Ground Plane Obstacle Detection under variable Camera Geometry Using a Predictive Stereo Matcher.

Stuart Cornell, John Porrill, John E W Mayhew.
Artificial Intelligence Vision Research Unit,
University of Sheffield, Sheffield, S10 2TN,England

Abstract

A scheme is proposed for ground plane obstacle detection under conditions of variable camera geometry. It uses a predictive stereo matcher implemented in the PILUT architecture described below, 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 the 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/right image points of features arising from the ground plane).

1 Introduction

The research described here is part of a project investigating adaptive control of a four degree of freedom stereo camera rig [1] 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. The component of the work to be described here is ground plane obstacle detection using a predictive stereo matcher to encode the disparity map of the ground plane (see Figure 1). For a vehicle operating in a limited operating enviroment such as a factory floor, simplest form of obstacle is a point in space which is

[1] The stereo camera rig used for this work comprises a 3-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 here. 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.

BMVC 1992 doi:10.5244/C.6.57

not on the ground plane. By reproducing the mapping of corresponding image points obtained from the stereo camera pair the obstacles can be detected by identifying violations in it. The complexity of this stereo correspondence problem is large as the disparity flow fields vary significantly with camera geometry. Also the disparity vectors have significant components in the X and Y directions presenting a complex non-linear problem of high dimensionality. Solutions to this have been obtained without the use of camera calibration or any dependency upon the dynamics of the camera rig. The approximation of simple local mappings to give more general global results has been extensively used throughout the work in the form of Parametrised Interpolated Look-Up Tables (PILUT). These have been implemented in the form of variably organised Neural Nets. They provide a constantly updateable result in a form which is easily combined into the subsumptive head control architecture of our mobile vehicle.

2 Disparity Fields

To illustrate the problem, Figure 2 shows ground plane disparity maps for different directions of gaze and elevation of the cameras (i.e. different viewing positions). For human vision, these viewing posiitons 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 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 straight ahead 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 was 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 and a centre of gravity process on the images. 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 sets so obtained were used to train the neural nets described below so as to generate the coordinates of the corresponding point in theone eye's view when given as input the retinal coordinates of the points in other eye. For the variable camera geometry methods the motor positions encoding camera states were also used as input data.

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 for points lying on the ground plane.

Figure 3 Shows real data traces (bottom left) of the ground plane and the obstacle, the predicted field for the same position (top), and the difference between the two. This clearly distinguishes the obstacle from the ground plane data points. [2] There are some important points to be made here.

[2]The obstacle was a book which stood about 3cm off the ground, and was placed such the head was fixated on position approx 1m in front of the vehicle and slightly to the left.

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.

It is important to note that the above are quite general points and are not an artefact of using a planar retina.

3 Projective Stereo Mapping

The convenience of using a planar retina such as we have in our camera rig, is that there is a relatively simple, but non-linear, 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 S, where $x_l.S = x_r$ up to a proportionality.(See figure 1)

The non-linearities 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 however, and the PILUT architecture described below 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).

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:

1. 13 sets of ground plane data were collected as described each for a different head position fixated on the ground.

2. at each position the coefficients of the PSM were found using the simple linear net training regime (See appendix); and

3. a least squares minimisation using Cholesky decomposition was used to find the best fitting quadratic surface for each of the 8 variable coefficients in the S matrix as a function of the head position parameters.

This method has been used extensively 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. See figure 5 for results.

4 PILUT's

The principal of the PILUT architecture at its most general is to use local linear approximations to multi-dimensional functions. It may be regarded as similar to the tensor-product 3D surface interpolation schemes used in computer graphics (and computer aided design) but in a PILUT the interpolating function is a local hyper-planar patch approximation. An alternative way to regard the architecture is as two levels of neural networks. The first is the indexing or parameterising 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 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 and 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 7 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. 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 stereo problem. In this situation the 7 dimensions are immediately reduce to 4 because we are dealing with just head positions which are fixations on the ground plane which make one verge redundant. It is also possible to use, as the indexing level, the coarsely-coded information of the position of only one of the eyes. The sensitivity of stereo to small differences can then be recaptured by using the retinal coordinates again as input to the second-level network that provides a local approximation of the full 7-dimensional hyper plane Figure 4.
This can be reduced to 4 because all the head positions are fixated on the ground plane so two verges are not needed and we can use only the left cameras X and Y because of the small differences between the left and the right images.

