PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Adaptive Control of a Four-Degrees-Of-Freedom Stereo Camera Head [and Discussion]

John E. W. Mayhew, Ying Zheng, Stuart Cornell and V. Torre

Phil. Trans. R. Soc. Lond. B 1992 **337**, 315-326
doi: 10.1098/rstb.1992.0109

The adaptive control of a four-degrees-of-freedom stereo camera head

JOHN E. W. MAYHEW, YING ZHENG AND STUART CORNELL

AI Vision Research Unit, University of Sheffield, Sheffield S10 2TN, U.K.

SUMMARY

The paper describes the use of biologically plausible neural network architectures to address some of the issues associated with the use of stereopsis under variable camera geometry. We report an implementation of a layered (subsumption) architecture for the adaptive control of microsaccadic tracking, and show experimental results demonstrating the use of lattice filter predictors for trajectory modelling. A rather simple, but seemingly adequate, neural network architecture for representing high-dimensional surface approximations (PILUTS) is evaluated as a method of encoding the predictive stereo mapping of the ground plane for different head positions.

1. INTRODUCTION

The research described here is part of a project investigating adaptive control of visual tracking and maintenance of fixation of a stereo camera rig† mounted on an autonomous vehicle. The research has strived for both psychological and physiological plausibility in both the specification of the particular competences involved, and the adaptive self-tuning methodologies used for their real-time implementation. Only two components of the research will be described in detail. First, the design of a module for the control of microsaccades‡ for tracking a moving target while the vehicle is also moving. Second, a neural net architecture for solving some problems associated with the use of stereopsis for ground plane obstacle detection under conditions of variable camera geometry.

Before turning to the two specific topics, the general nature of the project as a whole will be outlined to provide a context. This will be done by describing the main competences being sought, followed by some general remarks on the means sought for implementing them.

2. REQUIRED VISUAL MOTOR COMPETENCES

1. Fast eye saccades to fixate and start tracking the target, with concurrent slow head motion to centre the cameras, integrated with compensatory camera movements to null off the vehicle motion.

2. Compensatory head and camera movements that maintain fixation uner conditions of self-induced vehicle motion.

3. Automatic shunting of tracking motion from the camera vergence system to the head pan as the tracked target reaches the limits of the vergence range, augmented with the centring reflex above.

4. Feedforward predictive target trajectory modelling to provide zero-lag tracking control capable of maintaining fixation of a target under smooth motion, and yet capable of a fast adaptive reaction to discontinuities of target motion and other (statistically) non-stationary trajectories.

In all the above primitive reflexive behaviours the compensatory motions involve nonlinear dependencies on the current camera-head-vehicle system state.

† The stereo camera rig used for this work comprises a three-link kinematic chain, whose degrees of freedom are rotations around the following axes: (i) pan: a vertical axis corresponding to the 'neck'; (ii) tilt: an axis at right angles to the neck; and (iii) verge: each camera ('eye') can rotate independently around an axis at right angles to the tilt axis. The rig has been constructed so that the centres of rotation of the tilt and pan links coincide, and the centres of rotation of left and right verge and the tilt links coincide. It has been a principle of the project not to use measurements of the geometry of the 'head' either in the control of the head or in the development of the predictive stereo matcher to be described below. The length of the tilt link is approximately 12.5 cm for each eye (i.e. the head is about 25 cm wide); the length of the verge link (i.e. approximately how far the centre of rotation is from the focal centre of the camera) is 5 cm so that tilting the eye also produces a small translation. It is also of note that the right camera has been mounted with a 5 degree heterophoria and about 2.5 degrees of cyclotorsion. Stepper motors control the head and give a maximum saccade velocity of 50 deg s^{-1}.

‡ Microsaccades are generally used to refer to the very small saccades which, if they have any function at all, may be used to correct errors arising from drift during fixation of a stationary target (Carpenter 1988). We use the term microsaccadic tracking to describe a form of tracking which uses small vergence saccades (ranging in size from a few minutes of arc to two degrees), characterised by a fast movement stage, followed by a 80 ms image capture stage during which the eyes remain stationary. In humans, this form of tracking may not normally occur in isolation but seems to be an important component of pursuit movements (Carpenter 1988, p. 55).

Simple linear controllers are inadequate, although they can be exploited to provide the training data for adaptive predictive feedforward control and the basic starting competence for the system. Also, of course, they provide a fail-safe backup system.

A pilot system has been implemented that exhibits the above competences and some results from this system will be described here. In its continuing development, an important goal is to maintain stability and coherence between the various activities involved. Thus, for example, after the cameras have saccaded to fixate and start tracking an object, the slower moving head must be able to turn to 'centre' the cameras, but in doing so the compensatory camera motion signals must not interfere with the on-going tracking behaviour. Nor should any re-orienting of the vehicle interfere with the tracking, even though it requires compensatory motion on all the degrees of freedom of the camera-head system.

3. IMPLEMENTATION STRATEGIES

(*a*) *The development of self-tuning methods for recovering the eye-to-head transformation and the head-to-vehicle transformation*

These are necessary for the integration of geometry from multiple camera scans into a coherent spatial map of the scene, and for coherent control of the vehicle-head-camera system. A system for the maintenance of on-line calibration of the camera-head-vehicle (Thacker & Courtney 1992), and a neural network architecture for learning the inverse kinematics of the stereo camera head (Dean *et al.* 1991) have been described elsewhere.

(*b*) *The implementation of a high-level control architecture for the maintenance of selective attention and fixation strategies*

This is being achieved by developing modules within an architecture (ANIT: Architecture for Navigation and Intelligent Tracking) which takes as its starting point the subsumption ideas of Brooks (1986, 1991) but in which the encapsulated processes may implement rather more complex algorithms than those he uses in his modules. Nevertheless, the underlying idea is similar: to provide as far as possible an homogeneous inter-module communication structure to facilitate the wiring together of essentially autonomous concurrent behavioural processes. The philosophy is that the architecture should support behavioural parallelism over diverse spatial and temporal scales. This work is described elsewhere (Mayhew 1992; see also Abraham *et al.* 1991).

(*c*) *The exploitation of artificial neural net architectures as an implementation strategy*

By using and developing physiologically plausible network architectures, we believe that general principles of learning and adaptation in biological systems can be developed and evaluated.

4. CONTROL OF SACCADES

We have implemented a three-layered architecture for the control of the stereoscopic eye-saccade system of a stereo-camera head mounted on an autonomous vehicle. This system is shown in figure 1 as a functional block diagram. The three layers are as follows.

(*a*) *Level 0*

Level 0 is a proportional feedback controller enabling the head to foveate and track targets but requiring iteration through the vision system with the attendant heavy image processing overhead. The latter processing in the current implementation is a simple centre-of-gravity blob tracker. This rather crude level of image processing is driven by the real-time demands of the task and current equipment constraints. The image capture has three modes.

1. Tracking: a 3.75 degree square (64 × 64 pixels) 'foveal' region of interest (ROI) is used when a target has been located and is being tracked. The real-time demands of the task are evident from the fact the location of a small target in this fovea takes at least 20 ms (and more depending on the size of the blob).

