USC–SIPI REPORT #124

Stereo Matching Using a Neural Network

by

Yi-Tong Zhou and Rama Chellappa

March 1988

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 400
Los Angeles, CA 90089-2564 U.S.A.

Stereo Matching Using a Neural Network[1]

Y. T. Zhou and R. Chellappa

Signal and Image Processing Institute

Department of EE-Systems

University of Southern California

Abstract

A method for matching stereo images using a neural network is presented. Usually, the measurement primitives used for stereo matching are the intensity values, edges and linear features. Conventional methods based on such primitives suffer from amplitude bias, edge sparsity and noise distortion. We first fit a polynomial to find a smooth continuous intensity function in a window and estimate the first order intensity derivatives. Combination of smoothing and differentiation results in a window operator which functions very similar to the human eye in detecting the intensity changes. To give some insights into the resulting window operator, a theoretical analysis of the variances of the estimated derivatives is given. Since natural stereo images are usually digitized for the implementation on a digital computer, we consider the effect of spatial quantization on the estimation of the derivatives from natural images. A neural network is then employed to implement the

[1]This research work is partially supported by the AFOSR Contract No. F-49620-87-C-0007 and the AFOSR Grant No. 86-0196.

matching procedure under the epipolar, photometric and smoothness constraints based on the estimated first order derivatives. Owing to the dense intensity derivatives a dense array of disparities is generated with only a few iterations. This method does not require surface interpolation. Experimental results using random dot stereograms and natural images pairs are presented to demonstrate the efficacy of our method.

1 Introduction

Stereo matching is a primary means for recovering 3–D depth from two images taken from different viewpoints. The two central problems in stereo matching are to match the corresponding points and to obtain a depth map or disparity values between these points. In this paper we present a method for computing the disparities between the corresponding points in the two images recorded simultaneously from a pair of laterally displaced cameras based on the first order intensity derivatives. An implementation using a neural network is also given.

Basically, there exist two types of stereo matching methods: region based and feature based methods according to the nature of the measured primitives. The region based methods use the intensity values as the measurement primitives. A correlation technique or some simple modification is applied to certain local region around the pixel to evaluate the quality of matching. The region based methods usually suffer from the problems due to lack of local structures in homogeneous regions, amplitude bias between the images and

noise distortion. Recently, Barnard [1] applied a stochastic optimization approach for the stereo matching problem to overcome the difficulties due to homogeneous regions and noise distortion. Although this approach is different from the conventional region based methods, it still uses intensity values as the primitives with the aid of a smoothness constraint. Barnard's approach has several advantages: simple, suitable for parallel processing and a dense disparity map output. However, too many iterations, a common problem with the simulated annealing algorithm, makes it unattractive. It also suffers from the problem of amplitude bias between the two images.

The feature based methods use intensity edges or linear features (for example, see Grimson [2] and Medioni [3]) or intensity peaks which correspond to discontinuities in the first order derivatives of intensity [4]. The intensity edges are obtained using edge detectors such as the Marr–Hildreth edge detector [5] or the Nevatia–Babu line finder [6]. Since amplitude bias and small amount noise do not affect edge detection, feature based methods can handle natural images more efficiently. Owing to fewer measurement primitives to deal with, usually feature based methods are faster than the region based methods. However, a surface interpolation step has to be included. In order to obtain a smooth surface, several types of smoothing constraint techniques have been introduced [2]. The common problem in feature based methods is that if features are sparse, then surface interpolation step is difficult. Among the feature based methods, the Marr-Poggio-Grimson algorithm gives impressive results. But it is difficult to ensure continuity of disparity over

an area of the image. To overcome this problem, Grimson proposed a new version of their algorithm [7] including the figural continuity constraint [4] and other modifications. The figural continuity constraint is superior to the region continuity constraint. However, a occluding boundary or a sloping surface may cause problems. Another interesting approach is the integrated approach which consists of integrating matching, contour detection and surface interpolation steps [8]. The integrated approach uses no constraint other than the assumption of piecewise smoothness. A variety of stereo images were given in [8] to show the performance of this approach. Some problems of this approach reported by the authors are misplacement and missing of contours, and disparity errors due to inaccuracies of edge detection.

Julesz's example of random dot stereograms shows that stereo matching occurs very early in the visual process and is relatively independent of other forms of visual processing [9]. Early stereo process means that much more measurement primitives are used in matching. It seems that the region based methods are closer to the human stereo process than the edge based methods, because the intensity values are dense measurement primitives. However, region based methods suffer from the problems of amplitude bias and noise distortion, whereas human stereo process does not. The question then is what kind of measurement primitives human stereo process does use. Arguing that the amplitude bias can be eliminated by differential operation, the intensity derivatives are dense, and human visual system is sensitive to the intensity changes, the first order intensity

derivatives (simplest derivatives) may be considered as appropriate measurement primitives for the stereo matching problem. Noise distortion, which the first order derivatives are very sensitive to, can be reduced by some smoothing techniques such as a polynomial fitting technique. The first order intensity derivatives can be obtained by directly taking the derivative about the resulting continuous intensity function. Actually, the choice of window size is closely related to the theory of human visual system. There exits at least four independent channels containing different sized spatial filters in the early visual system [10, 11]. Combination of smoothing and differentiation results in a window operator which functions very similar to the human eye in detecting intensity changes. To give some insights into the resulting window operator, a theoretical analysis of the variances of the estimated derivatives is given. Since the natural stereo images are usually digitized for the implementation on a digital computer, we consider the effect of the spatial quantization on the estimation of the derivatives for the natural images.

Recently, many researchers have been using neural network for stereo matching based on either intensity values or edges [12, 13, 14, 15]. Early work on extracting the depth information from the random dot stereogram using neural network may be found in [16]. In this paper, we use a neural network with maximum evolution function to solve the stereo matching problem based on the first order intensity derivatives under the epipolar, photometric and smoothness constraints. We illustrate the usefulness of this approach by using the random dot stereograms and natural image pairs.

The organization of this paper is as follows: in Section 2, estimation of the first order intensity derivatives using discrete orthogonal polynomials is discussed, a theoretical analysis of the variances of the estimated derivatives and computational consideration of the effect of the spatial quantization error on the estimation of the derivatives from natural images are given. A neural network for stereo matching and model parameter estimation is described in Section 3. Computer simulations on the random dot stereograms and natural image pairs are presented in Section 4.

2 Estimation of the First Order Intensity Derivatives

Natural digital images usually are corrupted by certain amount of noise due to electronic imaging sensor, film granularity and quantization error. The derivatives obtained using a difference operator applied to digital images are not reliable. Since digital image comes about by sampling an analog image on an equally spaced lattice, a proper way to recover a smooth and continuous image surface is by a polynomial fitting technique. We first assume that, a point at the right image corresponding to a specified point in the left image lies somewhere on the corresponding epipolar line which is parallel to the row coordinate, i.e. in a horizontal direction, and second, in each neighborhood of image the underlying intensity function can be approximated by a fourth order polynomial. The first assumption

is also known as the epipolar constraint. With the help of this constraint, the first order intensity derivatives we need for matching are computed only for the horizontal direction. Under the second assumption, the intensity function in a window, centered at the point (i, j), of size $2\omega + 1$ is fitted by a polynomial of the form

$$g(i, j + y) = a_1 + a_2 y + a_3 y^2 + a_4 y^3 + a_5 y^4 \tag{1}$$

where y is lies in the range $-\omega$ to $+\omega$ and $\{a_i\}$ are coefficients. If the window size is 3, then a second order polynomial is sufficient to represent the intensity function. The first order intensity derivative at point (i, j) can easily be obtained by taking the derivative about $g(i, j + y)$ with respect to y and then setting $y = 0$

$$g'(i, j) \triangleq \frac{\partial g(i, j)}{\partial j} = \frac{dg(i, j + y)}{dy}|_{y=0} = a_2 \tag{2}$$

Thus, the estimation of the first order intensity derivatives is equivalent to determination of a_2.

2.1 Fitting Data Using the Chebyshev Polynomials

In order to estimate each coefficient independently, an orthogonal polynomial basis set is used. Several existing orthogonal polynomial basis sets can be found in [17, 18]. We use the discrete Chebyshev polynomial basis set, also used by Haralick for edge detection and topographic classification [19, 20]. The important property of using polynomials is that low order fits over a large window can reduce the effects of noise and give a smooth function.

Let a set of discrete Chebyshev polynomials be defined over an index set $\Omega = \{-\omega, -\omega+1, ..., \omega-1, \omega\}$, i.e. over a window of size $2\omega+1$, as

$$\begin{aligned} Ch_0(y) &= 1 \\ Ch_1(y) &= y \\ Ch_2(y) &= y^2 - q_2/q_0 \\ Ch_3(y) &= y^3 - (q_4/q_2)\, y \\ Ch_4(y) &= y^4 + [(q_2q_4 - q_0q_6)/(q_0q_4 - q_2^2)]\, y^2 + (q_2q_6 - q_4^2)/(q_0q_4 - q_2^2) \end{aligned} \tag{3}$$

where

$$q_n = \sum_{k\in\Omega} k^n.$$

With the window centered at point (i,j), the intensity function $g(i, j+y)$ for each $y \in \Omega$ can be obtained as

$$\hat{g}(i, j+y) = \sum_{m=0}^{4} d_m\, Ch_m(y) \tag{4}$$

where $\hat{g}(i, j+y)$ denotes the approximated continuous intensity function. For $\omega = 1$, only the first three Chebyshev polynomials are needed. By minimizing the least square error in estimation and taking advantage of the orthogonality of the polynomial set, the coefficients $\{d_m\}$ are obtained as

$$d_m = \frac{\sum_{y\in\Omega} Ch_m(y)\, g(i, j+y)}{\sum_{u\in\Omega} Ch_m^2(u)} \tag{5}$$

where $\{g(i, j+y)\}$ are the observed intensity values.

