

# Integrated Techniques for Self-organisation, Sampling, Habituation, and Motion-tracking in Visual Robotics Applications.

Mark W. Peters\* and Arcot Sowmya†  
School of Computer Science and Engineering  
University of New South Wales.

## Abstract

We summarise several techniques in use in our visual robotics research. Our aim is to develop robots that are thoroughly autonomous and adaptable. We describe a system that is independent of typical image derivation standards, sampling algorithms, *a priori* space, object, or motion models, yet is able to intelligently direct its attention to novel activity in a complex and changing environment.

## 1 Introduction

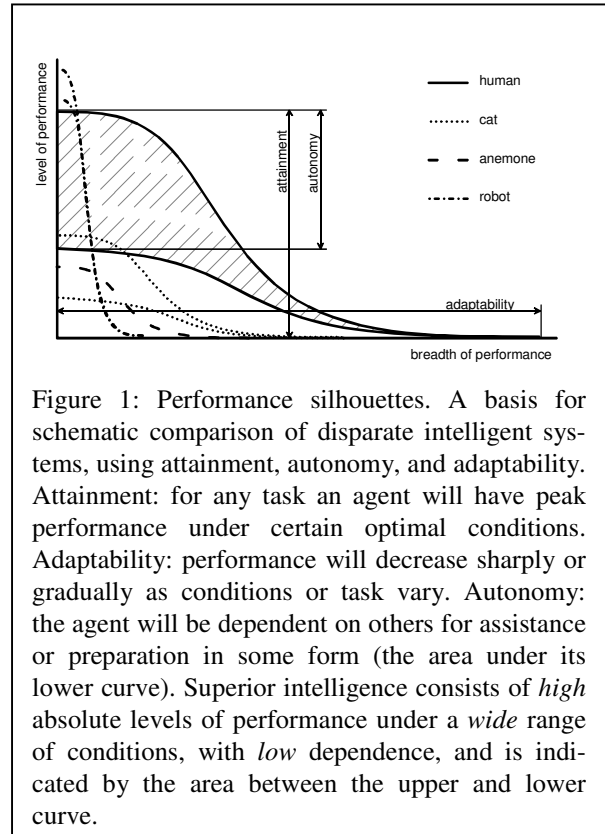
We have developed a visual robotic system known as WRAITH, whose ultimate purpose is to follow (in the sense of both ‘track’ and ‘understand’) human movement, and to do so under the most unfavourable conditions, both external and internal. This is a long-term project, and this paper represents a report on the principles of design and the progress of the system so far. Our main interest lies in the improvement of performance autonomy and robustness. In addressing this challenge we have often tried to emulate the important characteristics of biological systems, but this has not prevented us from adopting other methods when they provide an improvement over what biology has to offer.

## 2 Background

When we compare the general performance of robotic and biological solutions to the problems of dealing with a complex and changing environment we are struck by the superiority of the latter. The biological superiority has three salient features: attainment, autonomy and adaptability.

**Attainment:** animals achieve higher *absolute* levels of performance in many tasks (eg, flying through a forest, communicating with others, avoiding dangers).

**Autonomy:** animals generally learn most things themselves without having to be taught or be given multiple learning examples. Their ultimate attainment is far removed *relative* to what they receive from others, far further than that of robots. Without this autonomy, any intelligence is reproductive rather than productive.

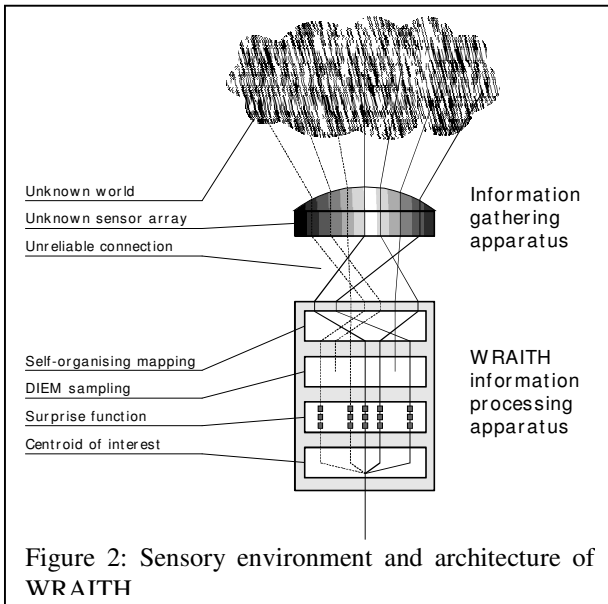


**Adaptability:** animals can thrive under a remarkable range of conditions by changing their behaviour, or physical characteristics.

