STEREO MATCHING USING ARTIFICIAL NEURAL NETWORKS

Goung Loung Zheng Tan

Dept. of Information & Control Engineering
Xi'an Jiaotong University (710049) P.R.C.

Comm. III

Abstract:

It is essential to combine the low-level vision with the high-level vision in the stereo matching problem. At the high-level vision, the stereo matching problem often attribute to the shape recognition of the feature. The algorithm at this level is robust but much more time is consumed. While at the low-level vision, the stereo matching problem often attribute to the area-based correlation algorithm. The reliability of the result at this level is not satisfied.

In this paper, a neural network is employeed to overcome the shortcoming of the traditional methods. This network consists of feature detecting layer, pattern matching layer, stereo fusion layer, compound dicision layer. The output of the pattern mathcing layer is fed back to itself and the fusion layer is guided by the pattern matching layer. The stereo mathing process is completed when the condition of the compound layer is satisfied.

Key words: neural networks, stereo matching, fusion, pattern recognition.

1. INTRODUCTION

Stereo vision has a wide application such as in robotics, automatic surveillance, remote sensing, medical imaging etc. The depth information in stereo vision depends on the retinal disparity, which is the difference between the location of the retinal image points in the two eyes. The traditional method te get the depth information from a pair of image is that : area based correlation method and feature based matching method. Each method has its shortcoming. Some researchers attempt to integrate the two methods but have not got a full success yet.

More and more evidence reveals that stereo vision is a complex problem. Using a computer-generated stereo random-dot stereogram, we can find that the forming of stereo vision does not depend on the recognition, or say the understanding of the object in the image. The correspondence is completed point by point in the stereo fusion process. But it is very difficult to describe the fusion process in certain a mathematical formular. And it is unavoidable to fall in local minimal point when a matching process is attributed to a optimum problem such as correlation which is employeed to solve the problem of similarity between the two images. That is to say, it is not necessary to form the stereo fusion on the basis of the understanding of the objects in the image. While on the other hand, we can find out the corresponding points in this way : we view a certain point in the left image and then we can throw the left image away , viewing the corresponding point in the right image, and vice versa. Here we recognize the feature of the point being viewed and there is no stereo fusion acted in the process.

The stereo vision, or say the stereo matching can be realised either at low-level vision or high-level vision. At the low-level vision, the stereo matching emphasize the parallism of the fusion process. And it is meaningfull only when the fusion condition is satisfied. That is to say the fusion activate only when the points being viewd is close enough in position to the candidate points. While at high-level vision, stereo matching emphasize the recognition of the feature point, the shape, the edge structure in the image. The feature representd the salient point consist of the high-frequency part in the image. The other points consists of low-frequency part in the image. The stereo fusion is formed mainly by the low frenquency part in the image and restricted by the shape, the structure of the image. According to the physiology, the high frequency part performs the rivalry while the low frequency part perform fusion. If the matching problem is realized by the combination of stereo fusion with the feature recognition, it's no doubt that the robustness of the algorithm will be improved and the computational amount will be reduced.

In this paper, we present a way to solve the correspondence problem by using the parallel mechanism and the computational power offered by the artificial neural networks. The neural networks consist of four layer : feature detecting layer, pattern matching layer, stereo fusion layer, the compound decision layer. While the pattern matching layer is working, stereo fusion layer acts simultaneously in each segment devided by the feature points. It act under the guidance of the pattern matching layer. The compound decision layer ensure the reliability of the result and adjust the network to avoid it fall into a local minimum point. The disgram of the neural network see fig. 1.

