Seeing Things

R Wilson, H Knutsson†,
University of Warwick,
Computer Science Department,
Coventry CV4 7AL,
U.K.
†University of Linköping,
Computer Vision Laboratory,
581 83 Linköping,
Sweden.
e-mail:rgw@dcs.warwick.ac.uk;†knutte@isy.liu.se

November 10, 1994

Index Terms

Active Vision, Invariance, Vision Theory.

Abstract

This paper is concerned with the problem of attaching meaningful symbols to aspects of the visible environment in machine and biological vision. It begins with a review of some of the arguments commonly used to support either the 'symbolic' or the 'behaviourist' approach to vision. Having explored these avenues without arriving at a satisfactory conclusion, we then present a novel argument, which starts from the question: given a functional description of a vision system, when could it be said to support a symbolic interpretation ? We argue that to attach symbols to a system, its behaviour must exhibit certain well defined regularities in its response to its visual input and these are best described in terms of *invariance* and *equivariance* to transformations which act in the world and induce corresponding changes of the vision system state. This approach is illustrated with a brief exploration of the problem of identifying and acquiring visual representations having these symmetry properties, which also highlights the advantages of using an 'active' model of vision.

1 Signals or Symbols ?

This paper is an attempt to deal with some of the issues in the continuing debate between adherents of the 'traditional' AI approach to vision (eg. [1, 2]) and those advocating a 'behaviourist', stimulus-response model [3] . While the debate is hardly new and is arguably part of the wider one between these approaches to AI in general (eg [4, 5]), it has been drawn into sharper focus in recent years by the upsurge of interest in neural networks in general and in particular by the *Active Vision* model of computer vision [6, 7], which was the central theme of the workshop on Computer Vision (Ruzenagaard, Denmark, 1992), in which we gave the first presentation of some of these ideas.

While there are several ways to characterise the dichotomy between the two approaches, it seems to us that the essential issue can be phrased in the question "Does an Active Vision system (AVS) need symbols ?", with the implied alternative that all that *needs* to be done in vision can be described adequately in terms of signal processing (we take it for granted that all will agree that any vision system must do *some* signal processing). The emphasis on the word 'active' is important and requires elucidation: in their definitions of Active Vision, both Bajcsy [6] and Aloimonos [7] stress what might be called *action for vision*, in other words, the vision system has at its disposal a set of actions by which it can acquire additional information. Typically this would involve at least motion of the image sensor (see also [8]), controlled by feedback from existing measurements, with the aim of gathering sufficient data for solution of a specific visual problem or problems. It also happens, however, that many applications of such vision systems involve *vision for action*: the vision system is part of the control processing for a robot or other mobile device, whose goal is not 'seeing', but 'acting' [3]. In this case, issues of real-time operation and other physical constraints are more significant. For our purposes, the significant feature in both cases is that the primary goal of the vision system is to produce an appropriate 'motor response' to its visual input, rather than to utter 'I see a cup' any time a picture of a cup is thrust before it - we cannot expect to get very far in our discussion of the need for symbols if we start by assuming that the output of the system has to be symbolic.

In this paper, we discuss those aspects of visual behaviour and visual coding which are relevant to the 'symbol question'. We then attempt to define symbols in a functional way, which depends on the relationship between internal and external states of a vision system. These ideas are illustrated using some simple model problems, in which the acquisition of visual representations using standard neural network techniques is the vehicle for an exploration of the main ideas in the paper. We have included as many examples as we can from current and past practice in the field, as well as from the relevant literature on biological vision, which continues to serve as an 'existence proof' that vision is at all possible. As a further tutorial aid, we introduce a rather trivial AVS - one whose level of sophistication is comparable, in biological terms, to the lowliest invertebrate. We hope that this will not only help to focus our discussions, but may also be useful as a tool for teaching and even exploring some of the other issues in Active Vision.

2 The Function of Vision

2.1 Vision as Inference

Marr's book begins with the 'plain man's' definition of "to see" as "to know what is where by looking" [1]. In fact, this might be taken as a definition of Active Vision, since looking is an active process. In any event, the definition carries with it symbolic baggage: however the information obtained visually is represented internally, it clearly is not the original source of the data; if I see a cup, I do not have a cup stored in my visual cortex, I have a symbol which represents a cup. Vision is therefore a process of inference, in which hypotheses about the state of the world are confirmed or rejected on the basis of evidence acquired visually. This is as true for machine vision as for biological vision (eg. [1, 2, 9, 10, 11, 12]). Thus in a recent survey article [13], Rosenfeld elaborates the Marr paradigm in a way which brings out the inferential nature of the vision process (a simplified version is given in Fig. 1) and makes the role of symbols clear. The unfortunate fact that the data upon which the inference is based are ambiguous in most cases of practical interest and, worse still, that the hyopthesis set is potentially unbounded, should not be allowed to obscure the fundamental point that any system which can be described as 'seeing' must use a symbolic description, in which a given symbol represents not just a subset of the input visual array, but more significantly corresponds to a subset of the input signal space. If we are working with single 2-D images of 256 $\times$ 256 pixels, then the visible object 'cup' will correspond to some well-defined subset V_c of the signal space $R^{256} \times R^{256}$

$$V_c \in R^{256} \times R^{256} \tag{1}$$

The main problem in vision is that no one has a very clear idea of what that subset is, but that is not a matter of principle. In principle, vision is an application area for statistical inference techniques [14, 6, 2]. Even when this is not explicitly stated, it is often implicit that *a priori* knowledge is being used to constrain solutions to problems, with the idea that some solutions are more likely than others (eg. [15, 16]).

There are certain practical difficulties in the way of applying the techniques of inference directly to the image data: $R^{256} \times R^{256}$ is a rather big place in which to look for subsets. This explains the relatively complex structure of the vision paradigm , compared with the conventional pattern recognition system components of feature extraction, followed by classification. Of course, even this hierarchical

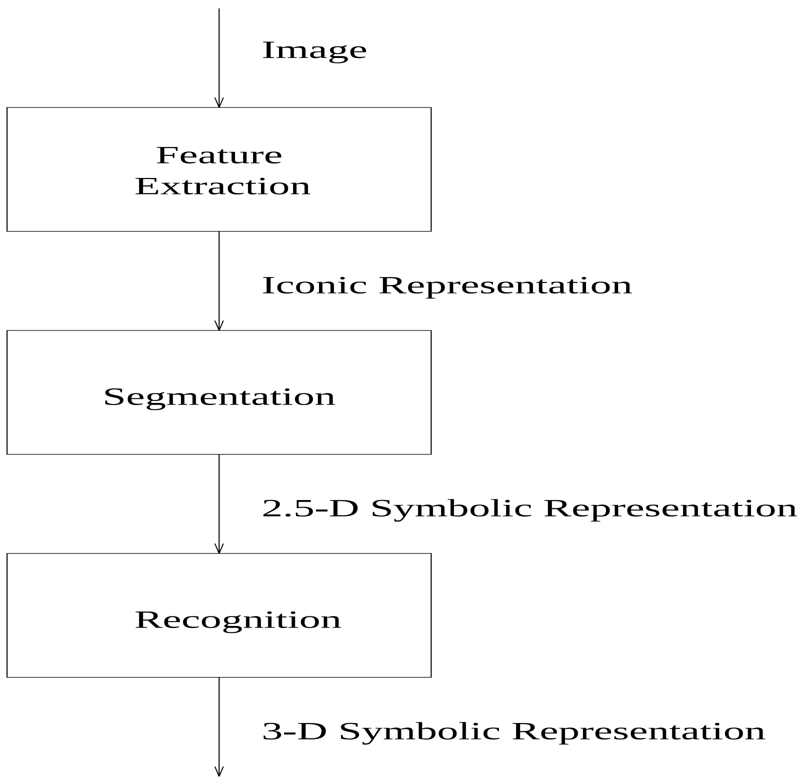

Figure 1: Structure of Symbolic Vision Paradigm.

approach has its limitations, particularly in its failure to address the control issues which are considered by many to be fundamental in vision [6, 17]. In this respect, one of the main attractions of Active Vision is that by the controlled gathering of additional information, many of the inference tasks are simplified [7]. Nonetheless, the combination of feature extraction and hypothesis testing remains the basic paradigm for vision, whether the processing is bottom-up, top-down or a combination of the two. This is as true today as it was 25 or so years ago, when the field was in its infancy. In this regard, it is interesting to note that there is clear evidence in biological systems of feedback paths from higher to lower visual centres [18].

Just as hypothesis testing strategies have become more subtle, so have the algorithms of feature extraction. From the work of Roberts [19] on luminance edge detection in 1965, this has evolved to the sophisticated techniques used today, such as regularization [15], Hough transforms [20, 21], multiresolution analysis [22, 23, 24], tensors [25] and tangent bundles [16], aimed at providing ever more effective ways of detecting edges and grouping them into potential boundary contours. Of course, in many systems, there is information available from shading, texture, motion and stereopsis (eg. [7]), but this is essentially a matter of degree, not of principle.

To illustrate the idea in an Active Vision context, consider the following simple simulation. The system consists of a 'cockroach', named Blatta Borealis (BB), in honour of its discovery in northern climes, which is able to explore the 3×3 window of its 2-D world centred on its current location. Its action is to seek the darkest place in the image in which to hide. In other words, it moves at each step in a direction which decreases intensity. The diagram below shows the result of a single simulation run, which terminates in this case in the absolute minimum of the luminance. The white crosses indicate the position of the cockroach at each step and the line shows its trajectory.

The symbol set used in this example consists of the 4 compass directions and a null symbol $\Sigma = \{N, S, E, W, \phi\}$, the 4 directions corresponding in the obvious way to the 4-neighbours of the current location $\vec{x} = (x, y)$. At each step, one symbol is selected on the basis of the luminance gradient $\vec{g}(\vec{x})$, which is estimated using a pair of gradient operators. If $u(\vec{x})$ is the luminance at $\vec{x}$, the gradient is given by

$$\vec{g}(\vec{x}) = \sum_{i=1}^{4} w_i \vec{\rho}_i (u(\vec{x} + \vec{\rho}_i) - u(\vec{x} - \vec{\rho}_i)) / \sum_j w_j \tag{2}$$

where the 4 displacement vectors $\vec{\rho}_i$ are given by

$$\vec{\rho}_1 = (1, 0) \quad \vec{\rho}_2 = (1, 1) \quad \vec{\rho}_3 = (0, 1) \quad \vec{\rho}_4 = (-1, 1) \tag{3}$$

and the weights are

$$w_i = \sqrt{2} / \|\vec{\rho}_i\|^2 \tag{4}$$

The mapping to symbols is simply

$$s(\vec{x}) = \begin{cases} N & \text{if } \|g_y(\vec{x})\| \geq \|g_x(\vec{x})\| \text{ and } g_y(\vec{x}) < 0 \\ S & \text{if } \|g_y(\vec{x})\| \geq \|g_x(\vec{x})\| \text{ and } g_y(\vec{x}) > 0 \\ E & \text{if } \|g_y(\vec{x})\| < \|g_x(\vec{x})\| \text{ and } g_x(\vec{x}) < 0 \\ W & \text{if } \|g_y(\vec{x})\| < \|g_x(\vec{x})\| \text{ and } g_x(\vec{x}) > 0 \\ \phi & \text{else} \end{cases} \tag{5}$$

The action taken by the cockroach is obvious: it moves to the pixel specified by the direction and if the direction is null, it remains where it is. It is not hard to see that, from an arbitrary starting point, the cockroach will eventually settle at the luminance minimum in whose basin of attraction that starting point lies. In the example shown, there is a single minimum, which is reached in some 120 steps from the given starting point. Why use a symbolic approach in such a simple problem ? There are three main reasons:

1. The set of symbols maps directly and $1-1$ onto an appropriate set of actions.
2. The mapping from luminance to symbols is chosen to ignore irrelevant features of the luminance: the symbols are *invariant* to both the mean luminance and to the steepness of the luminance gradient.
3. On the other hand, the symbols respect the natural symmetries of the problem. In the example of Fig. 2, the image array is 128×128 pixels, giving a group of symmetries generated by the 8 combinations of rotation by multiples of $\pi/2$ and reflection about the $y - axis$ and the 128 cyclic shifts in the horizontal and vertical directions. This large group of linear transformations $\boldsymbol{T} \in \mathcal{G}$ act on the image $\boldsymbol{u}(\vec{x})$ by changing the coordinates

$$\boldsymbol{T}\boldsymbol{u}(\vec{x}) \doteq \boldsymbol{u}(\boldsymbol{T}^{-1}\vec{x}) \quad \boldsymbol{T} \in \mathcal{G} \tag{6}$$

The symbols are transformed in the same way

$$\boldsymbol{T}s(\vec{x}) = s(\boldsymbol{T}^{-1}\vec{x}) \quad \boldsymbol{T} \in \mathcal{G} \tag{7}$$

2.2 Vision as a Mapping

It may be *possible* to treat vision as a symbolic process, but is it *necessary* or even useful to do so ? A behaviourist approach to vision is illustrated in Fig. 3 , in which vision is a function or mapping from the visual input to the response. If the image array is regarded as a vector $\boldsymbol{u} \in R^{256} \times R^{256}$ and the response is a vector $\boldsymbol{r} \in R^M$ then vision is defined as the function $\boldsymbol{f}$

$$\boldsymbol{r} = \boldsymbol{f}(\boldsymbol{u}), \quad \boldsymbol{f} : R^{256} \times R^{256} \mapsto R^M \tag{8}$$

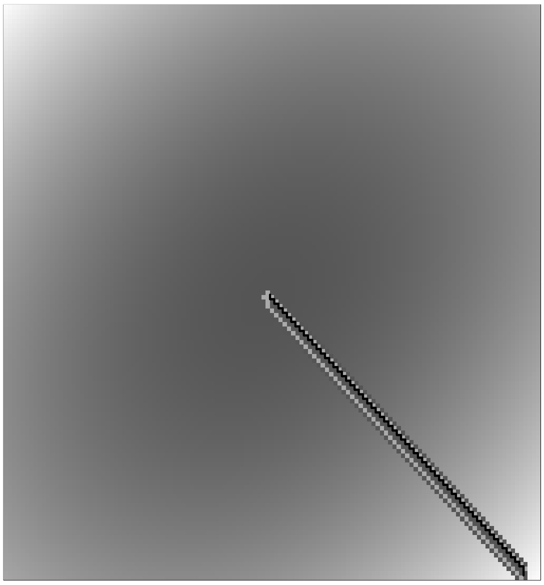

Figure 2: Simulated 'cockroach' locates the darkest point in its 2-D world. White crosses denote points on its trajectory, which are connected by the black line.