We can compare general performance in a schematic diagram where any agent’s achievement is represented by the area of its enclosed competences, its performance silhouette (see Figure 1). The silhouette is maximised by having a high upper edge (strong *attainment*), a low lower edge (strong *autonomy*, showing the individual can perform its own primary bootstrapping functions if necessary), and breadth (strong *adaptability*, the faculty of ranging behaviour over a broad repertoire). While complex animals like humans are relatively dependent compared to the anemone, they are highly autonomous compared to robots. While robots may attain extremes of performance, it is generally in limited domains, and with much assistance and careful set-up.

\*Address: Sch. of Computer Science and Engineering, UNSW, Sydney 2052, Australia. E-mail: markpeters@cse.unsw.edu.au

†Address: Sch. of Computer Science and Engineering, UNSW, Sydney 2052, Australia. E-mail: sowmya@cse.unsw.edu.au



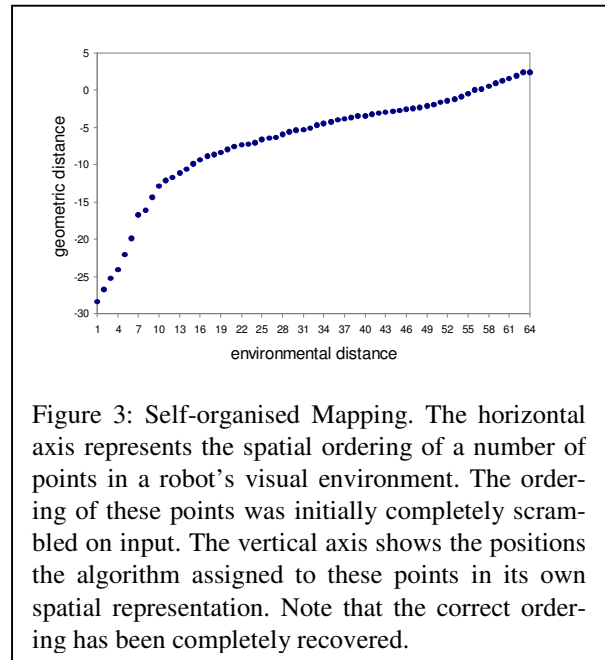
Not only do animals self-organise to an extraordinary degree (for example, starting as a single cell) but they can also make major adaptations to traumatic ‘reconfigurations’ such as the loss of a leg, or an eye, or even the sudden inversion of all their visual inputs [8]. Machines, by contrast, generally do not cope well with changes in configuration. If we intend to manufacture machines with a high degree of autonomy, the ability to perform robustly in the face of change to themselves, then they must adopt analogous adaptive processes.

More specifically, animals independently develop the ability to function and pursue goals in ever-changing spatial environments. Among other faculties, they develop spatial competence via *unsupervised incremental learning*. That is, they have neither learning aids such as reliable *a priori* test data nor supervisors to point out what works and what doesn’t. This basic spatial competence is very important; it is a prerequisite and basis of most intelligent behaviour. Developers are yet to create machines that can autonomously evolve both coherent vision and motor coordination [6]. To do so, we must solve a number of problems in novel ways, and our solutions must be labile enough to revise themselves whenever necessary. These are some of the challenges we are addressing.

### 3 Techniques

WRAITH’s current methods of operation can be broken down into four functions (self-organised mapping, DIEM sampling, the Surprise function, and the Centroid of Interest, see Figure 2), all of which we shall describe in some detail, giving the algorithms.

**Self-Organised Mapping.** Our assumption is that the apparatus that gathers information about the world, and



the apparatus that processes that information and subsequently selects the system’s reaction, may be physically and logically separated. There is therefore significant opportunity for information to be mis-formatted or even scrambled in the transmission from one apparatus to another. Even in robot systems designed as a unit there is ample room for decalibration, wear, broken protocols, etc. Ideally, a robot should retain or regain operational integrity despite all this. In order to qualify our processes of self-organisation and recovery we create the worst case, under which all information gathered by the anterior **monitoring system** is deliberately completely spatially unordered. If the posterior **processing system** can then incrementally and independently re-order its own inputs *without reference to what they represent*, purely on the basis of the information in the signals themselves, then it has the capacity to survive extreme reconfiguration, and will have demonstrated autonomy and adaptability [2].