2. Recovery: a 7.5 degree square ROI is used to relocate the target when it is temporarily lost during tracking.

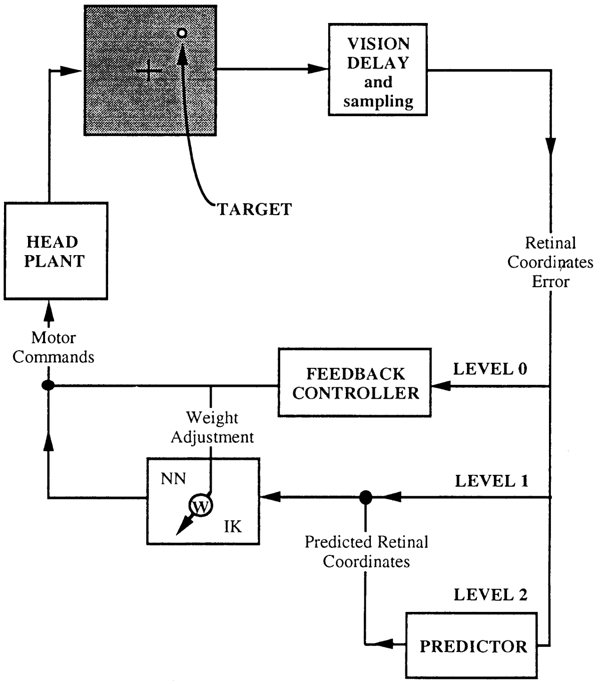

Figure 1. The philosophy of subsumption applied to a perceptual-motor task: three layers of competences for a microsaccadic tracking system able to maintain zero fixation error. Level 0 is the basic competence, a proportional feedback controller providing a stable starting point that is enhanced by the addition of two further layers of visuomotor competence. Level 1 subsumes the Level 0 competence, and improves it to provide a single saccade to achieve fixation of a stationary target. Level 2 subsumes both the lower level competences and augments them by providing zero fixation under conditions when the target is moving. See text for details.

3. Initialization: the full 30 degree square image (512×512) is used to locate the target at the start of tracking.

In the tracking mode, image processing is done concurrently and independently in the two images; in the recovery and initialization modes a sub-sampling strategy is used to locate a target in one image and then focus the search around the corresponding point in the other image.

The details of the implementation are unimportant but a principle may be worth elaborating. The real world is not the place to be 'lost in thought' analysing what a target might be before trying to track it. A tracking competence working on primitive, fast and even crude vision processing can provide the 'temporal glue' by which 'the thing you saw then is the object you recognize now'. Thus during tracking the foveal ROI is distributed as a continuous stream to another image processing system, completely independent of the tracking system, which samples the image stream at a very different and much slower rate. Currently this system is used only to display the images, but the direction of future system evolution is obvious.

(*b*) *Level 1*

This layer provides the feedforward inverse kinematics for saccadic eye movements allowing a ballistic movement to replace the 0-level control loop. The training data are provided by the feedback error signal from the 0-level controller.

It may seem a reasonable assumption that the angular rotation required to fixate is a simple linear function of the retinal distance of the target from the fovea. This is true under the conditions of Listing's Law (a special case of Donder's Law) but need not be the case for other rotation schemes. In the system used by our mechanical head, the motor commands for the saccade to fixate a target depend not only on the retinal coordinates in the eyes but are also dependent on the tilt and gaze angle of the eyes. The fact that changing the tilt of the eyes changes the orientation of the verge axes in three-dimensional world coordinates means that there is a nonlinear relationship between the x and y coordinates of a target and the rotations around the verge and tilt axes necessary to fixate it. This is particularly marked when the eyes are asymmetrically verged. Furthermore, this non-linearity is not simply a function of the eye-head geometry, but is also a function of the retinal coordinates.

We use neural network architectures (Dean *et al.* 1991; Mayhew *et al.* 1992) that learn to correct errors arising from a simple proportional controller of head position by exploiting an idea from Kawato *et al.* (1989). This is to exploit the error signal from a simple proportional controller as the training feedback for the neural networks. Though this principle is commonplace in the application of adaptive systems, Kawato and his colleagues have argued that it also occurs quite frequently in physiological systems. The desirability of generating an accurate ballistic action or a predictive control response is that it provides an adaptive advantage in those cases where iteration around a cortical control loop would provide unacceptable delay. Furthermore, the crude proportional controller not only provides the method for training but also a backup system should it be required.

(*c*) *Level 2*

This layer is an adaptive lattice filter which is used to track moving targets. The filter is trained using error feedback from previous saccades within the current tracking sequence, so that the filter learns to predict the future target position in the next image. This is used by the inverse kinematics module to generate the eye movement commands for the appropriate predictive succade.

For the Level 2 controller, we wished to develop a tracking prediction module with the following properties: (i) it should be as general as possible, making minimal assumptions about the complexity and stationary of the target trajectory; (ii) it should adapt in very few time steps or samples, both to the onset of motion and to any discontinuities in the trajectory, yet at the same time it should be robust over sequences of missing data such as frequently result from occlusions and low level image processing infelicities; and (iii) the implementation of the predictor should be not only computationally inexpensive, but also biologically plausible.

The use of multi-stage lattice filters (Goodwin & Sin 1984; Alexander 1986) for prediction is commonplace especially in the speech processing domain (Makoul, 1975). The general principle underlying their design is that the successive stages of the filter compute the partial correlations (or regressions) at different delays. We have explored several different adaptive algorithms for doing this. As expected, we found gradient methods of training inefficient compared to recursive data projection algorithms. However, an alternative method has been implemented (figure 2) that calculates the reflection coefficients of the lattice filter directly using a decaying running average of the smoothed partial correlations. By controlling the time constant of the estimator of the partial correlations, both the requirements of fast adaptive response and relative robustness to missing data can be satisfied. Because successive stages of the filter are orthogonal it is easily adapted on-line to the complexity of the signal by the simple expedient of adding or deleting stages of the lattice in response to variations in the partial correlations of the last stage (details in Zheng *et al.* 1992; results are shown here in figure 3).

An interesting and perhaps revealing issue is the biological plausibility of the lattice filter. Inspection of figure 2 shows that a single stage looks almost exactly like the Barlow & Levick (1965) model of a motion detector. There are two lines: one carries a delayed version of the other's input. Between them is a simple subtractive comparator. The additional feature, over and above the Barlow & Levick scheme, is a correla-