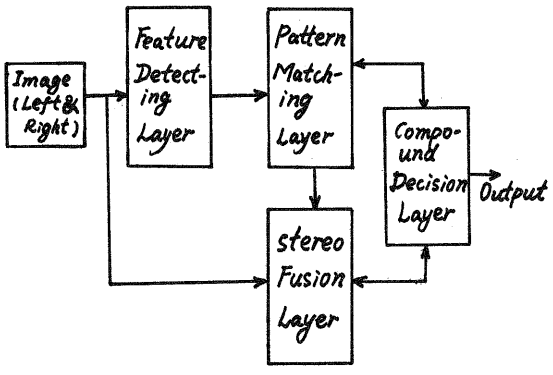

Fig. 1. The diagram of stereo matching neural networks

2. FEATURE DETECTING LAYER

In the common feature detecting problem, we always extract out the edges of the image in binary. The edges represent the discontinuties of the grey level in the image. But we loss so many important information in it such as the contrast of the edges to the background, the average value in a region with its neighbour. The feature we extracted out here is somewhat different from the conventional feature. The feature of every interest point is a group of six values in sequence such as: exist edge or not, the mean value with its eight neighbour points, mean square invariance with its eight neighbour points, medium value with its eight neighbour points..... The purpose of this is to describe the feature of interest points in every detail while not describe it in a binary value only to represent whether there is edge existed or not.

Here, we attribute the feature detecting process as a recognition process compared with a group of standard templete. This feature detecting layer is formed by a BP network. The structure of the network is shown in Fig. 2.

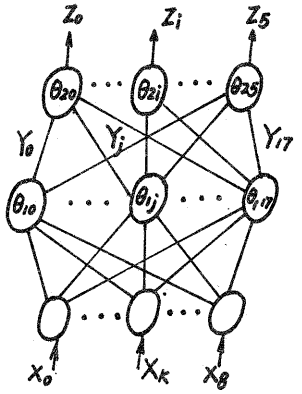

Fig. 2. Feature-detecting layer

The input node number : 9 X_k (k=0...8)
The hidden node number : 18 Y_j (j=0..17)
The output node number : 6 Z_i (i=0...5)
The calculation is done in parallel in the same layer but consequently from top to bottom between layers. The input of the network is the grey value in one 3x3 window.

The output of the hidden layer is:

$$Y_j = f_j\left(\sum_{k=0}^{8} W_{1kj} * X_k - \theta_{1j} \right) \qquad (1)$$

$$j = 0, 1, \ldots, 17$$

The output of the output-layer is :

$$Z_i = f_i\left(\sum_{j=0}^{17} W_{2ji} * Y_j - \theta_{2i} \right) \qquad (2)$$

where fi is a non-linear function

$$f_i(\alpha_i) = \frac{1}{1 + e^{-(\alpha_i - \theta_i)}} \qquad (3)$$

The network is trained according to the Back Propagation algorithm. The training step of the network is as followed :

step 1 :
Initiate W_{1kj}, W_{2ji}, θ_{1j}, θ_{2i} randomly with small non-zero value .
step 2 :
Input the templete image X_k, and the expected output value D_i. (D_i is got from the standard output from the templete image)
step 3 :
Shift the window along scanning line and calculate the output of the output-layer Z_i.
step 4 :
Adjust the weights and the threshold of the the network according the followed rule.
The weights and thresholds of the second layer:

$$W_{2ji}(t+1) = W_{2ji}(t) + \eta * \delta_{2i} * Y_{ji} + \alpha * (W_{2ji}(t) - W_{2ji}(t-1)) \qquad (4)$$

$$\theta_i(t+1) = \theta_i(t) - \eta * \delta_{2i} * C_{2i} \qquad (5)$$

where $\delta_{1i} = Z_i * (1-Z_i) * (D_i - Z_i)$,

C_{2i} is constant, i = 0...5 , j = 0...17.

The weights and the thresholds of the first layer:

$$W_{1kj}(t+1) = W_{1kj}(t) + \eta * \delta_{1j} * X_k + \alpha * (W_{1kj}(t) - W_{1kj}(t-1)) \qquad (6)$$

$$\theta_j(t+1) = \theta_j(t) - \eta * \theta_j(t) * C_{1j} \qquad (7)$$

where $\delta_{1j} = Y_i * \delta * W_{2j}$

C_{1j} is constant, j = 0...17 , k = 0....8

Step 5 :
Compare Z_{ij} with D_{ij}, if $|Z_{ij} - D_{ij}| < \varepsilon$, go to step 2; otherwise the training process ends.