Expanding (4) and comparing with terms in (1), the first order intensity derivative

coefficient a_2, is given by

$$\begin{aligned} a_2 &= d_1 - \frac{q_4}{q_2} d_3 \\ &= \sum_{y \in \Omega} M(y)\, g(i, j+y) \end{aligned} \tag{6}$$

where $M(y)$ is determined by

$$M(y) = \frac{Ch_1(y)}{\sum_{u \in \Omega} Ch_1^2(u)} - \frac{q_4\, Ch_3(y)}{q_2 \sum_{u \in \Omega} Ch_3^2(u)} \tag{7}$$

For $\omega = 1$, the second term in (7) is zero. From (6) one can see that $M(y)$ is a filter for detecting intensity changes.

2.2 Analysis of Filter $M(y)$

Basically, the filter $M(y)$ used for detecting intensity changes has to satisfy the following requirements. First, it should eliminate amplitude bias completely. Second, it should remove noise very efficiently.

For simplicity of notation, we rewrite (6) as

$$a_2 = M(j) * g(i, j) \tag{8}$$

where "$*$" denotes the convolution operator. Suppose that the image is corrupted by amplitude bias b and additive white noise $\{n_{i,j}\}$ with zero mean and variance σ_n^2. The observed image is

$$\tilde{g}(i, j) = g(i, j) + b + n(i, j). \tag{9}$$

where $\tilde{g}(i,j)$ and $g(i,j)$ are the corrupted and original intensity functions, respectively. Noting that the filter $M(j)$ is an anti-symmetric function of j, the amplitude bias b is completely eliminated after convolution operation. Therefore,

$$M(j) * \tilde{g}(i,j) = M(j) * (g(i,j) + n(i,j)). \tag{10}$$

The expected value of the filter output can be written as

$$\mathbf{E}\{M(j) * \tilde{g}(i,j)\} = M(j) * g(i,j). \tag{11}$$

Accordingly, the variance can be expressed as

$$\begin{aligned} &\mathbf{E}\{(M(j) * \tilde{g}(i,j) - \mathbf{E}\{M(j) * \tilde{g}(i,j)\})^2\} \\ &= \mathbf{E}\{(M(j) * n(i,j))^2\} \\ &= \sigma^2 \sum_{j \in \Omega} M^2(j) \end{aligned} \tag{12}$$

By using (7), it is straightforward to prove that

$$\sum_{j \in \Omega} M^2(j) = \frac{q_6}{q_6 q_2 - q_4^2} \tag{13}$$

where

$$q_i = \sum_{y \in \Omega} y^i.$$

Hence, the variance of the filter output is

$$\mathbf{E}\{(M(j) * \tilde{g}(i,j) - \mathbf{E}\{M(j) * \tilde{g}(i,j)\})^2\} = \frac{\sigma_n^2 q_6}{q_6 q_2 - q_4^2} \tag{14}$$

For large window size, $q_6 \gg q_2$.The variance can be approximated as

$$\mathbf{E}\{(M(j) * \tilde{g}(i,j) - \mathbf{E}\{M(j) * \tilde{g}(i,j)\})^2\} = \frac{\sigma_n^2}{q_2} \tag{15}$$

From (15), one can see that the variance becomes smaller and smaller as the window size increases. For instance, if the window size is 5, then the variance is $0.9\sigma^2$. If the window size is 11, then the variance is significantly reduced to $0.009\sigma^2$. However, large window causes some loss of local information due to smoothing which smears or erases local features. If one desires to retain local features, then a small window may be used, but more noise remains and the estimated intensity function is rough. Also in order to reduce effect of the spatial quantization error for the natural images, a window as small as size of 3 may be used, as discussed in the next section. The variance of the estimated derivatives using a 3×3 window is the same as that in (15). It appears that the choice of the window size is closely related to theory of human visual system. It is known [10, 11] that at least four different size channels exist in a human visual system. Marr suggested [21] that in order to detect intensity changes efficiently, the filter used should be first a differential operator, taking either a first or second order spatial derivative of the image and second be capable of being tuned to act at any appropriate scale.

The following examples show that by choosing a proper window size the effects of noise can be eliminated very efficiently. A 256×256 real image is used here.

Example 1: An amplitude bias of strength 20 and white Gaussian noise (30 dB SNR) were added to the image. A section of the image is shown in the Figure 1. The dashed and

solid lines in Figure 1(a) represent the original and noisy images, respectively. Obviously, there is no way to match these two image based on the noisy biased intensity values only. Figure 1(b) shows the estimated first order intensity derivatives from these two images using the polynomial method. The window size is 5, i.e. the index set is $\{-2,-1,0,1,2\}$.

Example 2: An amplitude bias of size 20 and white Gaussian noise corresponding to 20 dB SNR were added to the original image. Figure 2 shows a section of the image taken from the same location as in Example 1. Figure 1(a) gives the original and noisy biased images. Figure 1(b) shows the estimated first order intensity derivatives of these two images. Since noise in this case is large, a large window of size 11 was used to reduce its effect. One can see that the derivatives of two images are matched very well.

2.3 Computational Consideration for Natural Images

For the implementation on a digital computer, the natural stereo images must be digitized both spatially and in the amplitude. Under the perspective projection, the natural stereo pair images, that is, the left and right images, can not be matched very well at sample points because of the spatial quantization error. The spatial quantization error affects the intensity function as well as the derivatives. In this section we consider the effect of spatial quantization error on the estimation of the intensity derivatives. Similar results also hold for edge detection. A recent discussion about the problems of the quantization error in

stereo matching can be found in [22].

For analysis purposes, a typical camera configuration system similar to that used by Horn [23] is given in Figure 3. Assume two cameras are rigidly attached to each other so that their optical axes are parallel and separated by a distance d. The focal length of the lens is represented by f which takes a negative value in the world coordinate system $OXYZ$. The origin of a right handed coordinate system of the world is located midway between the camera lens centers. The positive Z-axis is directed along the camera optical axes. The baseline connecting the lens centers is assumed to be perpendicular to the optical axes and oriented along the Y-axis. Let the coordinate systems of the left and right image plane be $o_L x_L y_L$ and $o_R x_R y_R$, respectively. Then a point in the world, (X, Y, Z), projects into the left and right image planes at

$$(x_L, y_L) = (\frac{f\,X}{Z}, \frac{f\,(Y + \frac{d}{2})}{Z}) \tag{16}$$

and

$$(x_R, y_R) = (\frac{f\,X}{Z}, \frac{f\,(Y - \frac{d}{2})}{Z}) \tag{17}$$

respectively. Disparity $D_{X,Y}$ can then be defined as

$$D_{X,Y} \triangleq y_R - y_L = -f\,d\,\frac{1}{Z} \tag{18}$$

Suppose we sample the left image uniformly at line $X_L = X_R = X_0$, a set of equally-spaced points $\{..., L_{-2}, L_{-1}, L_0, L_1, L_2, ...\}$ are obtained at

$$\{..., (x_0, y_{L_{-2}}), (x_0, y_{L_{-1}}), (x_0, y_{L_0}), (x_0, y_{L_1}), (x_0, y_{L_2}), ..., \}$$

The corresponding object points $\{..., P_{-2}, P_{-1}, P_0, P_1, P_2, ...\}$ are located at

$$\{..., (X_0, Y_{-2}, Z_{-2}), (X_0, Y_{-1}, Z_{-1}), (X_0, Y_0, Z_0), (X_0, Y_1, Z_1), (X_0, Y_2, Z_2), ...\}$$

on the surface. These object also project into the right image plane at

$$\{..., (x_0, y_{R_{-2}}), (x_0, y_{R_{-1}}), (x_{R_0}, y_{R_0}), (x_0, y_{R_1}), (x_0, y_{R_2}), ..., \}$$

When the object surface is not parallel to the image plane, the corresponding points on the surface are unequally-spaced. Consequently, the image points in the right image plane are also unequally-spaced which means that the image points do not match the sample points everywhere if the right image is uniformly sampled. This phenomenon is shown in Figure 3.

We assume that in the left image plane, the sample points match the image points exactly, and in the right image only the image point R_0 matches the sample point as illustrated in Figure 3 and other image points may not match the sample points. Thus

$$y_{L_i} - y_{L_0} = y_{L_i^s} - y_{L_0^s}$$

and

$$y_{L_i} - y_{L_0} = y_{R_i^s} - y_{R_0}$$

where the "s" denotes the sample point. The spatial quantization error i.e. the distance

between the sample point and the corresponding image point can be calculated as

$$\eta_i = \begin{cases} (y_{R_i} - y_{R_0}) - (y_{R_i'} - y_{R_0}) = (y_{R_i} - y_{R_0}) - (y_{L_i} - y_{L_0}) = f\,d\,(\frac{1}{Z_0} - \frac{1}{Z_i}), & i > 0 \\ (y_{R_0} - y_{R_i}) - (y_{R_0} - y_{R_i'}) = (y_{R_0} - y_{R_i}) - (y_{L_0} - y_{L_i}) = f\,d\,(\frac{1}{Z_i} - \frac{1}{Z_0}), & i < 0 \end{cases} \tag{19}$$

Obviously,

$$\eta_i > \eta_{i-1}, \quad i > 0,$$

and

$$\eta_i < \eta_{i-1}, \quad i < 0.$$

This shows that the spatial quantization error depends on the coordinate Z, focal length f and the distance d between the cameras. If the object surface is parallel to the image plane, then the sample points will match the corresponding image points perfectly because

$$Z_0 = Z_i, \quad \forall\ i.$$

An interesting aspect of (19) is that by definition of disparity in (18) the spatial quantization error is exactly equal to the difference of the disparities between the points P_0 and P_i. Therefore, stereo matching algorithms using intensity value as the measurement primitives can not detect such a difference if the sample interval is twice as large as the spatial quantization error.