The main problem in vision is not, therefore, identification of subsets of $R^{256} \times R^{256}$ but identification of the mapping $\boldsymbol{f}$. This also is a nontrivial problem, but not as a matter of principle. There are indeed several powerful general theorems on 3-layer neural networks as universal mappings (eg. [27, 28, 29]). These show that any continuous function such as $\boldsymbol{f}$ can be approximated by such a network, as can certain discontinuous ones.

While it may go against the grain for some to accept this paradigm, it has arguably a longer history than symbolic approaches to vision, one which goes back at at least to the work of Pitts and McCulloch on modelling the reflex control of eye movements [30] and building on relevant biological literature (eg. [31, 32]) to the current high level of interest in neural networks for visuomotor control (eg. [33, 34, 36]). There are several reasons for preferring this approach to conventional symbolic vision:

1. By embedding the visual response in a metric space, an error measure on the system's performance may be defined. This makes it feasible to use the methods of functional approximation with a neural network based on error minimization. It removes at a stroke one of the major weaknesses of the 'standard model': the difficulty of quantifying the performance of the various system components [6, 2].
2. Equally importantly, the functional approach emphasises continuity of response, ensuring that small perturbations of the input produce only small

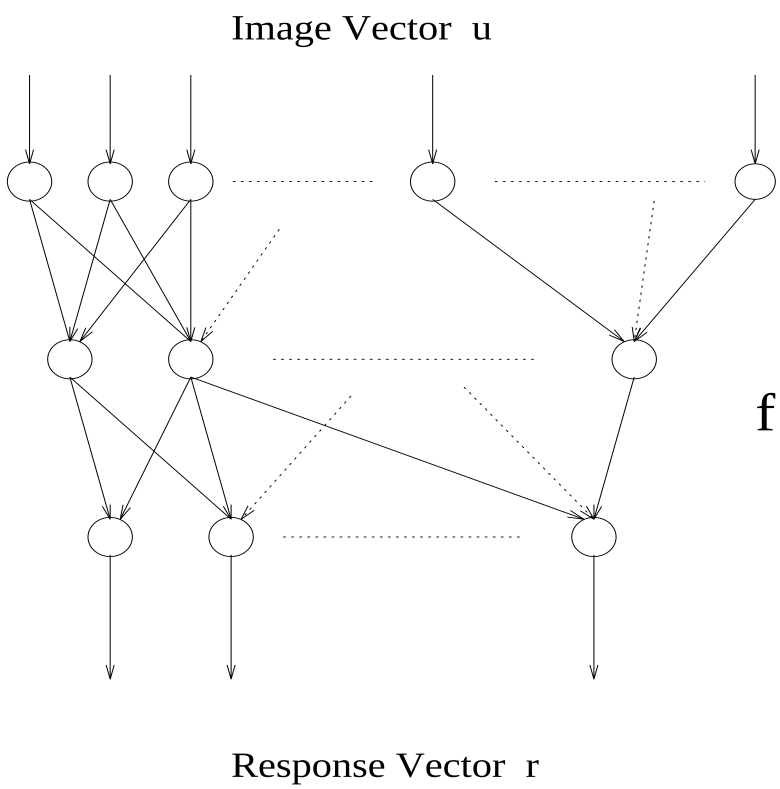

Figure 3: Behaviourist Approach to Vision.

effects on the output. This gives the system noise immunity and underlies its ability to generalise from patterns in a training set to novel stimuli [37].

3. Precisely because they have been inspired by biological systems, neural networks have a computational style which is much closer to that used by biological systems [37, 4]. This raises opportunities for cross-fertilization between artificial and natural vision research, whose benefits are already becoming apparent in a number of studies of visuomotor coordination [34, 33, 35].

These properties would be advantageous in any perceptual system, but in Active Vision, they are decisive: only functional methods can be expected to perform adequately inside the control loop of an AVS. As an illustration, consider the functional approach to the cockroach problem, BBV (Blatta Borealis Volans). This also makes use of the gradient, but unlike BB uses the update rule

$$\vec{x}_n = \vec{x}_{n-1} - \delta \vec{g}(\vec{x}_{n-1}) \tag{9}$$

where δ is a suitably chosen step size. Since the coordinates are discrete, positions have to be rounded to the nearest pixel, but this is a limitation imposed by the problem domain. The result of the corresponding simulation run to Fig. 2 is shown in Fig. 4, which illustrates that BBV does indeed 'fly' or at least hop to the minimum in less than a dozen iterations. A further improvement on BB is the respect for a full range of rotations of the image, as opposed to the 4 rotation subgroup. In short, this is a superior system.

3 Visually Guided Behaviour

3.1 Discontinuity in Behaviour

The argument about continuity of response is specious, for two main reasons:

1. The decision boundaries in the symbolic system can be made 'soft' in a number of ways. For example, any effective inference technique will give an estimate of the probability that a given symbol is correct [16, 12]. The symbolic output can either be represented by this probability distribution, which will generally be a continuous function of the input, or the symbol can be selected according to the probabilities [14]. Both methods give a form of continuity in the mapping from inputs to symbols.
2. In any case, there are many situations where the system needs to 'jump' to a distinct state, based on its observation of its surroundings. To take a well studied example, consider a rather simple biological AVS - the toad.

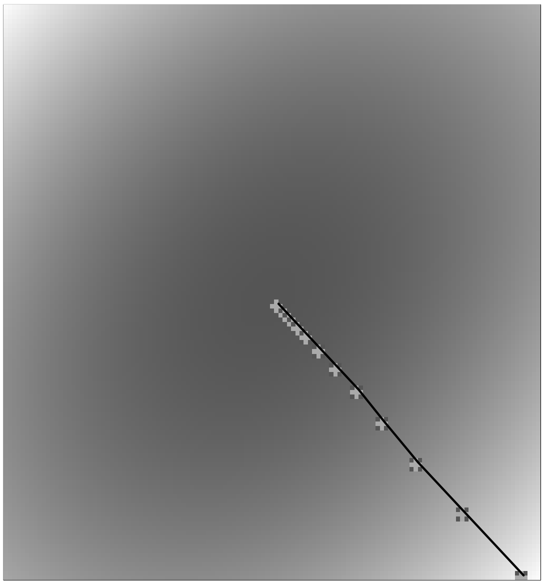

Figure 4: Continuous functional approach to the cockroach problem.

Ewert showed that depending on the size and orientation of a visual target pattern, toads will either orient towards it, with a view to eating it, or veer away, with a view to avoiding being eaten [32]: toads have at least two quite distinct visually guided behaviours - Prey (P) or Flee (F). There is no behaviour which is $0.8P + 0.2F$ (in ambiguous cases, they may not move). Evidently, the toad is attaching symbols to its visual input and these symbols, P and F, determine which type of response the toad will adopt. There are two senses in which these symbols act:

(a) They define sets of input patterns which are equivalent, in the sense that each pattern in a given set gives rise to the same symbol. This is the *invariance* property of symbols.

(b) They define sets of actions which are equivalent. The symbol is an initiator of the response, but its subsequent development depends on future inputs - if the prey or predator moves in response to the toad's action, it will have to adjust its response accordingly, if it is to succeed.

This second aspect of symbols deserves further comment. The major distinguishing feature of an AVS is that it changes its input, whether by sensor motion or other means. This implies that, given the state of the system at time t, it is not generally possible to predict with certainty its state at time $t + 1$. Consequently, there will be a multitude of paths from a given initial state to a specified *goal* state, such as

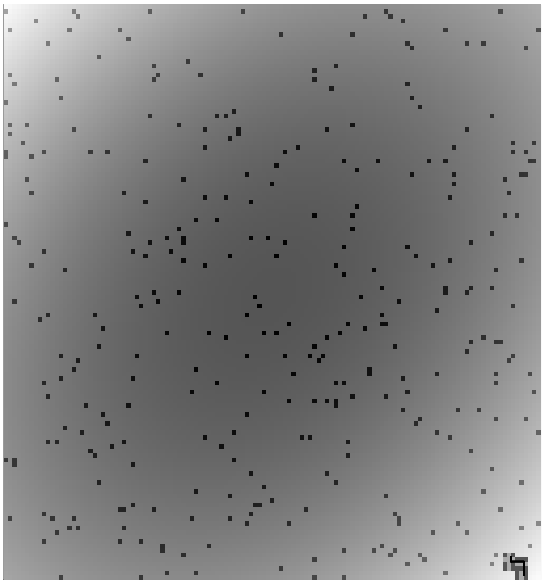

Figure 5: The cockroach gets trapped in a 'black hole'.

eating or not being eaten. Thus action require symbols as surely as perception does - a symbol for an action must constrain the evolution of the system state sufficiently for the goal to be reached, but not so tightly that it cannot take the unexpected into account. These two properties define the *semantics* of the symbols: the mapping between symbols and the system states, which we may assume include its input.

As an illustration of this idea, consider the following modification of the BB problem, illustrated in Fig. 5. In this problem, BB's world is peppered with randomly placed 'black holes', which are deep enough to attract its attention, but too small to hide in. As can be seen from Fig. 5, BB soon gets trapped by a black hole into a limit cycle of '1 step forward -1 step back'. In order to overcome the problem, we must introduce jumps into BB's behaviour: if there is no detectable black hole in the window, then carry on as before; otherwise, ignore that part of the gradient estimate due to the black hole. This is easily accomplished by modifying the weights w_i of (4) via

$$w_i = \begin{cases} \sqrt{2}/\|\rho_i\|^2 & \text{if } |g_i(\vec{x})| < t \\ 0 & \text{else} \end{cases} \tag{10}$$

This effectively introduces a 'black hole' symbol, which attaches to the individual gradient components $g_i(\vec{x})$, which is used to adapt the 'motor response' symbols to the perceived environment. The success of the strategy is illustrated in Fig. 6, which restores the response of BB to its original form, despite the 'noisy' environment. To acknowledge the sophistication of this behaviour, this evolved form will be called BBS (Blatta Borealis Sapiens). Albeit simple, this example illustrates a general principle of

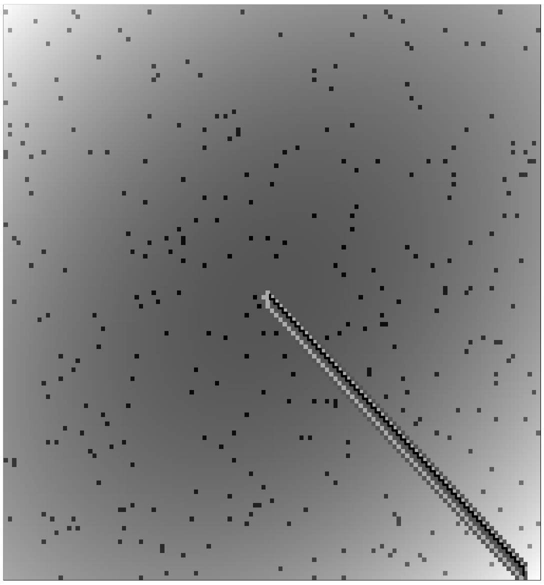

Figure 6: By using a simple 'black hole avoidance' procedure, the smart cockroach avoids the traps.

symbols in behaviour, which is expressed in Fig. 7. In this system, symbols are used to switch the behaviour of a linear dynamical system mapping inputs to outputs, so that it can adapt its response to discontinuities in the environment, such as the presence or absence of a predator. It is evident that it has the same 'universal mapping' property as 3-layer neural networks [29]. In effect, symbols in this system represent control strategies or programs, between which the system switches according to its perceived environment and goals. Attaching symbols to behaviours in this way ensures that the symbols represent something 'real' to the system: the visual input is then organised or grouped according to the needs of the system and not to the apparently arbitrary metric of the vector space in which the input vectors are defined. In this view, symbols simplify both the description and the implementation of the control tasks confronting the system. In more realistic examples, this reduction of complexity is essential in understanding the system behaviour.

3.2 Jumping without Symbols

Of course frogs and toads jump, but that says nothing about their use of symbols; on that definition, a thermostat is permanently engaged in a monologue of exclamations of the form "I am too hot"/"I am too cold". Such an interpretation is wrong on two grounds:

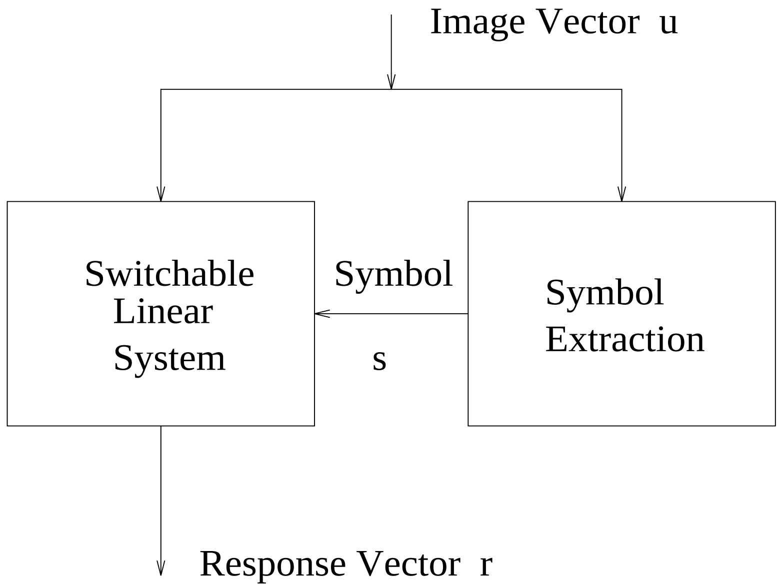

Figure 7: Symbols can be used to switch between modes of a linear system to model arbitrary nonlinear behaviour.

1. A discontinuous signal is not a symbol; it is a discontinuous signal. Even if a system uses such signals to switch between behaviours, there is no reason to imbue them with symbolic content. On that basis, presumably *any* signal which is constant across some subset of $R^N \times R^N$ can be regarded as a symbol, simply because it can be regarded as the characteristic function of the set. Most such sets have no conceivable *meaning* - the only semantics that can be built out of them is set theory !