The coefficients of the local interpolating hyper plane 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) 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 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 parameterisation is the CMAC (Albus 1976). 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 maybe 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 customised, 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 and 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. The latter information may be determined from stereo itself or from the oculomotor system.

5 Results

Figure 5 shows results of the different methods for a fixed head position: looking ahead with symmetrical vergence. The results show that:

1. a simple linear net is unable to capture the disparity mapping;

2. there is little difference between the optimal PSM network and a local PILUT using a 3 by 3 planar tessellation; and

3. 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.

Figure 5 shows the change in distribution of errors for non-ground plane points. A clear distinction between the different heights is easy to discern. Note also the that a 4cm obstacle does not cause twice the distribution shift of a 2cm one. This illustrates further the non-linearities involved.

6 Summary

The paper has described part of a project to explore the use of a biologically plausible neural network architectures as part of the system to exploit stereopsis under the variable camera geometry of a four degree of freedom stereo camera rig. A rather simple, but seemingly, adequate neural network architecture for representing high dimensional surface approximations (PILUTs) was evaluated as a method of encoding the projective stereo mapping of the ground plane for different head positions. This has been shown to be succesful as a primitive Ground Plane Obstacle detection device, and we are pursuing an analysis of these results to determine more sophisticated ways of using the predicted mappings.

References

[1] J. Albus, "A new approach to manipulator control: The cerebellar model articulation controller (CMAC)", Trans. ASME - J. Dyn. Syst. Meas. Control, 1975,vol. 97, pp 220-227.

[2] J. Albus,"Data storage in the cerebellar model articulation controllelr (CMAC)", Trans. ASME - J Dyn. Syst. Meas. Control, 1975,vol 97, pp 228-233.

[3] P. Dean, J.E.W. Mayhew, N. Thacker, & P.M. Langdon, "Saccade control in a simulated robot camera-head system: neural net architectures for efficient learning of inverse kinematics.", Biological Cybernetics,1991, 66, 27-36.

[4] F. Girosi, T. Poggio, (1989) "Representation properties of networks: Kolmogorov's theorem is irrelevant", Neural Computation, 1989, vol. 1, no. 4 pp 465-469.

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

[6] H.A. Mallot, E. Schulze, & K. Storjohann, "Neural network strategies for robot navigation.", Proc. n'Euro,In G. Dreyfus & L. Personnaz (Ed.), 1988 ,Paris:

[7] J.E.W. Mayhew, P. Dean,P. Langdon, "Artifical neural networks for the kinematic control of a stereo camera head", (in preparation),1992

[8] J.E.W. Mayhew, H.C. Longuet-Higgins, "A computational model of binocular depth perception.", Nature, 1982, 297 (5865) 376-379.

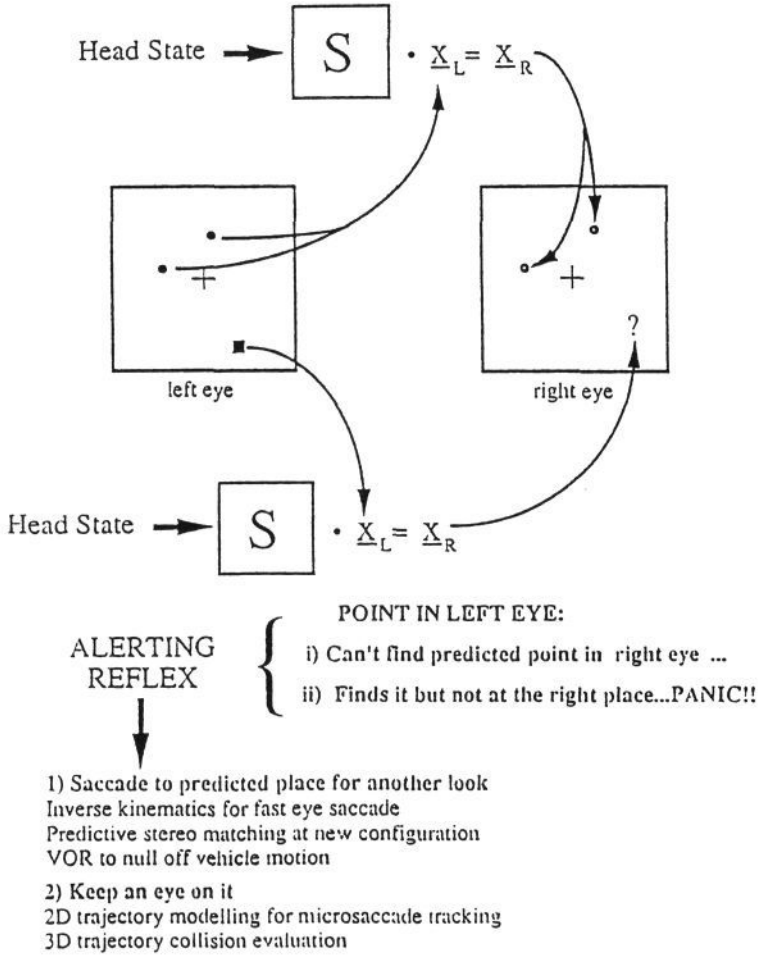
[9] T. Poggio,F. Girosi, "A theory of networks for approximation and learning.",A.I. MEMO NO. 1140. ,Atifical Inteligence Laboratory, Massachusetts Institute of Technology,1989

