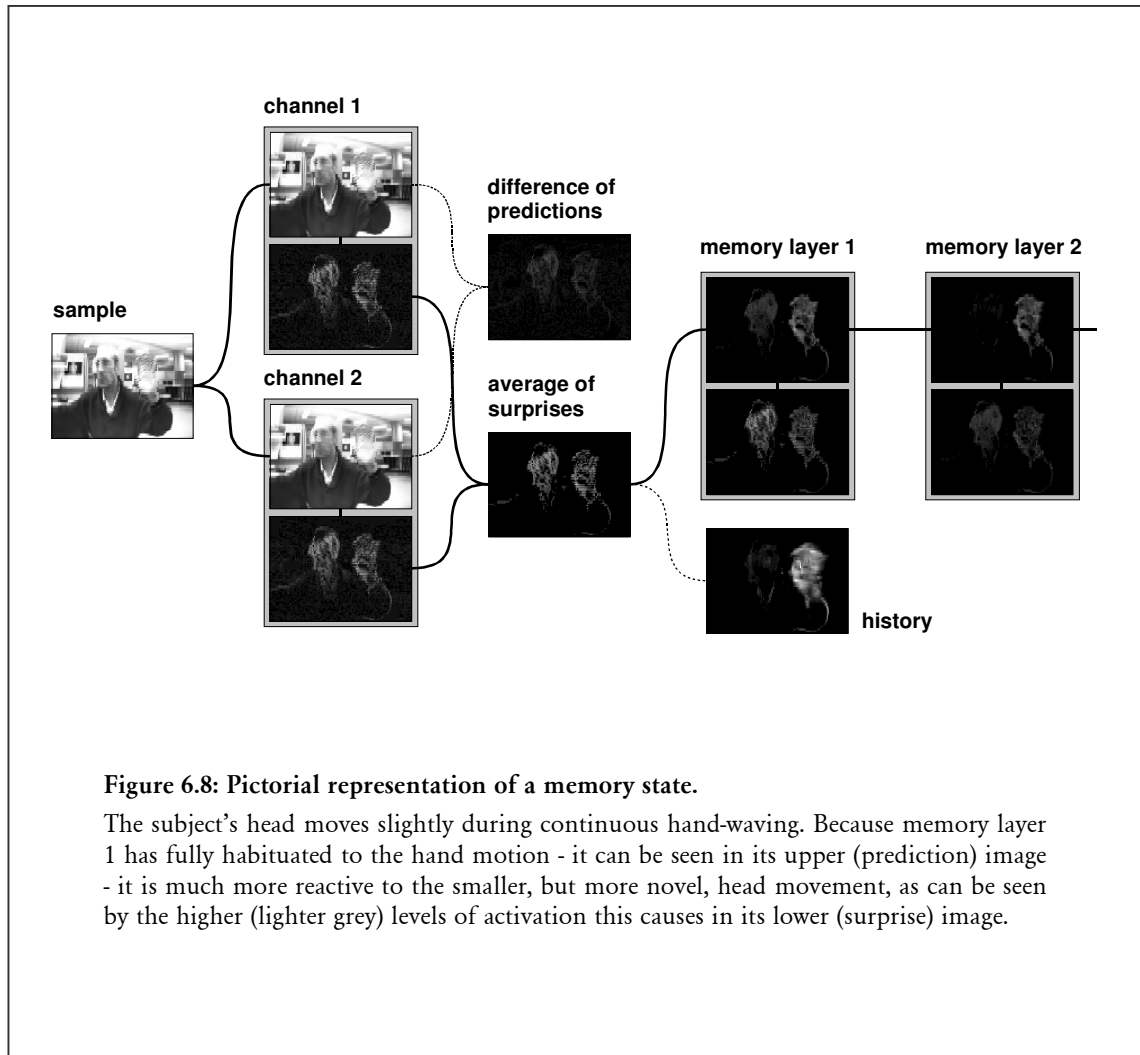
where  $m$  is the number of memory layers, each of which ( $l$ ) has a weight  $l_w$ . This is shown in **Figure 6.6** and the integrated memory unit now looks like **Figure 6.7**.

These processes are shown pictorially in **Figure 6.8** where the four memory units of **Figure 6.6** are represented by four two-image blocks in each of which the upper image is



made from prediction ( $D$ ) values and the lower image is made from surprise ( $G$ ) values. The images show the state of the system at an instant of time. It can be seen that in memory layer 2, surprise  $G$  values activated by the movement of the subject's head are stronger than those caused by the movement of his hand. This is because the hand has been waving almost continuously (evidenced by high levels in both the *history* image and memory layer 2's prediction  $D$  image), whereas the head has until this moment been quite still. In other words the system is no longer attracted by the frantic hand, and has just transferred its attention to the less intense but more novel head movement.