In the actual application, we take $\alpha = 0.1$.

After the training process is completed, the network can be applied to detect the feature of the interest point. The feature of one interest point is denoted as $F_{i,j}$, where i means the position of the interest point in the image. j means the type of the feature. Here, j is from 0 to 5.

3. PATTERN RECOGNITION LAYER

It is not an easy task to find the corresponding points between the left image and the right image, especially when a number of interest points occurs in one image but does not occur in another image. Therefore, only a number of interest points in left image may find corresponding interest points in right image and vice versa. Each interest point in the matching process should satisfy the uniqueness constraint.

In this paper, we are supposed that the left image and the right image has been rectifyed after relative orientation, so that the search of corresponding interest points can be done alone the corresponding epipolar line. The epipolar lines are parallel to each other. So we match the corresponding interest points in one dimension. The structure of the continual edges is reflect in the mutual restriction between the adjacent epipolar line. That is to say, if continual edges occur across two epipolar lines, the

salient points at the edge should act simutaneously in the matching process of two adjacent epipolar lines. This can be controlled by software in the simulation process and can be performed in the real neural network. So that the structure information of the image act in hidden way in one dimension in the matching process.

As opposed in the feature detecting layer, the feature of one interest point consist of six parameters. The feature of an interest point is denoted by Fij, where i denote the type of the feature, j denote the input position of the interest points. The number of the interest point in left image and right image is Nl, Nr respectively. The model for pattern matching layer see Fig. 3.

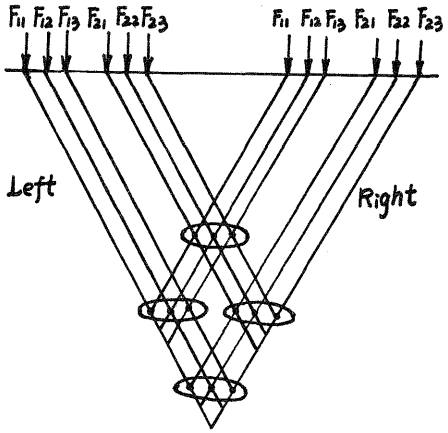

Fig. 3. The Model for Pattern Matching Layer

The intersection denoted by circles represent the possible matching element. They occur at intersection which bring the same type of feature together. All matching elements corresponding to a common position make a group for that position. And mutual inhibitory connections are defined between these groups in the same way as for the stereo matching network. Here, a group of feature in solid circle inhibit the other group of feature in the two oblique direction along the line of vision.

When using the neural network, the pattern matching problem can be formulated as the minimisation of a cost function (constrained optimisation). The cost function we adopted for the solution of the pattern matching problem is as follows :

$$E = -(1/2)\sum_{i=1}^{Nl}\sum_{k=1}^{Nr}\sum_{j=1}^{Nl}\sum_{l=0}^{Nr}\sum_{m=1}^{M} T_{ijklm}\, V_{ikm}\, V_{jlm} - \sum_{i=1}^{Nl}\sum_{k=1}^{Nr}\sum_{m=1}^{M} I_{ikm}\, V_{ikm} \qquad (8)$$

where M is the feature number of one interest point, Where V_{ikm} and V_{jlm} represent the binary state of ik and jl neurons respectively, which can be either 1(active) or 0(inactive). T_{ijklm} is the interconnection strength between the two neurons, I_{ikm} is the initial input to each neuron. A change in the state of neuron ik by ΔV_{ikm} cause an energy change of ΔE_{ik}.

