STEREO-DISPARITY ESTIMATION USING A SUPERVISED NEURAL NETWORK

Y. V. Venkatesh, B. S. Venkatesh and A. Jaya Kumar
Department of Electrical Engineering
Indian Institute of Science
Bangalore 560012, INDIA
Phone: 0091 80 2293 2572
Fax: 0091 80 2360 0764
E-mail: yvvele@ee.iisc.ernet.in

Abstract.

We deal with the problem of determining disparity in gray-level stereoimage-pairs, by treating it as a nonlinear classification problem, and invoking Marr and Poggio's [1] neighborhood criterion. To this end, we propose the application of an artificial neural network (ANN). The main contribution of the paper is believed to be the use of neurons which are trained to be disparity selective, and thereby dispensing with the standard assumptions made about the neighborhood.

The disparity estimates so obtained for random-dot and natural stereoimage-pairs are comparable to those found in the literature. Whereas Khotanzad *et al.* [3] used a multi-layer perceptron (MLP) in order to learn the constraints of a cooperative stereo algorithm for binary, random-dot stereograms, we employ a single layer ANN. Further, in our scheme, the ANN weights adapt themselves to the neighborhood, and are able to learn the constraints successfully.

INTRODUCTION

When images of objects at different depths (in the physical world) are captured by a stereocamera-pair, each point on an object is uniquely represented by a specific pixel on each image; and the position of the pixel depends on the depth of the physical point in the scene. The displacement in the positions of these pixels **corresponding** to the same physical point is called *disparity*. And the physical depth of the point on the object can be extracted from this disparity. Therefore, the problem of stereoimage-pair analysis amounts to searching for corresponding pixels in the left and right images (hence the name, **correspondence problem** (CP), which is considered to be *not completely* solved).

0-7803-8608-6/04/$20.00 ©2004 IEEE

Many assumptions and constraints have been proposed to simplify the computational approach to the CP. For instance, the cameras are assumed to be set up in such a way that the displacement of pixels (in the two images) takes place along the same horizontal ('epipolar') line. Hence the search area (for solving the CP) is constrained to be along the epipolar line (hence the name, *epipolar constraint*).

Marr and Poggio [4] formulated the CP based on the following three constraints:

- Compatibility: a similarity measure is used for matching any two pixels.
- Uniqueness: a pixel in one image can be matched to only one (and hence unique) pixel in the other,
- Continuity: objects in the world have, in general, a smooth and continuously varying disparity.

Based on these constraints, Marr and Poggio [4] propose two methods to solve the CP. The first method involves an iterative process, using the constraints as excitation and inhibition, and is called the *cooperative stereo* algorithm. In the second method, image features, like *zero crossings*, are used to match the corresponding points in the images. The present paper is motivated by the cooperative stereo algorithm.

Marr and Poggio [1] propose a cooperative algorithm in order to extract disparity from binary random-dot-stereograms. Zitnick and Kanade [5] extend this area-based approach to gray level images, and invoke the above three constraints to rectify the matches iteratively. For binary random-dot stereo-pairs, Khotanzad *et al.* [3] suggest a non-iterative method of solving the same problem using a neural network. They make use of a multi-layer perceptron (MLP) along with a neighborhood constraint, as in [1], in order to learn the excitatory and inhibitory relations between the neighboring pixels.

The main contribution of the present paper is an extension of the method proposed in [3] to gray scale images, but using a **single layer perceptron**. Recent work by Henkel [2], suggests that coherence property can be exploited in the attempt to estimate disparity in a stereo-image pair. Note that Marr and Poggio [1] do not deal with coherence because of the neighborhood constraints.

It should be added here that we do not make any assumptions on the neighborhoods. On the other hand, we use the network to prune the neighborhood by *itself*.

The outline of this paper is as follows. In Sec. 2, we formulate the correspondence problem using the compatibility matrix; and propose a neural architecture in Sec. 3. We present the experimental results on random-dot and natural stereo pair in Sec. 4. Finally, we conclude the paper in Sec. 5.