To quantify how WRAITH behaves under different memory configurations, a scene containing two 5 cm discs with alternating black and white quadrants, attached to small motors set 12 cm apart, was set up (see **Figure 6.9**). **Figure 6.10** plots the focus of attention of three layers of memory units. The short-term plot is that of the two parallel units named *channel1* and *channel2*, the medium-term plot is that of memory layer 1, and the long-term plot is that of memory layer 2. The plot is actually the horizontal  $x$  co-ordinate of these three foci, plotted here vertically against time. The upper trace and lower trace represent approximately 31 seconds and 38 seconds, respectively. In each trace the grey bars correspond to the times that each disc was spinning. In the upper trace the first disc



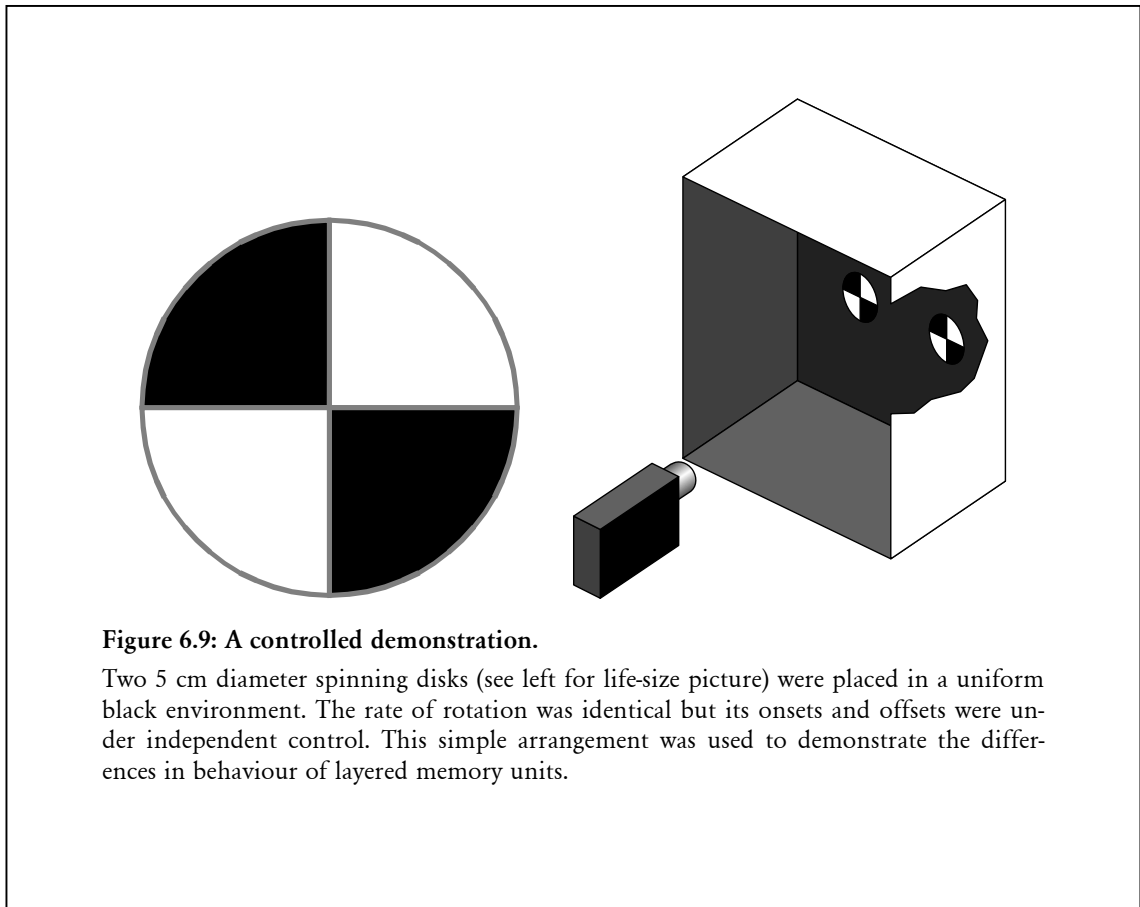
**Figure 6.8: Pictorial representation of a memory state.**

The subject's head moves slightly during continuous hand-waving. Because memory layer 1 has fully habituated to the hand motion - it can be seen in its upper (prediction) image - it is much more reactive to the smaller, but more novel, head movement, as can be seen by the higher (lighter grey) levels of activation this causes in its lower (surprise) image.

starts to spin after about a second, and soon attracts the attention of all three memory layers. It can be seen that the short-term memory is highly reactive. The second and third layers are slower to react.

