# Behavioural Categorisation: Behaviour makes up for bad vision

Emmet Spier[1]

[1]Centre for Computational Neuroscience and Robotics
Department of Informatics, Sussex University
emmet@sussex.ac.uk

**Abstract**

The performance of a mobile robot with a vision system is assessed in an everyday object categorisation task. The ability of the robot to arrive at a specified object, *behavioural categorisation*, is compared to the moment to moment results from the computations of its vision system, here called perceptual classification. It is found that the mobile robot using the vision system is significantly more accurate at behavioural categorisation than the underlying performance of the visual system's perceptual categorisation. This result is discussed as supporting the hypothesis that embodied systems using real time algorithms find that 'fast, cheap' visual systems are sufficient for their needs.

## Introduction

Animals use vision, along with the rest of their senses, in the control of their behaviour. The creation of internal states that correlate with the presence of objects in the world - the goal of most of computer vision research, and for which there is enticing neuroscientific data (Logothetis et al. 1995) - is not the purpose of vision in the animal. Rather, animals use these products of their visual processing in order to achieve the tasks they set upon. Thus an assessment of the performance of a particular artificial vision system should be judged in terms of the overall performance of the complete system incorporating a vision capacity. Certainly for many industrial classification tasks working with controlled positioning and lighting the recognition ability of the visual system will be the main arbiter of performance. However, as Ballard (1991) observed under the label of *active vision*, a visual system need not rely on computational algorithms alone in order to function; moving the camera can make the difficult easy (e.g. Borotschnig et al., 2000).

When a vision system is integrated as part of the control system of a mobile robot then the internal states of the system are irrelevant to how we judge the performance of the robot. Harvey, Husbands and Cliff (1994) used artificial evolution to discover a network controller for a visually guided robot that can behaviourally distinguish between a triangle and a square. The controller needed to use only two pixels from its visual field to discriminate between a triangle and a square. The discrimination ability of the two-pixel-vision robot was derived from its ability to move, and the use of this self-controlled visual information in concert with its control system to aid its decision. The performance of the visual system is inseparable from the robot's behaviour. Brooks (1991) called such a robot *situated* and Clark (1997) describes the research program as 'embodied cognitive science' to emphasise that animals and robots possess a body whose controlled behaviour will permit their (artificial) brains to solve the 'agent's' problems in a more interactive manner, often reducing the task's complexity and therefore requiring a 'cheaper brain'. More recent embodied mobile robot experiments have also provided further evidence that control architectures combining 'minimal pixel' vision and behaviour can achieve proportionately impressive results (e.g. Scheier, Pfeifer and Kunyioshi, 1988; Marocco and Floreano, 2002). Mobile robots exploiting high resolution sensors in the control of their behaviour have followed two possible strategies, either to make strong assumptions about the visual scene (e.g. the robot Polly, Horswill, 1993, recognised people if they waggled their foot), or to mark the objects of interest in the scene with easy to detect features (e.g. the robot Nomad, Krichmar and Edelman, 2002, uses objects marked with horizontal and vertical lines and an edge detecting filter for discrimination).

This conference paper provides a preliminary investigation of the effect of integrating a mobile robot with a visual system that uses a high resolution image and possesses the acuity and capacity to discriminate between arbitrary everyday objects. We believe that this is the first behaviour-based robot with this capacity. In the next section the perceptual categorisation model incorporated into the robot's control system is described followed by a description of the robot and its experimental environment. The results of an experiment that compares the performance of the visual system alone (*perceptual categorisation*) with the performance of the robot incorporating the visual system (*behavioural categorisation*) are provided and discussed.