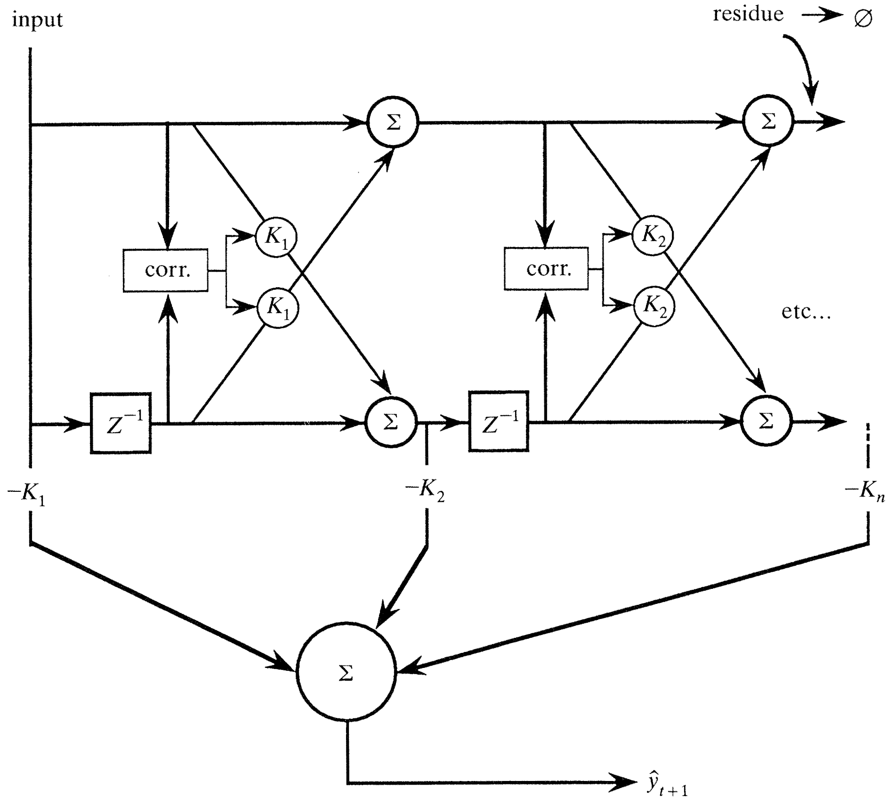

Figure 2. The lattice filter implementation of a transversal filter. The filter uses the current and past observations to form a prediction of the future output. It makes no assumption about the signal being linear and finite dimensional, it simply uses an auto-regressive model for the structure of the filter (which may not be optimal) and chooses the coefficients to minimize the mean square prediction error. K_1, K_2, etc. are the reflection or partial correlation coefficients computed between the top and bottom 'delay' lines. K_1 is generally negative, and initialized to -1 to give rapid convergence. Z^{-1} is a delay operator. Successive stages are delayed by increments of the sampling interval. The one step ahead prediction is given by summating the negated reflection coefficients K_1, K_2 . . . K_n.

tor for adjusting the gain of the subtraction operation, which raises the question as to whether a similar process might be implemented in biological systems. All components of each stage of the lattice filter would seem to be readily capable of implementation in real neurons.

We have evaluated the lattice filter predictor in several modes.

1. Relative mode: visual target prediction using fixation error feedback. The filter was used to generate predictions of the future retinal coordinates of the target with respect to the fovea. The prediction was then used (via the kinematics) to move the head to a position which nulled off the predicted retinal error. The predictor has no access to the actual target trajectory but must estimate it in the context of its own saccades and the measured retinal errors. Four independent filters are used, one to track each of the retinal coordinates of the target in the left and right images.

2. Absolute mode: motor state prediction using fixation error feedback. The filter was used to generate predictions of the future motor states which would foveate the target. The predictor has access to the absolute motor states at which the image was taken, and the error measured in retinal coordinates is converted via the inverse kinematics to motor commands. Three filters are required: one for each of the verge motors and the other to control the tilt.

3. Image capture modes: serial and pipeline. The above tracking task can be broken into the following four stages: image capture, image processing, prediction and inverse kinematics, and head motion.

Pipelining is a form of parallelism which is appropriate when a repetitive activity consists of a sequence of stages. The strategy is to overlap the processing of the stages so that while stage n is being processed, stages $n-1$ and $n-2$, etc. of successive instances of the action are processed concurrently. Figure 4 shows how it is possible to pipeline the components of the microsaccadic tracking task.

The advantage of pipelining is clear: it increases through-put of a processing stream. Here, the important difference from serial processing is that in pipeline mode the target-locked image sampling frequency is maximised. Furthermore, while maintaining the same sampling frequency or image capture rate, it is possible to treble the amount of time available for the image processing and inverse kinematic stages. Also, because the number of head motion stages has doubled, the maximum target velocity can be increased proportionately. From this it follows that a pipeline tracker is much less vulnerable to variations in processing demands than a tracker operating in

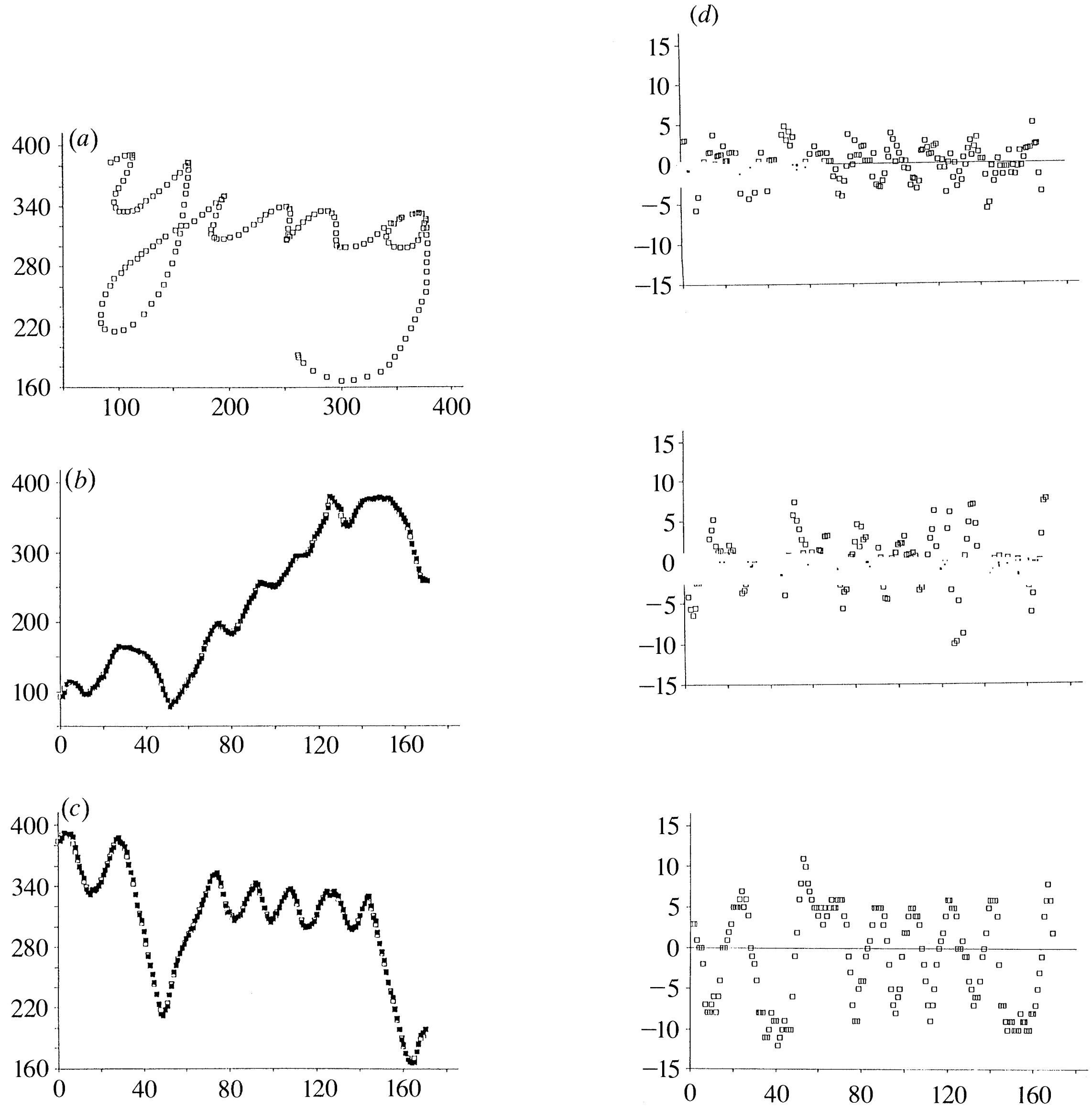

Figure 3. Experimental results comparing a lattice filter, a kalman filter and a non-predictive tracker. The path followed by the target-to-be-tracked is shown in (*a*): these data were obtained by keeping the head still and collecting, as a time series, the retinal coordinates of a small light source tracing out a trajectory on the floor in front of the vehile whose *x* and *y* coordinates are shown in (*b*) and (*c*) respectively. The performance of the three filters in tracking the *x* coordinates are shown in (*d*) as plots of the tracking errors in pixels.

serial mode. There is some potential for oscillation because the sequence involves a two-step lag but this danger is reduced by using the lattice filter to generate 2-step ahead predictions. This stabilizes the system and reduces the tracking errors.