About 1.8 seconds later the second disc starts to spin. The short-term memory simply centres its attention precisely halfway between the discs. The medium-term memory is momentarily attracted close to the second disc, but also soon centres its attention. Meanwhile the long-term memory, though slower to react, takes a hard look at the second disc before also finally centring.

A more interesting phenomenon occurs at 13.4 seconds when the second disc suddenly ceases to spin. The short-term reactive memory is now free to return to the first disc, but long-term memory, having fully adapted to the motion of the second disc, is now dramatically attracted to the sudden cessation of rotation, despite the fact that *all motion is on the other side of its visual field*. The same goes for the medium-term memory, though to a



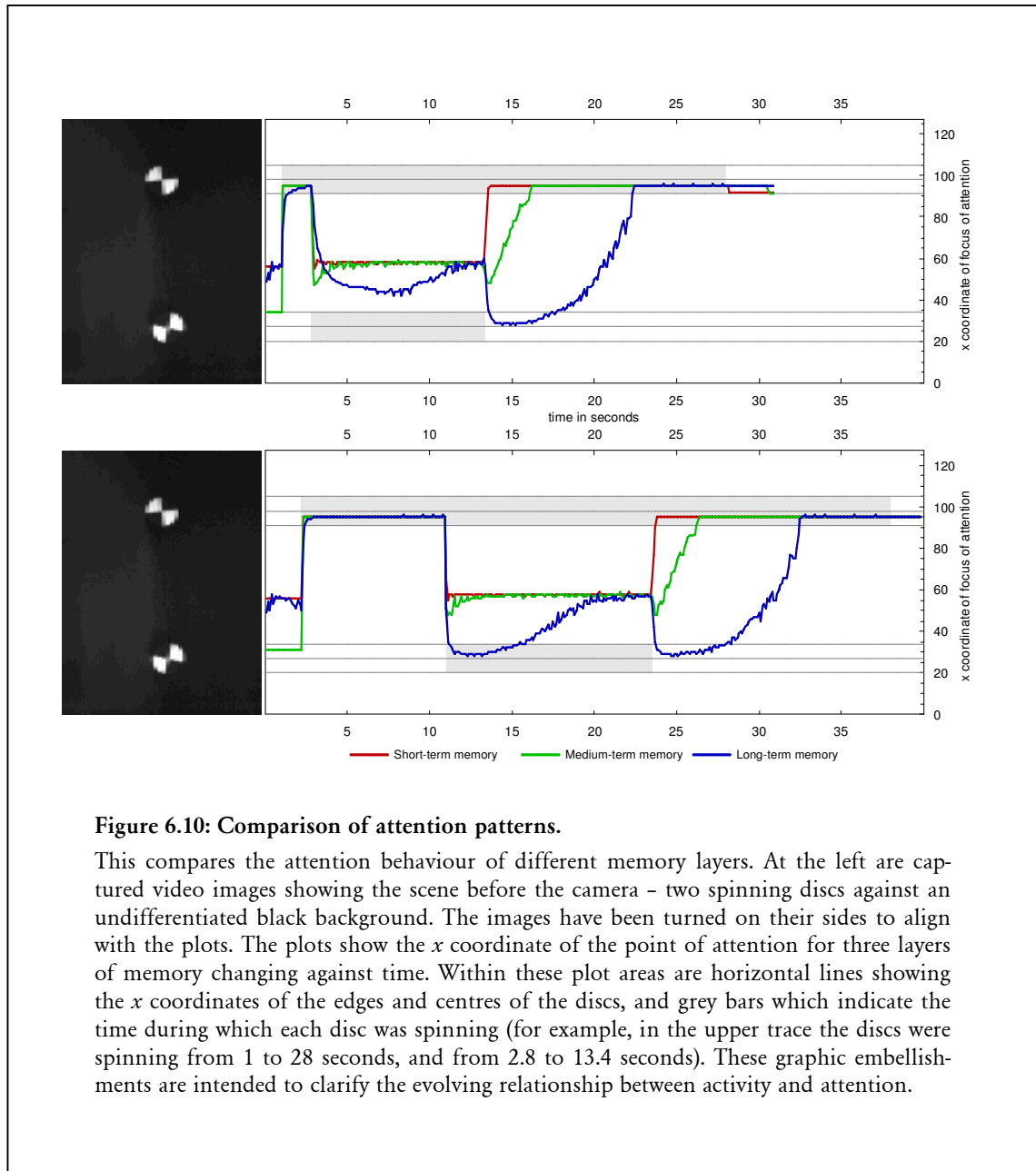
**Figure 6.9: A controlled demonstration.**