2. It is doubtful whether the system need be discontinuous to achieve this sort of behaviour - it is enough that its behavioural modes correspond to distinct attractors, towards which the *unforced* system will tend to move, ie. in the absence of external stimuli (eg. [38]). The symbols in Fig. 7 could be replaced by continuous *context* vectors which determine the system dynamics. This use of contextual control signals was first proposed as a useful approach to vision by Granlund [39].

In the case of BBV, it is interesting to note that BBV still performs adequately in the presence of black holes, although it takes rather longer on average to converge (Fig. 8). A simple continuous adaptive rule is again based on modifying the weights

$$w_i(\vec{x}) = \sqrt{2}\exp(-\gamma g_i^2(\vec{x}))\sum_j \exp(-\gamma g_j^2(\vec{x}))/\|\rho_i\|^2 \tag{11}$$

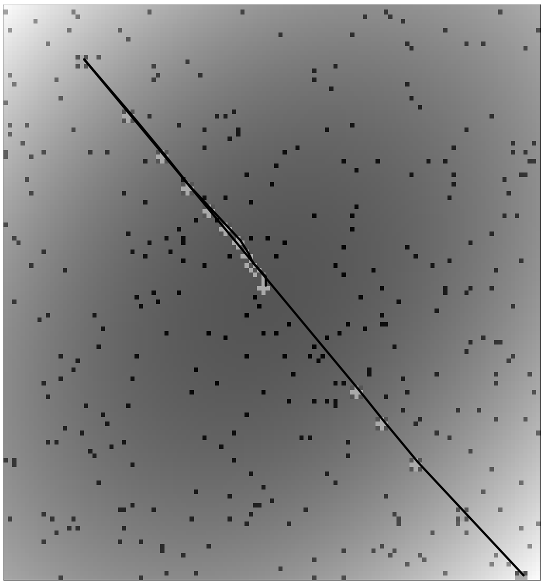

Figure 8: The continuous cockroach is less troubled by black holes, but they do upset its trajectory.

leading to a system BBVS, whose performance in noise is as good as that of BBV on clean data, as shown in Fig. 9. Symbols serve no useful purpose in describing this behaviour, which is essentially that of a 3-layer network (Fig. 3).

4 Visual Coding and Memory

4.1 A Rate-Distortion Problem ?

It may be possible to achieve functional equivalence in a system using continuous data, but this ignores an important part of the cost of such a system - the quantity of information it requires. The amount of information generated by BB is rather easy to measure: it is 2 bits per pixel(bpp), since the 4 direction symbols occur with equal probability and the termination symbol occurs with vanishingly small probability on any family of smooth surfaces showing isotropy statistically. Its continuous equivalent, BBV, generates a potentially unbounded amount of information, on the other hand, because its data are continuous. In real animals, which need to store and process such low level representations, the quantity of information involved is clearly a significant factor. The idea that low level vision serves a coding function is one which is both widespread in the literature on biological vision [40, 41] and is implicit or explicit in several works on neural networks for low level vision [42, 43, 44]. And what is the output of a coder, if not symbols ?

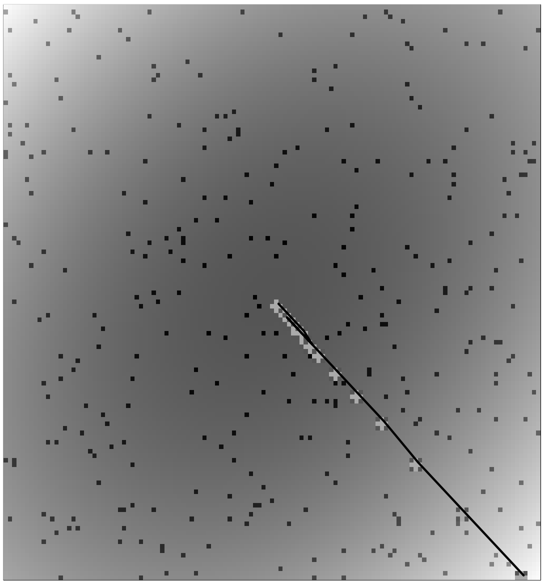

Figure 9: Avoidance procedure improves the continuous cockroach performance.

The branch of information theory dealing with this problem is rate-distortion (R-D) theory [45], which answers the question "What is the *minimum* average amount of information we need to preserve from an information source in order to reconstruct the source outputs with some prescribed average distortion level ?". More precisely, if the source outputs are image vectors $\boldsymbol{u} \in R^N \times R^N$, the coder can be defined as a *stochastic* mapping to vectors $\boldsymbol{v}$ in a reproduction alphabet, governed by a conditional probability density $q(\boldsymbol{v}|\boldsymbol{u})$. A *distortion function*, $d(\boldsymbol{u}, \boldsymbol{v})$, is defined as a distance measure on pairs $\boldsymbol{u}, \boldsymbol{v}$ of source outputs and reproductions. The rate-distortion problem is, given the source output density $p(\boldsymbol{u})$, to find the conditional density $q(./.)$ for which the mutual information $I(q)$

$$I(q) = \int\int d\boldsymbol{u} d\boldsymbol{v} p(\boldsymbol{u}) q(\boldsymbol{v}|\boldsymbol{u}) \log_2[q(\boldsymbol{v}|\boldsymbol{u})/r(\boldsymbol{v})] \tag{12}$$

is minimum, subject to the constraint on average distortion

$$D(q) = \int\int d\boldsymbol{u} d\boldsymbol{v} p(\boldsymbol{u}) q(\boldsymbol{v}|\boldsymbol{u}) d(\boldsymbol{u}, \boldsymbol{v}) = D \tag{13}$$

where

$$r(\boldsymbol{v}) = \int d\boldsymbol{u} p(\boldsymbol{u}) q(\boldsymbol{v}|\boldsymbol{u}) \tag{14}$$

is the probability density for the reproduction. The rate-distortion function $R(D)$ is thus defined [45] as

$$R(D) = \inf_{q:D(q)=D} I(q) \tag{15}$$

This can be solved at least formally by the use of variational methods, but concrete results generally require the use of a tractable form of distortion function, such as mean square error (m.s.e.)

$$d(\boldsymbol{u}, \boldsymbol{v}) = N^{-2} \sum_i \sum_j (u_{ij} - v_{ij})^2 \tag{16}$$

and a simple density $p(\boldsymbol{u})$. It is nonetheless interesting to note that for a correlated Gaussian vector source and m.s.e. error criterion, the optimum coder uses a Karhunen-Loève (KL) transform [46], which happens to be the solution to which Sanger's network converges [43] and to be a special case of the Linsker model [42]. On the other hand, the Laplacian pyramid scheme [47] obviously draws some inspiration from the measured responses of retinal ganglion cells [48], while the more complex coder devised by Daugman is based on the so called Gabor representation [44, 49], whose similarity to the responses of simple cortical neurons was noted by Marcelja [50] and also informed the original work of Granlund [51]. There are therefore numerous ways in which the 'coding metaphor' has been found productive.

In all of these cases, the coder consists of a linear transformation, whose primary effect is to redistribute the energy in the image so that most of the energy is in relatively few transformed components, followed by quantization, which has the effect of assigning most bits to the transformed signal components of highest variance and fewest to those with least [46]. From the neural perspective, an appropriate form of quantizer is a so called threshold coder, which *shares* the information it produces between an *address* component and a *magnitude* component, just as biological neurons do [46].

For such a formulation of vision to be feasible, both the source density $p(\boldsymbol{u})$ and the distortion function $d(\boldsymbol{u}, \boldsymbol{v})$ must be known. Now the source statistics can be acquired *implicitly* through the use of a neural network, as the work of Linsker [42] and Sanger [43] shows. The choice of a distortion measure appears to be more of a problem, to which there are two possible answers:

1. The cost function does not have to be a simple error measure such as m.s.e.. Indeed, $\boldsymbol{v}$ does not even have to be a vector such as $\boldsymbol{u}$. For the theory to be applicable, all that is required is an appropriate form of cost function on the joint space of source and coder outputs [45]. In the case of BB, for example, it is possible to define the cost between the quantized gradient and the original gradient in terms of the actions which they cause. One suitable form of cost is the *excess path length* BBV uses when it uses a quantized gradient, compared with that used when the gradient is not quantized, in getting from its initial position to its terminal point. Let $\vec{x}_n(m)$ be the nth point on the trajectory of BBV when its gradient is quantized to m bits accuracy

$$\vec{x}_n(m) = \vec{x}_{n-1}(m) - \delta Q_m(\vec{g}(\vec{x}_{n-1}(m)) \tag{17}$$

where $Q_m(.)$ is the mapping defined by an m-bit quantizer. The total path length from an initial point $\vec{x}_0$ is then a function of the quantizer resolution m

$$l(m) = \sum_n \|\vec{x}_n(m) - \vec{x}_{n-1}(m)\| \tag{18}$$

where $\vec{x}_0(m) = \vec{x}_0$. An 'action based' cost function for the gradient representation is just the average of the excess path length

$$e(m) = l(m) - \|\vec{x}_\infty(m) - \vec{x}_0(m)\| \tag{19}$$

over an appropriate ensemble of surfaces and initial points. The distortion-rate curve of Fig. 10 was produced by averaging $e(m)$ over an ensemble of 50 randomly selected initial points on the surface of Fig. 2, with the gradient quantized using a scalar quantizer, ie. for the two components of the gradient,

$$Q_m(g_a(\vec{x}) = 2^m Q(2^{-m} g_a(\vec{x})), \quad a = x, y \tag{20}$$

where $Q(.)$ is a 6-bit uniform quantizer whose range exactly covers the range of gradients in the image. The entropy of the quantizer output, estimated from the histogram over the entire ensemble, was used as the abscissa in Fig. 10. As can be seen, the error is negligible for quantizer resolutions $m \geq 4$, the spatial quantization of the array being the dominant factor.

The point of this example is not to advocate the use of path length as a cost measure in Active Vision; it may be appropriate for some applications. It is just to emphasise that from 'average cost $\neq$ m.s.e.' it does not follow that 'R-D theory no use'.

2. Such a use of R-D theory with an application specific cost measure might be regarded as a *general* solution to a *specific* problem [7]. On the other hand, it is reasonable to ask how the theory might be applied to the *general* problem of vision, given that the visual system can be applied to a wide variety of tasks, each with its distinctive costs. To take a specific example, visual memory can be used for a huge variety of tasks, from navigation to object recognition. Which code should be used for visual memory ? The most robust solution may well be to use a *general*, image domain distortion measure, ie. $\boldsymbol{v} \in R^N \times R^N$: a specific solution to a general problem. While m.s.e. has its weaknesses in this respect, some other relatively simple measure may be found. Indeed, the whole motivation for regarding vision as an appropriate domain for scientific investigation is surely that it *is* general - it is not simply part of a control loop.

Although these two approaches are contradictory to some extent, it may be possible to combine a robust and general *low level* cost function with application specific *high level* measures, such as path length, in an overall solution.

In this connection, it is interesting to note that the R-D function for a *product* source, whose components are statistically independent, ie.

$$p(\boldsymbol{u}) = \prod_i p(\boldsymbol{u}_i) \tag{21}$$

and a *sum* distortion measure

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sum_i d(\boldsymbol{u}_i, \boldsymbol{v}_i) \tag{22}$$

is just the sum of the R-D functions for the individual components $\boldsymbol{u}_i$. This could be an explanation for the parallel coding of *form*, *colour* and *motion* found in human vision [18, 52]: at least to a first approximation, these properties of visual objects are independent and can reasonably be assumed to have separate distortion criteria attached to them.

The above discussion has focused on the *source coding* function in vision. This is only part of the story, however. Source coding tells us *what* to represent, it does not say *how* to represent it: it is a significant, but weak constraint on the form of visual representation. One problem it does not address is the *channel coding* problem, which in the present context may be formulated as "What form of representation minimises the risk of 'channel errors' in its translation by subsequent processes ?". In other words, are some forms of representation more susceptible than others to ambiguity ? The answer is surely that there must be a premium on selecting internal representations which are as robust as possible. In this connection, it is worth recalling the results of some early experiments of Shannon on the per character entropy rate of sources of standard English text. He found that given a sufficient left context, the rate is about 1 bit/character, using a 27 character alphabet [53]. Since displayed text normally employs a bit array of at least 7×5 in size, this implies that in the visual input, only 1/35th of the information is actually used. The consequence is great tolerance to errors. Perhaps one of the main reasons why we tend to equate discrete signals with symbols is that the systems we have developed to transmit and store information are highly redundant - the mapping between signals and the symbols that they express is virtually unambiguous. Unfortunately, the mapping between visual signals and the symbols they express is (i) out of our control, by and large, and (ii) extremely ambiguous. This does not mean that we should throw away the symbols.

4.2 A Control Problem

Arguments from information theory have an appeal partly because of the very abstractness and generality of the theory - the 'Information Theory, Photosynthesis and

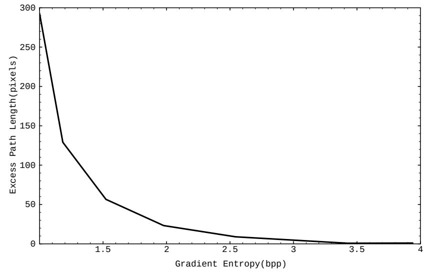

Figure 10: Rate-distortion curve for the gradient estimator used in the cockroach problem. The distortion function is the additional path length travelled by a cockroach using a quantized gradient. The information rate is the entropy of the quantised gradient.

Religion' problem [54]. This has made it all too easy to misapply information theory.

Suppose for the sake of argument, however, that such is not the case here. What can R-D theory say about vision ? One thing it cannot say is that vision must use symbols: R-D theory applies equally to any communication system, whether its input and output are continuous or discrete [45]. Similarly, the neural networks of Linsker and Sanger store information as continuous variables, both the 'short term' neural activations and the 'long term' unit weights are continuous, as is the case in many networks employing some form of Hebbian learning. More fundamentally, how is it possible to categorize the visual input without considering the system response ? To take a simple example, suppose we have a system which tracks a moving object, so that whenever the object is displaced by a vector $\vec{x}$ from one time frame to the next, the system responds with a corresponding movement $\delta\vec{y} = \vec{x}$. If we only examine the input $\vec{x}$, we shall not reach any conclusion about what the system 'thinks' is important, but if we look at the combination of input and response, we can see there is a definite and simple relation, to which we can attach meaning.