$$\Delta E_{ik} = -\left[\sum_{j=1}^{Nl}\sum_{l=1}^{Nr}\sum_{m=1}^{M} T_{ijklm}\, V_{ikm}\, V_{jlm} + I_{ikm}\right]\Delta V_{ikm} \qquad (9)$$

The equation above describing the Dynamics of the network was shown by Hopfield to be always negative with a stochastic updating rule.

$$V_{ik} \to 0 \quad \text{if} \quad \left[\sum_{j=1}^{Nl}\sum_{l=1}^{Nr}\sum_{m=1}^{M} T_{ijklm}\, V_{ikm}\, V_{jlm} + I_{ikm}\right] > 0$$

$$V_{ik} \to 1 \quad \text{if} \quad \left[\sum_{j=1}^{Nl}\sum_{l=1}^{Nr}\sum_{m=1}^{M} T_{ijklm}\, V_{ikm}\, V_{jlm} + I_{ikm}\right] < 0$$

$$\text{no change if} \quad \left[\sum_{j=1}^{Nl}\sum_{l=1}^{Nr}\sum_{m=1}^{M} T_{ijklm}\, V_{ikm}\, V_{jlm} + I_{ikm}\right] = 0$$

The interconnection of the neural cell is indicated as Fig. 4.

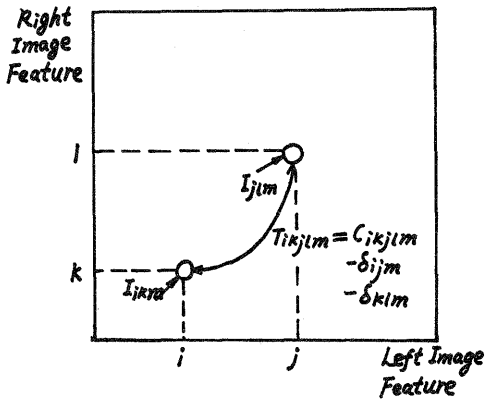

Fig. 4. Interconnection of neural cell

The deformation of the cost function for the stereo corresponding given below is minimized :

$$E = -\sum_{i=1}^{Nl}\sum_{k=1}^{Nr}\sum_{j=1}^{Nl}\sum_{l=0}^{Nr}\sum_{m=1}^{M} C_{ikjlm}\, P_{ikm}\, P_{jlm} + \sum_{i=1}^{Nl}\sum_{m=1}^{M}\left(1 - \sum_{k=1}^{Nr}\sum_{m=1}^{M} P_{ikm}\right)^2 + \sum_{k=1}^{Nr}\sum_{m=1}^{M}\left(1 - \sum_{i=1}^{Nl}\sum_{m=1}^{M} P_{ikm}\right)^2 \qquad (10)$$

The first term in (10) represent the degree of compatibility of a match between a pair of points (k, l) in the right image, while the second and third terms tend to enforce the uniquess constraint where the probabilities in each epipolar line should add up to 1. The compatibility measure is given by

$$C_{ikjlm} = \frac{2}{1 + e^{\lambda(x-\theta)}} - 1$$

$$X = W_1\,|\Delta d| + W_2\,|\Delta D|$$

where Δd is the differendce in the disparities of the matched points pairs (i, k) and (j, l). ΔD is

the difference between the distance from i to j and the difference from k to l.

The synaptic connection weight between two neurons is defined as:

$$T_{ikjlm} = (C_{ikjlm} - \delta_{ijm} - \delta_{klm})$$

where $\delta_{ijm} = 1$ if i=j, otherwise 0; $\delta_{klm} = 1$ if l=k, otherwise 0. The deformation of the cost function to the Lyapunov function of a Hopfield network with the neuron is defined as $V_{ik} = P_{ik}$, $V_{jl} = P_{jl}$.