Figure 5 shows the effect of using the filter to model the trajectory and the advantages over a simple non-predictive pipelined tracker in terms of the off-fovea retinal error. Figure 6 shows the value of the pipeline as a way of reducing the demand on the visual processing while maintaining a high rate of target locked image sampling. We have simulated the effect of increased complexity of visual processing by including a two-frame (80 ms) delay in the target location process (other experiments in which the area of foveal ROI was increased gave similar results, see figure 6*c*). There is no important change in the stability of the sampling rate (8.33 Hz) of the pipelined tracking performance with the noise (caused an occasional missed frame) staying constant at around 12%. In contrast, the serial tracking sampling rate is reduced by the delay (from 6.25 Hz to 4.16 Hz) and the noise increased to about 34%. The lag for the serial predictor with this delay is the same as for the pipeline (240 ms), and has surprisingly little effect on the overall performance. There are two reasons for this. The predictions, which enable the tracking to continue until the target is lost for three successive samples; and the automatic increase in the size of ROI when signalled by the predictor which enables the tracker to claw itself back on course albeit at the expense of increased processing delay. Tracking errors are often critically close to the boundary of the (small) foveal tracking window, and on several occasions the lattice

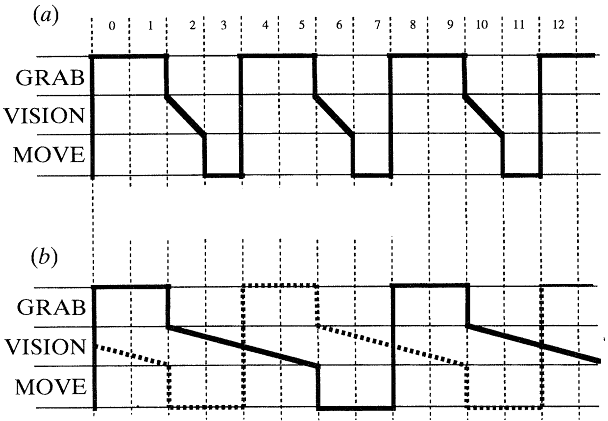

Figure 4. Serial and pipeline modes for microsaccadic tracking. Time is shown left-to-right quantized into 12 chunks, each of 40 ms (set by the image frame grabbing rate). The vertical axis depicts the image capture (GRAB), image processing (VISION) and head movement (MOVE) stages. For clarity of exposition, it is assumed that the target velocity is constant and, at the maximum consonant with the time allotted to the head movement stage, is sufficient to maintain tracking without 'slipping' a frame. To minimise blur, the image capture stage is triggered by completion of the head movement stage. In (*a*) (serial mode), the thick line shows which stage is in operation at any given time. In (*b*) (pipeline mode), the thick line and the dotted line show the simultaneous operation of different stages.

(*a*) Serial mode: The maximum sampling rate is 6.25 Hz, the samples have a 160 ms lag, and at a maximum head velocity of 50 deg s^{-1} the retinal target velocity is 200 pixels s^{-1}. This is equivalent to a displacement across the image of half the 'foveal ROI' per sample. The critical feature is that to maintain this rate the demand on visual processing stage is maintained at 40 ms. This provides serious constraints on both the size of the region that can be processed and the complexity of the algorithms that can be used to support the tracking. (*b*) Pipeline mode: The sampling rate is maintained at 6.25 Hz. The samples lag by 320 ms, and the maximum allowed average velocity is 400 pixels s^{-1}. The feature is that if the lag does not cause instability (a function of the complexity of the target trajectory), the pipelining mode increases the time allowed for image processing by a factor of four. This provides an important buffer for maintaining the temporal stability of the tracking sample frequency.

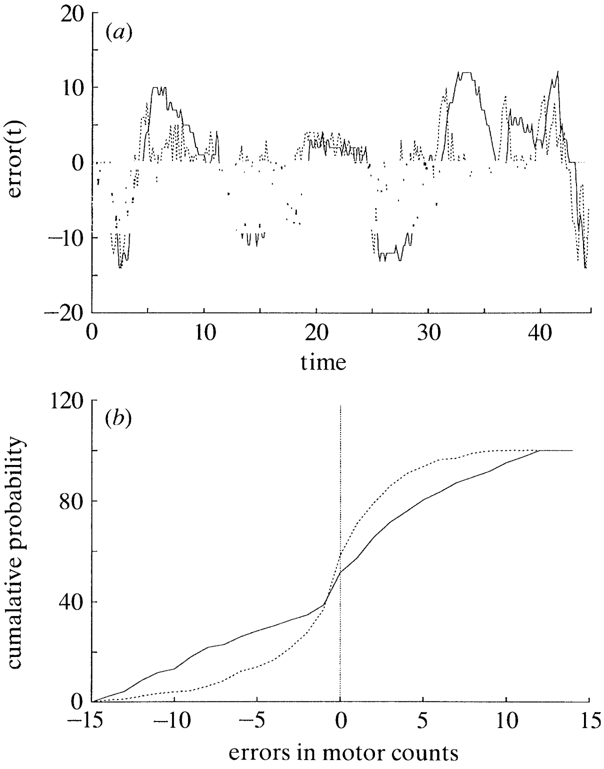

Figure 5. The improvement in microsaccadic tracking performance provided by a four-stage, two-step-ahead lattice predictor (absolute mode using vision error feed back) compared with a non-predictive tracker, both running in pipeline mode. The target was a small light source moved by a robot arm over the ground plane in front of the vehicle. Solid line, non-predictive tracker; dotted line, four-step ahead predictor. (*a*) Fixation errors (represented as left verge motor counts). (*b*) Normalized cumulative frequency distribution of the fixation errors. The important point to be noted is the rapidity with which the errors of the predictive tracker quickly return to near-zero after major error excursions caused by a sudden change in the direction of target trajectory. The predictor rapidly learns the trajectory, the non-predictive tracker is one step behind. The difference between the two modes is statistically very marked as can be seen from (*b*). The r.m.s. errors are: predictor 4.16, non-predictive tracker 6.84. Each motor count corresponds to about 7.7 min visual arc so the residual tracking errors of the predictor are very small and roughly of the same order as human vision (Carpenter 1988, p. 125).

filter predictor 'invented' the target position when it was lost, which is of course one of the duties of a predictor. We conclude that the microsaccadic tracking architecture as a whole appears robust and stable.

In conclusion, pipelining and the resulting maintenance of the maximum image sampling frequency is potentially important, perhaps less for its effects on the tracking performance per se, but because a pipelined tracker can support other concurrently operating vision processes which themselves require high sampling frequencies with limited temporal noise (e.g. modelling object deformations as a time series, or building a model of the target trajectory in world geometry coordinates in order to evaluate the risk of collision). That the tracking competence can be subsumed by other vision processes is an important consideration for the long-term development and evolution of the system.