The central argument, in the context of Active Vision, is whether there should be some level of vision for which a task-independent, image domain error measure is appropriate. On the existence or usefulness of such a layer, the main argument for 'vision as coding' rests. The importance of making a distinction between active and passive observers has long been recognised among perceptual researchers [55], and it is thought that active vision is necessary for the development of visually guided behaviour [56]:

1. That active vision plays a central role in learning was elegantly demonstrated by Held and Hein 1963 [57]. They reared kittens in the dark until they were 8-12 weeks old. From that age on, the kittens' sole visual exposure was received in a specially designed apparatus where one (active) kitten could move about freely while a second (passive) kitten was carried around in a gondola copying the motions of the active kitten, thus making sure that the two kittens received the same visual stimuli. The kittens were later tested in a series of experiments involving visually guided behaviour. As can be expected, the active kitten tested normal but the passive kitten had no ability to make proper use of visual information.
2. An extension of this work was done by Hein and Gower 1970 [58]. They repeated the above experiment, but this time each kitten received both active and passive visual exposure. One eye was used for active vision the other for passive vision. When tested the kittens behaved normally when using the 'active' eye but their visually guided behaviour was lost when only the 'passive' eye was used.
3. The necessity of active vision for adaption and learning in humans was also demonstated by Held and Hein 1958 [59] in a number of optical rearrangement experiments. The results of the experiments have been verified several times [60]. In these experiments the subjects view the world through a pair of goggles that optically displaced or rotated the field of view. In one experiment, the observer viewed his hand under three different conditions. One was a no-movement condition, the second was a passive-movement condition, in which the observer's arm was moved back and forth by the experimenter. The third was an active-movement condition, in which the observer moved his hand. There was considerable adaption to the distortion produced by the goggles under the active-movement condition, whereas in the other conditions there was not.
4. In another experiment observers wearing goggles either walked around for about one hour or were wheeled around in a wheelchair for an hour over the same path. They were then tested to determine the amount of perceptual change that had taken place. Adaptation to the distortion produced by the goggles occured in the active vision condition, but not in the passive vision condition [61, 62].

While these experiments cannot be conclusive, they indicate strongly that activity plays a decisive role in visual learning. Even in humans, who might reasonably be assumed to have the greatest need for generality in visual processing, it seems that the association of visual stimulus with activity is essential for successful learning. It seems that closing the *system-environment loop* is necessary for learning and that the appropriate metaphor for an Active Vision System is therefore a *control* one, rather than a *coding* one [36].

The implications for system design are profound - 'understanding' of the world is impossible to achieve by studying the world in isolation, but can be attained only by integrating the effect on the world caused by the system's response into the analysis. 'Understanding' then no longer means being able to decompose the world into bits and pieces (ie symbols), but that the system is able to predict changes in the world caused by its actions. Thus, understanding the world gives the system the potential to drive the world towards a desired state. It would seem then that what an Active Vision System needs is not a set of image domain symbols, but rather task-dependent control strategies that force the (world+system) to converge towards desired states.

5 Signals and Symbols

While other arguments could of course be adduced in favour of one side of the debate or another, we feel that we have done enough to demonstrate that for every conventional argument in favour of symbolic vision, we can come up with a counter which at the very least raises doubts about its validity. Perhaps we have been looking at the problem from the wrong point of view. Suppose, for example , that we have a purely functional description of an AVS, but one in which we make explicit the presence of stored information by using a state-space description of the form

$$\boldsymbol{r}(t) = \boldsymbol{f}[\boldsymbol{u}(t), \boldsymbol{\phi}(t)] \tag{23}$$

where

$$\boldsymbol{\phi}(t+1) = \boldsymbol{g}[\boldsymbol{u}(t), \boldsymbol{\phi}(t)] \tag{24}$$

is the state at time $t+1$ and the other terms are as in (8) and the functions $\boldsymbol{f}[.], \boldsymbol{g}[.]$ are nonlinear mappings. The state vector $\boldsymbol{\phi}(t) \in \boldsymbol{V}$ will contain all internal variables, as well as 'visible' parts such as the system's position and orientation; if learning is involved, it will also include memory variables such as connection strengths. The significant point about the description is that it 'opens up the black box' of conventional functional approaches to external inspection: we are free to measure any state variable at any time. Are there any circumstances in which we would be justified in attributing symbols to such a system ?

5.1 What are Symbols ?

To put it differently, do certain types of mapping $\boldsymbol{f}[.], \boldsymbol{g}[.]$ admit of a symbolic interpretation ? It could be that any such mapping does or that only certain classes do -

in a sense, the more restricted the class is, the more useful the concept of symbol. By the same token, we take no comfort at all from the observation that any computable function can be implemented on what Newell calls a *physical symbol system* [63]. We take it for granted that the mappings could be implemented on a general purpose digital computer or on analogue hardware. The real issue is whether symbols would help us in (i) identifying and (ii) describing the mappings - in seeing inside the black box (or Chinese Room [5]). How would we recognise a symbol if we saw one ?

It is probably useful to start with a simpler concept, about which there ought to be correspondingly fewer doubts: a *measurement*. To be specific, what are the minimum requirements we would place on one of the state variables, ϕ_m, say, for it to be identified as a measurement of average luminance ? We believe there are two main requirements:

1. The state variable should vary systematically (preferably linearly) with the average luminance.
2. It should not vary with any other change in the visible environment.

The second of these properties is a familiar one: a measurement should be invariant to transformations of the input other than those which change the quantity measured. The first was alluded to briefly in section 1: a measurement should be *equivariant* to transformations affecting the quantity measured. There has recently been a great deal of work on symmetries in general nonlinear mappings of the form of (24), in which the concepts of invariance and equivariance have played a central role [65]. Note that we have said nothing about the details of how ϕ_m is to be computed, but we have specified how it is to behave with respect to the environment. In other words, if we decided it would be useful to have such a state variable, we would have to train the neural network explicitly to produce it; given some unit in a previously trained network, we could decide whether it had such a function by observation of its response to changes in the input. It would help us to understand what was going on in the network to arrive at such a conclusion.

In our view, measurement is the simplest example of a process with symbolic content: measurements are typically scalars, whose *magnitude* varies in the prescribed way with respect to changes of input; symbols are in general based on vector quantities, but they are defined by the invariance and equivariance of those quantitites with respect to transformations acting on the world state. These transformations include not only the obvious coordinate transforms caused by relative motions, but also changes in luminance, the presence of other objects and indeed the 'motions' from one object to another.

To understand the relationship between the various transformations with which we have to deal in vision, consider Fig. 11, which shows the three parts of a typical vision scenario: the 'world' or environment in which the system is located and

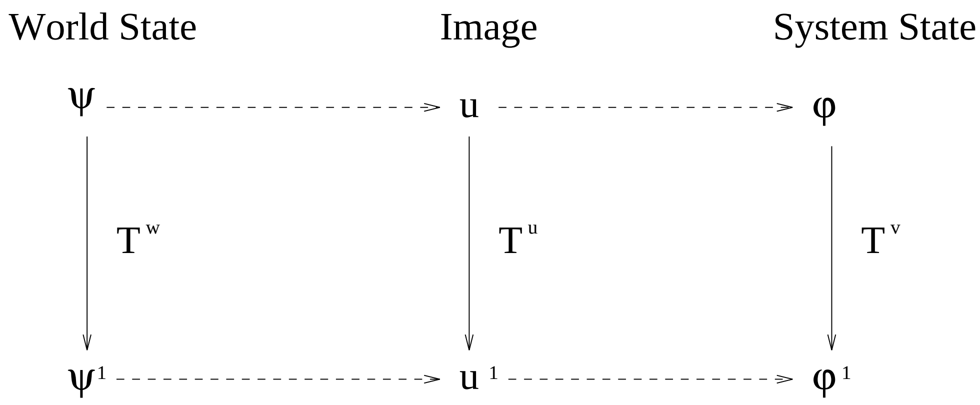

Figure 11: Transformations in vision. Changes in the 'world' state induce changes in the image, to which the system responds. Vision must solve the inverse problem of identifying the change in world state from that in the image.

which contains visible objects; the image by which the system acquires its information about the world and the system state vector, which contains the 'knowledge' the system has about the world and which determines its response to its current input. Transformations act on each of these components: a change in a visible object, such as a translation or rotation, *induces* a corresponding change in the image and this in turn leads to a change in the system state. If we can discover, by examining the way the system state changes in response to changes in the world, a systematic dependence of certain state variables on particular objects or object properties, then we may reasonably infer that the system is using symbols for those objects/properties. The problem is made vastly more difficult by the presence of the intermediate step - the formation of the image - a process which introduces singularity and consequent uncertainties in the communication between *outer* and *inner* worlds [64, 7].

In an AVS model, the picture is apparently complicated by the existence of feedback: actions must affect world state (Fig. 12). Thus if $\boldsymbol{\psi}(t)$ is the current world state, then

$$\boldsymbol{\psi}(t+1) = \boldsymbol{h}[\boldsymbol{\psi}(t), \boldsymbol{r}(t)] \tag{25}$$

where the response vector $\boldsymbol{r}(t)$ is, as in (23) above, the system output at time t.

In one sense, however, the AVS model has one obvious benefit over the conventional one: the presence of an output coupled to the environment means that not only is it possible to learn by associating input and output - stimulus and response - but it may also be the case that feedback from the environment can function as a form of 'supervision', albeit a noisy one. The idea of using vision to provide feedback has been studied in a variety of contexts, including control of eye movements [66], coordination of hand and eye movement [67, 68] and object recognition and manipulation [69]. The precise form of the feedback and learning mechanisms vary, but

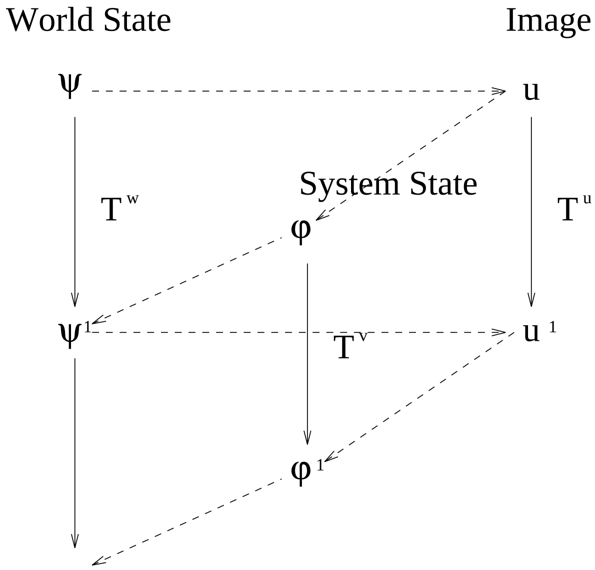

Figure 12: Transformations in active vision. Changes in the 'world' state induce changes in the image, to which the system responds, inducing a consequent transformation in world state.

the general principle is the same: specific motor signals result in motions producing specific changes in visual input.

We can illustrate invariance and equivariance in a way which draws on this principle and which also illustrates a major difference between active and passive vision. Consider the four element pattern of Fig. 13 in its four configurations, which correspond to (i) the original pattern, (ii) swapping left and right elements on each row, (iii) swapping the two rows and (iv) swapping both rows and columns. Although there are only four distinct patterns, this suffices because the four transformations form a group which is the direct product of two factor groups, namely the subgroup $\mathcal{T}_1 = \{\boldsymbol{I}, \boldsymbol{T}_1\}$, where the first element is the identity and the second swaps columns, and the subroup $\mathcal{T}_2 = \{\boldsymbol{I}, \boldsymbol{T}_2\}$, where the second transformation is the row swap. If the pattern is written as a vector $\boldsymbol{v} \in R^4$, the matrix representations of the transformations are just

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{26}$$

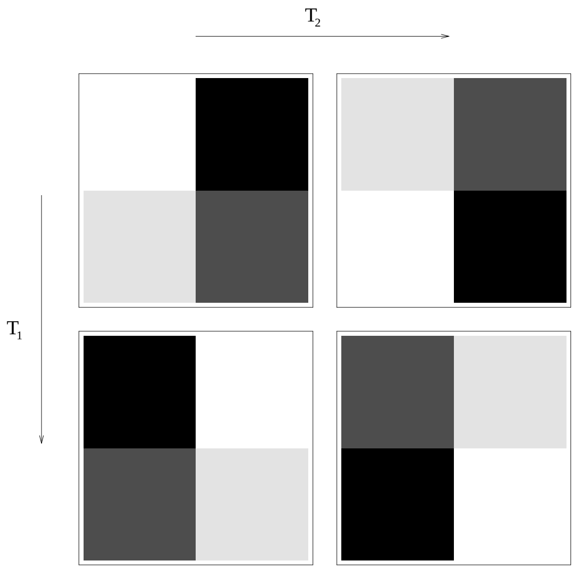

Figure 13: A four component pattern and transformations formed by row and column swaps.

$$\boldsymbol{T}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{27}$$

and

$$\boldsymbol{T}_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \tag{28}$$

It is not hard to see that these operations commute and hence that the four products form a group which represents the actions illustrated in Fig. 13.