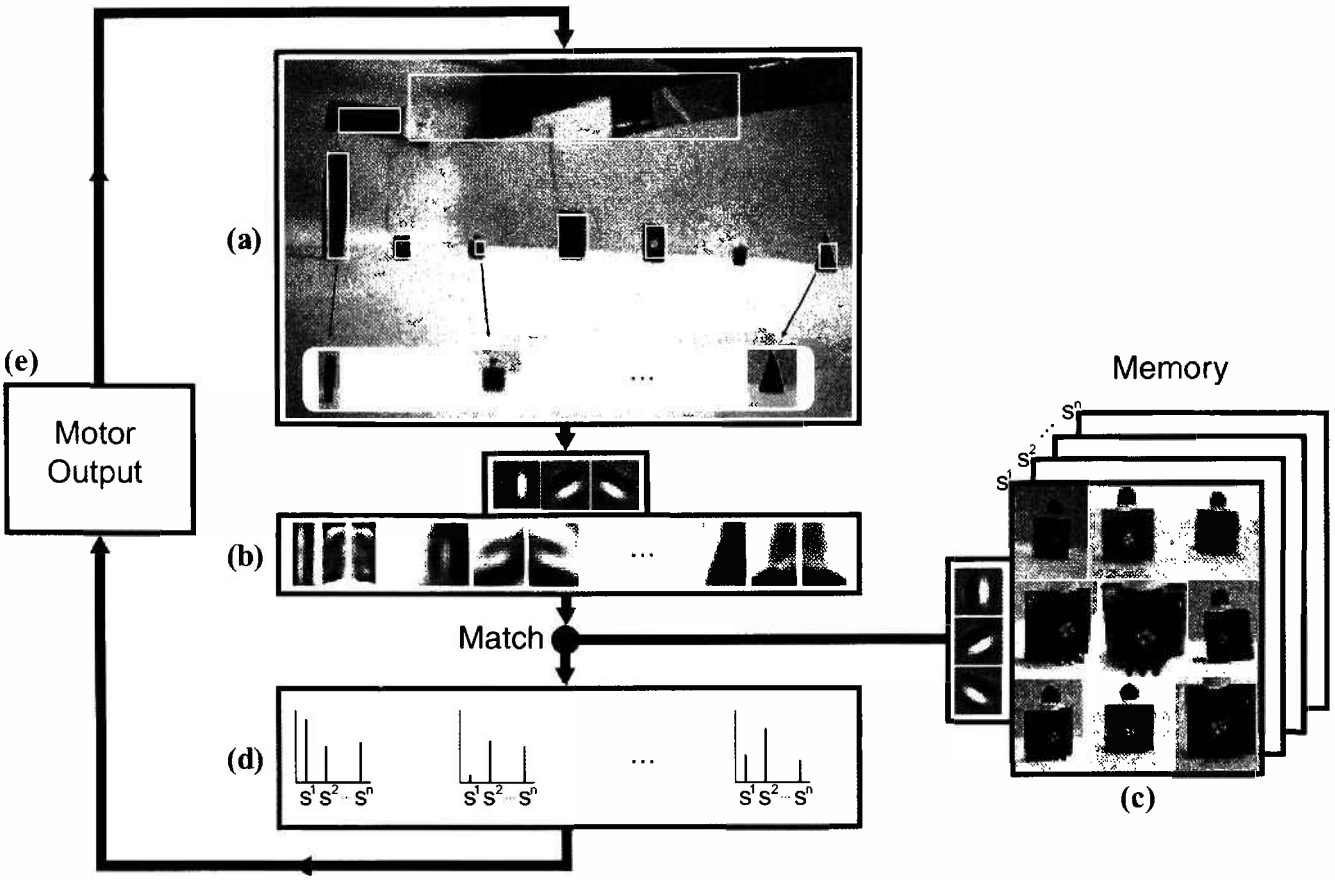


Figure 1: Implemented system diagram. (a) A $760 \times 570$ image is segmented into rectangular sections $I_k$ scaled to a maximum dimension of 40 pixels, (b) the $I_k$ are convolved with size $21 \times 21$ Gabor filters at angles of $\pi$, $\frac{\pi}{3}$ and $\frac{2\pi}{3}$, (c) A memory of stored classes $S^i$ of snapshots (the complete set used for the light sensor class shown here) are compared with the results of stage (b) providing a set of measures of similarity (d) between the blobs $I_k$ and the classes $S^i$. (e) Motor commands are sent to the robot according to whether the closest $I_k$ to target $S^*$ is to the left or right of the centre of the $760 \times 570$ image.