The concrete convergence programm in the (10) equlity is proved difficult. We use an approximate energy change as followed. The mateamatical proof refers to [2].

$$\Delta E_{ikm} = - [\sum_{j=1}^{N_l} \sum_{l=0}^{N_r} \sum_{m=1}^{M} (C_{ikjlm} - \delta_{ijm} - \delta_{klm}) \cdot P_{jl} + 2] \Delta P_{ikm} \qquad (11)$$

According to the Hopfield updating rule

$$P_{ik} \to 0, \text{ if } | \sum_{j=1}^{N_l} \sum_{l=0}^{N_r} \sum_{m=1}^{M} (C_{ikjlm} - \delta_{ijm} - \delta_{klm}) P_{jl} + 2 | < 0$$

$$P_{ik} \to 1 \text{ if } | \sum_{j=1}^{N_l} \sum_{l=0}^{N_r} \sum_{m=1}^{M} (C_{ikjlm} - \delta_{ijm} - \delta_{klm}) P_{jl} + 2 | > 0$$

$$\text{no change if } | \sum_{j=1}^{N_l} \sum_{l=0}^{N_r} \sum_{m=1}^{M} (C_{ikjlm} - \delta_{ijm} - \delta_{klm}) P_{jl} + 2 | = 0$$

The optimal solution is completed when the Hopfield network is at its minimum energy point. However, it may settle down into one of the many locally stable state. So we cannot only rely on the stable point in the Hopfield network to get a full satisfication in the stereo matching process. We adopt stereo fusion layer for our further decisive basis. Another reason for the stereo fusion layer is that the interest points is so sparse that the result of the matching result cannot reconstruct the real surface of the object. Only when the pattern recognition layer and the stereo fusion layer convergence simultaneously, the result of the system is reliable. In the next section, we will discuss the stereo fusion layer.

The discrete (binary output) state was chosen in the pattern matching layer rather than the continuous value because of its simplicity in computational complexity. However, using a discrete Hopfield network, a number of local minima may not be avoided owing to the discontinuity of energy function caused by the discontinual interest points.

4. STEREO FUSION LAYER

The function of the stereo fusion layer is: It match the other points which are not the interest points. It perform the minimum of a energy fucnction which is based on the stereo fusion criterion. The stereo fusion is completed in local segments which is confined by the interest points. The surface of this local area is smooth for there is not salient point in this segment and so that this network may not fall into a local minimum point. The calculation of different segments is in separate and parallel way. There is no relation between the different segments for the depth may be discontinual at the interest point. But the difference of disparity between the neibour points in one segment should be very small, owing to the object rigidity and surface smoothness.

This layer is formed by another Hopfield network proposed by Y.S. Zhang [4]. The stereo fusion is assumed along the epipolar line. The energy funcion is given by :

$$E = \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=-D}^{D} [P_l(i, j) - P_r(i \oplus k, j)]^2 V_{i,j,k} + \lambda/2 \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=-D}^{D} \sum_{s \in S} (V_{i,j,k} - V_{i,m \circ s,k})^2 \qquad (12)$$

By comparising with the standard Hopfield network in two dimensional application:

$$E = -(1/2) \sum_{i=1}^{N_r} \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{l=1}^{N_c} \sum_{k=-D}^{D} \sum_{n=-D}^{D} T_{ijklmn} V_{ijk} V_{lmn} - \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=-D}^{D} I_{ijk} V_{imn} \qquad (13)$$

We can get

$$T_{ijkimn} = - 8\lambda \delta_{ij} \delta_{jm} \delta_{kn} + 2\lambda \sum_{s \in S} \delta_{il} \delta_{jms} \delta_{kn}$$

$$I_{i,j,k} = - [P_l(i, j) - P_r(i \oplus k, j)]^2$$

Where 2D+1 is the maximum disparity, S is an index set for four nearest neighbours at points (i, j), Nr and Nc is the image window row and column size, respectively. More detailed convergence of the network refers to the paper by Y.S. Zhang [4].