The algorithm proceeds by iteratively refining the relative assigned positions (in the representation) of inputs (signals) from the outside world. Each input (e.g.,  $C$ ) is geometrically located according to distances (e.g.,  $AC$ ,  $BC$ ) calculated on the basis of differences in behaviour of inputs (e.g.,  $ab$ ,  $ac$ ,  $bc$ ) and a conversion function  $f$ .

$$AC = \frac{AB \cdot f(ac)}{f(ab)}, \quad BC = \frac{AB \cdot f(bc)}{f(ab)}$$

The assumption behind this algorithm is that points lying close together in environmental space create pixel input values that are close together in brightness and behave similarly over time. Over time this causes the scrambled internal representation to correct itself and become a monotonic mapping of the outside world (see

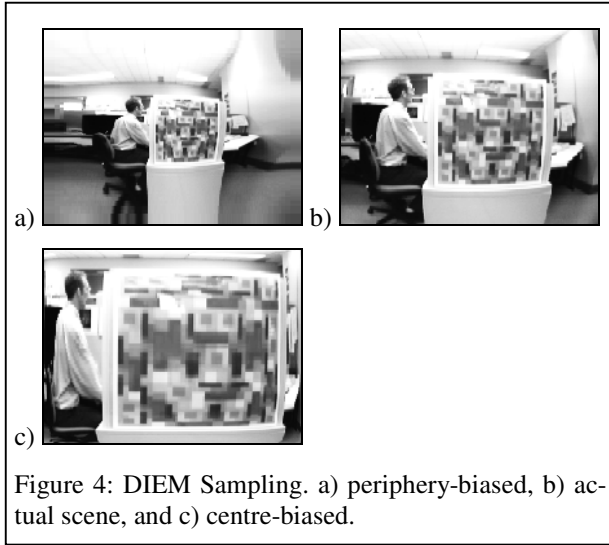


Figure 3).

**DIEM Sampling.** Once such robust and adaptive processes have organised the information available to the system, it then has the opportunity to be selective about which information to monitor. There are times when a system may need to focus attention in a frontal foveal area, when a moving object has already been detected and must now be observed, for example. At other times the system may not be able to afford such focus, but must spread its attention as wide as possible, selectively monitoring the periphery of its visual field. This would be the case when no object had yet been detected, but the system needs to remain alert in case a moving object enters the visual field from an unforeseen direction, or when a very fast moving object is being tracked. DIEM stands for Dimensionally Independent Exponential Mapping, and refers to the fact that sampling densities can be varied independent in x and y dimensions [5] (see Figure 4).

Given an image of width  $W$  (measured in pixels) and height  $H$ , from which we wish to sample  $w$  points in the horizontal dimension and  $h$  points in the vertical dimension, each pair of original image data point co-ordinates  $(x, y)$  is given by:

$$x = \frac{(W-1)}{2} \left( \frac{2s_w}{w-1} \right)^{\gamma_w}, \quad y = \frac{(H-1)}{2} \left( \frac{2s_h}{h-1} \right)^{\gamma_h}$$

where  $s_w$  and  $s_h$  are the indices of the sample (or co-ordinates of the derived sample image). This sampling system is quite distinct from those currently in use elsewhere [7].

**Surprise Function.** The third step in our process controls selectivity of motion events. For example, though an object might appear, disappear, and reappear continuously, after a while it may be unproductive to focus the system's attention on that object. It may be more

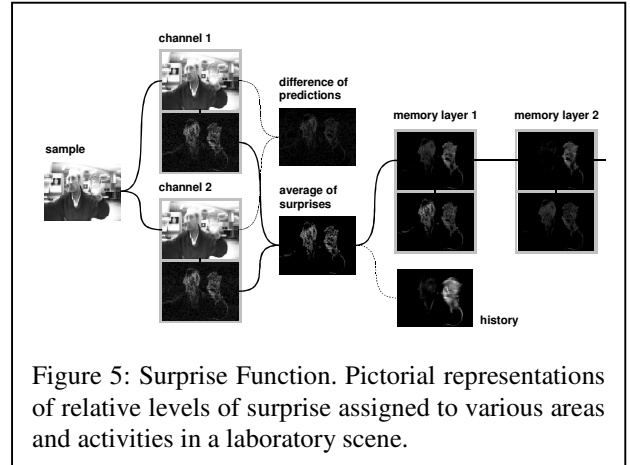


Figure 5: Surprise Function. Pictorial representations of relative levels of surprise assigned to various areas and activities in a laboratory scene.

advantageous to simply note that its motion is repetitive, thus releasing attention to search for other moving objects. And if later the first object suddenly stops moving then we may also want to have attention revert to it for a moment. This faculty prevents the system becoming unduly fixated on relatively uninteresting objects, such as cooling fans, when there are more interesting objects such as people in the vicinity [1, 4, 5].

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

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

Varying  $R$  adjusts the relative weight given to more recent values of  $S$ . If  $1 - R$  is small then the value of  $M$  will be very large compared to  $S$ , due to its rapid accumulative effect. So, as we wish to compare  $S$  to  $M$  to derive a surprise value  $U$ , we first need to renormalise  $M$ . To do this we use another constant,  $P$ , which we derive from  $R$ :

$$P = 1 / (1 - R).$$

$P$  measures the persistence of memory. Now:

$$D = M / P$$

where  $D$  (prediction) is an exponentially decaying moving average of  $S$ . Then:

$$U = |S - D|$$

Thus  $U$  (surprise) corresponds to the 'figural' content set against the 'ground' of  $P$ .