5. NEURAL NETWORK ARCHITECTURES FOR STEREOPSIS UNDER VARIABLE CAMERA GEOMETRY

The problems we have been addressing under this section of the project are as follows: (i) generating the inverse kinematics needed for vergence saccades; (ii) finding epipolar geometry for generic stereo matching; and (iii) detecting obstacles by discovering violations

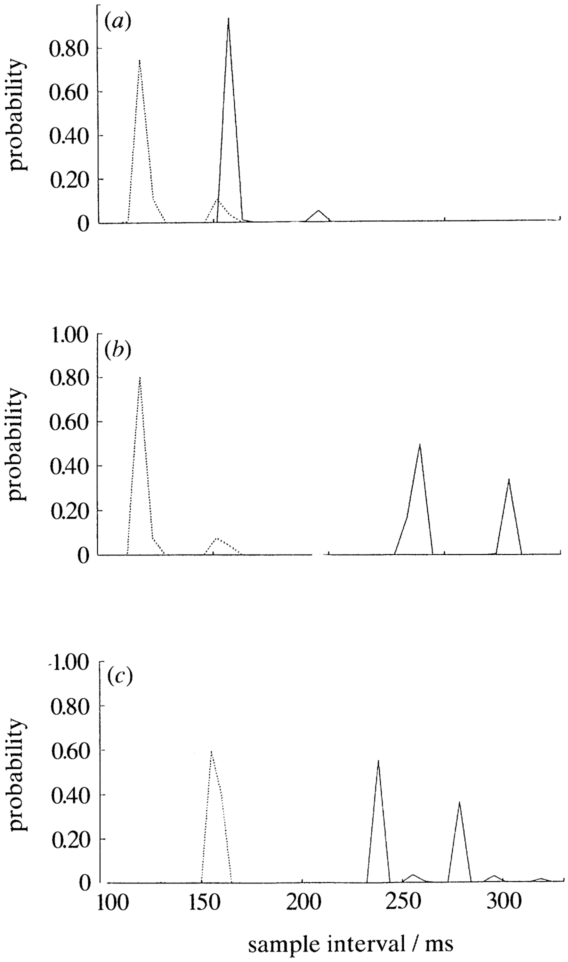

Figure 6. The effect of increasing the image processing demand: comparison between pipeline (dotted lines) and serial (solid lines) modes supported by lattice predictors. The graphs show the probabilities of a given interval between successive movements of the cameras. The target velocity was such that tracking could be maintained in both modes. Increased complexity in visual processing is simulated by adding in (*a*) no delay, and in (*b*) a 80 ms delay to the vision processing stage. Two points should be noted: (i) the period of the pipeline frequency remains similar under the two conditions, and the performance is unaffected by the increased demand; and (ii) tracking in the serial mode is still maintained though at a reduced rate and with increased jitter (frame slip). (*c*) Increasing the size of the image ROI by a factor of four reduces the sampling rate for the pipeline tracker from 8.33 Hz to 6.25 Hz (comparable to the best performance of the serial tracker). The serial tracker 'jittered' between 4.16 Hz and 3.57 Hz. In (*c*) the speed of the target was increased by 50%.

of disparities predicted for the ground plane. Solutions to these problems have all been implemented as self-organizing tables of neural nets whose inputs are retinal coordinates and camera motor positions, and which use local interpolation to recover the required geometrical information. This common architecture will be referred to as a PILUT (Parameterised Interpolating Look-Up Table) and its use will be illustrated in the context of a biologically plausible predictive stereo matching system for ground plane obstacle detection.

6. THE PILUT ARCHITECTURE

At its most general, the principle of the architecture is to use local linear approximations to multi-dimensional functions. It may be regarded as similar to the tensor-product three-dimensional surface interpolation schemes used in computer graphics (and computer aided design) but in a PILUT the interpolating function is a local hyperplanar patch approximation. An alternative way to regard the architecture is as two levels of neural networks. The first is the indexing or parameterizing network: it is coded relatively coarsely and generally has few dimensions, often simply serving to act as a blending function for the local piece-wise approximations carried out by the second level. The latter is constructed on demand in a particular context and has higher resolution inputs and, in general, more dimensions than the indexing level, possibly including the indexing dimensions at a higher resolution.

It is generally appreciated that the phase space trajectories of multi-dimensional systems lie on sub-manifolds which locally may be of a very much lower dimensionality than the system (Potts & Broomhead 1991). This is because, in general, the physics just does not allow the full combinatorial explosion to occur. The PILUT architecture is an attempt to provide a similar reduction in dimensionality while at the same time allowing high resolution local approximations to the full dimensional surface.

There are seven dimensions in the problems under consideration here: three for the degrees of freedom of the cameras in the saccade system (head tilt, left and right verge), and four for the x and y retinal coordinates of a target in the left and right images (figure 7). At first sight it appears there is little potential for a reduction in these dimensions. The insight, however, is to recognise constraints provided by the different stereo problems. For example, in the inverse kinematics problem the seven dimensions are immediately reduced to four by exploiting the high correlation between the positions of the eyes, and similarly between the positions of targets on the retinae. It is thus possible to use, as the indexing level, the coarsely coded information of the position of one of the eyes. The exquisite sensitivity of stereo to small differences can then be recaptured by using the full resolution of both camera motors and target position states as input to the second-level network that provides a local approximation of the full seven-dimensional hyperplane (see figure 7).

The coefficients of the local interpolating hyperplane are stored in a matrix (it is a simple linear net). When the network is accessed through the coarse indexing scheme, a composite matrix is formed by blending together the matrices in a region of the indexing parameter values. Two blending schemes have been explored, both biologically plausible.

One method uses radial basis functions (RBFS; figure 7) to populate the indexing parameter space with a number of centres positioned according to the coarse indexing scheme. Associated with each of the centres is a gaussian weighting function, in addition to a linear approximation to the surface. On indexing the

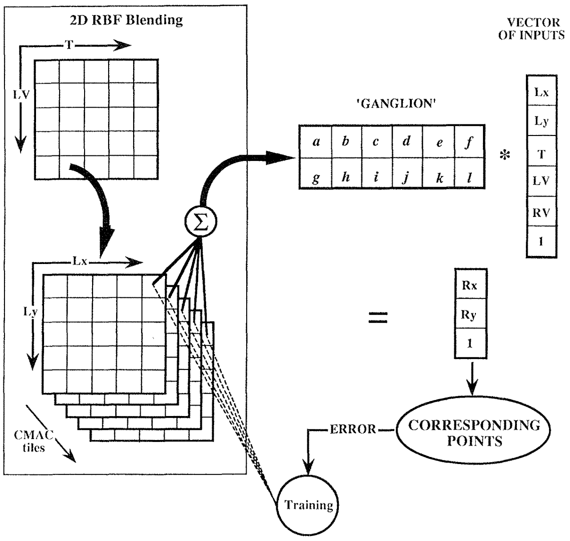

Figure 7. The PILUT architecture applied to stereopsis under variable camera geometry. The principle is to project the higher dimensional space onto a subspace, possibly a subset of the original dimensions, then to use a coarse coding scheme of the subspace, and full dimensional linear interpolation schemes to encode the hyperplanar approximation of the surface up to the required resolution. The simple structure of the stereo geometry makes the choice of the subspace obvious and there is no need to apply statistical techniques to recover the principle components. The figure shows a heterogeneous architecture which uses two different two-dimensional schemes for blending the coarse indexing dimensions. The indexing dimensions are: (i) the head tilt (T) and left verge (LV), and (ii) the x and y retinal coordinates of the left image (Lx,Ly). The idea that is fundamental to the architecture is that the local interpolating linear nets are created on-line by blending together, in appropriate proportions, the coefficients from nets in a neighbourhood around the input to the four dimensional indexing space. This composite net is used to provide a local linear approximation to the multi-dimensional surface. It is indexed using the full resolution available on those dimensions. The composite local interpolating linear net is trained using the simple delta rule, and the update in the coefficients (weights) distributed among the stored nets in proportion to their contribution. In the problem domain in which we have used the PILUT architecture we have found that the above heterogeneous architecture provides a marginally better approximation to the full dimensional surface than either a four-dimensional CMAC, or a four-dimensional RBF blending scheme used alone, although the analysis of the dynamics of the PILUT architectures is currently at a preliminary stage.