## The Perceptual Categorisation Model

An image processing method for a mobile robot would experience the same three dimensional object at various degrees of rotation, perspective, illumination, scale and translation. A traditional (though, in general, neither solved nor automatic) approach to this problem is to attempt to build a 'CAD' style three dimensional model of the input image (after Marr, 1982). For over a decade an increasingly productive alternative approach has been to calculate similarity measures of the input against a class of multiple snapshot-like views. This method was first described using radial basis functions with vertex data input (Poggio and Edelman, 1990) and later using receptive field input (Edelman and Duvdevani-Bar, 1997), though other methods of image approximation can also be used, for instance eigenvectors (Murase and Nayar, 1995). Such 'view' methods provide invariance to rotation and perspective, and a simple version exploiting receptive fields that offer a degree of illumination invariance is described below. Further, invariance to scale and translation can be achieved through enabling the visual system with the capacity for (virtual) saccades towards every identified visual blob (group of pixels). A crude method of blob segmentation (in general a challenging and completely unsolved issue) was used that thresholded the pixel values of an image scene possessing a highly contrasting background.

Figure 1 is a schematic of the complete classification system. An input image is first segmented into rectangular blobs (a) using both (i) the result of thresholding the distance between each input image pixel and the input image's average colour, and (ii) the result of an adaptive thresholding of the distance between each input pixel and that pixel's local average colour. The first method works well with uniform illumination (a reasonable approximation when close to objects, or within a carefully lit room) and the second provides a good level of robustness to varying light conditions across the input image when objects are of a similar size (the

threshold adjusted to maintain a minimum and maximum blob pixel size). Monochrome rectangular cut outs from the input image are made containing each identified contiguous pixel area in the threshold images (the virtual saccades, figure 1 has identified nine) then scaled to a common size. A resulting cut out image ($I_k$) offers the visual system a reasonable invariance to object translation (from the saccade) and changes of size (with the loss of the ability to discern large from small) within the input image.

The activity of the visual system's low level receptive fields to each $I_k$ can now be computed (b). The products of the convolution of an input image with a collection of Gabor functions of different phases is seen as good first approximation to the receptive fields found in the primary visual cortex (V1), sensitivity to local edges of various orientations (Marcelja, 1980). A Gabor function convolution also effectively removes the input image's non-zero bias (mean illumination) from the resultant receptive field activity. Figure 1b shows the results of each $I_k$ under three convolutions using Gabor functions with rotations of 0, $\frac{\pi}{3}$ and $\frac{2\pi}{3}$, for formal purposes we describe the collection of convolutions of $I_k$ as a real valued vector $g(I_k)$. Howell and Buxton (1998) found this to be a good representation for image matching and unlike eigenspace methods (Murase and Nayar, 1995) it does not require the recalculation of the eigenspace when new objects are learnt.

Each object class $S^i$ is a collection of stored $I_k$ snapshots $S^i{}_j$, figure 1c. The selection of snapshots within any $S^i$ is under human control, their automatic recruitment needing to be addressed separately from here. The similarity measure between an $I_k$ and $S^i$ is a distance metric defined as

$$d^i(I_k) = \text{argmin}_j[\| g(I_k) - g(S^i{}_j) \|^2 \alpha(I_k, S^i{}_j)]$$

where $\alpha(I_k, S^i{}_j)$ is a scale factor which is equals 1 when the aspect ratios of $I_k$ and $S^i{}_j$ are the same and increases in proportion to their difference. (An alternative method (Edelman and Duvdevani-Bar, 1997) is to use a radial basis function with centres at each $g(S^i{}_j)$ but here we are only interested in the best match.) Figure 1d shows a sketch of the $d^i$ calculated for each of the $I_k$ identified in (a). These $d^i$ constitute the instantaneous categorisation assessments of the visual system; the smallest $d^i$ for each $I_k$ identifying the best fit, the *perceptual categorisation* of that area of the visual space. Using the best fit is equivalent to mutual inhibition between the $S^i$ class responses and controls for the possibility that within any particular visual scene some $I_k$ generally responds with smaller $d^i$ values than other $I_k$.