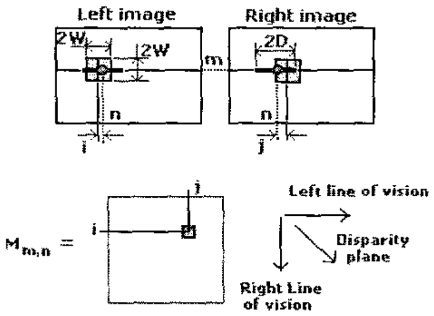


Figure 1: Structure of compatibility matrix

PROBLEM FORMULATION

In Marr and Poggio's [1] cooperative algorithm, a pixel in the line segment from the left image is compared with each pixel in the corresponding line segment of the right image of the stereo pair — *this is the assumption of epipolar geometry* — in order to arrive at a compatibility matrix. Since Marr and Poggio [1] deal with binary random-dot stereo pairs, they use the XNOR operator in order to compare the 2 pixels: *when there is a match, the matrix element is set to 1, else 0.*

Similar logic is extended to build, for gray-level images, a compatibility matrix, $M_{m,n}(i,j)$ (for each pixel (m,n)), using some match-measures, like correlation or squared differences of windows centered at $(m, n+i)^{th}$ and $(m, n+j)^{th}$ pixel in the left and right images. For simplicity, we use the sum of the absolute pixel-by-pixel difference of the windows centered at $(m, n+i)$ and $(m, n+j)^{th}$ pixels in the left and right images, respectively, for $M_{m,n}$ i.e., if I_L and I_R are the left and right images, then the compatibility matrix at $(m,n)^{th}$ pixel is given by

$$M_{m,n}(i,j) = \sum_{p=-W}^{W} \sum_{q=-W}^{W} |I_L(m+p, n+i+q) - I_R(m+p, n+j+q)| \tag{1}$$

where $i, j = 1, 2, ..., (2D-1), 2D$; $D > D_{max}$, assuming that disparity ranges from $-D_{max}$ to D_{max}; $2W$ is the window size; and $2D$ is the segment size.

The matrix elements are normalized so that the values lie between 0 and 1. Hence a low value of the matrix element implies a better match. In order that the matrix embeds the correct match within itself, the size of the segment considered must be greater than the maximum disparity that exists in the segment ($D > D_{max}$). In fact, it is fixed at a value higher than the highest disparity value in the image. As a consequence, the disparity value of $(m,n)^{th}$ pixel can be determined from $M_{m,n}$.

We consider only the interactions along the horizontal direction, i.e., only horizontal line segments are used to construct the compatibility matrix. This procedure can be extended to include vertical neighborhoods in order to get a 3D-compatibility volume [1].

According to [1], at a match point (i, j) only the neighborhood along the line of vision (and in the same disparity plane) affects the disparity value. In the compatibility matrix, the lines of vision lie along the rows and columns, i.e., $M_{m,n}(i+k, j)$ and $M_{m,n}(i, j+k)$ (see Fig. 1), and the disparity plane lies along the direction of the principal diagonal, i.e. $M_{m,n}(i+k, j+k)$. The uniqueness condition gives rise to inhibition along the lines of vision and excitation along the same disparity. In [3], the authors use these constrained neighborhoods in order to (a) extract the training parameters, and (b) learn the excitatory and inhibitory effects. But [2] suggests that disparity can be extracted from the compatibility matrix using the coherence property which predicts that the match values cluster along the different disparity planes, i.e., along $M_{m,n}(i+k, j-k)$. It is likely to be advantageous to consider all the match points, without any constrained neighborhoods, so that the network should be able to prune the neighborhood by itself.

NEURAL ARCHITECTURE

In order to train a neural network to learn the constraints embedded within the compatibility matrix, we need to present to the network inputs and the desired outputs. The implication is that we need a stereo pair, I_L and I_R, and its true disparity map, I_D, such that $I_D(m, n)$ gives the disparity of the corresponding pixel $I_L(m, n)$. Hence to train the network, the compatibility matrix $M_{m,n}$ is fed as the input, and the output forced to take on the value of $I_D(m, n)$.

We employ a linear network which is trained in a supervised manner using the gradient descent learning method. The network (Fig. 2) has a fully-connected, $4D^2$ input nodes and $2D_{max}$ number of output nodes, one for each disparity plane. The compatibility matrix ($2D$x$2D$) is fed as a ($4D^2$x1) vector, x, to the network, and the node corresponding to the expected disparity value is made high. i.e., if d_i is the desired output vector, then for (m, n)th pixel,

$$d_i = \begin{cases} 1 & : \text{ if } i = I_D(m,n) \\ 0 & : \text{ elsewhere} \end{cases} \tag{2}$$

where $i = 1, 2, ..., 2D_{max}$

Let $w_{i,j}$ be the weight connecting jth input node to the ith output node. Since we consider a linear node, the output of the ith output node, o_i is

$$o_i = \left(\sum_{j=1}^{4D^2} w_{i,j} * x_j \right) \tag{3}$$

where $i = 1, 2, ..., 2D_{max}$ and x_j is the jth element of the vector constructed from $M_{m,n}$.

We make use of the quadratic error function, E, for gradient descent defined as

$$E = \frac{1}{2} \sum_{i=1}^{2D_{max}} (o_i - d_i)^2 \tag{4}$$

where d_i is as given by eqn.2.