[10] T. Poggio, & F. Girosi, "Networks for approximation and learning.", Proceedings of the IEEE, 78(9),1990, 1481-1497.

[11] T. Poggio, & F. Girosi, "Regularization algorithms for leaning that are equivalent to multilayer networks.", Science,1990, 247, 978-982.

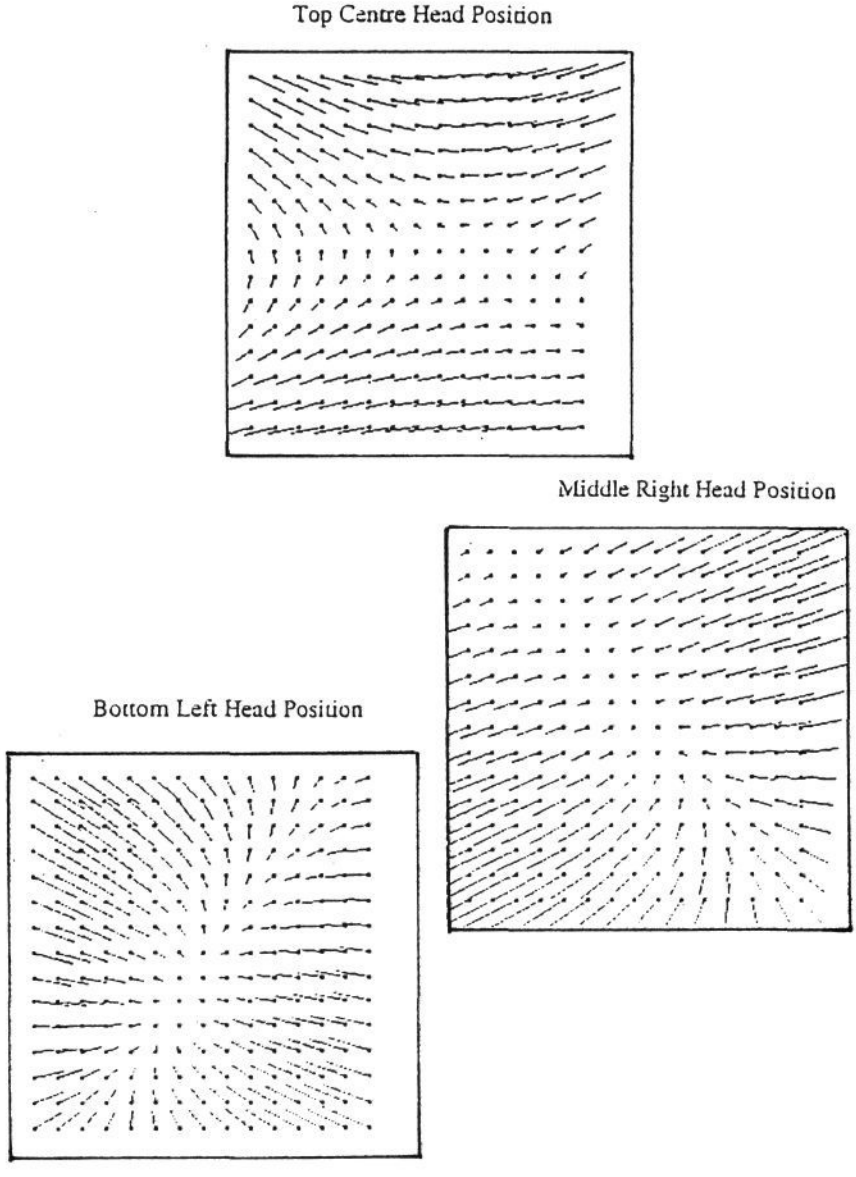
[12] M.A.S. Potts, D.S. Broomhead, "Time series prediction with a radial basis function neural network.", Adaptive Signal processing, Simon Haykin (Ed), Proceedings of SPIE 1565,1991, 255-266