## Robot Design and Experimental Apparatus

Figure 2 shows the robot used for the reported experiment. The top section comprises a 24MHz 68332 microcontroller mounted in a Lego compatible box. Attached to this box, from left to right, was a wireless video transmitter, a wireless serial transceiver (Abacom Technologies), a 40 character liquid crystal display and a fixed focus CCD video camera with a 62° field of view. The bottom section comprised of two Lego motor, motor controller and gear-chain sets and a front whisker touch sensor responsive to a pressure of 1$\mu$N. The Ni-MH rechargeable cell (seen inserted below the serial transceiver) contained sufficient energy to power all attached devices for one hour.

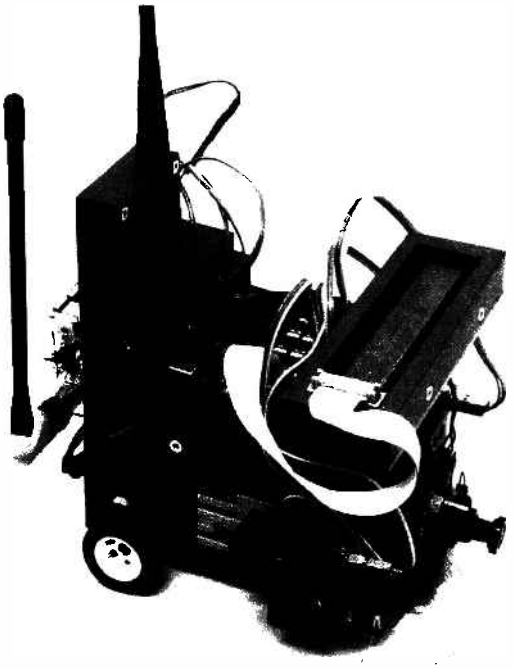

Figure 2: The robot used, 16cm×12cm×15cm excluding aerials, weighing 800g and supported by two independently driven wheels and a coaster attached front centre. The maximum speed used for this experiment was 2.5cms$^{-1}$.

The robot microcontroller handled the motor control and touch sensor monitoring locally with all other computation carried out on a remote workstation (Pentium 4, 1.7GHz) connected via two radio linked channels. The video channel comprised a standard video surveillance broadcast device receiving its signal from the robot's video camera and broadcasting to a remote monitor unit which provided a composite video signal output that was connected to a video capture card (FlashBus MV Lite, Integral Technologies) installed in the workstation. The capture card was programmed to transfer both a monochrome and colour version of the video signal by direct memory access to the workstation's memory at 24 frames per second without significant processor interaction. The robot command channel ran wirelessly over an error corrected 38400 baud serial connection exchanging sensor data from the robot and commands from the workstation 30 times per second. With nine classes comprising a total of 66 snapshots and a captured frame containing ten segmented

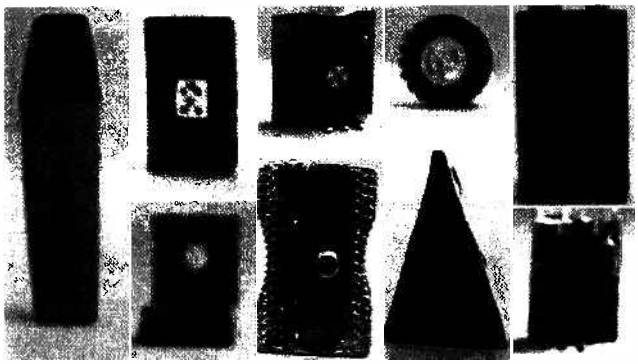

Figure 3: The objects, clockwise from left: blue marker pen, rubber, light sensor, wheel, battery, IR sensor, triangle, pencil sharpener, Lego right angle.

blobs the perceptual model implemented on the workstation would run at around 14Hz.

The experimental arena comprised a floor area of 90cm×120cm covered with matt yellow paper rising to a height of 20cm at the edges. To the right of the arena was an external window providing strong side illumination during the day (top centre in Figure 1a) and otherwise lit from a normal office tubular fluorescent light fixed to the ceiling. Figure 3 shows the nine objects used for the experiment, the criteria for selection of these objects was that they should have a width between 1cm and 4cm and neither be yellow nor possess a monochrome intensity close to the yellow arena covering's own intensity.