$U$  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  $S$ . It is therefore subject to any noise carried by  $S$  and does not, in itself, provide any temporal smoothing. Fortunately, we already have a form of internal smoothing in  $D$ . Conse-

quently, the first memory units in our visual system are implemented in parallel, and the  $D$  outputs used to produce a difference  $d$ , that is

$$d = |channel1.D - channel2.D|$$

As this value is dependent only on departure from a temporally and spatially local norm, it is directly responsible for very useful behaviour: a tendency to turn towards the unusual, no matter what the context. It does this by assigning an implicit interest measure to all visible directions.

**Centroid of Interest.** Finally, having ranked various areas of motion using a range of interest levels, the system must act upon its information. We now describe how the system incrementally calculates a new direction of gaze. It does this without any object model, with no prior information about the scene or the domain of operation, and with no segmentation of the scene. It derives a single unambiguous direction of gaze from any scene, taking into account all the levels of interest it has developed [3]. The centroid of interest is:

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

where  $n$  is the number of memory units, each of which ( $u$ ), has an  $x$  co-ordinate, a  $y$  co-ordinate, and a difference  $d$ , and where  $m$  is the number of memory layers, each of which ( $l$ ) has a weight  $g$ .

The result of these techniques is a system that can mechanically reposition its camera approximately three times per second, following the motion of people walking around a room, or out in the street. Once a walking person, for example, sits down to type at a keyboard, and becomes relatively still (moving, but in a repetitive way), WRAITH is automatically able to shift attention to something else in the room. Secondly, if part of the room contains a high level of activity, such as areas near doors, it will learn to pay less attention to such an area, but be highly reactive and focussed if movement is detected in a less-used area of the room, such as a corner. This is a result of its ability to autonomously develop a notion of 'normality' in any scene, and to react most strongly when that pattern of normality is broken. It has already been used this way as a security device in our own laboratories, taking photographs of people when they move into areas not normally used at night.

## 4 Discussion

Robots that are not prone to conclude that the world fundamentally changes just because their sensory apparatus does are clearly desirable. Enabling robots to self-

organise their sensory arrays and build their own representations allows them to avoid this obvious error. This characteristic also brings other benefits, as it overcomes a number of data fusion problems, where the integration of multiple cameras, for example, needs to be fused into a single representation. These techniques make no distinction between the source of incoming signals, as they are effectively unaware of any sources anyway. They are all able to work without standard orthogonal image arrays upon which most systems are dependent.

Finally, though it is obvious that actuators cannot be self-calibrated without feedback, it is also obvious that sensory systems do not need to provide feedback, they are sufficient in themselves to organise themselves. They consequently have an independence that actuator systems cannot have, and should therefore be the starting point of any research directed at developing robots with well-developed performance silhouettes.

## 5 References

- [1] M. W. Peters, "Towards artificial forms of intelligence, creativity, and surprise." Proceedings of the 20<sup>th</sup> Meeting of the Cognitive Science Society, pp. 836-841, Madison, Wisconsin, 1998.
- [2] M. W. Peters, "Spatial competence via self-organisation: an intersect of perception and development." Proceedings of the 20<sup>th</sup> Meeting of the Cognitive Science Society, pp. 830-835, Madison, Wisconsin, 1998.
- [3] M. W. Peters & Sowmya, A., "Active vision and adaptive learning." Proceedings of Intelligent Robots and Computer Vision XV, SPIE Volume 2904, pp. 413-424, Boston, Massachusetts, 1996.
- [4] M. W. Peters & Sowmya, A., "WRAITH: ringing the changes in a changing world." Proceedings of the Fourth Conference of the Australasian Cognitive Science Society, Newcastle, Australia, 1997.
- [5] M. W. Peters & Sowmya, A., "A real-time variable sampling technique for active vision: DIEM." Proceedings of the International Conference on Pattern Recognition, pp. 316-321, Brisbane, Australia, 1998.
- [6] P. N. Prokopowicz, "The Development of Perceptual integration Across Eye Movements in Visual Robots." PhD dissertation, Northwestern University, 1994.
- [7] M. Vincze & C. F. R. Weiman, "A general relationship for optimal tracking performance." Proceedings of Intelligent Robots and Computer Vision XV, SPIE Volume 2904, pp. 402-412, Boston, Massachusetts, 1996.
- [8] R. B. Welch "Perceptual modification: adapting to altered sensory environments" Academic Press, New York, 1978.