5. THE COMPOUND DECISION LAYER

The pattern matching layer match the interest points while the stereo fusion layer match the other points according to the stereo fusion criterion. The stereo fusion area is confined by the interest points so that the stereo fusion process is guided by the pattern matching layer. While the pattern matching layer and the stereo fusion may fall into local minimum points in the convergence process so that a compound decision layer is employeed to complete cooperative decision to enforce the reliability of the matching result.

Fig. 5 shows the possible matching cell between the left image and right image. We are supposed that there are two possible set of correspondence :

$$i \longleftrightarrow l, \; j \longleftrightarrow m, \; k \longleftrightarrow n$$

or $i \longleftrightarrow l, \; j \longleftrightarrow n.$

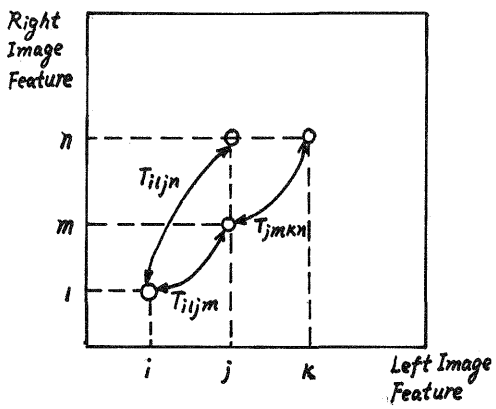

Fig. 5. Possible matching cell

The criterion of this layer is :

if $E_{iljm} < E_{iljn}$,
the correct correspondence should be

$i \longleftrightarrow l,\ j \longleftrightarrow m;$

otherwise the correct correspondence should be

$i \longleftrightarrow l,\ j \longleftrightarrow n$

where the E_{iljm} and E_{iljn} is the energy value at the stable point in the stereo fusion layer when the the correspondence is $j \longleftrightarrow m$ or $j \longleftrightarrow n$, respectively.

6. SIMULATION AND CONCLUSION

The preliminary simulation results shows that the internal power of this stereo matching system. The actual processing in human binocular vision is in parallel. Here, we make use of multi-feature rather than the single feature of the interest points to improve the correctness of matching result. Under the guidance of pattern matching layer, the possible candidate area in the stereo fusion layer is greatly reduced . The compound decision layer enforce the correctness by making a cooperative decision from the pattern matching layer result and the stereo fusion layer result. Although each individual neuron is slow, the network as a whole is very powerful, owing to the parallelism of the network.

The prospect of the parallelism of neural network in stereo vision is attractive. How to fully make use of the coorelative nature between the neibouring epipolar lines to apply our network in two dimension is to be researched further. The convergence process of the network will be faster and the algoritm will be more robust in two dimension.

REFERENCE

[1] U.R. Dhond and J.K. Aggarwal, "Structure from Stereo-A Review," IEEE Trans. Syst., Man, Cybern., Vol. 19 Nov. 1989.
[2] N. M. Nasrabadi and Chang Y. Choo, "Hopfield Network for Stereo Vision," IEEE Trans. Neural Networks, Vol. 3. Jan. 1992.
[3] Y. S. Zhang, "Automatic Computation of Relative Orientation Using Neural Networks" ISPRS Proceeding of Symp. Vol. 28. 1990 Wuhan, China.
[4] Y. S. Zhang "A Neural Networks Approach to Stereo Matching" ISPRS Proceeding of Symp. Vol. 28. 1990 Wuhan, China.
[5] J. Hopfield, "Neural Network and Physical system with emergent collective computational abilities," Proc. Nat. Acad. Sci., Vol. 79, April 1982.
[6] J. Hopfield and D. W. Tank, "Neural Computation of decision in optimization problems," Biol. Cybe. Vol. 52 Jan. 1985.
[7] Z. Tan and G. Loung, "Reconstructing 3-D Model from 2-D Images," Proc. of 1th Asia/Pacific Inter. Intru., Measu., Auto. Contr.