## Experiment and Results

The performance of the robot and its perceptual model was assessed in an object categorisation task. The robot control system as specified in figure 1 would move forward, left or right depending on the whether the blob $I_k$ best matching target class $S^*$, that moment's particular perceptual categorisation, was in the centre, left or right of the input image. If there was no blob whose best match was class $S^*$ for ten perceptual cycles then alternating and increasing ballistic turns to the left or right would be made as part of a sweeping search strategy.

The criteria for the robot's *behavioural categorisation* of an object was defined to be the single object in the robot's field of view when the robot was sufficiently close to the object that only it could be seen in the image.

Each trial an object was selected as the target class $S^*$ and specified to the control system. The control system was pre-loaded with the nine $S^i$ for the objects shown in figure 3, the particular $S^i_j$ chosen manually during training trials as a sparse sampling that provided good performance. The nine objects were arranged 9cm apart in a randomised order tracing an arc of a circle with radius 50cm. The robot was placed 50cm from the objects and arranged so that the centre

| Object | Trials | Mean Accuracy % | Arrival % |
|---|---|---|---|
| blue marker pen | 11 | 67.92 | 90.91 |
| IR sensor | 15 | 59.65 | 93.33 |
| Lego right angle | 11 | 89.29 | 100.00 |
| light sensor | 12 | 72.28 | 91.67 |
| pencil sharpener | 13 | 63.90 | 84.62 |
| rubber | 10 | 95.40 | 100.00 |
| triangle | 12 | 72.14 | 100.00 |
| wheel | 7 | 77.86 | 100.00 |
| battery | 7 | 85.16 | 100.00 |

Table 1: Summary data of the robot's performance at the perceptual categorisation (mean accuracy %) and behavioural categorisation (arrival %) of each object.

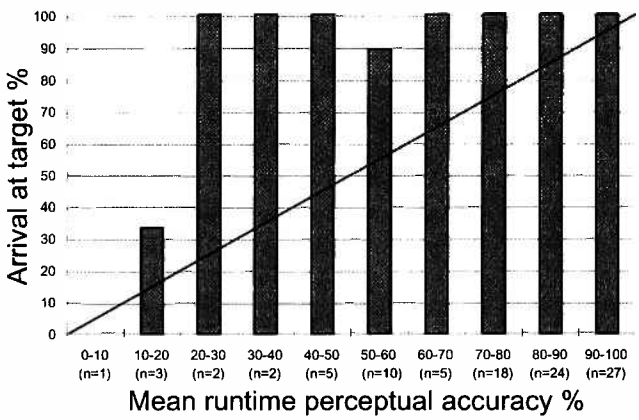


Figure 4: A histogram showing the behavioural categorisation accuracy of the robot for varying perceptual categorisation accuracy (grouped into 10% bins, lower bound inclusive, with the number of trials noted). The continuous line marks the situation where behavioural categorisation performance would be equal to perceptual categorisation performance. Data above the line indicates that behavioural categorisation is more accurate than perceptual categorisation.

of its view field was to the left or right of the target object. The trial was then initiated and terminated when the robot had made a behavioural categorisation.

A complete log of the perceptual system's input and computations during the trial was stored. This permitted, during post trial analysis, a judgement to be made concerning the accuracy of each cycle of perceptual categorisation made by the control system. Every cycle's frame was tagged with a label 'correct' or 'incorrect' according to a human judgment that either the system correctly identified $S^i$, or the system identified an object judged not be to $S^i$ or did not identify any object as $S^i$.