network, a composite local approximation is constructed by adding together the coefficients associated with the individual centres in proportion to the distance they are from the input. During the training phase the errors are propagated back as for simple linear networks, to adjust the coefficients of each matrix associated with each centre in proportion to its contribution to the composite network.

The second architecture we have explored for blending or interpolating across the parameterization is the CMAC (Albus 1976; figure 7). The CMAC is generally used for the representation of continuous multi-dimensional scalar functions. We choose to use the CMAC architecture to carry the coefficients of the matrices. The CMAC uses a coarse-coding strategy for the discretisation of the parameter space, and movement in the parameters may be regarded as equivalent to the discrete differentiation of the function at the resolution of the coarse coding. As for the RBF network, when accessed the CMAC builds a composite matrix by integrating the individual matrices indexed by the different layers of the coarse coding. During training the coefficients of the individual matrices are adjusted using the usual gradient descent methods.

Both these architectures have been used separately and in conjunction with each other as the indexing and blending level of the PILUTS, and details of the implementation and training of them may be found elsewhere (Mayhew *et al.* 1992). The appeal of the PILUT architecture is that it is in principle simple, readily customized, local and hence stable, easily trained and biologically plausible. Its disadvantages are that it is potentially expensive in memory. It seems to be a simplification of the Hyper Basis Function network representation proposed by Poggio & Girosi (1989, 1990), and a similar idea recently proposed by Lane *et al.* (1991). Physiologically one can regard the LUTS and interpolating nets as a matrix of receptive fields whose configuration or kernel is modulated by eye position information, determined from stereo itself (Mayhew & Longuet-Higgins 1982) or from the oculomotor system.

7. GROUND PLANE OBSTACLE DETECTION USING A PREDICTIVE STEREO MATCHER

Figure 8 shows a scheme for ground plane obstacle detection under conditions of variable camera geometry. It uses a predictive stereo matcher implemented in the PILUT architecture describe above, in which is encoded the disparity map of the ground plane for the different viewing positions required to scan the work space. The research is an extension of Mallot *et al.*'s (1989) scheme for ground plane obstacle detection which begins with an inverse perspective mapping of the left and right images that transforms the image locations of all points arising from the ground plane so that they have zero disparity: simple differencing of the resulting images then permits ready detection of obstacles. The essence of this physiologically inspired method is to exploit knowledge of the prevailing camera geometry (to find epipolar lines) and the expectation of a ground plane (to predict the locations along epipolars of corresponding left and right image points of features arising from the ground plane). We extend this approach by developing neural net methods for computing the epipolar geometry and corresponding ground plane points for variable camera geometry. To illustrate the problem, figure 9 shows ground plane disparity maps for different directions of gaze and elevation of the cameras (i.e. different viewing positions). For human vision, these disparity fields would roughly correspond to looking at the bottom left and right corners and the central fold of an open book lying on a table at about the

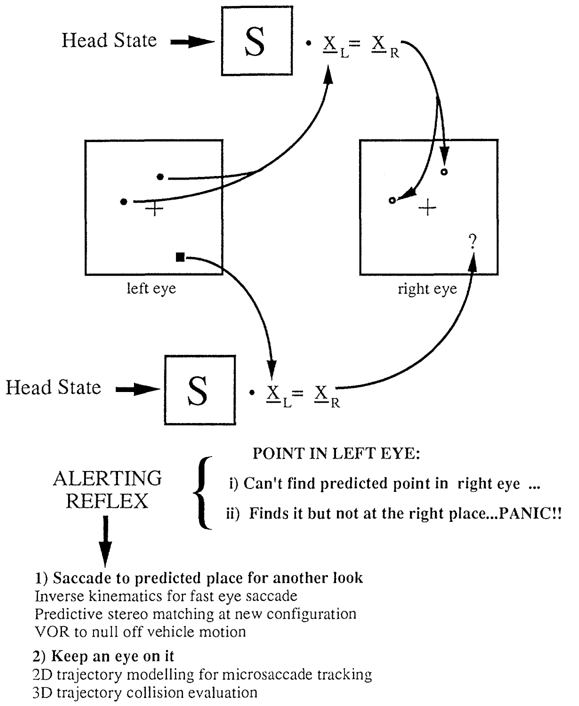
Head State
S
• $\underline{X}_L = \underline{X}_R$
left eye
right eye
?
Head State
S
• $\underline{X}_L = \underline{X}_R$
ALERTING REFLEX
POINT IN LEFT EYE:
i) Can't find predicted point in right eye ...
ii) Finds it but not at the right place...PANIC!!
1) Saccade to predicted place for another look
Inverse kinematics for fast eye saccade
Predictive stereo matching at new configuration
VOR to null off vehicle motion
2) Keep an eye on it
2D trajectory modelling for microsaccade tracking
3D trajectory collision evaluation

Figure 8. Ground plane obstacle detection under variable camera geometry. The scheme uses a predictive stereo matcher that encodes, for each head state, the map from the left eye to the right eye of corresponding points lying on the ground plane. Points that deviate from their predicted coordinates, by more than allowed by the error model, are subject to further inspection as potential targets.

normal reading distance. The disparities are the output of local networks serving regions of the motor states of our vehicle when the cameras are directed left, right and straightahead at points at approximately the same distance away on the ground plane (in fact, about 150 cm; cameras about 75 cm above the ground plane).

The data used to train the nets to deliver (predict) these ground disparities were collected by moving a small light source around on the floor of the laboratory, and, with the head still, tracking the light stereoscopically in real-time using a small ROI window. Apart from being in keeping with the tracking theme of the paper, this procedure offered a simple temporal solution to the stereo correspondence problem which obviated the need for a sophisticated stereo algorithm, with considerable benefits in reducing training time while developing the PILUT. The data set so obtained was used to train a neural net able to generate the coordinates of the corresponding point in the one eye's view when given as input the retinal coordinates of the points in other eye's view, and of course the motor positions encoding camera states.

The interpolation achieved by the nets is brought out in the figures as they show a mapping from a grid of retinal locations in the left eye to the corresponding locations in the right eye that are predicted to be generated by points lying on the ground plane (obviously, the training data were much more haphazardly distributed). It is possible to think of the different maps as showing what would result from moving the head from side to side and up and down while maintaining fixation on the same point on the floor. While you do this it might be worth reflecting on the extensions to the outflow theory (Carpenter 1988) that are required to make the world appear as stable as it does.