We further assume that the incident illumination and absorption characteristics of the object surface are roughly constant, and the surface orientation and the distance to two cameras are almost same. Therefore, the left and right image planes receive the

same amounts of light which means the intensity functions of conjugate image points are almost same. Expand the intensity function $g(x_0, y_{R_i'})$ as a Taylor series about the point $(x_0, y_{R_i'}) = (x_0, y_{R_i})$

$$g(x_0, y_{R_i'}) = \begin{cases} g(x_0, y_{R_i} - \eta_i) &= g(x_0, y_{R_i}) - \eta_i \, g'(x_0, y_R)|_{y_R = y_{R_i}} + O(\eta_i^2), \quad i > 0 \\ g(x_0, y_{R_i} + \eta_i) &= g(x_0, y_{R_i}) + \eta_i \, g'(x_0, y_R)|_{y_R = y_{R_i}} + O(\eta_i^2), \quad i < 0 \end{cases} \tag{20}$$

where $g'(x_0, y_R)|_{y_R = y_{R_i}}$ is the derivative of intensity function at the point (x_0, y_{R_i}). By using the sampled intensity function to estimate the first order derivative of the intensity function $g(x_0, y_{R_0})$, (6) becomes

$$\tilde{g'}(x_0, y_{R_0}) = \sum_{y \in \Omega} M(y) \, g(x_0, (y_{R_0} + y)^s) \tag{21}$$

where the "$\sim$" denotes the estimate of the intensity derivative using the sampled intensity functions.

Replacing the sampled intensity functions in (21) by (20), we have

$$\tilde{g'}(x_0, y_R)|_{y_R = y_{R_0}} \simeq \sum_{i \in \Omega} M(i) \, g(x_{R_0}, y_{R_i}) + \sum_{i \in \Omega} u(i) \, \eta_i \, M(i) \, g'(x_0, y_R)|_{y_R = y_{R_i}} \tag{22}$$

where $u(i)$ is a step function

$$u(i) = \begin{cases} 1, & i > 0 \\ -1, & i < 0. \end{cases}$$

Clearly, the first term in the right side of (22) is equal to (6) which means it is a correct estimate, and the second term is an estimation error caused by the spatial quantization

error. Since the spatial quantization error is proportional to f and d, and is inversely proportional to Z, the estimation error will be small when the camera is close enough and/or the object is far enough. When the surface is not parallel to the image planes and the object is close to the camera, using a large window to estimate the derivatives will give a large error due to the accumulated quantization error. Hence, a small window is preferred if the object is close to the camera. As proposed in [24], the smallest channel in the human visual system contains a filter with a central diameter of 1.5′, roughly corresponding to 4 pixels. Therefore, considering the effects of noise distortion and the spatial quantization error, a filter $m(y)$ with size of $3-7$ pixels is the proper one for the natural stereo images.

In fact, (22) can be considered as a general form for both derivative estimation and edge detection as most of the edge detection algorithms can be considered as a window operation followed by appropriate thresholding. Owing to the output error of window operation, the edge detector may miss an edge, give a false edge or shift the edge. In other words, the edge output also suffers from the spatial quantization error.

Noting that the filter $m(y)$ is anti-symmetric function of y and assuming that the derivatives at sample points are the same, (22) can be simplified as

$$\tilde{g}'(x_0, y_r)|_{y_r=y_{r_0}} \simeq g'(x_0, y_r)|_{y_r=y_{r_0}}[1 - \sum_{i=1}^{\omega} m(i)\ (\eta_i + \eta_{-i})]. \tag{23}$$

Substituting (19) and then (18), we finally have

$$\tilde{g}'(x_0, y_r)|_{y_r=y_{r_0}} \simeq g'(x_0, y_r)|_{y_r=y_{r_0}}[1 - \sum_{i=1}^{\omega} m(i)\ (D_i - D_{-i})] \tag{24}$$

The estimate of the derivatives may be either larger or smaller than the true value which depends on the orientation of the object surface.

3 A Neural Network for Matching

3.1 A Neural Network

We use a discrete neural network containing binary neurons for representing the disparity values between the two images. The model consists of $N_r \times N_c \times D$ mutually interconnected neurons, where D is the maximum disparity, N_r and N_c are the image row and column sizes, respectively. Let $V = \{v_{i,j,k}, 1 \leq i \leq N_r, 1 \leq j \leq N_c, 0 \leq k \leq D\}$ be a binary state set of the neural network with $v_{i,j,k}$ (1 for firing and 0 for resting) denoting the state of the (i,j,k)th neuron. Especially, when the neuron $v_{i,j,k}$ is 1, this means that the disparity value is k at the point (i,j). Every point is represented by $D+1$ mutually exclusive neurons, i.e. only one neuron is firing and others are resting, due to the uniqueness constraint of the matching problem. Let $T_{i,j,k;l,m,n}$ denote the strength (possibly negative) of the interconnection between neuron (i,j,k) and neuron (l,m,n). We require symmetry

$$T_{i,j,k;l,m,n} = T_{l,m,n;i,j,k} \quad for \quad 1 \leq i,l \leq N_r \;\; 1 \leq j,m \leq N_c \;\; and \;\; 0 \leq k,n \leq D$$

We also insist that the neurons have self-feedback, i.e. $T_{i,j,k;i,j,k} \neq 0$. In this model, each neuron (i,j,k) randomly and asynchronously (or synchronously) receives inputs from all

neurons and a bias input

$$u_{i,j,k} = \sum_{l=1}^{N_r} \sum_{m=1}^{N_c} \sum_{n=0}^{D} T_{i,j,k;l,m,n} v_{l,m,n} + I_{i,j,k} \tag{25}$$

Each $u_{i,j,k}$ is fed back to corresponding neurons after maximum evolution

$$v_{i,j,k} = g(u_{i,j,k}) \tag{26}$$

where $g(x_{i,j,k})$ is a nonlinear maximum evolution function whose form is taken as

$$g(x_{i,j,k}) = \begin{cases} 1 & if \;\; x_{i,j,k} = max(x_{i,j,l}; l = 0, 1, ..., D). \\ 0 & otherwise. \end{cases} \tag{27}$$

In the asynchronous updating case, the state of each neuron is updated by using the latest information about other neurons. While, in the synchronous updating case, the information received by the neuron at time t is about previous states of other neurons at time $t-1$. The uniqueness of matching problem is ensured by a batch updating scheme–$D+1$ neurons $\{v_{i,j,0}, ...v_{i,j,D}\}$ at site (i,j) are updated at each step simultaneously.

3.2 Estimation of Model Parameters

The neural model parameters, the interconnection strengths and the bias inputs, can be determined in terms of the energy function of the neural network. As defined in [25], the energy function of the neural network can be written as

$$E = -\frac{1}{2} \sum_{i=1}^{N_r} \sum_{l=1}^{N_r} \sum_{j=1}^{N_c} \sum_{m=1}^{N_c} \sum_{k=0}^{D} \sum_{n=0}^{D} T_{i,j,k;l,m,n} \, v_{i,j,k} \, v_{l,m,n} - \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=0}^{D} I_{i,j,k} \, v_{i,j,k} \tag{28}$$

In order to use the spontaneous energy–minimization process of the neural network, we reformulate the stereo matching problem under the epipolar assumption as one of minimizing an error function with constraints defined as

$$E = \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=0}^{D} \left(g_L'(i,j) - g_R'(i, j \oplus k)\right)^2 v_{i,j,k} + \frac{\lambda}{2} \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=0}^{D} \sum_{s \in S} \left(v_{i,j,k} - v_{(i,j)\oplus s,k}\right)^2 \quad (29)$$

where $\{g_L'(\cdot)\}$ and $\{g_R'(\cdot)\}$ are the first order intensity derivatives of the left and right images, respectively, S is an index set excluding $(0,0)$ for all neighbors in a $\Gamma \times \Gamma$ window centered at point (i,j), λ is a constant and the symbol $\oplus$ denotes that

$$f_{a \oplus b} = \begin{cases} f_{a+b} & if \;\; 0 \le a+b \le N_c, N_r \\ 0 & otherwise \end{cases}$$

The first term called the photometric constraint in (29) is to seek disparity values such that all regions of two images are matched as close as possible in a least squares sense. Meanwhile, the second term is the smoothness constraint on the solution. The constant λ determines the relative importance of the two terms to achieve the best results.

By taking $\Gamma = 5$ and comparing the terms in the expansion of (29) with the corresponding terms in (28), we can determine the interconnection strengths and bias inputs as

$$T_{i,j,k;l,m,n} = -48\lambda \delta_{i,l} \delta_{j,m} \delta_{k,n} + 2\lambda \sum_{s \in S} \delta_{(i,j),(l,m)\oplus s} \delta_{k,n} \quad (30)$$

and

$$I_{i,j,k} = -\left(g_L'(i,j) - g_R'(i, j \oplus k)\right)^2 \quad (31)$$

where $\delta_{a,b}$ is the Dirac delta function. The size of the smoothing window used in (30) is 5. However, one can choose either larger or smaller window. From (30) one can see that the self-connection $T_{i,j,k;i,j,k}$ is not zero which requires self-feedback for neurons.