Table 1 shows a summary of the trials carried out to measure the robot's ability to recognise each of the objects tested (over 3000 perceptual cycles). Here we note that while

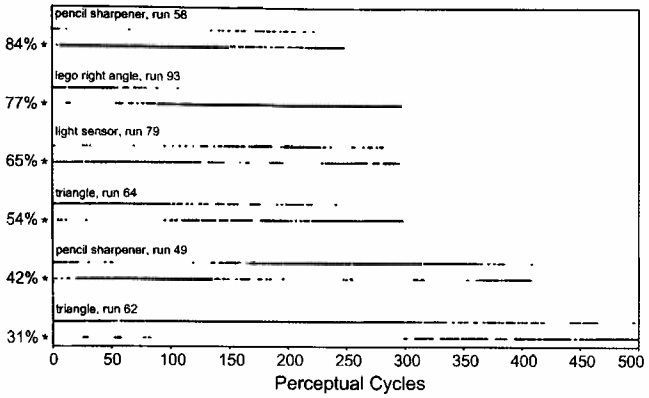


Figure 5: The perceptual categorisation accuracy during specific trials. The row marked with a '$*$' records cycles when the perceptual categorisation was 'correct' and its sibling records the 'incorrect', the specific percentage accuracy recorded on the left of the figure. The trials were randomly selected from their 10% bins.

the mean perceptual categorisation accuracy (mean accuracy %) for each object is mediocre the behavioural categorisation accuracy (arrival %) is much improved. A better understanding of the performance of the system can be gained from consideration of figure 4. This histogram considers the behavioural categorisation accuracy of the robot for sets of trials grouped together with a similar perceptual accuracy. When a particular histogram bar is higher than the diagonal line shown in the figure then the behavioural categorisation accuracy is better than the perceptual categorisation accuracy; that is, the robot arrives at the target object S* more reliably than would be expected from consideration only of the performance of the perceptual model.

Figure 4 indicates that for trials with overall perceptual categorisation accuracy of 60% (rather average although comparable to state of the art systems using an equivalent number of views per class, Murase and Nayar, 1995) the robot successfully achieves 100% accurate behavioural categorisation. As the data collected is categorical it is only possible to use non-parametric statistical tests. A Mann-Whitney U test was carried out within each bin to test the hypothesis that the robot arrival success was different from the perceptual accuracy. For bins having a mean perceptual accuracy greater 20% the arrival success was significantly different ($p < 0.001$). It should be noted that the number of trials in the bins recording perceptual categorisation accuracy below 40% are too few to make firm statements however the trend of significantly better behavioural categorisation than perceptual categorisation is continued.

## Discussion

What was it that the embodied robot system provided that enabled it to so significantly improve upon the performance of the perceptual categorisation system? Figure 5 discounts some of the possible trivial explanations, that the segmentation or classification aspects of the perceptual classification system are distance sensitive. If the segmentation system only worked within a particular distance range (nearby for instance where further away the 'objects' would not be identified in the visual scene to be categorised) then it would be expected that during every trial the perceptual categorisation accuracy would have a fixed period of error, and therefore the mean value of perceptual accuracy would be reduced. A similar argument can be made for the perceptual classification system being accurate only within a certain range. However, as can be seen from the 84% run there does not appear to be a consistent period of error, and from consideration of the other runs this judgement appears to remain, each run (of a selected successively lower performance) having its own characteristic periods of error.

One possible explanation, relying on the embodiment of the robot, can be based upon the assumption that when the perceptual system makes an error, this error can be broken down into two types. An error could be systematic or random. A systematic error would occur when the perceptual system incorrectly categorised an object outside the class S* as being part of that class because the object's $I_k$ matched a particular view within S*. A random error would occur when through noise in the segmentation algorithm, blurring through movement of the robot, periods when the actual $I_k$ of the target object falls in between two views in S*, and other sources meant that the target object was not matched correctly by the perceptual system. If a significant proportion of the errors were random errors (and this needs to the tested by a future experiment), these errors would be expected to be distributed uniformly to the left and the right of the robot and over a course of a trial they would cancel out. The consequence of this cancelling out of random errors would be that the *effective* perceptual accuracy, as a result of being part of a control system performing behavioural categorisation, would be higher than the recorded perceptual accuracy. A further prediction can be made under this assumption, it would be expected that the behavioural recognition accuracy of the robot would decrease if, because of the arrangement of objects, the proportion of systematic errors increased.