[13] N.A. Thacker, J.E.W. Mayhew, "Optimal combination of stereo camera calibration from arbitrary stereo images.", Image and Vision Computing (feb 1991). vol 9 no 1 27-32.



- Figure 1. Ground plane obstacle detection under variable camera geometry. The scheme uses a predictive stereo matcher which 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 is allowed by the error model are subject to further inspection as potential targets.

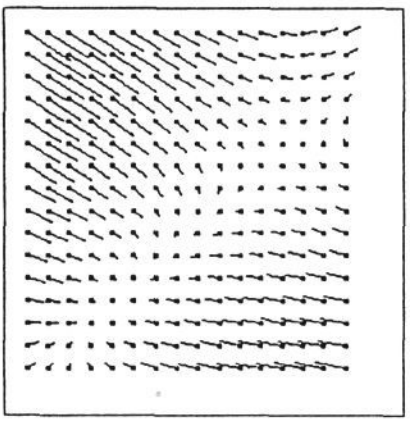
Left to Right Disparity Maps for the Whole Retina



• Figure 2. 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).

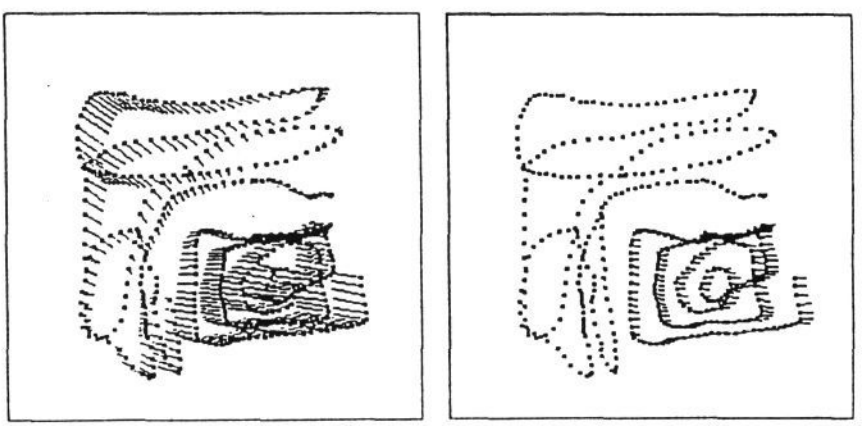
Effect of a 3cm. Obstacle on Disparity Maps for the Bottom Left Position