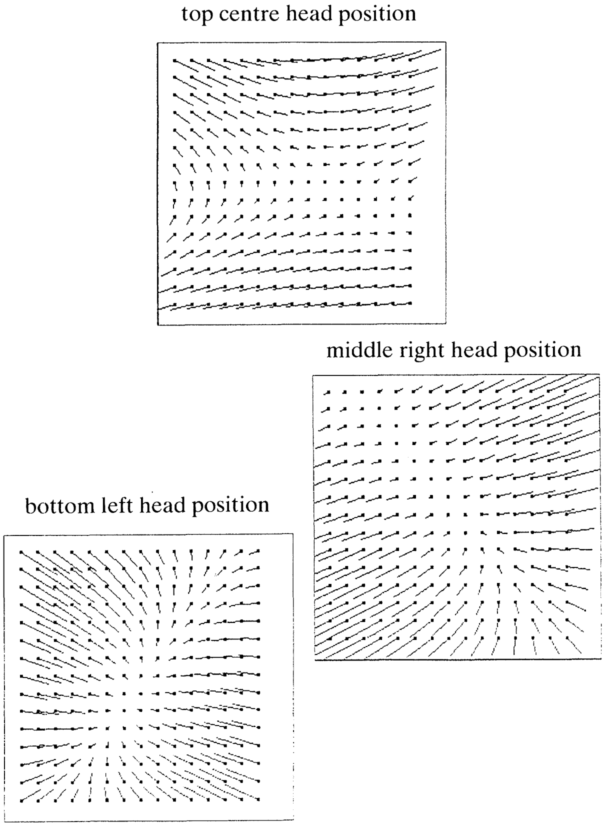

Figure 9. Ground plane disparity maps for different head positions: looking ahead, to the left and to the right. The sampling spacing corresponds to approximately one degree of visual angle (see text for details). Note that changing the eye position radically changes the pattern of disparities, the disparities contain both vertical and horizontal components, and often the vertical components are of the order of a degree (the particular pattern of vertical disparities is determined by the position on the retina, and the camera geometry, and is independent of scene structure to first order).

There are some important points to be made.

1. The disparities involved are in general large. The visual angle subtended by the images is almost exactly 30 degrees. Thus the eccentricity of the points lying towards the periphery are not at all excessive, and yet the disparities are often more than a degree in magnitude.

2. There are both vertical and horizontal components to the disparities. The vertical components are often very large, and at some locations, larger than the horizontal disparities.

3. The pattern of disparities is markedly affected by changing the direction of gaze.

4. They illustrate the scale of the inverse kinematics problem!

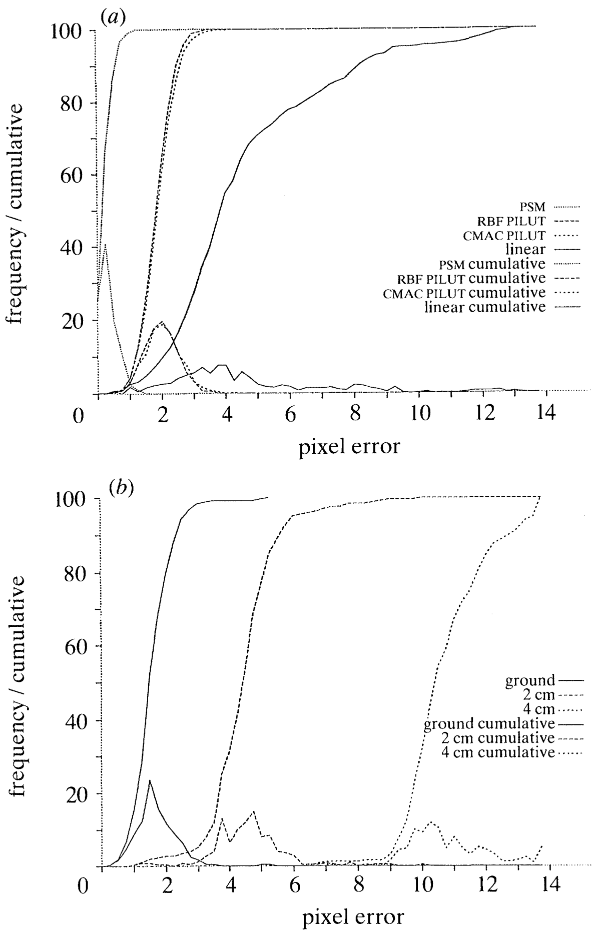

Figure 10. Representative experimental results evaluating the ground plane stereo matcher at a single head position (we find no difference in performance dependent on the magnitude of the asymmetry of vergence). (*a*) Error distributions shown as normalised frequency and cumulative distributions in pixels of disparity for a single linear net, a PSM, a PILUT 3 × 3 blended RBF, and a PILUT 3 × 3 blended CMAC. Performance of the linear net is clearly inadequate whereas the stereo mapping is solved both by the PSM and the PILUTS. (*b*) Superimposed normalised frequency and cumulative error distributions for the PILUTS at a different head position from the data shown in (*a*). The clearly distinguishable distributions correspond to the ground plane, and two obstacles, one two cm and the other 4 cm high. A simple statistical decision measure would be sufficient to trigger an alerting reflex.

It is important to note that the above are quite general points and are not an artefact of using a planar retina. The convenience of using a planar retina is that there is a relatively simple, but nonlinear, projective relationship (termed here the Projective Stereo Mapping, PSM) between the positions of corresponding retinal points when a planar surface is viewed. This relationship may be represented as the homogeneous projective matrix, where $x_1 \cdot S = x_r$ up to a proportionality. The nonlinearities arise from the division with the coefficients in the bottom row of the S matrix. The coefficients of the S matrix are a function of the cross products of the retinal coordinates, and can be found by solving a simple linear least squares problem given the coordinates of corresponding points as the input data. Thus a simple linear net can be used to estimate the coefficients but a division must be performed to use them.

The PSM is applicable only to planar retinae: we use it as a benchmark for evaluating the performance of the PILUT architecture described above. The latter assumes only that the function can be locally approximated by a blending of planar patches and is therefore more general (but, in the case of planar retinae, necessarily sub-optimal). Figure 10 shows results for a fixed head position: looking ahead with symmetrical vergence. The results show that: (i) a simple linar net is unable to capture the disparity mapping; (ii) there is little difference between the optimal PSM network and a local PILUT using a 3 by 3 planar tessellation; and (iii) the same PILUT can be used to detect small obstacles lying on the floor about a metre and a half away, as they show up as departures from the disparities predicted for the ground plane. It is well known that errors in stereoscopic depth vary as the square of the viewing distance. Hence, though the system can detect obstacles as small as a centimetre high when fairly nearby, the resolution rapidly decreases at greater distances.

A PILUT for the stereoscopic ground plane mapping that makes no concession to biological plausibility but is economical both in storage and in training overhead was created as follows: (i) ground plane data were collected at a small number of selected head positions; (ii) at each position the coefficients of the PSM were found using a simple linear net training regime; and (iii) a standard algorithm was used to find the best fitting quadratic surface for each of the eight variable coefficients in the S matrix as a function of the head position parameters. We have used this 'engineering solution' PILUT not only to prove the principle but also as a source of training and test data for experiments on the different variations of other PILUT architectures.