This set can be seen as a simple illustration of a set of patterns representing 'what' and 'where' independently. For example, we might regard the patterns $\boldsymbol{v}, \boldsymbol{T}_1\boldsymbol{v}$ as representing two configurations ('views') of the same object, ie. one 'what' and two 'wheres', but regard the pattern $\boldsymbol{T}_2\boldsymbol{u}$ as a different object, but in the same configuration as the object $\boldsymbol{u}$, whatever the choice of $\boldsymbol{u}$. It may seem a rather small world, with four states, but it is sufficient to illustrate invariance and equivariance. Suppose, for example, that we can find some *symbol* $\boldsymbol{f}(\boldsymbol{u})$, where $\boldsymbol{u} \in \{\boldsymbol{v}, \boldsymbol{T}_1\boldsymbol{v}, \boldsymbol{T}_2\boldsymbol{v}, \boldsymbol{T}_2\boldsymbol{T}_1\boldsymbol{v}\}$

which is *invariant* to the 'what' subgroup $\{\boldsymbol{I}, \boldsymbol{T}_2\}$. This is a function of the pattern elements which is unchanged if they are row-swapped. One obvious choice is the vector defined by

$$\boldsymbol{f}_1(\boldsymbol{u}) = (< \boldsymbol{u}, \boldsymbol{e}_0 >, < \boldsymbol{u}, \boldsymbol{e}_2 >)^T \tag{29}$$

where $< ., . >$ denotes inner product and the four vectors $\boldsymbol{e}_i, 0 \leq i < 4$ are the unit vectors

$$\boldsymbol{e}_0 = \frac{1}{2}(1, 1, 1, 1)^T \tag{30}$$

$$\boldsymbol{e}_1 = \frac{1}{2}(1, 1, -1, -1)^T \tag{31}$$

$$\boldsymbol{e}_2 = \frac{1}{2}(1, -1, 1, -1)^T \tag{32}$$

and

$$\boldsymbol{e}_3 = \frac{1}{2}(1, -1, -1, 1)^T \tag{33}$$

Those familiar with the Haar transform will recognise these vectors. More importantly, they are eigenvectors of the transformations $\boldsymbol{T}_1, \boldsymbol{T}_2$ and hence provide an irreducible representation of the group $\mathcal{T}_1 \otimes \mathcal{T}_2$. To show that the function $\boldsymbol{f}_1(\boldsymbol{u})$ is invariant to row swaps, it suffices to note that

$$\boldsymbol{T}_2 \boldsymbol{e}_i = (-1)^i \boldsymbol{e}_i, \quad 0 \leq i < 4 \tag{34}$$

and hence that

$$\boldsymbol{f}_1(\boldsymbol{T}_2\boldsymbol{u}) = (< \boldsymbol{T}_2\boldsymbol{u}, \boldsymbol{e}_0 >, < \boldsymbol{T}_2\boldsymbol{u}, \boldsymbol{e}_2 >)^T = \boldsymbol{f}_1(\boldsymbol{u}) \tag{35}$$

where the right equality follows from (34) and the observation that each $\boldsymbol{T}_i$ has a symmetric matrix representation. Similarly, the function $\boldsymbol{f}_2(\boldsymbol{u})$

$$\boldsymbol{f}_2(\boldsymbol{T}_1\boldsymbol{u}) = (< \boldsymbol{u}, \boldsymbol{e}_0 >, < \boldsymbol{u}, \boldsymbol{e}_1 >)^T \tag{36}$$

is invariant to column swaps. Since these two functions are invariant to the respective subgroups, it follows that *any* function of the form $\boldsymbol{g}(\boldsymbol{f}_i(\boldsymbol{u}))$ inherits the invariance. Moreover, the two functions also exhibit equivariance, in the sense that there is a representation of the subgroup $\mathcal{T}_i$ on the $2 - D$ subspace spanned by the respective eigenvectors, with matrix representation

$$\boldsymbol{I}' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{37}$$

and

$$\boldsymbol{T}'_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{38}$$

which is the same for both subgroups. In other words,

$$\boldsymbol{f}_i(\boldsymbol{T}_i\boldsymbol{u}) = \boldsymbol{T}'_1\boldsymbol{f}_i(\boldsymbol{u}), \quad i = 1, 2 \tag{39}$$

Thus the functions $\boldsymbol{f}_i(\boldsymbol{u})$ exhibit the invariance and equivariance properties required to separate 'what' from 'where' and provide a simple example of an ideal symbol system. It is less obvious, however, whether such a representation of the signal space can be acquired through passive 'observation'. This clearly requires an unsupervised learning algorithm, such as Sanger's generalisation of the Oja eigenvector estimator [43], which uses successive approximation to find eigenvectors of the input signal covariance matrix. In the present case, if we present such a network with a training sequence consisting of all four transformations of vectors selected at random, then the covariance matrix will have the form

$$\boldsymbol{R}_u = \frac{1}{4}(\boldsymbol{R}_v + \boldsymbol{T}_1\boldsymbol{R}_v\boldsymbol{T}_1 + \boldsymbol{T}_2\boldsymbol{R}_v\boldsymbol{T}_2 + \boldsymbol{T}_1\boldsymbol{T}_2\boldsymbol{R}_v\boldsymbol{T}_2\boldsymbol{T}_1) \tag{40}$$

where

$$\boldsymbol{R}_v = E\boldsymbol{v}\boldsymbol{v}^T \tag{41}$$

is the covariance matrix of the random vectors. It follows immediately from (40) that the covariance $\boldsymbol{R}_u$ is invariant to the transformations $\boldsymbol{T}_1, \boldsymbol{T}_2$ and hence that its eigenvectors are just those of (30)-(33). Its eigenvalues depend of course on the signal statistics - a fact which can be significant in estimating them from the training data. The four eigenvector estimates at training cycle n are given by

$$\boldsymbol{e}_i(n) =< (\boldsymbol{v}(n) - \hat{\boldsymbol{v}}_i(n)), \boldsymbol{e}_i(n) > (\boldsymbol{v}(n) - \hat{\boldsymbol{v}}_i(n)) \tag{42}$$

where

$$\hat{\boldsymbol{v}}_i(n) = \hat{\boldsymbol{v}}_{i-1}(n) + < (\boldsymbol{v}(n) - \hat{\boldsymbol{v}}_{i-1}(n)), \boldsymbol{e}_{i-1}(n) > \boldsymbol{e}_{i-1}(n) \tag{43}$$

is the estimate of the current input vector $\boldsymbol{v}(n)$ based on the first i estimated eigenvectors and ensures orthogonality of the estimated vectors. At each step, the vectors $\boldsymbol{e}_i(n)$ are normalised. This algorithm was used to produce the estimated eigenvectors shown in Fig. 14, which are correct to about 0.1% and were produced in a training sequence consisting of 10 cycles of each of the four transformations of a set of 1000 randomly chosen vectors. Thus a conventional eigenvector hunter can locate the relevant subspaces of the signal space to derive the symbols. Correspondingly, a network which transformed the signal into the eigenvector coordinates, ie.

$$\boldsymbol{y} = \boldsymbol{E}\boldsymbol{u} \tag{44}$$

where $\boldsymbol{E}$ is the matrix whose rows are eigenvectors $\boldsymbol{e}_i, 0 \leq i < 4$, would have an output $\boldsymbol{y}$ which displayed the desired invariance and equivariance. Note, moreover, that the eigenvectors are unique in this respect - no other basis has the right invariance and

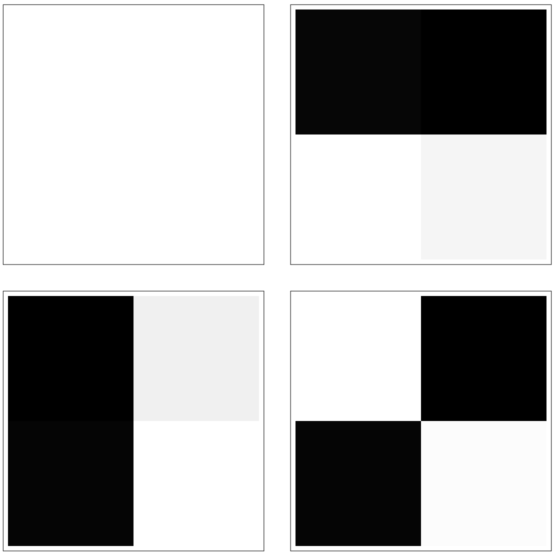

Figure 14: Eigenvector estimates found by the Sanger algorithm, closely matching the eigenvectors of the covariance matrix.

equivariance properties - and that they also represent the most efficient way to code the input [43, 42].

Now consider an appropriate model for the active case. A simple way to express this is to assume that a network is required whose output (response) is invariant/equivariant to the two subgroups $\mathcal{T}_i, i = 1, 2$. A 4-input 2-output single layer perceptron, was presented with a training sequence consisting of 1000 cycles of each of the four transformations of a randomly chosen vector and target outputs given by the four columns of the matrix

$$\boldsymbol{Y} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \tag{45}$$

This corresponds to representations of $\boldsymbol{T}_i, i = 1, 2$, given by

$$\boldsymbol{T}_1'' = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{46}$$

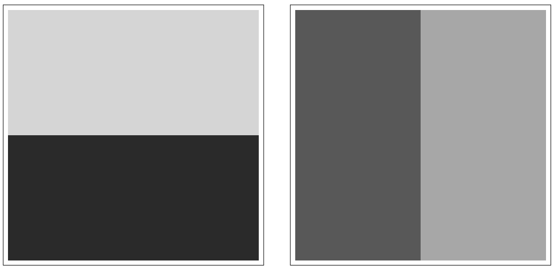

Figure 15: Weight vectors for the two unit perceptron after training. The weights closely match the two relevant eigenvectors. The left image is the weight vector for the 'what' unit and the right that for the 'where' unit.

and

$$\boldsymbol{T}''_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \tag{47}$$

In effect, the first unit is a symbol for 'what' and the second one for 'where', according to our previous definition. The result of the training is shown in Fig. 15, which shows the weight vectors for the two units. It is not too hard to see that they are just the eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2$, which are indeed the correct $1-D$ subspaces of the signal space. The residual m.s. output error was less than 1% in this case.

Although both 'passive' and 'active' models show the right properties in the ideal case, this example is defective as an abstraction of the vision problem. The missing component is the singularity of the process converting the 'real world' of the visible environment into images. The singularity has several sources: the projection from $3-D$ to $2-D$, noise and the windowing of the signal - the world is only partly visible. In the present example, we shall focus on the windowing problem, which like the $3-D$ to $2-D$ problem is actually a projection problem, in that only a subspace of the signal space is observable. Consider, therefore, the same problem as before, with the group consisting of the direct product of row and column swap subgroups, but suppose now that, while the same transformations affect the 'real' signal, the observable signal is just the subspace consisting of the first three signal components. In other words, the 'retina' of the model system has three input units, not four, giving the input signal

$$\boldsymbol{u}' = \boldsymbol{I}_3 \boldsymbol{u} \tag{48}$$

where

$$\boldsymbol{I}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{49}$$

The crucial feature of the windowing operator $\boldsymbol{I}_3$ is that it does not commute with the swaps $\boldsymbol{T}_i, i = 1, 2$, as is easily verified

$$\boldsymbol{I}_3 \boldsymbol{T}_i \neq \boldsymbol{T}_i \boldsymbol{I}_3, \quad i = 1, 2 \tag{50}$$

This has the effect of destroying the eigenstructure of the covariance matrix $\boldsymbol{R}_u$ of (40), replacing the eigenvectors by those shown in Fig. 16, which unsurprisingly have a zero in their last component.

The presence of a zero is not in itself the problem, but the fact that the eigenevectors no longer represent invariant subspaces of the transformation operators $\boldsymbol{T}_i, i = 1, 2$, most certainly is: no combination of the projections of the signal onto these eigenvectors will give a function having invariance or equivariance to the transformations. Indeed, there is no subspace of the $3 - D$ windowed signal which is invariant to the transformations and so *no* polynomial function of the input components which is invariant. Thus a symmetry property of the 'real world' can be and generally will be lost in perception. This is a consequence of the lack of commutation between the windowing operator and the transformations - a close relation of the uncertainty principle [12]. In fact it is formally identical to the commutation relations defining the prolate spheroidal sequences. In other words, although this is a simple example, it illustrates the fundamental problem that 'what' and 'where' operations do not commute at the signal level [12], despite our unshakable belief that in reality they do: we could place an arbitrary object at an arbitrary position, assuming the Laws of Nature do not forbid it.

Now consider the 2-output perceptron, which has two advantages over the unsupervised net: it has a sigmoid nonlinearity in its output and it is supervised. It may not come as a surprise to find that it does indeed learn to represent 'what' and 'where' just about as effectively as before, but it is nonetheless significant that it does. To illustrate the point, in Fig. 17, the total squared output error is shown as a function of training cycle number for both the original and windowed inputs, using exactly

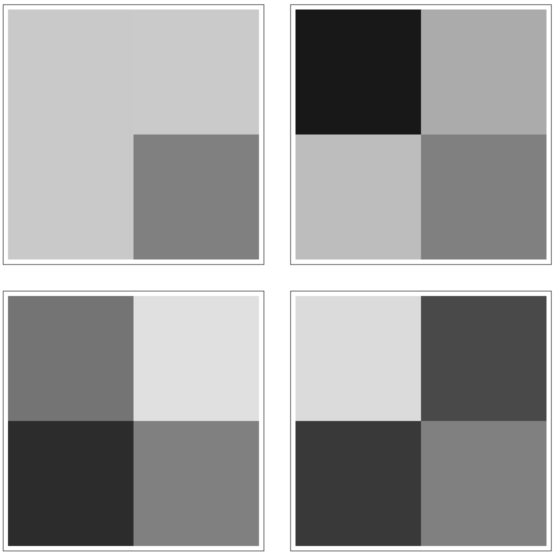

Figure 16: Eigenvector estimates found by the Sanger algorithm, after windowing to three nonzero components.

the same initial conditions and training sequence. There is a slightly larger error for the windowed input case, but the performance is virtually unaffected: the average error in the output is still less than 1%. In other words, if actions and visual input are associated and if the actions predictably cause specific transformations of the environment, then the limitations inherent in the process of vision can be overcome: action can indeed help vision.

To try to get a more realistic model, it is worth examining the growing body of work on visuomotor coordination. One general approach to learning such skills, which has received considerable attention is known as *feedback-error-learning* (eg. [71, 69]), which is based on the idea that a feedback control system can be used to generate an error signal, which is in a suitable form for training a feedforward (open loop) controller. This has the obvious attraction that by minimising the error, the system acquires a model of the dynamics which by definition minimises the amount of information needed from feedback, so that, in the limit, it can virtually act in an 'open-loop' fashion, maximising its speed of response to motor commands, for example. In practice, the accuracy of the model will be affected by sensor noise and

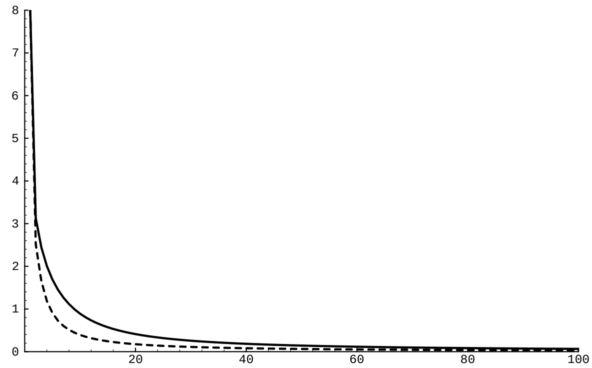

Figure 17: Output error plotted against training cycle number for the 4-input (dashed) and 3-input (solid) perceptrons.

variability of the environment, necessitating some level of feedback after learning. Nonetheless, there is clearly a sense in which the reduction in feedback variance corresponds to the replacement of conscious effort by effortless competence, which is typical of human learning, especially in visuomotor tasks.