Two 5 cm diameter spinning disks (see left for life-size picture) were placed in a uniform black environment. The rate of rotation was identical but its onsets and offsets were under independent control. This simple arrangement was used to demonstrate the differences in behaviour of layered memory units.

lesser extent. After 23 seconds all foci are attracted to the sole spinning disc, where they remain once that disc too finally stops after 28 seconds.

In the lower trace the sequence of onsets and offsets is the same, except that now the memory is given more time to adapt to the rotation of the first disc. Instead of only 1.8 seconds, it remains spinning for about 8.7 seconds before the second disc starts. By its reaction, it can be seen that the long-term memory has fully adapted to the constant rotation of the first disc, so it moves fully to the centre of the second disc when it starts, and remains preoccupied with it for several seconds before eventually accommodating to that motion too.

WRAITH's behaviour in response to any particular stimulus is predicated on its internal state. Consequently when the second disc starts turning in the upper trace, attention is diverted towards it, but not far enough to actually foveate it. This is because the onset of motion follows too soon after the onset of the first disc, which is still exerting some hold on WRAITH's attention. In the lower trace, WRAITH has had more time to accommo-



**Figure 6.10: Comparison of attention patterns.**

This compares the attention behaviour of different memory layers. At the left are captured video images showing the scene before the camera – two spinning discs against an undifferentiated black background. The images have been turned on their sides to align with the plots. The plots show the  $x$  coordinate of the point of attention for three layers of memory changing against time. Within these plot areas are horizontal lines showing the  $x$  coordinates of the edges and centres of the discs, and grey bars which indicate the time during which each disc was spinning (for example, in the upper trace the discs were spinning from 1 to 28 seconds, and from 2.8 to 13.4 seconds). These graphic embellishments are intended to clarify the evolving relationship between activity and attention.

date to the activity of the first disc, and is therefore much more drawn to the second disc when it starts.

#### 6.1.5.4 Developing the perceptual machinery capable of segmentation

WRAITH achieves these results by passing information through three layers, each applying its own SURPRISE function. In **Table 3.1** it can be seen that this is not sufficient to go beyond level four (Physical: differentiation of motion types) and therefore the best that we can hope for with the current arrangement is the development of a range of responses for control purposes.

Despite the vast gulf that lies between current performance and full understanding of movement, this approach contains a valuable principle, that of *demand-driven* recursive auto-generation of higher-level perceptual layers. This may provide a sketch of the secret of the neo-cortex that Crick (1994) wondered about. This approach is also interesting by virtue of its ability to perform hierarchical segmentation of the world, and the fact that it can also automatically generate new higher-level reinforcement functions that can train it to seek out complexity that it was once *incapable of perceiving*. These three important characteristics will now be discussed separately:

#### 6.1.5.4.1 Demand-driven recursive auto-generation of higher-level perceptual layers

Let us first examine *how* new layers might be generated, then the question of *when*. Given that deeper memory layers are structurally similar to early memory, the task of generating them is very straightforward: it involves no more than copying existing structures. The new memory must then also inherit constant values for *retention* and *weight*. These may be tuned to give the greatest increase in information content. Little more is required.

The time at which new layers are generated will depend on the ability of the earlier layers to fully accommodate (note the ambiguity of this term) incoming signals. If the current deepest memory layer is persistently experiencing high levels of surprise then it is clearly inadequate to the task of accommodating all signals from the world. This high level of excitation is the cue to generate a further layer of memory. The deeper memory has the job of detecting or accommodating patterns in the part of the signal that previous layers have been unable to accommodate. The more complex the world, the more complex the perceptual machinery generated from it.

The system is now able to use this way of reacting according to the time and memory it has available. If there is no undue pressure to act, then all memory may be used, but if there is immediate danger of collision, say, then higher level memory is superfluous. This indicates that memory searches should probably proceed in order from the earliest to the deepest layers, though the situation regarding biological systems may be more complicated than this.

It is feasible that deeper memory layers may atrophy if earlier layers are perfectly capable of fully accommodating all signals. By extension, there is no reason why deeper memory layers cannot pass in and out of existence according to the vagaries of the prevailing environment.