3.3 Stereo Matching

Stereo matching is carried out by neuron evaluation. Once the parameters $T_{i,j,k;l,m,n}$ and $I_{i,j,k}$ are obtained using (30) and (31), each neuron can randomly and asynchronously (or synchronously) evaluate its state and readjust accordingly using (25) and (26). The synchronous updating scheme can be implemented in parallel, while the asynchronous updating scheme can be sequentially implemented without loss of generality. Another updating scheme called the hybrid updating scheme is that some neurons are synchronously updated and others are asynchronously updated. For natural stereo images, we will use this hybrid neural network to overcome the difficulty of lack of local structures in homogeneous regions.

The initial state of the neurons were set as

$$v_{i,j,k} = \begin{cases} 1 & if \;\; I_{i,j,k} = max(I_{i,j,l}; l = 0, 1, ..., D). \\ 0 & otherwise \end{cases} \tag{32}$$

where $I_{i,j,k}$ is the bias input.

However, this neural network has self-feedback, i.e. $T_{i,j,k;i,j,k} \neq 0$, as a result the energy function E does not always decrease monotonically with a transition. This is explained as

follows. Since we are using a batch updating scheme, $(D+1)$ neurons $\{v_{i,j,k}; k = 0, ..., D\}$ corresponding to the image point (i, j) are simultaneously updated at each step. However, at most two of the $(D+1)$ neurons change their state at each step. Define the state changes $\Delta v_{i,j,k}$ and $\Delta v_{i,j,k'}$ of neurons (i, j, k) and (i, j, k') and energy change ΔE as

$$\Delta v_{i,j,k} = v_{i,j,k}^{new} - v_{i,j,k}^{old}$$

$$\Delta v_{i,j,k'} = v_{i,j,k'}^{new} - v_{i,j,k'}^{old}$$

and

$$\Delta E = E^{new} - E^{old}$$

Consider the energy function

$$E = -\frac{1}{2}\sum_{i=1}^{N_r}\sum_{l=1}^{N_r}\sum_{j=1}^{N_c}\sum_{m=1}^{N_c}\sum_{k=0}^{D}\sum_{n=0}^{D} T_{i,j,k;l,m,n}\, v_{i,j,k}\, v_{l,m,n} - \sum_{i=1}^{N_r}\sum_{j=1}^{N_c}\sum_{k=0}^{D} I_{i,j,k}\, v_{i,j,k}. \qquad (33)$$

The change ΔE due to a changes $\Delta v_{i,j,k}$ and $\Delta v_{i,j,k'}$ given by

$$\begin{aligned}\Delta E \;=\; & -(\sum_{l=1}^{N_r}\sum_{m=1}^{N_c}\sum_{n=0}^{D} T_{i,j,k;l,m,n}\, v_{l,m,n} + I_{i,j,k}\,)\,\Delta v_{i,j,k} - \frac{1}{2}\,T_{i,j,k;i,j,k}\,(\Delta v_{i,j,k})^2 \\ & -(\sum_{l=1}^{N_r}\sum_{m=1}^{N_c}\sum_{n=0}^{D} T_{i,j,k';l,m,n}\, v_{l,m,n} + I_{i,j,k'}\,)\,\Delta v_{i,j,k'} - \frac{1}{2}\,T_{i,j,k';i,j,k'}\,(\Delta v_{i,j,k'})^2 \\ & -T_{i,j,k;i,j,k'}(\Delta v_{i,j,k} v_{i,j,k'}^{new} + \Delta v_{i,j,k'} v_{i,j,k}^{new}) \qquad (34)\end{aligned}$$

is not always negative. For instance, since

$$u_{i,j,k} = \sum_{l=1}^{N_r}\sum_{m=1}^{N_c}\sum_{n=0}^{D} T_{i,j,k;l,m,n}\, v_{l,m,n} + I_{i,j,k}$$

and

$$u_{i,j,k'} = \sum_{l=1}^{N_r} \sum_{m=1}^{N_c} \sum_{n=0}^{D} T_{i,j,k';l,m,n}\, v_{l,m,n} + I_{i,j,k'}$$

if

$$v_{i,j,k}^{old} = 0, \qquad v_{i,j,k'}^{old} = 1,$$

$$u_{i,j,k} > u_{i,j,k'}$$

and the maximum evolution function is as in (27), then

$$v_{i,j,k}^{new} = 1, \qquad v_{i,j,k'}^{new} = 0$$

and

$$\Delta v_{i,j,k} = 1, \qquad \Delta v_{i,j,k'} = -1.$$

Noting that

$$T_{i,j,k;i,j,k'} = 0 \quad if \quad k \neq k',$$

(34) can be simplified as

$$\Delta E = (u_{i,j,k'} - u_{i,j,k}) - \frac{1}{2}\,(T_{i,j,k;i,j,k} + T_{i,j,k';i,j,k'}) \tag{35}$$

Thus, the first term in (35) is negative. But

$$T_{i,j,k;i,j,k} + T_{i,j,k';i,j,k'} = -96\,\lambda < 0$$

leading to

$$-\frac{1}{2}\,(T_{i,j,k;i,j,k} + T_{i,j,k';i,j,k'}) > 0$$

When the first term is less than the second term in (34), then $\Delta E > 0$ (we have observed this in our experiments), which means E is not a Lyapunov function and hence the network is unstable. Consequently, the convergence of the network is not guaranteed [26].

To ensure convergence of the network probably to a local minimum, we have designed a deterministic decision rule. The rule is to take a new state $v_{i,j,k}^{new}$ and $v_{i,j,k'}$ of neurons (i,j,k) and (i,j,k') if the energy change ΔE due to state changes $\Delta v_{i,j,k}$ and $\Delta v_{i,j,k'}$ is less than zero. If ΔE due to state change is > 0, no state change is affected. A stochastic decision rule can also be used to obtain a globally optimal solution [27, 28].

The stereo matching algorithm can then be summarized as

1. Set the initial state of the neurons.

2. Update the state of all neurons randomly and asynchronously (or synchronously) according to the decision rule.

3. Check the energy function; if energy does not change anymore, stop; otherwise, go back to step 2.

4 Experimental Results

A variety of images including random dot stereograms and natural stereo image pairs were tested using our algorithm. A 5×5 (i.e. $\Gamma = 5$) smoothing window was used for all images.

4.1 Random Dot Stereograms

The random dot stereograms were created by the pseudo random number generating method described in [29]. Each dot consists of only one element. All the following random dot stereograms are of size 128×128 and in the form of a three level "wedding cake". The background plane has zero disparity and each successive layer plane has additional two elements of disparity. In order to implement this algorithm more efficiently on a conventional computer, we make the following simplifications. Since only one of $D+1$ neurons is firing at each point, we used one neuron lying in the range 0 to D to represent the disparity value instead of $D+1$ neurons. From (30) one can see that the interconnections between the neurons are local (a $\Gamma \times \Gamma$ neighborhood) and have the same structure for all neurons. Therefore, for $\Gamma = 5$ we used a 5×5 window for computing $U_{i,j,k}$ and energy function E instead of a $N_r N_c(D+1) \times N_r N_c(D+1)$ interconnection strength matrix. The simplified algorithm greatly reduces the space complexity by increasing the program complexity little. Therefore, it is very fast and efficient.

Figure 4 shows a 10% random dot stereogram. Intensity values of the white and black elements are 255 and 0, respectively. Figure 4(c) is the resulting disparity map after 10 iterations using asynchronous updating scheme. When the synchronous updating scheme is used, 23 iterations are needed. The disparity values are encoded as intensity values with the brightest value denoting the maximum disparity value. We used $\lambda = 20$, $D = 6$ and $\omega = 2$ (i.e. window size was 5). Note that the disparity map is dense.

A similar test was run on the decorrelated stereogram [9]. The original stereogram is 50% density random dots. In the left image, 20% of the dots were decorrelated at random. By setting $\lambda = 2800$, $D = 6$ and $\omega = 2$, a dense disparity map in Figure 5(c) was obtained after 12 asynchronous iterations. The same result can be obtained after 19 synchronous iterations. .

Another type of perturbation is Gaussian white noise [29]. Figures 6(a) and 6(b) show a pair of multi gray level random dot images with intensity value in the range $(0 - 255)$. Gaussian white noise corresponding to 5 dB SNR was added to the left image. The SNR is defined as

$$SNR = 10 \log_{10} \frac{\sigma_o^2}{\sigma_n^2}$$

where σ_o^2 and σ_n^2 are the variances of original left image and noise. The parameters were set as $\lambda = 450$, $D = 6$ and $\omega = 2$. Only 6 asynchronous iterations were needed to get the final result in Figure 6(c). Using synchronous updating scheme needed 9 iterations to get the same result.

As expected, both the synchronous and asynchronous updating schemes work very well, although the latter takes more iterations. The synchronous updating scheme is suitable for parallel processing.

4.2 Natural Stereo Images

Two stereo pairs of natural images, the Renault part and the Pentagon images, are used to test our algorithm. All images are of size 256×256. Since natural stereo images may not satisfy the epipolar constraint, small alignment corrections in the vertical direction are needed. A hybrid updating scheme was used for both the Renault and the Pentagon image pairs. The image is segmented into homogeneous and inhomogeneous regions by using a local variance criterion. Homogeneous region is defined as a smooth region with the small local variances. In homogeneous image regions, the corresponding neurons are updated sequentially, while other neurons corresponding to inhomogeneous regions are updated in parallel. Since the first derivatives of the intensity function in homogeneous regions are small, the inputs are small and the neurons intend to take the same state as their neighbors because of the smoothness constraint. No doubt, the neurons near the boundary will be first affected by the neighbors corresponding to inhomogeneous regions. As the neurons corresponding to homogeneous region are sequentially updated, they will all be affected by the boundary conditions which means surfaces in homogeneous regions can be interpolated.