Although this particular model of learning is more directed at motor control than vision, it does indicate that, for example, task-oriented object discrimination is possible [69]. A more general setting for discussing the relationship between action and vision is provided by the model of Fig. 18, which adapts the standard state-space control paradigm to active vision. The principle of operation of such systems is that the state estimator uses the sensor output to estimate the 'plant state', which in this case is that part of the world state relevant to the system; the estimated plant state is then used through the control law to derive a suitable control signal for the motor system, generating a response [70]. There is of course no question that such a system has advantages over an open loop system in the control of actions; it is less obvious whether it may have advantages in terms purely of vision. This is the problem we wish to address now.

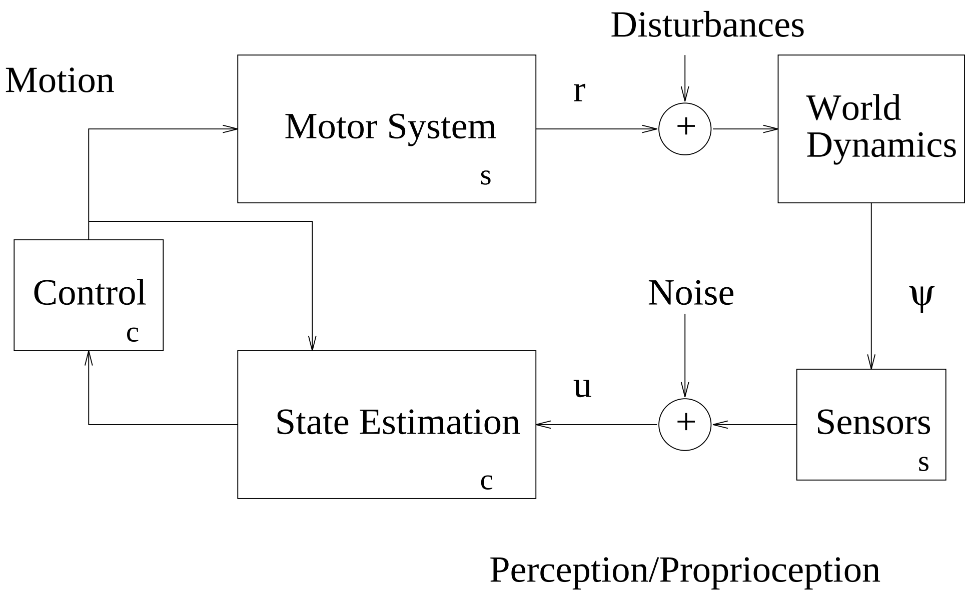

Figure 18: Adaptation of state-space control system to active vision. Note: functions labelled s are subcortical; c are cortical.

In other words, we are interested in state estimation: state estimation is a symbolic act, according to our definition. In effect, state estimation implies modelling of the 'system+world', requiring an internal representation with the proper invariance and equivariance properties. A system description such as Fig. 18 sets constraints on what can be going on 'inside the box': not every feedback system will have state variables ϕ_i which correspond to 'plant' states ψ_i, in the way which is demanded by this system. In particular, there will need to be estimates of the system's position and configuration with respect to the environment. It is apparent, however, that the strict invariance and equivariance of group theory will have to be weakened to account for the fact that noise and disturbances affect the state estimation. We are thus led to the concepts of *ϵ-invariance and ϵ-equivariance*: a random mapping $\boldsymbol{f} : R^n \to R^m$ is ϵ-invariant with respect to a set $\mathcal{T}$ of transformations, selected according to some probability distribution $P(\boldsymbol{T})$, if the mean squared error over both the transformations and the mapping 'noise',

$$E||\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{T}\boldsymbol{x})||^2 \leq \epsilon E||\boldsymbol{f}(\boldsymbol{x})||^2 \tag{51}$$

while a mapping $\boldsymbol{f} : R^n \to R^m$ is ϵ-equivariant with respect to a set $\mathcal{T}$ of transformations if

$$E||\boldsymbol{T}\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{T}\boldsymbol{x})||^2 \leq \epsilon E||\boldsymbol{f}(\boldsymbol{x})||^2 \tag{52}$$

Note that the transformation will generally have a different representation in the output space, but this does not prevent its being classed as equivariant. The idea behind these definitions is just to give a 'mean square error' measure which is appropriate in the present context. It is perhaps worth noting that the rate-distortion arguments of section 4.1 support the weakening of invariance and equivariance in this way, on the grounds of efficiency. As a concrete example, the associator whose weight vectors are illustrated in Fig. 15 is ϵ-invariant/equivariant at the level of 0.03% to the respective row and column swaps; the corresponding figure for the eigenvector approximations is 1%.

Thus, if the 'external' state of the system, $\boldsymbol{\psi}$, is changed to $\boldsymbol{\psi}' = \boldsymbol{T}\boldsymbol{\psi}$, where the transformation $\boldsymbol{T}$ is selected according to the distribution $P(\boldsymbol{T})$ and there exists a subspace of the 'internal' system state-space, with state vector $\boldsymbol{\phi}_1 = \boldsymbol{\Pi}\ \boldsymbol{\phi}$, such that the mapping $\boldsymbol{f}$

$$\boldsymbol{\phi}_1 = \boldsymbol{f}(\boldsymbol{\psi}) \tag{53}$$

is ϵ-equivariant, ie.

$$E||\boldsymbol{f}(\boldsymbol{T}\boldsymbol{\psi}) - \boldsymbol{T}\boldsymbol{\phi}_1||^2 < \epsilon E||\boldsymbol{\phi}_1||^2 \tag{54}$$

then the model of Fig. 18 can be applied meaningfully to the system and correspondingly, it can be said to use symbols for the external state $\boldsymbol{\psi}$. Of course, one or two difficulties attend the translation of this definition into practice. In the first place, both the transformations which act on the external state and their internal representation take finite time to operate. This can be simulated with a discrete time system, in which, say, $\boldsymbol{\psi}(t+1) = \boldsymbol{T}\boldsymbol{\psi}(t)$ and, similarly, $\boldsymbol{\phi}_1(t+1) = \boldsymbol{T}\boldsymbol{\phi}_1(t)$. More significantly, we hit the following problems:

1. The system requires an accurate model of the 'plant', ie. the dynamics of the interaction between motor command signals and world state. This is a complex problem, made more difficult in biological systems by the changes that occur in the dynamics as an animal grows. Thus, in higher vertebrates, there must be a degree of adaptivity in the modelling - the model must be at least partially learned. This is a system identification problem.

2. Even with an accurate plant model, state estimation is only effective if the external state variables are *observable*. An informal definition of observability is that a system is *observable* if, given a sufficiently long observation of the sensor output, we can deduce what the state was at any given time [70]. Observability requires that the sensor output is sufficiently rich that any state variable projects unambiguously onto the sensor output space.

We do not have to think too long about this to see that observability is likely to be a problem. In the first place, any sensor must have a finite window, so that there are no direct measurements of many of the state variables $\psi_i(t)$ at any given time. This is just the singularity problem which was examined in the above example.

Secondly, state variables such as the system's position are not directly observable at any time: for the most part, system motions act differentially; from the visual input, only velocity can be estimated. In effect, the 'world dynamics' include an integrator for such state variables, introducing an unknown constant of integration into the estimation. We might be able to cope partially with the first problem by 'extending the window' using memory, but if we cannot solve the second problem, then this will be impossible. To state the problem succinctly: how do we know where we are, when the only measurements we have are velocity measurements ?

The answer is that we must use a *map* of our world, that is, we must be able to translate certain *unique features* of the (visible part of the) state vector $\boldsymbol{\psi}$ into a direct measurement of position. This is a (perhaps the) fundamental perceptual problem confronting animals: no animal can be born with a map of its environment - it must acquire it during development and it must acquire it without direct measurement of position. It hardly needs saying that such a map is a symbolic entity, but how can it be acquired ? To answer this question, we have devised a 1-D simulation, which nonetheless illustrates some of the major issues in map-building and which highlights the difference between active and passive vision in a way which is relevant to the discussion in section 4. In particular, we are concerned to discover whether it might be easier to learn a map if you move yourself around than if you are wheeled around by someone else: does action really help vision in acquiring a map ? We start by representing the system state as a vector $\boldsymbol{\psi} \in R^N$ selected from the set of N delta functions $\boldsymbol{e}_i$, $e_{ij} = \delta_{ij}$, $0 \leq i, j < N$, each of which corresponds to a different 1-D 'position' i. Each position in the world contains a distinct pattern, corresponding to the visual input at that location, $\boldsymbol{u}(\boldsymbol{\psi})$. The map-building task is to learn to associate the visual pattern $\boldsymbol{u}(\boldsymbol{e}_i)$ with the position vector $\boldsymbol{e}_i$. Consider, then, a pattern association task in which each of N patterns $\boldsymbol{u}_i \in R^M$ is to be associated with a delta function $\boldsymbol{e}_i$, which represents the $1 - D$ 'position' of the pattern. To simplify computation, we choose a pattern dimensionality $M = N$ and choose the patterns to be orthonormal: $\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$, $0 \leq i, j < N$. With these choices, the problem is one which is readily solved using a linear associator, whose matrix form is just

$$\boldsymbol{A} = \boldsymbol{U}^T \tag{55}$$

where $\boldsymbol{U}$ is the matrix whose columns are the pattern vectors. This is an almost trivial problem - a 1-level, linear 'perceptron', whose coefficients can be learned using the delta rule [37]

$$\boldsymbol{A}(n) = \boldsymbol{A}(n-1) + \alpha(n)(\boldsymbol{F} - \boldsymbol{A}\boldsymbol{U})\boldsymbol{U}^T \tag{56}$$

where the n is the training cycle number, $\alpha(n)$ the learning coefficient and the matrix $\boldsymbol{F}$ is the matrix whose columns are the target patterns. In the ideal case where $\boldsymbol{F} = \boldsymbol{E}$, the targets are noiseless and the associator can be learned in a single training cycle. Note that, once training is complete, the mapping from the state $\boldsymbol{\psi}$ to associator

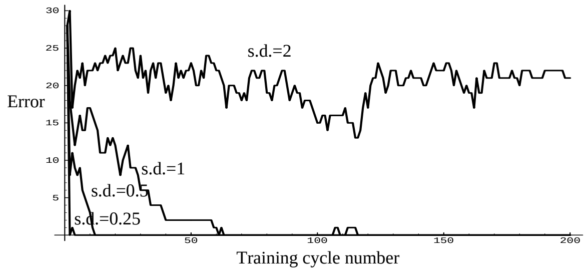

Figure 19: Error as a function of training cycle number for various noise standard deviations.

output $\boldsymbol{A u}_i$ is equivariant to 'rotations mod$N$', ie. if $\boldsymbol{T}_{i-j}\boldsymbol{e}_i = \boldsymbol{e}_j$, then

$$\boldsymbol{A u}(\boldsymbol{T}_j \boldsymbol{e}_i) = \boldsymbol{T}_j \boldsymbol{A u}(\boldsymbol{e}_i), \quad 0 \leq i, j < N \tag{57}$$

This example shows that even a simple system can be regarded as using symbols because it has an internal representation of the 'real world position' which satisfies the equivariance requirement. We have reduced the problem to this form so that we can focus on the relevant issue in the present context, namely that we do not have access to the 'target patterns' $\boldsymbol{e}_i$ during training. The best that can be expected is a noisy estimate of position based on velocity measurements. The effect of noise on learning is illustrated in Fig. 19 and 20. These are based on a noisy target matrix $\boldsymbol{F}$ whose columns are obtained by shifting the corresponding columns of $\boldsymbol{E}$ by an amount $j_i(n)$ which varies randomly from one training cycle to the next and which obeys a normal density, quantized to integers

$$\boldsymbol{f}_i(n) = \boldsymbol{e}_{i+j_i(n)} \;\; 0 \leq i < N \tag{58}$$

where the index $i + j_i(n)$ is calculated mod$N$ and the probability distribution of the shift $j_i(n)$ is

$$P_{j_i}(k) = \frac{1}{\sqrt{2\pi}\sigma} \int_{k-1}^{k} dx \; \exp[-x^2/2\sigma^2] \tag{59}$$

Note that in this problem, noise is wholly detrimental to the system's performance: the noise-free system performs perfectly after one training cycle. This is a different situation from those in which a degree of noise in the input training data can be

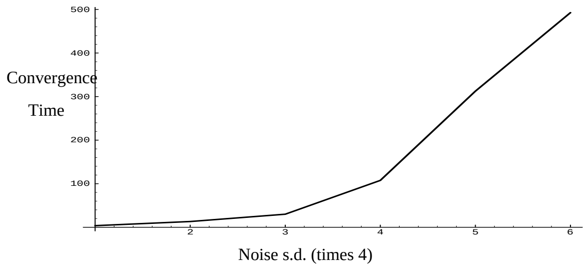

Figure 20: Convergence time vs. noise s.d.

used to compensate for an inadequate training set [72]. Fig. 19 shows the error as a function of cycle number. The first observation to make is that, in order to get meaningful results, it is necessary to use a learning coefficient which guarantees stochastic convergence, ie. $\alpha(n) = n^{-1}$: with a constant learning coefficient, there is no convergence in general. With this choice of learning coefficient, however, there is convergence of the linear associator to a form in which a given input pattern $\boldsymbol{u}_i$ produces an output vector with a maximum at the true position i and a spread which reflects the noise standard deviation σ. To produce the results shown in the figures, a maximum detector was applied to the output of the linear associator, to give an output on cycle n

$$\hat{\boldsymbol{e}}_i = \boldsymbol{e}_{i_m} \quad v_{ii_m} = \max_j v_{ij} \tag{60}$$

where

$$\boldsymbol{v}_i = \boldsymbol{A}\boldsymbol{u}_i \tag{61}$$

Such 'winner-take-all' circuitry is typically implemented by lateral inhibition, a ubiquitous feature of biological perceptual machinery. More importantly, it gives us a way to compare results and talk about convergence, since in the limit, the modified (nonlinear) associator will now function correctly, regardless of the noise variance. More precisely, for any noise s.d. and error $\epsilon > 0$, there exists a cycle number n such that the system is almost certainly ϵ-equivariant after n training cycles.