Now according to gradient descent, the change in the weight $w_{i,j}$ in the nth iteration is given by,

$$\Delta w_{i,j}^{(n)} = -\eta \frac{\partial E}{\partial w_{i,j}^{(n)}} \tag{5}$$

and,

$$\frac{\partial E}{\partial w_{i,j}^{(n)}} = (o_i - d_i) \frac{\partial o_i}{\partial w_{i,j}^{(n)}} \tag{6}$$

$$\frac{\partial o_i}{\partial w_{i,j}^{(n)}} = x_j \tag{7}$$

where $i, p = 1, 2, ..., 2D_{max}$; $j = 1, 2, ..., 4D^2$. and η is a pre-defined learning constant.

Hence the update equation is given by

$$w_{i,j}^{(n+1)} = w_{i,j}^{(n)} + \Delta w_{i,j}^{(n)} \tag{8}$$

We train the network using (a) the above equation, and (b) I_L, I_R and I_D. While testing the network, with some other stereo pairs, the output disparity value is calculated as the weighted average of the neurons output. Therefore, the disparity of the (m, n)th pixel is given by,

$$I_D(m, n) = \frac{\sum_{k=1}^{2D_{max}} (k - D_{max}) * o_k}{\sum_{k=1}^{2D_{max}} o_k} \tag{9}$$

where o is the output of the network, for the input x, constructed from the compatibility matrix $M_{m,n}$. Since the group response is considered, we get disparity with sub-pixel accuracy.

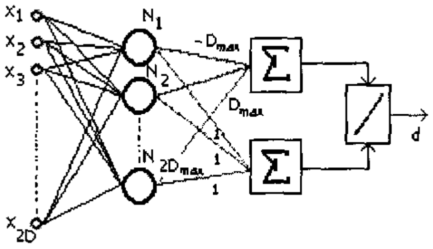
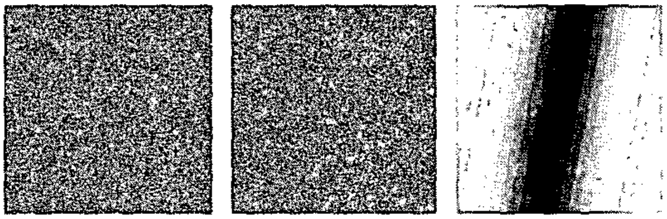


Figure 2: Neural network structure

EXPERIMENTAL RESULTS

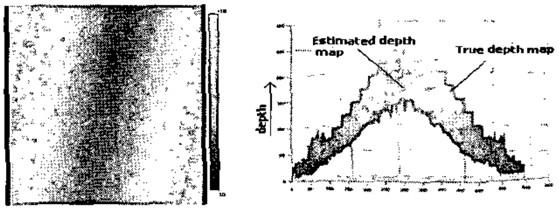
In order to train the neural network, we generate random-dot and textured stereo pairs from known disparity maps. Figure 3 shows a gray scale random-dot stereo pair (256x256), and the corresponding true disparity map, ranging from -5 to $+5$ pixels. The network was trained with a segment size of 20 (i.e., $D = 10$) and window size 6 (i.e., $W = 3$). The result is shown in Fig. 6 along with the 3D depth plot in Fig. 7. It is to be noted that the result obtained with $D = 5$ is comparable to that for $D = 10$. The continuity constraint is questionable, since a window with $D = 5$ will not have any neighbor from the same disparity plane for -5 and $+5$ disparity values, but the network is still able to recognize them.

Figure 5 shows the plot of weights as gray level image(negative) of different disparity nodes. Note that it has excitatory weights along the principal diagonal, but the inhibitory weights seem to be spread out on both sides, and do not seem to support the uniqueness constraint. When the disparity value increases, the excitation shifts away from the principal diagonal, and down-wards. For zero disparity, the excitation lies on the principal diagonal. Observe that the excitation decreases from the center, and there is high inhibition on both the sides.



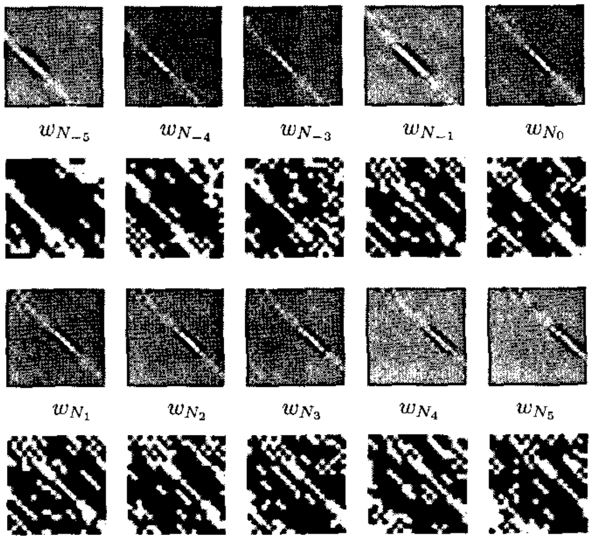
(a) (b) (c)

Figure 3: (a) and (b) Random-dot stereo pair and (c) its true disparity map ($D_{max} = 5$).