For the Renault images, the parameters were set as $\lambda = 12$, $D = 13$ and $\omega = 1$. The threshold for the local variances was set to 1.0. The local variance was computed in a 5×5 window. About 72 iterations were required. Since a discrete network was used, the disparity only takes integer number. A simple smoothing technique was applied to the nonzero portions of the disparity surface. Figures 7(a) and 7(b) show the left and right

Renault part images. The final result is given in Figure 7(c), while (d) shows the smoothed version of (c) by using a 9×9 mean filter. Figures 8 and 9 give the plots of the unsmoothed and smoothed disparity surfaces corresponding to Figures 7(c) and 7(d), respectively.

Figures 10(a) and 10(b) show the left and right Pentagon images. By choosing parameters $\lambda = 10$, $D = 4$, $\omega = 1$ and the local variance threshold 0.01, a disparity map was generated after 51 iterations. The window for computing the local variance is of size 5×5. Figures 10(c) and 10(d) give the unsmoothed and smoothed disparity maps, respectively. A 13×13 filter was used for smoothing. The plots of Figures 7(c) and 7(d) are shown in Figures 11 and 12, respectively.

5 Acknowledgement

The authors are thankful to Mr. Andres Huertas of the Institute of Robotics and Intelligent Systems at the University of Southern California for providing the Pentagon and Renault part image data.

References

[1] S. T. Barnard, "A Stochastic Approach to Stereo Vision", In *Proc. Fifth National Conf. on Artificial Intelligence*, Philadelphia, PA, August 1986.

[2] W. E. L. Grimson, *From Images to Surfaces*, The MIT press, Cambridge, MA, 1981.

[3] G. Medioni and R. Nevatia, "Segment-Based Stereo Matching", *Computer Vision, Graph. and Image Processing*, vol. 31, pp. 2–18, 1985.

[4] J. E. W. Mayhew and J. P. Frisby, "Psychophysical and Computational Studies towards a Theory of Human Stereopsis", *Artificial Intelligence*, vol. 17, pp. 349–385, August. 1981.

[5] D. Marr and E. C. Hildreth, "Theory of Edge Detection", *Proc. Royal Society of London*, vol. B-207, pp. 187–217, February 1980.

[6] R. Nevatia and K. R. Babu, "Linear Feature Extraction and Description", *Computer Graph. and Image Processing*, vol. 13, pp. 257–269, 1980.

[7] W. E. L. Grimson, "Computational Experiments with a Feature Based Stereo Algorithm", *IEEE Trans. on Patt. Anal. and Mach. Intel.*, vol. PAMI-7, pp. 17–34, January 1985.

[8] W. Hoff and N. Ahuja, "Extracting Surfaces from Stereo Images: An Integrated Approach", In *Proc. First International Conf. on Computer Vision*, pp. 284–294, London, England, June 1987.

[9] B. Julesz, *Foundations of Cyclopean Perception*, The University of Chicago Press, Chicago, IL, 1971.

[10] H. R. Wilson and S. C. Giese, "Threshold Visibility of Frequency Grating Patterns", *Vision Research*, 17, pp. 1177–1190, 1977.

[11] H. R. Wilson and J. R. Bergen, "A Four Mechanism Model for Threshold Spatial Vision", *Vision Research*, 19, pp. 19–32, 1979.

[12] N. M. Grzywacz and A. L.Yuille, "Motion Correspondence and Analog Networks", In *Proc. Conf. on Neural Networks for Computing*, pp. 200–205, American Institute of Physics, Snowbird, UT, 1986.

[13] C. V. Stewar and C. R. Dyer, "A Connectionist Model for Stereo Vision", In *Proc. IEEE First Annual Intl. Conf. on Neural Networks*, San Diego, CA, June 1987.

[14] G. Z. Sun, H. H. Chen, and Y. C. Lee, "Learning Stereopsis with Neural Networks", In *Proc. IEEE First Annual Intl. Conf. on Neural Networks*, San Diego, CA, June 1987.

[15] A. F. Gmitro and G. R. Gindi, "Optical Neurocomputer for Implementation of the Marr–Poggio Stereo Algorithm", In *Proc. IEEE First Annual Intl. Conf. on Neural Networks*, San Diego, CA, June 1987.

[16] P.Dev, "Perception of Depth Surfaces in Random-dot Stereogram: a Neural Model", *Int. J. Man-Machine Studies*, 7, pp. 511–528, 1975.

[17] G. G. Lorentz, *Approximation of functions*, Holt, Rinehart and Winston, New York, 1966.

[18] P. Beckmann, *Orthogonal Polynomials for Engineers and Physicists*, Golem, Boulder, CO, 1973.

[19] R. M. Haralick, "Digital Step Edges from Zero Crossings of Second Directional Derivatives", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-6, pp. 58–68, January 1984.

[20] T. J. Laffey, R. M. Haralick, and L. T. Watson, "Topographic Classification of Digital Image Intensity Surfaces", In *The proc. IEEE Workshop on Computer Vision: Theory and Control*, Rindge, New Hampshire, August 1982.

[21] D. Marr, *Vision*, W. H. Freeman and Company, New York, 1982.

[22] S. D. Blostein and T. Huang, "Quantization Errors in Stereo Triangulation", In *Proc. First International Conf. on Computer Vision*, pp. 325–334, London, England, June 1987.

[23] B. K. P. Horn, *Robot Vision*, The MIT Press, Cambridge, Massachusetts, 1986.

[24] D. Marr, T. Poggio, and E. C. Hildreth, "The Smallest Channel in Early Human Vision", *J. Opt. Soc. Am.*, vol. 90, pp. 868–870, 1979.

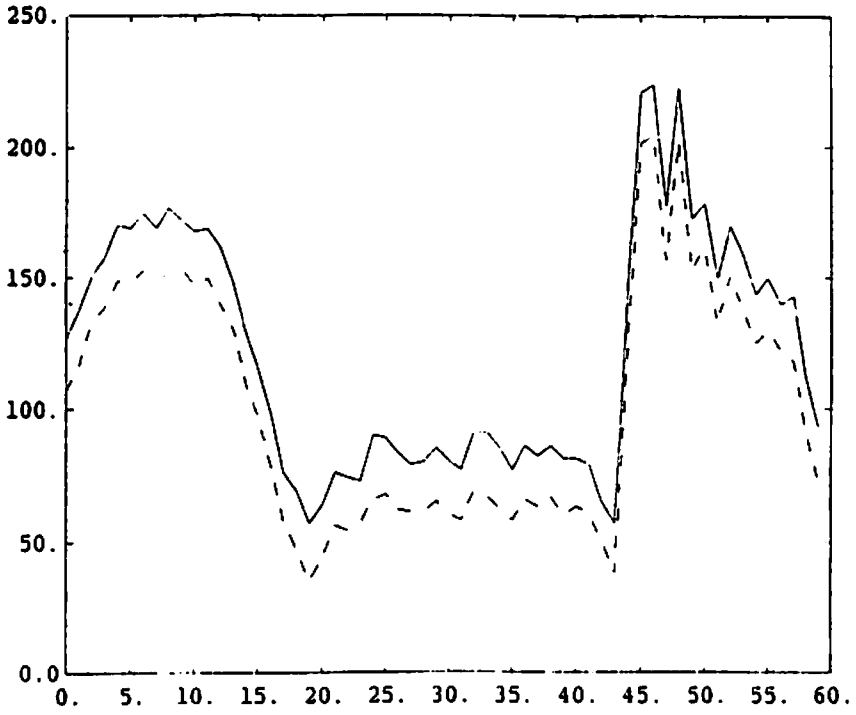
[25] J. J. Hopfield and D. W. Tank, "Neural Computation of Decisions in Optimization Problems", *Biological Cybernetics*, vol. 52, pp. 141–152, 1985.

[26] J. P. LaSalle, *The Stability and Control of Discrete Processes*, Springer-Verlag, New York, New York, 1986.

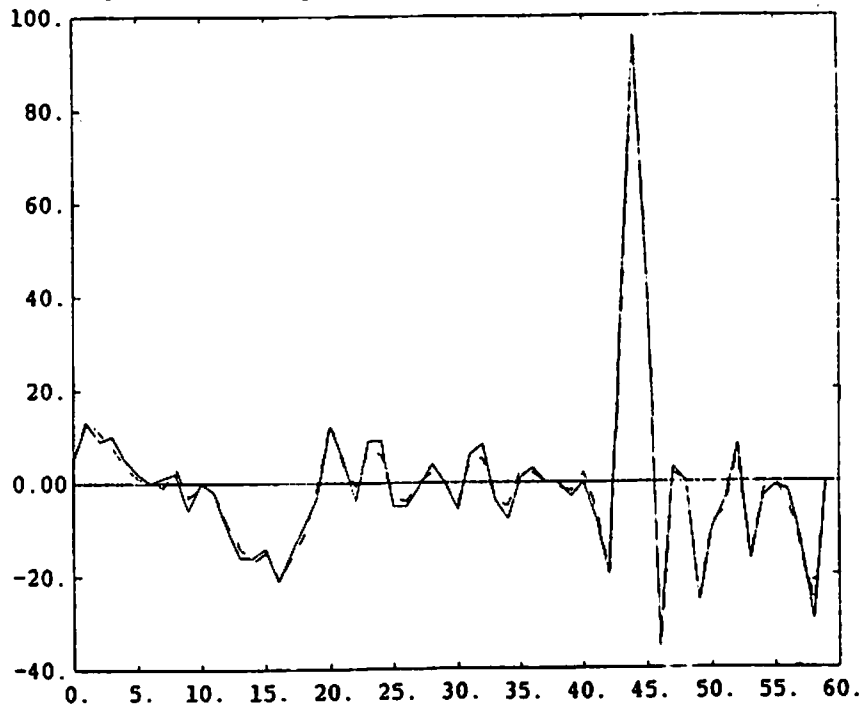
[27] Y. T. Zhou, R. Chellappa, and B. K. Jenkins, "A Novel Approach to Image Restoration Based on a Neural Network", In *Proc. IEEE First Annual Intl. Conf. on Neural Networks*, San Diego, CA, June 1987.