Fig. 19 demonstrates that the time to convergence is highly dependent on σ, a fact which is confirmed by Fig. 20, which shows the convergence time as a function of σ, averaged over 10 runs. It appears that the time to learn the association increases as a power of σ. This has two causes: the normal curve gets flatter as σ increases and the

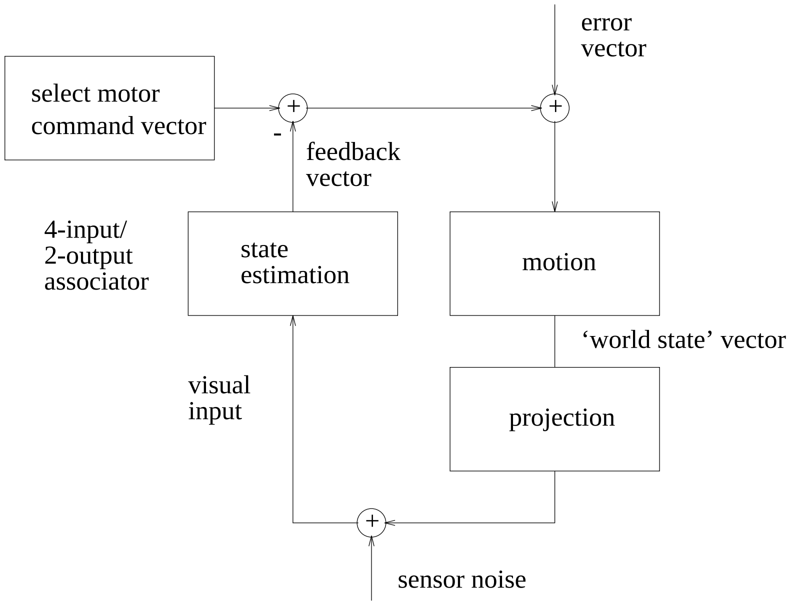

Figure 21: System used in test of visual learning and control.

variance of the associator output also increases. Both factors lead to an increase in the number of training cycles needed to detect the maximum reliably. The conclusion is that, while it is possible to acquire a map from noisy position estimates, the less noise there is the better - an active system, with direct motion information, is likely to be markedly superior to one which relies on velocities estimated from visual input.

A final example combines these ideas in the context of a control problem which requires 'visual' learning. The system is illustrated in Fig. 21, which bears an obvious resemblance to the state controller of Fig. 18. The transformations and state space are taken from the first example: there are exactly four world states, related to each other by the row and column swaps. The 'visual input' consists of a projection onto $3-D$ of the $4-D$ state (cf. (48)). The transformation of the state vector is controlled by a $2-D$ binary vector, which in general is the sum $(\bmod 2)$ of the motor command vector, the feedback vector and an error vector representing dynamical disturbances. On the nth iteration, the motor command vector, $\boldsymbol{c}(n)$, is selected randomly from the set of four binary vectors, with equal probability for each of the four. The disturbance vector, $\boldsymbol{e}(n)$, is also selected at random, with errors in the two vector components being independent and equiprobable, to give a specified error rate, giving a motor vector

$$\boldsymbol{r}(n) = \boldsymbol{c}(n) + \boldsymbol{e}(n) \tag{62}$$

where the addition is mod2. The overall control task is to control the configuration of the system, so that the world state corresponds to the one selected by the motor

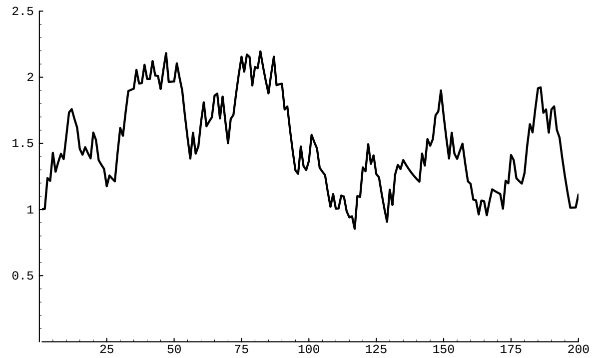

Figure 22: Average squared error for open loop system.

command, despite the effect of the errors. To demonstrate that this is not a suitable problem for an open loop system, Fig. 22 shows the squared error between the desired world state and the actual state, averaged with a simple recursive filter, for the open loop system with an error rate of 0.1. This is not an acceptable performance. In order to achieve satisfactory performance, the state must be estimated from the visual input and the error used to initiate a corrective motion, using a state estimate which is a binary vector of the form

$$\boldsymbol{\phi}(n) = \boldsymbol{f}(\boldsymbol{I}_3\boldsymbol{\psi}(n)) \tag{63}$$

where $\boldsymbol{\psi}(n) = \boldsymbol{T}_{r(n)}\boldsymbol{v}$ is the world state at time n and is determined by the transformation selected by the motor vector $\boldsymbol{r}(n)$. As before, the vector $\boldsymbol{v}$ is selected at random at the start of the test. The state estimate $\boldsymbol{\phi}(n)$ is used to derive a correction vector $\boldsymbol{\gamma}(n)$ by subtracting it from the desired state $\boldsymbol{c}(n)$

$$\boldsymbol{\gamma}(n) = \boldsymbol{c}(n) + \boldsymbol{\phi}(n) \tag{64}$$

The correction is then added to the motor vector $\boldsymbol{r}(n)$. It is not hard to see that, if the nonlinear associator $\boldsymbol{f}(.)$ produces a reliable estimate of the actual state $\boldsymbol{r}(n)$,

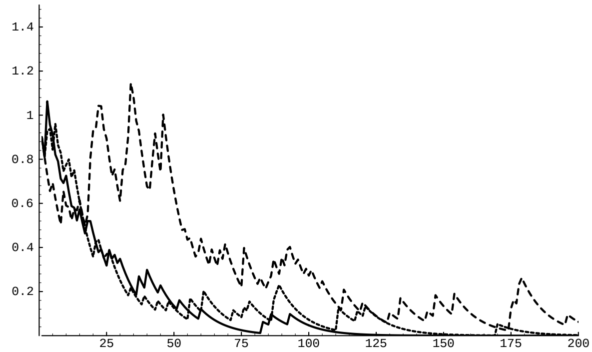

Figure 23: Average squared error for feedback system during learning for error rates of 0.1 (solid), 0.2 (dotted) and 0.4 (dashed).

then the correction of (64) will lead to the desired final state

$$\boldsymbol{c}(n) = \boldsymbol{r}(n) + \boldsymbol{\gamma}(n) \tag{65}$$

The visual learning task is, therefore, to associate each input vector with the corresponding motor command vector, which requires a 4-input 2-output associator (cf. (45)) which is robust to errors - in effect, the environment is acting as a noisy supervisor. Note that, in training, it is necessary to replace the discontinuous nonlinearity of $\boldsymbol{f}(.)$ by a sigmoid function. To show that such learning is effective, Fig. 23 shows the averaged error, but this time *after* application of the feedback correction of (63),(64). The result is that even at the higher error rates, the system still learns to control its configuration, regardless of the projection from $4-D$ to $3-D$ and at low error rates, learning is rapid. Fig. 24 shows that even the addition of normally distributed sensor noise at a SNR of 10dB leaves the system approximately preserving the necessary invariance and equivariance.

The approximate invariance and equivariance are illustrated in Fig. 25, which shows the response of the associator to the four transformations after training with

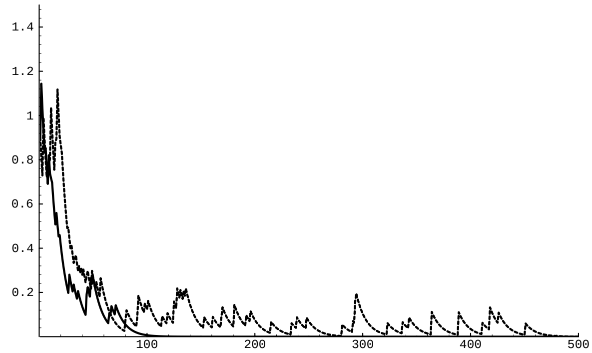

Figure 24: Average squared error for feedback system during learning, for error rate of 0.1 and no sensor noise (solid) or sensor noise with signal-noise ratio of 10 (dotted).

noisy data and an error rate of 0.1. In this figure, each point represents the two output components in response to input patterns distorted by noise at an SNR of 10dB, with the hard limiting nonlinearity replaced by a logistic function (as in training). Although the noise scatters the reponses of both units, 94.5% of responses lie in the correct quadrant, giving the correct system response. In this case, the output shows invariance and equivariance at the level of 10%, which mirrors the input SNR.

While this problem is tiny in scale compared with any realistic visual control problem and the dynamics have been greatly simplified, the example brings together the various elements discussed above into a structure which is directly applicable to large scale problems and shows that action can help visual learning.

6 Conclusions

In this paper, we have tried to tackle the perennial problem of finding a meaningful and useful definition of symbols in vision. We explored some of the conventional

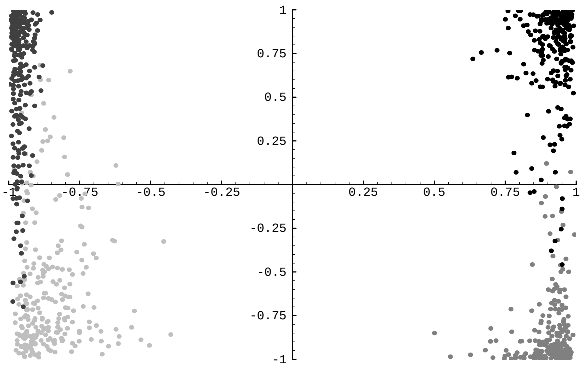

Figure 25: Response of 4-input 2-output associator after training to row and column swaps, showing the separation between the responses to the four transformations.

ideas of symbols and showed that none of them is completely satisfactory. As an alternative, we have proposed an *operational definition* based on symmetry properties and shown that it can lead to insights into the problem of inferring the state of the 'world' from a necessarily limited set of measurements. We have introduced the concept of *equivariance* to emphasise that a symbol, to be meaningful, must stand in a well defined relation to the state of the world, which can be best described in terms of the action of the transformations which affect visible objects. We gave a simple example to illustrate that active vision has the potential for overcoming the fundamental limitations imposed by the projection operations which are inherent in vision and which inevitably destroy the symmetries which the $3-D$ world manifests. Despite its obvious simplicity, this example is formally identical, in terms of the destruction of symmetry caused by windowing, to that which occurs when objects in translational motion are seen through a spatially limited window such as the eyes. That example was also used to introduce the idea that we regard a change of object as a transformation having the same type of decription as motion. While this is a less familiar idea than that of motion, it is inescapable in a state-space description, whether that operates at the level of signals or objects. Indeed, it would seem that the only consistent way to separate changes of object from changes of object position is that in the latter case, but not in the former, it is possible to restore the original view of the object by a motion of the observer. We are tempted to speculate that our commonsense model of the world, inhabited by objects obeying the laws of motion, cannot be derived by vision alone, for the reason that the motions have no simple representation in the signal space; it is only when a system is able to *act* upon its environment that the natural symmetry is restored. This seems to us to be about as close as one can get to explaining why there is an apparent gulf between traditional symbolic methods, which are based on the 'Newtonian' ideas of identity and motion, and connectionist methods, which are bound to the 'quantum' world of the signal space. Clearly, both types of description are necessary and each has a proper place within the representations employed by an AVS. What is also clear is that trying to use 'classical' methods in a 'passive' vision system is an enterprise which is likely to fail. The experience of the last thirty years of vision research supports this view.

We went on to propose a state-space model of active vision, in which visual signals provide feedback, leading to the view of perception as *state estimation*. The problem of map-building was used to illustrate this idea and to show that even in this essentially perceptual task, active systems can have a considerable advantage, in terms of learning time. Moreover, it could be argued that many learning tasks can be modelled in this way - the ability to 'navigate around' some conceptual domain is an appealing metaphor for that which we commonly call knowledge. As a final example, these ideas were put together in a simple visual control task, which showed that the learning of a state estimator based on visual input is effective in motor control: the system learns and uses a symbolic representation of the world state to control movement in the presence of errors. In this connection, it is interesting to speculate on what vision has to tell us about the more general use of symbols. It is clear that concepts such as invariance and equivariance make sense in an essentially geometric domain, such as vision, but their usefulness in more general contexts seems questionable: could they help to explain what is going on in the Chinese Room ? It may be that, just as we were obliged in the Active Vision framework, to introduce a world model into our analysis, so as that world model becomes more complex, there will be symbols of a more abstract nature definable in the same terms as those we have discussed. We are in no position to make general prognostications on this subject, but perhaps it is appropriate to quote someone who gave some thought to such matters - Wittgenstein: "A *name* signifies only what is an *element* of reality. What cannot be destroyed; what remains the same in all changes." [78].

References

[1] D. Marr, *Vision*, San Francisco, Freeman, 1982.

[2] A.R. Hanson and E.M. Riseman, "From Image Measurements to Object Hypotheses", *COINS Tech. Rept. 87-129*, Univ. of Mass., 1987.

[3] R.A. Brooks, "Intelligence without Reason", *Proc. 12th Int'l. Joint Conf. on AI*, 1991.

[4] P.M. Churchland and P.S. Churchland, "Could a Machine Think ?", *Sci. Amer.*, 262, pp. 26-33, Jan. 1990.

[5] J.R. Searle, "Is the Brain's Mind a Computer Program ?", *Sci. Amer.*, 262, pp. 20-25, Jan. 1990.

[6] R. Bajcsy, "Active Perception", *Proc. IEEE*, 76, 8, pp. 996-1005, Aug. 1988.

[7] J. Aloimonos, I. Weiss and A. Bandopadhyay, "Active Vision", *Int'l. Jnl. Comput. Vision*, 2, 1988.