#### 6.1.5.4.2 Hierarchical segmentation of the world

Activities that we call ordinary (a highly contextual term) can be accommodated in a finite number of perceptual layers. Watching people working on a production line may fail to excite more than, say, six layers in our kind of perceptual system. Extraordinary patterns, on the other hand, are defined as those that exceed some arbitrary number of activation levels. 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, and performing a natural bottom-up hierarchical segmentation. If interrogated, such a system would be able to rate the novelty of the current activity at several different levels. This is an important prerequisite to intelligent behaviour.

Pattern intervals are complex multi-level phenomena. Thibadeau (1986) noted the importance of identifying the moments when particular actions start and finish. We attempt to make a system react to these special moments both adaptively, and *without a priori knowledge in any form*, so that future higher functions, which might have specific recognition-based tasks, may receive representations that have already conveniently carved the world at its joints.

#### 6.1.5.4.3 Automatic generation of new higher-level reinforcement functions

In order to qualify as useful, the automatic generation of more sophisticated perceptual machinery, and the enhancements to perception it brings, must be able to modify the general behaviour of the system. We therefore need to find some kind of reinforcement or reward function to help the system evaluate competing states, and decide which it prefers. For such an exercise we must effectively set up goals (actually, establishing goals and establishing a reinforcement function amount to much the same thing). Let our primary goal be to garner information from the world. To do this well we must constantly seek out situations in which information is ‘dense’. Temporally, this means avoiding situations that are similar to the past, and being selective in any particular spatial situation, so that the system focuses on only the most novel inputs.

Assuming that a system were able to seek out novelty, what would be the effect internally? The effect would be that deeper memory layers would be generated, and would have relatively high levels of surprise. If this is a characteristic of the situation we desire, then we need only find a way to use it as a reinforcement function. This could hardly be simpler.

We can easily calculate the overall surprise at each layer, and can weight the deeper layers so that they contribute more reinforcement. The reinforcement is then reduced to a single metric that can be applied to the system's motor behaviour. This provides the curiosity drive.

Using absolute stimulation levels in this role may seem far 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, as well as many other developmental processes outside the nervous system. Indeed, in the case of the neuron it may be that the *only* available reinforcement function is stimulation, as Hebb (1949) suggested.

### 6.1.6 Conclusion

The main contribution of SURPRISE is that it provides an adaptive response that has the following properties:

- It can operate with stochastic spatial arrangements, such as those produced by JIGSAW.

- It can be linked using lateral inhibition to generate smoothed or unsmoothed motion representations.

- It can be linked serially, with dynamic generation and atrophy of layers, to provide different levels of response.

- It can provide edge detection when combined with egomotion.

It is interesting that, if systems using similar principles to those of WRAITH can produce hierarchical and modifiable responses to represent specific kinds and levels of activity in the world, then very little learning, in any conventional sense (i.e., machine learning), 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 in turn take on the learning reinforcement function. In applying the activation level of deeper memory layers to the reinforcement function we effectively give the system a drive for additional stimulus. If systems are to achieve a high degree of autonomy, to the extent that they are able to set their own learning agendas as humans do, then such a drive will be indispensable.

There is good evidence of active novelty-seeking primary behaviour in infants as young as four weeks. Eimas *et al.* (1971) connected a rubber nipple to a tape recorder so that the infants' sucking action made it play. The faster they sucked, the louder it played. The experiment showed that infants were able to associate sucking with change in the variety of sounds they heard. Having established this conditioning, when the tape played a unvaried sequence such as *ba ba ba* the sucking slowed down, but when the sound changed to *pa pa pa*, for instance, infants sucked harder to hear the new sounds. The change (i.e., novelty) acted as a reward and triggered the very response the infants had learnt was responsible for increasing change.

SURPRISE applied recursively as reinforcement would be an effective way of achieving curiosity in artificial systems. It enables us to build a vertically integrated solution to the four questions posed at the start of this chapter (how to direct attention to novel or 'interesting' events in the world, how to separate novel or interesting information from everything else, how to perform this segmentation of the world at multiple levels, and how to develop perceptual machinery capable of this). Whether this particular algorithm is powerful enough to take on bigger problems remains unknown, but the more important point is that the problem of creating self-replicating cognitive structures that are able to derive their own higher-level learning functions has been broached.

SURPRISE contributes to the field of machine vision a context-independent, matrix-independent, level-independent method of measuring interest in image sequences. It is a single, simple function that performs a number of important roles in the processing of data. They are smoothing to reduce the effects of noise, motion detection, edge detection via egomotion, and accommodation to constant change. SURPRISE is versatile in being applicable in both parallel and series, and at different resolution levels.