[28] Y. T. Zhou, R. Chellappa, A. Vaid, and B. K. Jenkins, "Image Restoration Using a Neural Network", To appear in IEEE Trans. Acoust, Speech and Signal Processing, July, 1988.

[29] B. Julesz, "Binocular Depth Perception of Computer-Generated Patterns", *Bell System Technical J.*, vol. 39, pp. 1125–1162, Sept. 1960.

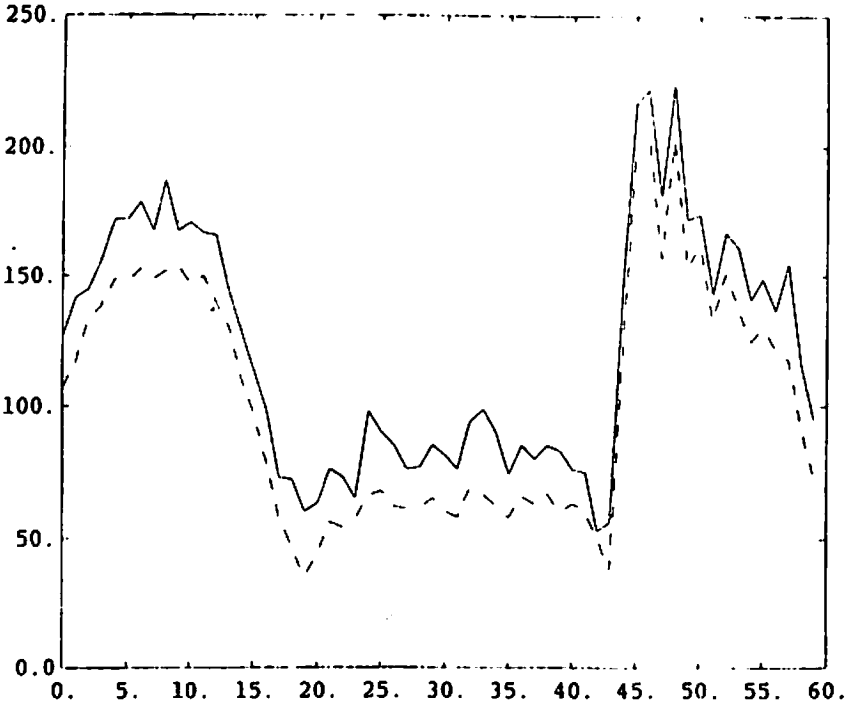


(a) Intensity values of original and noisy images.

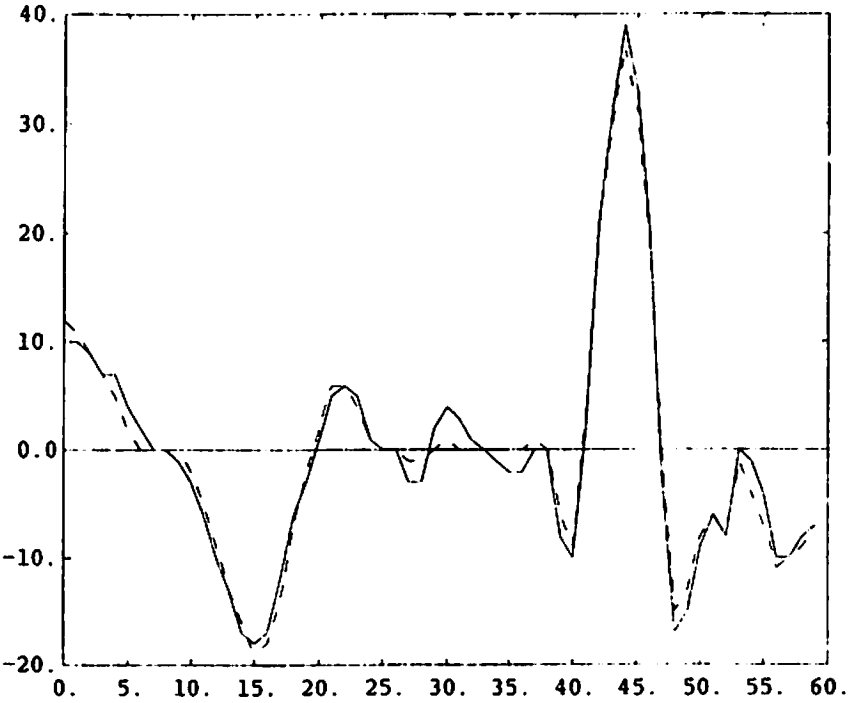


(b) First order derivatives of intensity values of original and noisy images.

Figure 1: A section of a real image with amplitude bias 20 and 30 dB noise. The original image is represented by the dashed line and the noisy image is represented by the solid line.



(a) Intensity values of original and noisy images.



(b) First order derivatives of intensity values of original and noisy images.

Figure 2: A section of a real image with amplitude bias 20 and 20 dB noise. The original image is represented by the dashed line and the noisy image is represented by the solid line.

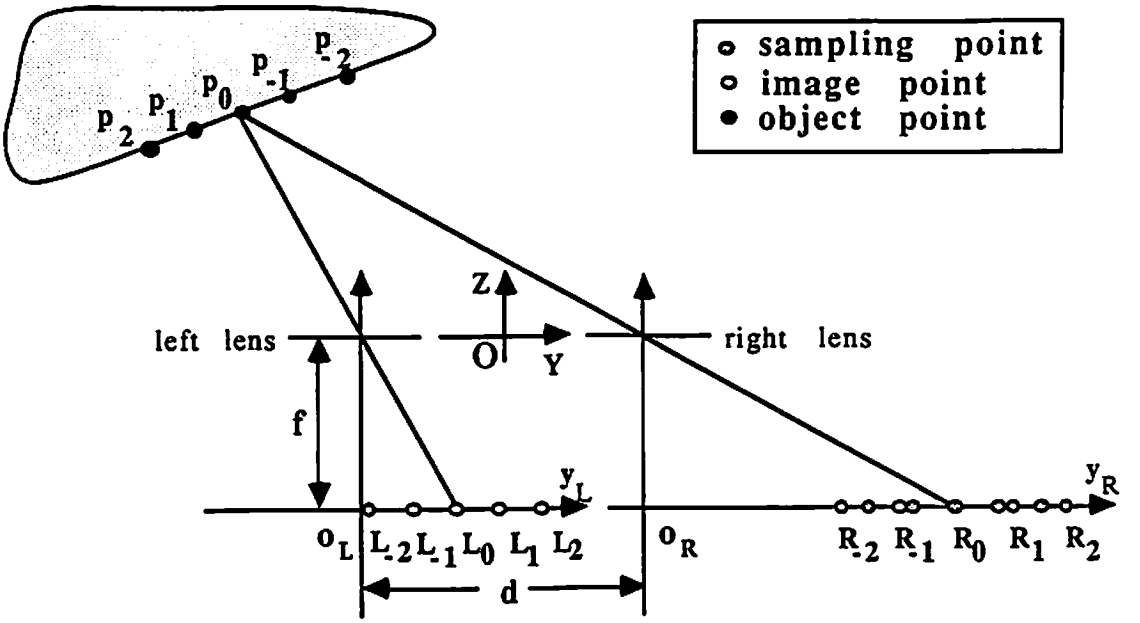
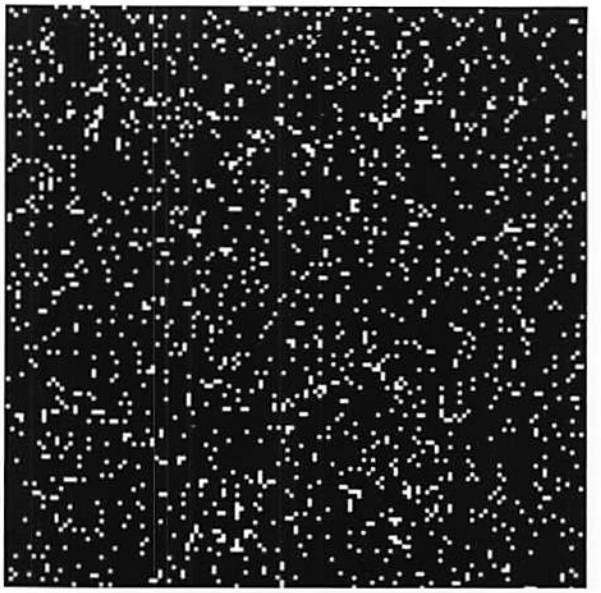
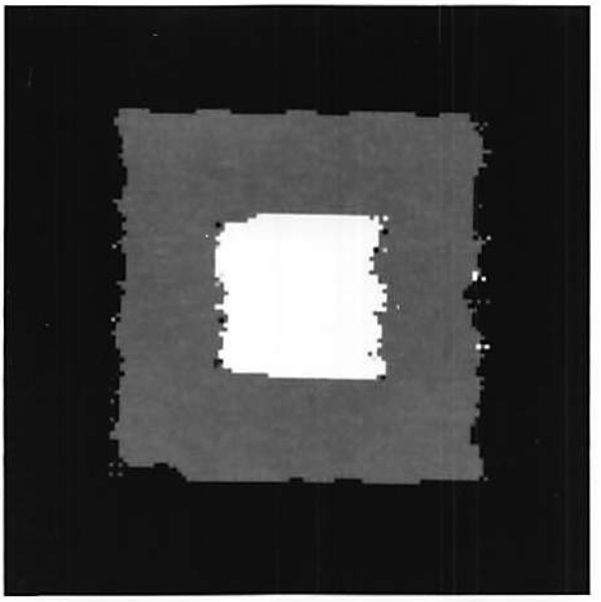


Figure 3: Camera geometry for stereo photography.