[8] E. Krotkov, *Exploratory Visual Sensing for Determining Spatial Layout with an Agile Camera System*, Ph.D. Thesis, Univ. of Pennsylvania, 1987.

[9] H.G. Barrow and J.M. Tenenbaum, "Computational Vision", *Proc. IEEE*, 69, pp. 572-595, 1981.

[10] S. Coren, C. Porac and L.M. Ward, *Sensation and Perception*, ch. 1, New York, Academic Pr., 1979.

[11] G.W. Humphreys and V. Bruce, *Visual Cognition*, London, Lawrence Erlbaum, 1989.

[12] R. Wilson and H. Knutsson, "Uncertainty and Inference in the Visual System", *IEEE Trans. Sys. Man Cybern.*, 18, pp. 305-311, April 1988.

[13] A. Rosenfeld, "Computer Vision: Basic Principles", *Proc. IEEE*, 76, pp. 863-868, Aug. 1988.

[14] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images", *IEEE Trans. Patt. Anal. and Machine Intell.*, 6, pp. 721-741, Nov. 1984.

[15] M. Bertero, T.A. Poggio and V. Torre, "Ill-Posed Problems in Early Vision", *Proc. IEEE*, 76, pp. 869-889, Aug. 1988.

[16] R.A. Hummel and S.W. Zucker,"On the Foundations of Relaxation Labelling Processes",*IEEE Trans. Patt. Anal. and Machine Intell.*, 5, pp.267-287, 1983.

[17] B.A. Draper, R.T. Collins, J. Brolio, A.R. Hanson and E.M Riseman,"The Schema System",*Int'l. Jnl. Comput. Vision*, 2, pp. 209-250, 1989.

[18] S. Zeki,"The Visual Image in Mind and Brain",*Sci. Amer.*, 267, pp.42-51, Sept. 1992.

[19] L.G. Roberts,"Machine Perception of Three-Dimensional Solids", in *Optical and Electro-Optical Inform. Process.*, ed. J.T. Tippet, Cambridge, MIT, 1965.

[20] D.H. Ballard,"Generalizing the Hough Transform to Detect Arbitrary Shapes", *Patt. Recogn.*, 13, pp.111-122, 1981.

[21] J. Princen, J. Illingworth and J. Kittler,"A Hierarchical Approach to Line Extraction Based on the Hough Transform",*Comput. Vision, Graph. and Image Process.*, 52, pp.57-77, 1990.

[22] E.P. Simoncelli, W.T. Freeman, E.H. Adelson and D.J. Heeger,"Shiftable Multiscale Transforms",*IEEE Trans. Inform. Th.*, 38, pp.587-607, Mar. 1992.

[23] S. Mallat and W.L. Hwang,"Singularity Detection and Processing with Wavelets", *IEEE Trans. Inform. Th.*, 38, pp.617-643, Mar. 1992.

[24] R. Wilson, A.D. Calway and E.R.S. Pearson,"A Generalized Wavelet Transform for Fourier Analysis: the Multiresolution Fourier Transform and its Application to Image and Audio Analysis",*IEEE Trans. Inform. Th.*, 38, pp.674-690, Mar. 1992.

[25] H. Knutsson,"Representing Local Image Structure Using Tensors",*Proc. 6th Scand. Conf. on Image Anal.*, pp. 244-251, Finland, 1989.

[26] R.C. Gonzalez and P. Wintz,*Digital Image Processing*, 2nd Ed., Reading, Addison-Wesley, 1987.

[27] G. Cybenko,"Approximation by Superpositions of a Sigmoid Function", *Mathematics of Control, Signals and Systems*, 2, pp.303-314, 1989.

[28] Y. Ito,"Approximation of Functions on a Compact Set by Finite Sums of a Sigmoid Function without Scaling",*Neural Networks*, 4, pp.817-826, 1991.

[29] K. Hornik, M. Stinchcombe and H. White,"Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks",*Neural Networks*, 3, pp. 551-560, 1990.

[30] W.H. Pitts and W.S. McCulloch,"How We Know Universals: the Perception of Auditory and Visual Forms",*Bull. Math. Biophysics*, 9, pp.127-147, 1947.

[31] D. Ingle,"Visual Releasers of Prey-catching Behaviour in Frogs and Toads", *Brain, Behaviour and Evol.*, 1., pp.500-518, 1968.

[32] J-P. Ewert,"The Neural Basis of Visually Guided Behaviour", in *Recent Progress in Perception*, ed. R. Held and W. Richards, San Francisco, Freeman, 1976.

[33] A. Pellionisz and R. Llinas,"Tensor network Theory of the Metaorganization of Functional Geometries in the Central Nervous System",*Neuroscience*, 16, pp.245-273, 1985.

[34] M.A. Arbib,"Modelling Neural Mechanisms of Visuomotor Co-ordination in Frog and Toad",*COINS Tech. Rept. 82-1*, Univ. of Mass., 1982.

[35] D.B. Tweed and T. Vilis,"The Superior Colliculus and Spatiotemporal Translation in the Saccadic System",*Neural Networks*, 3, pp.75-86, 1990.

[36] R.A. Brooks,"A Hardware Retargetable Distributed Layered Architecture for Mobile Robot Control",*Proc. IEEE Int'l. Conf. on Robotics*, pp. 106-110, 1987.

[37] D.E. Rumelhart, J. McClelland and the PDP group,*Parallel Distributed Processing vol. 1*, Cambridge, MIT Pr. 1987.

[38] J.J. Hopfield,"Neurons with Graded Responses Have Collective Computation Properties like Those of Two-State Neurons",*Proc. Nat. Acad. Sci.*, pp.3088-3092, 1984.

[39] G.H. Granlund and H.E. Knutsson,"Contrast of Structured and Homogeneous Representations", in *Physical and Biological Processing of Images*, Ed. O.J. Braddick and A.C. Sleigh, Berlin, pp.282-303, 1983.

[40] F. Ratliff,"Contour and Contrast", in *Recent Progress in Perception*, ed. R. Held and W. Richards, San Francisco, Freeman, 1976.

[41] R.N. Haber and M. Hershenson,*The Psychology of Visual Perception*, 2nd Ed., ch. 4, New York, Holt, Rhinehart and Winston, 1980.

[42] R. Linsker,"Self-Organization in a Perceptual Network",*IEEE Computer*, pp. 105-117, 1988.

[43] T.D. Sanger,"Optimal Unsupervised Learning in a Single Layer Linear Feedforward Neural Network",*Neural Networks*, 2, pp. 459-473, 1989.

[44] J.G. Daugman,"Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression",*IEEE Trans. Acous., Speech and Signal Process.*, 36, pp.1169-1179, 1988.

[45] T. Berger,*Rate-Distortion Theory*, Englewood Cliffs, Prentice Hall, 1971.

[46] A.K. Jain,*Fundamentals of Digital Image Processing*, Englewood Cliffs, Prentice Hall, 1986.

[47] P.J. Burt and E.H. Adelson,"The Laplacian Pyramid as a Compact Image Code", *IEEE Trans. Commun.*, 31, pp.532-540, 1983.

[48] C. Enroth-Cugell and J.G. Robson,"The Contrast sensitivity of Retinal Ganglion Cells of the Cat",*Jnl. Physiol.*, 187, pp.517-522, 1966.

[49] D. Gabor,"Theory of Communication",*Proc. IEE*, 93, pp.429-441, 1946.

[50] S. Marcelja,"Mathematical Description of the Responses of Simple Cortical Cells",*Jnl. Opt. Soc. Amer.*, 70, pp.1297-1300, 1980.

[51] G.H. Granlund,"In Search of a General Picture Processing Operator",*Comput. Graph. and Image Proc.*, 8, pp.155-173, 1978.

[52] M.S. Livingstone,"Artt, Illusion and the Visual System",*Sci. amer.*, 258, pp.68-75, 1988.

[53] C.E. Shannon,"Prediction and Entropy of Printed English",*Bell Sys. Tech. Jnl.*, 30, pp.50-64, 1951.

[54] J. Horgan,"Profile: Claude E. Shannon",*Sci. Amer.*, 262, pp.16-17, 1990.

[55] E. von Holst and H. Mittelstadt,"Das Reafferenzprincip",*Naturwissenschaften*, 37, pp.464-476, 1950.

[56] A. Hein,"The development of visually guided behavior",in *Visual Coding and Adaptability*, Ed. C. Harris, Hillsdale, Lawrence Erlbaum, pp.51-68, 1980.

[57] R. Held and A. Hein,"Movement-produced stimulation in the development of visually guided behavior",*Jnl. of Comparative and Physiological Psychology*, 56, pp. 872–876, 1963.

[58] A. Hein and E.C. Gower,"Development and segmentation of visually controlled movement by selective exposure during rearing",*Jnl. of Comparative and Physiological Psychology*, 73, pp.181-187, 1970.

[59] R. Held and A. Hein,"Adaptation of disarranged hand-eye coordination contingent upon re-afferent stimulation",*Perceptual and Motor Skills*, 8, pp.87-90, 1958.

[60] H.L. Pick and J.C. Hay," passive test of the Held reafference hypothesis", *Perceptual and Motor Skills*, 20, pp.1070-1072, 1965.

[61] R. Held and J. Bossom,"Neonatal deprivation and adult rearrangement. Complementary techniques for analyzing plastic sensory-motor coordinations", *Jnl. of Comparative and Physiological Psychology*, pp.33-37, 1961.

[62] G. Mikaelian and R. Held,"Two types of adaptation to an optically-rotated visual field",*American Jnl. of Psychology*, 77, pp.257-263, 1964.

[63] A. Newell,"Physical Symbol Systems",*Cognitive Science*, 4, pp.135-183, 1980.

[64] S.D. Whitehead and D.H. Ballard,"Learning to Perceive and Act",*Tech. Rept. 331*, Univ. of Rochester, June 1990.

[65] M. Golubitsky, I. Stewart and D.G. Schaeffer,*Singularities and Groups in Bifurcation Theory, vol. II*, Berlin, Springer-Verlag, 1988.

[66] S. Grossberg and M. Kuperstein,*Neural Dynamics of Adaptive Sensory-Motor Control*, Oxford, Pergamon, 1989.

[67] M. Kuperstein,"Neural Model of adaptive Hand-Eye Coordination for Single Postures",*Science*, 239, pp.1308-1311, 1988.

[68] D. Bullock and S. Grossberg,"VITE and FLETE: Neural Models for Trajectory Formation and Postural Control", in *Volitional Action*, Ed. W.A. Hershberger, Amsterdam, Elsevier, 1989.

[69] H. Gomi and M Kawato,"Recognition of Manipulated Objects by Motor Learning with Modular Architecture Networks",*Neural Networks*, 6, pp. 485-497, 1993

[70] G.F. Franklin and J.D. Powell,*Digital Control of Dynamic Systems*, Reading, Addison-Wesley, 1980.

[71] T. Miller, R.S. Sutton and P.J. Werbos (Eds.),*Neural Networks for Control*, Cambridge, MA, MIT Press, 1990.

[72] J. Sietsma and R.J.F. Dow,"Creating Artificial Neural Networks that Generalize",*Neural Networks*, 4, pp.67-79, 1991.

[73] A. Papoulis,*Signal Analysis*, New York, McGraw-Hill, 1977.

[74] D. Slepian,"Prolate Spheroidal Wavefunctions, Fourier Analysis and Uncertainty - V: the Discrete Case",*Bell Sys. Tech. Jnl.*, 57, pp.1317-1430, 1978.

[75] R. Wilson and G.H. Granlund,"The Uncertainty Principle in Image Processing", *IEEE Trans. Patt. Anal. and Machine Intell.*, 6. pp. 758-767, 1984.

[76] W.C. Hoffman,"The Lie Algebra of Visual Perception",*Jnl. Math. Psychol.*, 3, pp.65-98, 1966.

[77] W.C. Hoffman,"Higher Visual Perception as Prolongation of the Basic Lie Transformation Group",*Mathematical Biosciences*, 6, pp.437-471, 1970.

[78] L. Wittgenstein,"Philosophical Investigations", London, Blackwell, 1958.

Figure Captions

1. Structure of Symbolic Vision Paradigm.
2. Simulated 'cockroach' locates the darkest point in its 2-D world. White crosses denote points on its trajectory, which are connected by the black line.
3. Behaviourist Approach to Vision.
4. Continuous functional approach to the cockroach problem.
5. The cockroach gets trapped in a 'black hole'.
6. By using a simple 'black hole avoidance' procedure, the smart cockroach avoids the traps.
7. Symbols can be used to switch between modes of a linear system to model arbitrary nonlinear behaviour.
8. The continuous cockroach is less troubled by black holes, but they do upset its trajectory.
9. Avoidance procedure improves the continuous cockroach performa nce.
10. Rate-distortion curve for the gradient estimator used in the cockroach problem. The distortion function is the additional path length travelled by a cockroach using a quantized gradient. The information rate is the entropy of the quantised gradient.
11. Transformations in vision. Changes in the 'world' state induce changes in the image, to which the system responds. Vision must solve the inverse problem of identifying the change in world state from that in the image.
12. Transformations in active vision. Changes in the 'world' state induce changes in the image, to which the system responds, inducing a consequent transformation in world state.
13. A four component pattern and transformations formed by row and column swaps.
14. Eigenvector estimates found by the Sanger algorithm, closely matching the eigenvectors of the covariance matrix.
15. Weight vectors for the two unit perceptron after training. The weights closely match the two relevant eigenvectors. The left image is the weight vector for the 'what' unit and the right that for the 'where' unit.
16. Eigenvector estimates found by the Sanger algorithm, after windowing to three nonzero components.
17. Output error plotted against training cycle number for the 4-input (dashed) and 3-input (solid) pereceptrons.

18. Adaptation of state-space control system to active vision. Note: functions labelled s are subcortical; c are cortical.
19. Error as a function of training cycle number for various noise standard deviations.
20. Convergence time vs. noise s.d.
21. System used in test of visual learning and control.
22. Average squared error for open loop system.
23. Average squared error for feedback system during learning for error rates of 0.1 (solid), 0.2 (dotted) and 0.4 (dashed).
24. Average squared error for feedback system during learning, for error rate of 0.1 and no sensor noise (solid) or sensor noise with signal-noise ratio of 10 (dotted).
25. Response of 4-input 2-output associator after training to row and column swaps, showing the separation between the responses to the four transformations.