Ground Plane Data

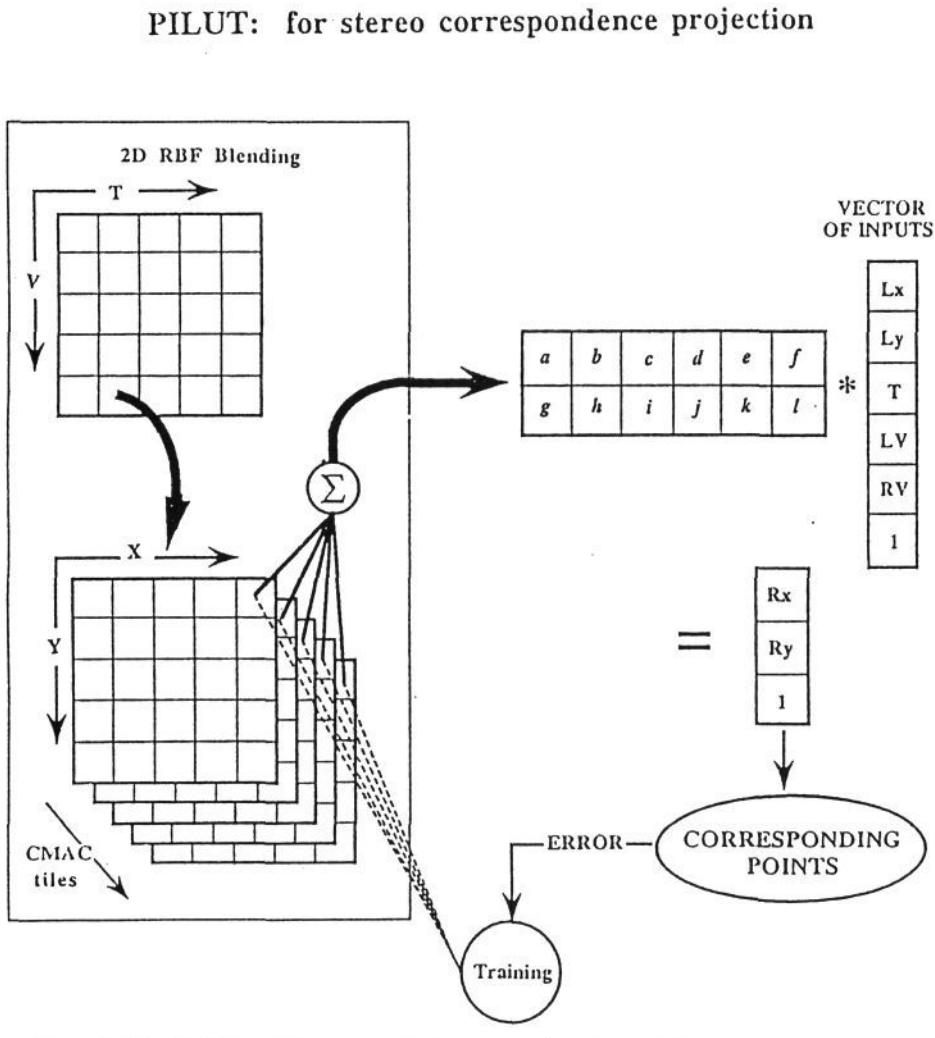


Ground Plane and Obstacle

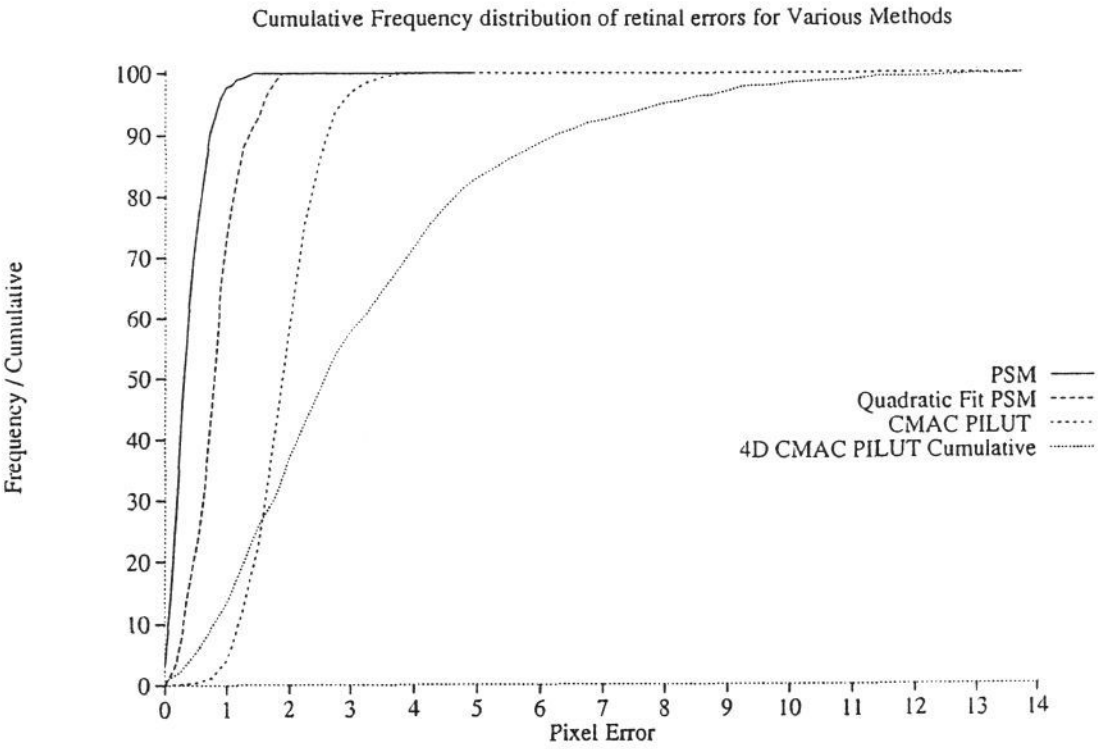
Differences from predicted map for ground plane