(a) Left image.

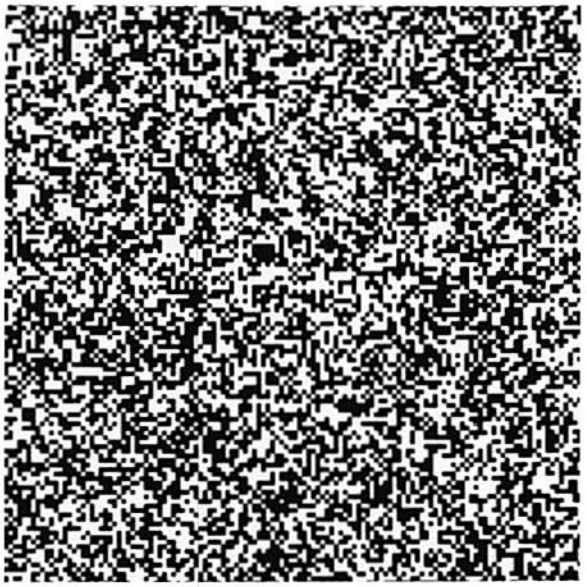


(b) Right image.

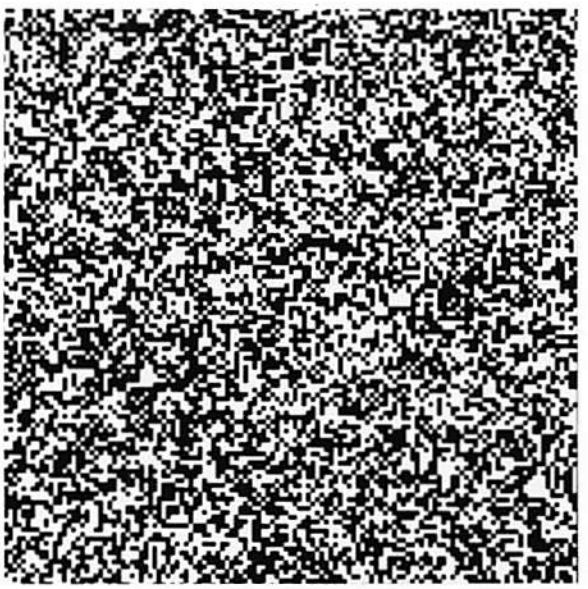


(c) Disparity map represented by an intensity image.

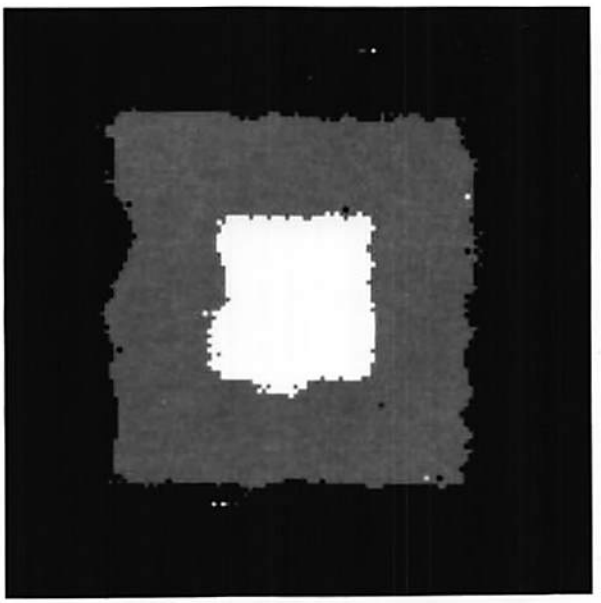
Figure 4: A 10% density random dot stereogram.



(a) Left image.

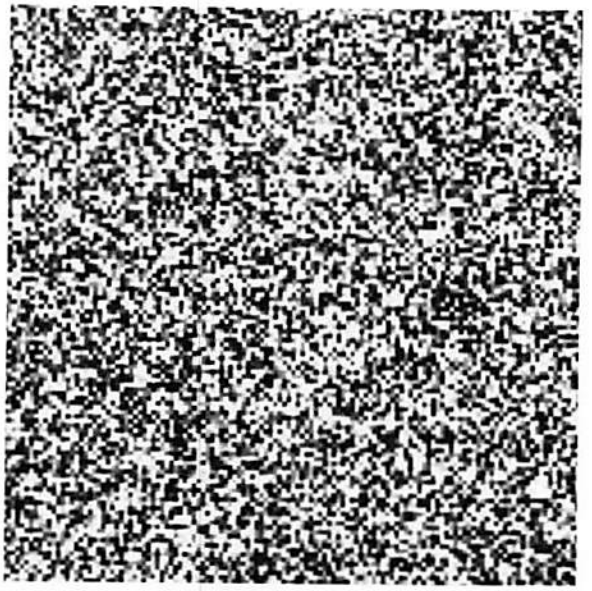


(b) Right image.

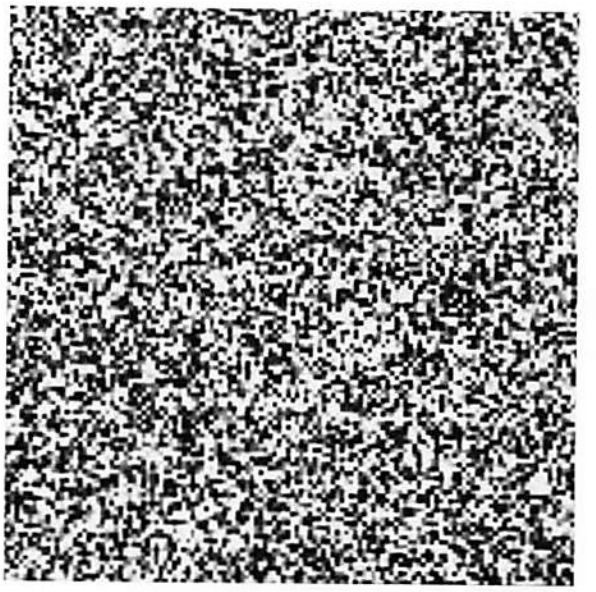


(c) Disparity map represented by an intensity image.

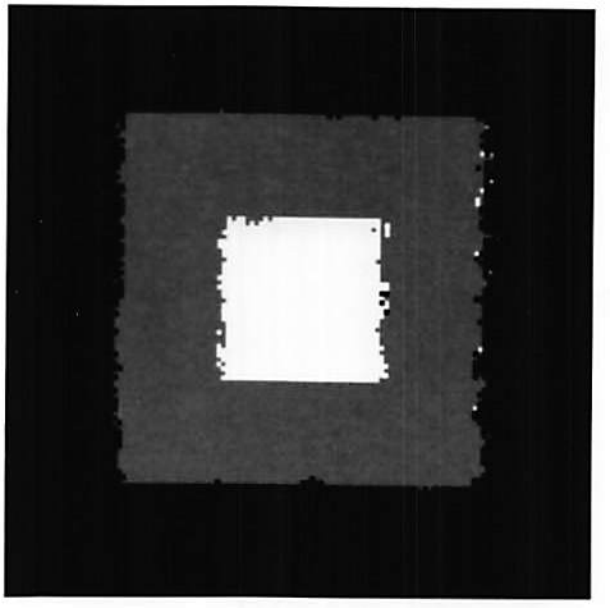
Figure 5: A 50% density random dot stereogram. In the left image, 20% of the dots were decorrelated at random.



(a) Left image.



(b) Right image.

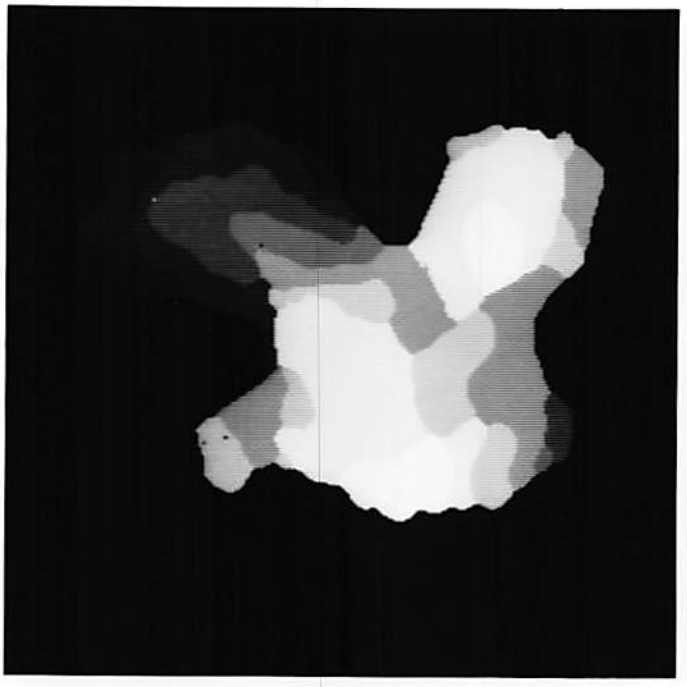


(c) Disparity map represented by an intensity image.