(a) (b)

Figure 4: (a) Estimated disparity map and (b) the 3D plot of estimated and the true disparity of the random dot stereo pair

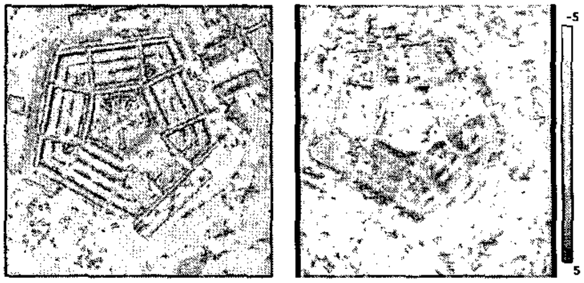


$w_{N_{-5}}$ $w_{N_{-4}}$ $w_{N_{-3}}$ $w_{N_{-1}}$ w_{N_0}

w_{N_1} w_{N_2} w_{N_3} w_{N_4} w_{N_5}

Figure 5: 2D image plot(negative) of weights of different disparity nodes, and their corresponding binary excitation(white)-inhibition(black) plots for $D = 10$ and $D_{max} = 5$

The network has been tested on the pentagon stereo pair. The results are shown in Fig. 6, and the corresponding 3D plot in Fig. 6. Observe that the disparity map is noisy compared to the results obtained from cooperative algorithms by Kanade *et al.* [5]. This shortcoming is perhaps due to the lack of cooperation between the neurons corresponding to different compatibility matrices. The network does not seem to perform well when the window under consideration has insufficient patterns in it. See Fig. 8, where the camera is misclassified.



(a) (b)

Figure 6: (a) Left image of pentagon stereo pair and (b) the estimated disparity map, with $D = 5$ and $W = 3$

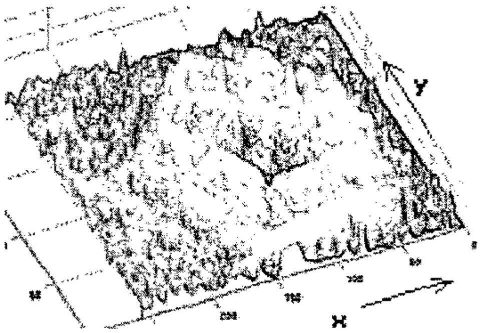
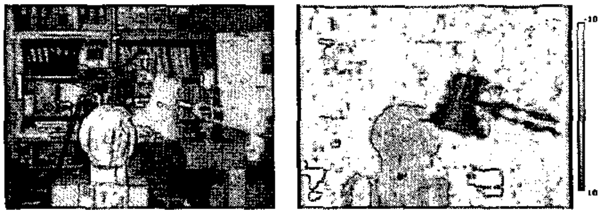


Figure 7: 3D plot of the estimated disparity map



(a) (b)

Figure 8: Right image of the tsukuba stereo pair

Even though the network works well for synthetic images, it fails to give comparably good results for noisy images. But this can be improved by constructing the 3D compatibility-matrix M, taking into account the neighborhood along the vertical direction as well, and by using better similarity measures to construct M.

Since the method is non-iterative, it is fast; and it can be implemented in parallel, since the compatibility matrix for each pixel can be constructed separately.

CONCLUSIONS

Based on the concept of cooperative stereo. a new method has been proposed, using an ANN, for stereo disparity estimation. Its performance compares favorably with that of recent algorithms (on stereo-image pair analysis) in the literature. Some illustrations are given.

REFERENCES

[1] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283–287, October 1976.

[2] R. D. Henkel, "A simple and fast neural network approach to stereo vision," *Proc. of NIPS'97 in Denver, MIT Press, Cambridge*, pp. 808–814, 1998.

[3] Alireza Khotanzad, Amol Bokil, and Ying-Wung Lee, "Stereopsis by constraint learning feed-forward neural networks," *IEEE Transactions on Neural Networks*, vol. 4, no. 2, pp. 332–342, March 1993.

[4] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proceedings of Royal Society London B*, vol. 204, pp. 301–328, 1979.

[5] C. Lawrence Zitnick and Takeo Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 675–684, July 2000.