## Conclusion

It is functionally advantageous for an agent to have as accurate as possible classification ability towards relevant objects in its environment. However, an agent's actual ability, as with ever other, constitutes a trade-off between the competing demands of other systems and constraints on the agent's resources be they, amongst others, computational power, heat dissipation, glucose or tissue capacitance. This work provides a demonstration that an agent through its behaviour can reduce the demand it makes upon one part of its nervous system while possibly increasing the demand on an-

other. This flexibility offers a space in which brain or control system design can be optimised for its resource use, thus permitting 'cheaper' solutions than that which would be available if only a disembodied system was used.

The experiment presented provides a demonstration of how agent's which behave in real time do not necessarily require the products of sensory information processing to be highly accuracy. Any particular momentary error the robot makes can be corrected in the next and successive perceptual cycles. Spier and McFarland (1997) showed that sequences of simple reflex decisions can offer better performance than a more complex static calculation. Here we see a relatively computationally cheap vision system (six vector distance calculations per object per image and naturally parallelisable) can provide near perfect performance when situated by a mobile robot.

## Acknowledgements

The author would like to thank Bill Bigge for his design and technical support of the robot hardware, Ian Macinnes for the robot development environment, Edgar Bermudez Contreras for help with the post trial analysis, the anonymous reviewers for their comments, and appreciates conversations about this work with John Anderson, Chrisantha Fernando, John Howell and Andrew Philippides. This work has been supported in part by a grant from the Nuffield Foundation (NAL/00669/G).

## References

Ballard, D. 1991. Animate Vision. Artificial Intelligence 48:57–86.

Borotschnig, H., Paletta, L., Prantl, M. and Pinz, A. 2000. Appearance-based active object recognition. Image and Vision Computing 18:715–727.

Brooks, R. A. 1991. Intelligence Without Reason. In Proceedings of the 12th international conference on artifical intelligence (IJCAI–91), 569–595.

Clark 1997. Being There. Putting brain, body and world together again. MIT Press.

Edelman, S. and Duvdevani–Bar, S. 1997. A model of visual recognition and categorization. Phil. Trans. R. Soc. Lond. B 352:1191–1202.

Harvey, I., Husbands, P. and Cliff, D. 1994. Seeing the Light: Artificial Evolution, Real Vision. In Proceedings of the Third Conference on the Simulation of Adaptive Behaviour., 392–401. MIT Press.

Horswill, I. 1993. Polly: A Vision-Based Artificial Agent. In Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93).

Howell, A. J. and Buxton, H. 1998. Learning Identity with Radial Basis function Networks. Neurocomputing 20:15–34.

Krichmar, J. L. and Edelman, G. M. 2002. Machine Psychology: Autonomous Behavior, Perceptual Categorization and Conditioning in a Brain-Based Device. Cerebral Cortex 12:818-830.

Logothetis, N.–K., Pauls, J. and Poggio, T. 1995. Shape recognition in the inferior temporal cortex of monkeys. Current Biology 5:552–563.

Marcelja, S. 1980. Mathematical description of the responses of simple cortical cells. Journal of the Optical Society of America 70:1297–1300.

Marocco, D. and Floreano, D. 2002. Active Vision and Feature Selection in Evolutionary Behavioural Systems. In Proceedings of the Seventh Conference on the Simulation of Adaptive Behaviour (SAB2002), Edinburgh, UK. MIT Press.

Marr, D. 1982. Vision. W.H. Freeman and Company, New York.

Murase, H. and Nayar, S. K. 1995. Visual Larning and Recogntion of 3–D Objects from Appearance. International Journal of Computer Vision 14:5–24.

Poggio, T. and Edelman, S. 1990. A network that learns to recognize three-dimensional objects. Nature 343:263–266.

Scheier, C., Pfeifer, R. and Kunyioshi, Y. 1998. Embedded Neural Networks: Exploiting Constraints. Neural Networks 11:1551–1569.

Spier, E. and McFarland, D. 1997. Possibly Optimal Decision Making under Self-sufficiency and Autonomy. Journal of theoretical biology 189:317–331.