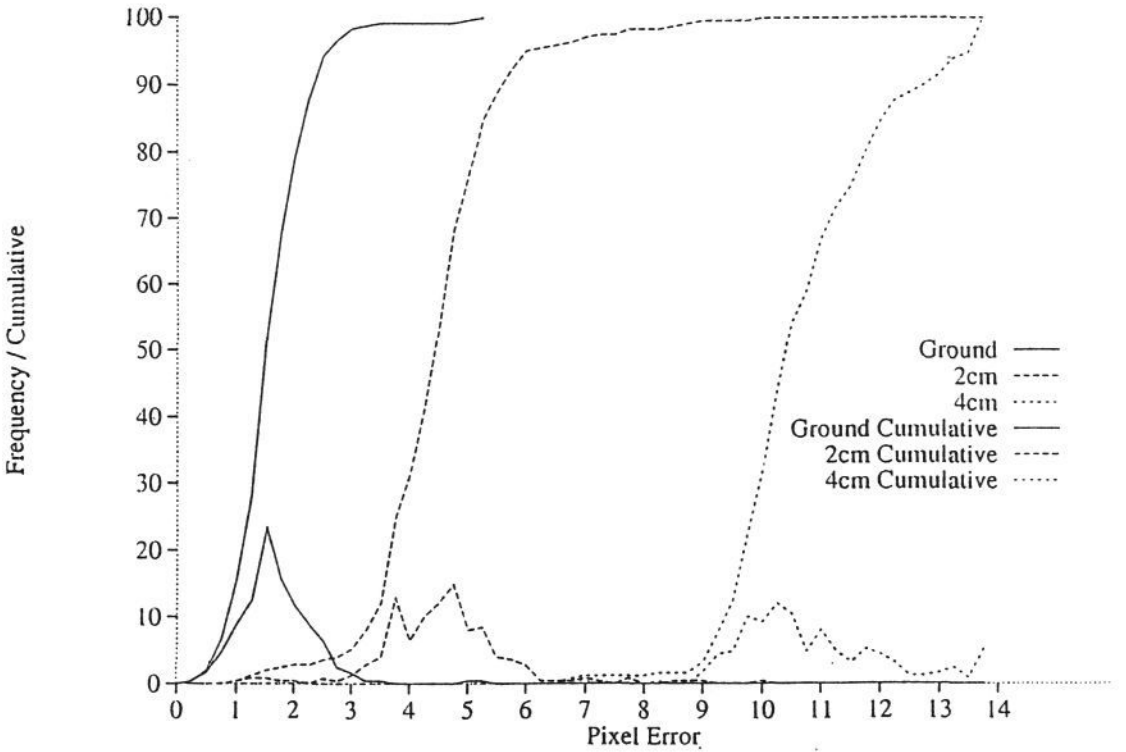
• Figure 3. Disparity Maps for obstacle detection: top: the predicited disparity map for a fixed head position. left: the actual disparity data for ground and obstacle. right: actual - predicted showing disparity error for obstacle points. 3cm obstacle at distance of 1m approx.



• Figure 4. 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.



• Figure 5. a) Representative experimental results evaluating the ground plane stereo predictor 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 3x3 blended RBF, and a PILUT 3x3 blended CMAC all as compared to a data set for that position. 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 PI-LUTs at a different head position from the data shown in a. The clearly distinguishable distributions correspond to the ground plane, and two obstacles, one 2 cm and the other 4 cm high. A simple statistical decision measure would be sufficient to trigger an alerting reflex in the control architecture.