Ying Zheng is supported by a MOD-RSRE grant awarded to John P. Frisby, Steve A. Billings and John E. W. Mayhew. Stuart Cornell is supported by Grant No. GR/G 61724 SERC-ACME awarded to Mayhew, Frisby and Chris Brown. Much of the pilot work was funded by Grant No. SPG 825592 Cognitive Science-HCI Initiative awarded to Mayhew, Paul Dean and Frisby. The authors wish to acknowledge their debt to all the members of AIVRU. In particular: Patrick Courtney for his heroics in implementing the image capture system, Stephen Pollard and John Porrill for the TINA VISION processing environment and valuable discussion, and John Frisby both for his role in the management of AIVRU ('it's rather like herding cats. . .') and his contribution to the writing of the paper.

REFERENCES

Abraham, I., Bedworth, M., Booth, C., Booth, D., Bounds, D. & Mayhew, J. 1991 'ATTITUDE project: final report. *RIPR technical report no. 2000/3/91*. RSRE (South), St. Andrews Road, Malvern, Worcestershire.

Albus, J. 1975 A new approach to manipulator control: the cerebellar model articulation controller (CMAC). *Trans. ASME-J. Dyn. Syst. Meas. Control* **97**, 220–227.

Albus, J. 1975 Data storage in the cerebellar model articulation controller (CMAC), '*Trans. ASME-J. Dyn. Syst. Meas. Control* **97**, 228–233.

Alexander, S.T. 1986 *Adaptive signal processing*. New York: Springer-Verlag.

Barlow, H.B. & Levick, W. 1965 The mechanism of directionally sensitive units in the rabbit's retina. *Physiol., Lond.* **178**, 477–504.

Brooks, R.A. 1986 A layered intelligent control system for a mobile robot. In *Robotics research: The Third International Symposium*, pp. 365–372. MIT Press.

Brooks, R.A. 1991 Intelligence without representation. *Artif. Intell.* **47**, 139–159.

Brooks, R.A. 1991 New approaches to robotics. *Science, Wash.* **253**, 1227–1232.

Carpenter, R.H.S. 1988 *Movements of the eyes*. London: Pion.

Dean, P., Mayhew, J.E.W., Thacker, N. & Langdon, P.M. 1991 Saccade control in a simulated robot camera-head system: neural net architectures for efficient learning of inverse kinematics. *Biol. Cyber.* **66**, 27–36.

Girosi, F. & Poggio, T. 1989 Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Comput.* **1** (4), 465–469.

Goodwin, G.C. & Sin, K.S. 1984 *Adaptive filtering, prediction and control*. New Jersey: Prentice-Hall.

Kawato, M. 1989 Neural network models for formation and control of multijoint arm trajectory. In *Neural programming* (ed. M. Ito), pp. 189–201. (Taniguchi Symposia on Brain Sciences No 12.) *Karger, Basel: Japan Scientific Society Press.*

Lane, S.H., Flax, M.G., Handelman, D.A. & Gelfand, J.J. 1991 Function approximation using multi-layered neural networks with B-spline receptive field functions. *CSL Report* **47**, 1–37.

Makhoul, J. 1975 Linear prediction: a tutorial review. *Proc. IEEE*, **63** (4).

Mallot, H.A., Schulze, E. & Storjohann, K. 1988 Neural network strategies for robot navigation. In *Proc. n'Euro* (ed. G. Dreyfus & L. Personnaz). Paris.

Mayhew, J.E.W. 1992 ANIT: Architecture for navigation and intelligent tracking. (In preparation.)

Mayhew, J.E.W., Dean, P. & Langdon, P. 1992 Artifical neural networks for the kinematic control of a stereo camera head. (In preparation.)

Mayhew, J.E.W. & Longuet-Higgins, H.C. 1982 A computational model of binocular depth perception. *Nature, Lond.* **297** (5865), 376–379.

Poggio, T. & Girosi, F. 1989 A theory of networks for approximation and learning, *A.I. MEMO No. 1140.* Massachusetts Institute of Technology: Artifical Inteligence Laboratory.

Poggio, T. & Girosi, F. 1990 Networks for approximation and learning. *Proc. IEEE* **78** (9), 1481–1497.

Poggio, T. & Girosi, F. 1990 Regularization algorithms for leaning that are equivalent to multilayer networks. *Science, Wash.* **247**, 978–982.

Potts, M.A.S. & Broomhead, D.S. 1991 Time series prediction with a radial basis function neural network. In *Adaptive signal processing* (ed. Simon Haykin) (*Proc. SPIE* **1565**) pp. 255–266.

Thacker, N.A. & Courtney, P. 1992 Online stereo camera calibration. *AIVRU Memo No. 62*. University of Sheffield: A1 Vision Research Unit.

Zheng, Y., Mayhew, J.E.W., Billings, S.A. & Frisby, J.P. 1992 Lattice predictor for 3D vision and intelligent tracking. *AIVRU Memo No. 67*. University of Sheffield: A1 Vision Research Unit.

Zheng, Y. 1992 Stereo vergence tracking. (In preparation.)

Discussion

V. Torre (*Departmento Di Fisica, Universita di Genova, Genova, Italy*). There seems to be some controversy concerning the effect of eye rotations on the disparity field. Could Professor Mayhew please elaborate on this in the light of the work he has just presented?

J. E. W. Mayhew. The problem arises I think because the difference between the concepts of optic array and the measured disparity field is frequently confused.

If we define the optic array to be the lines of sight through the nodal points of the eye and assume that the eye rotates around this point, then clearly the optic array cannot be changed by any rotation of the eye (and as is well known, this is why it is impossible to recover information about the three-dimensional structure of the scene from rotations alone). Similarly if we assume another eye, it too will have associated with it a different optic array, which cannot be changed by rotating that eye. Furthermore, the angle between the lines of sight which derive from the same point in the world will subtend at the nodal point of the eyes an angle (referred to as binocular parallax) that cannot be changed by the rotations of the eyes. This is a good thing because it implies that there is information about the three-dimensional structure of the world which is in principle available to the stereo mechanism independently of which point in the world the eyes fixate.

The problems arise when one considers the issues involved in measuring binocular parallax (the angle between the corresponding lines of sight) of the point in the world. The physiological mechanism for this measures the retinal disparity, by which is generally meant the difference in retinal position in the left and right eyes of the two lines of sight derived from the same three-dimensional point. Of course it is now simple to see how moving the eyes from fixating one point to fixating another must affect the pattern of measured disparities on the retina. Subject to physiological abnormalities, every point in the world that the eyes fixate gives rise to a zero disparity measurement. Whereas in general, when the eyes fixate another different point, the previously fixated point will give rise to a measured retinal disparity which is a function of both the difference in three-dimensional positions of the points, and the directions and magnitudes of the rotations of the eyes which were needed to change fixation.

Of course if were feasible for a visual system to use a method of measuring disparities of corresponding image points which was independent of where the eyes were looking then it could be argued because eye rotations in such a system were irrelevant, such disparity fields of the type shown in figure 9 are a misrepresentation of the problem that is solved by brain.

As it is, the eyes do seem to use the fovea to fixate selected targets, the physiological evidence suggests that disparities are measured by the binocular cells serving somewhat similar retinal regions, and the mathematics of the geometry of stereo projection consonant with such assumptions will predict almost exactly the sort of disparity fields illustrated. It is important to emphasize that the flow fields in question are not simulated data, but are generated by the adaptive biologically plausible neural network architectures described in more detail in the paper. Furthermore, methods for the reliable extraction of information about both the scene structure and eye position from such disparity flow fields are well known.