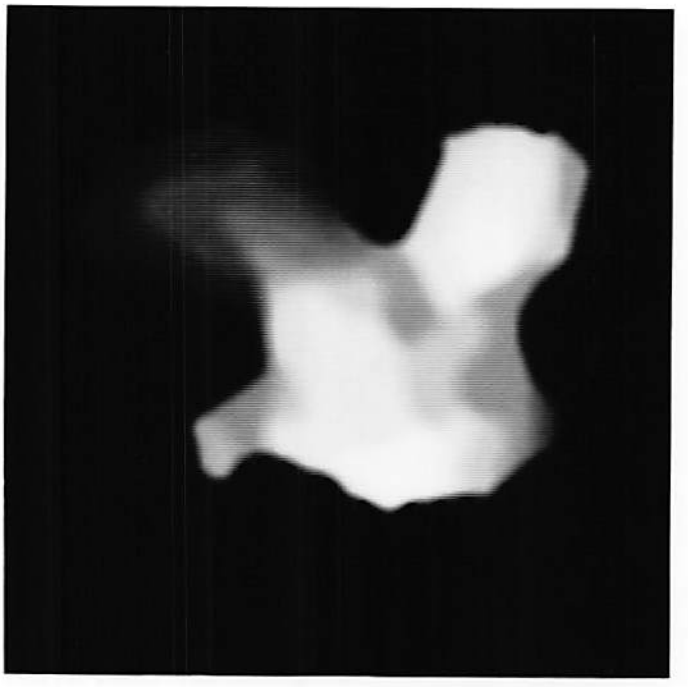
Figure 6: A multilevel random dot stereogram. In left image, Gaussian white noise corresponding to 5 dB SNR has been added.

(a) Left image.

(b) Right image.



(c) Disparity map.



(d) Smoothed disparity map.

Figure 7: The Renault part images. The disparity maps are represented by intensity images.

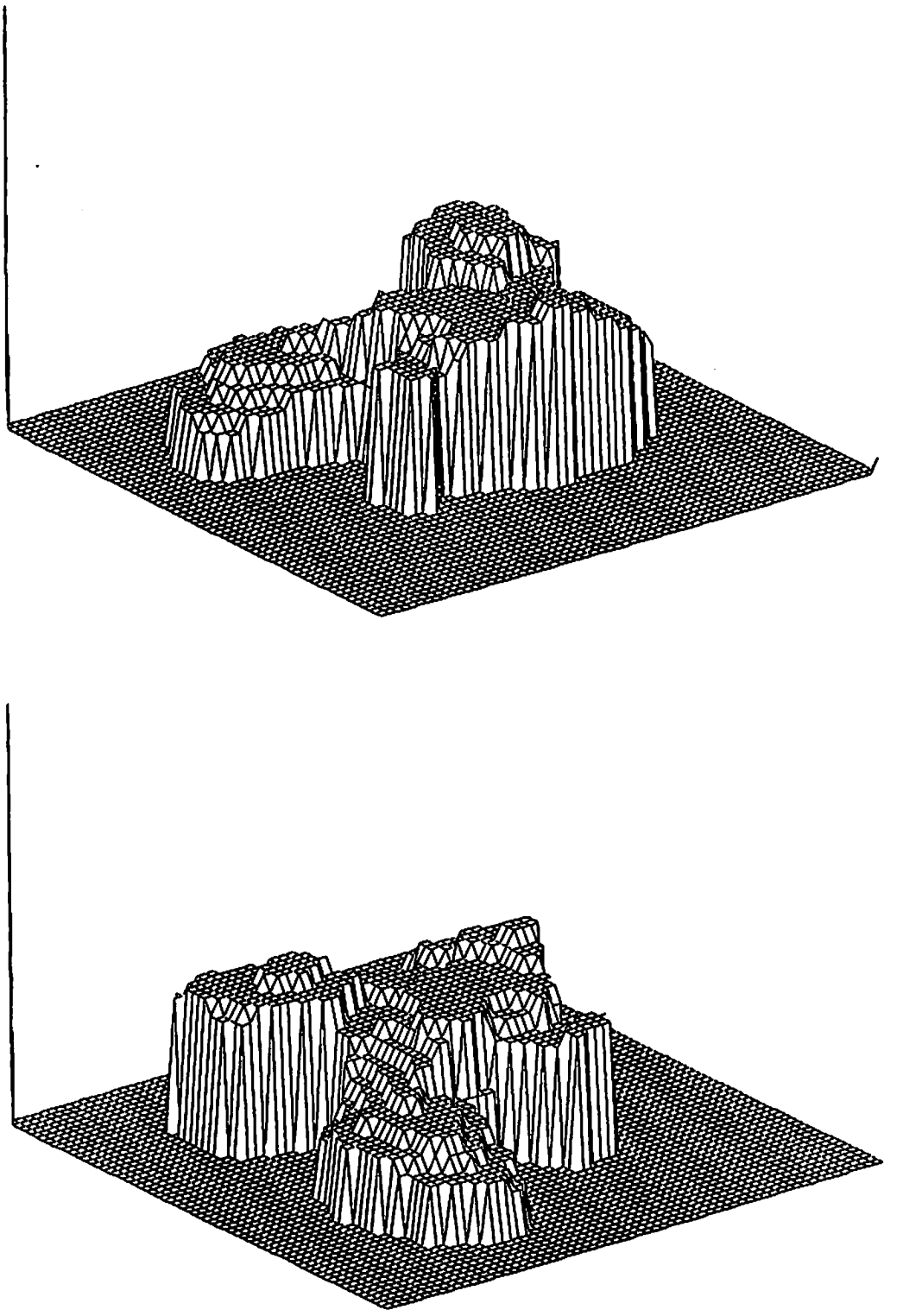


Figure 8: 3-D plots of the disparity map for the Renault part images.

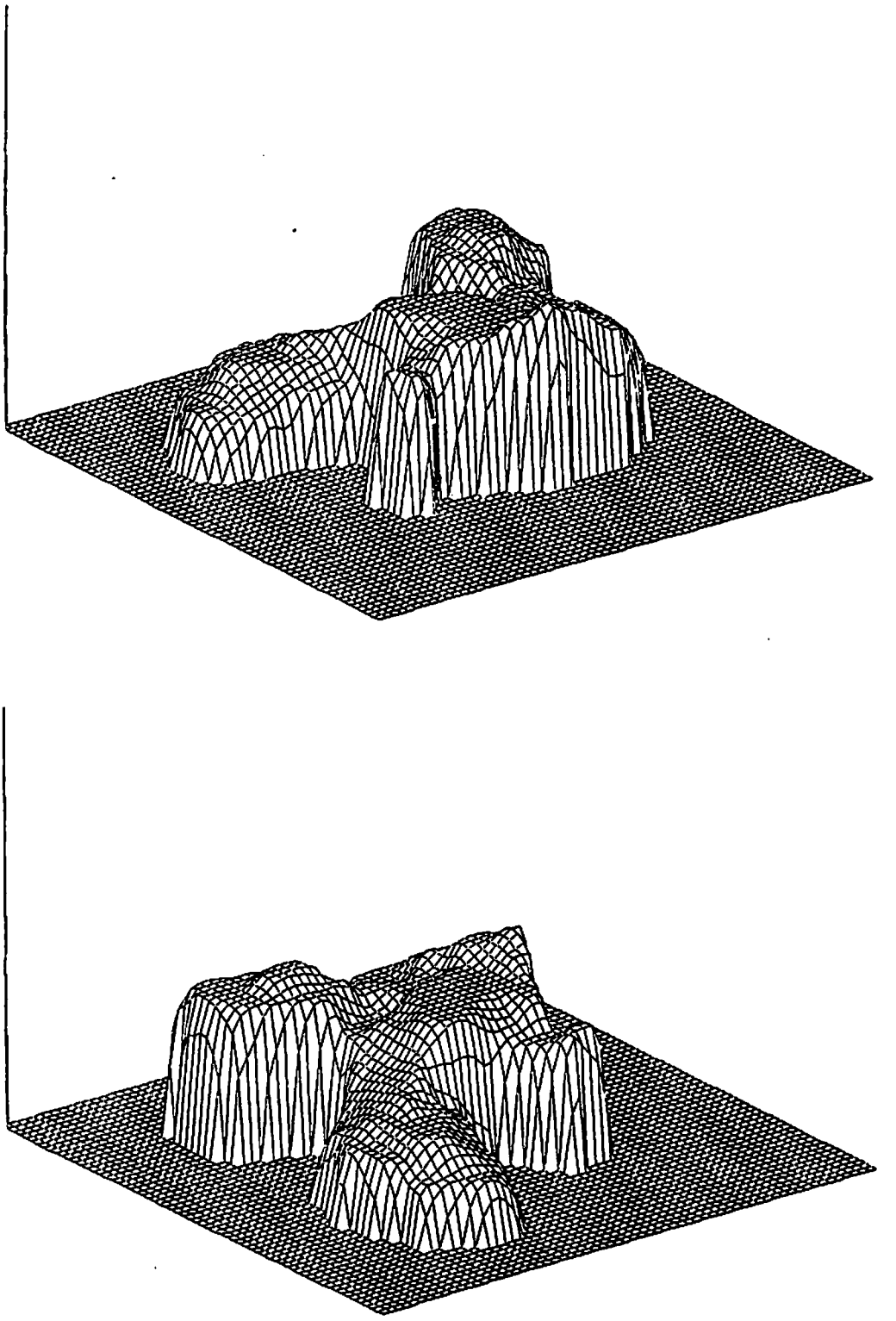
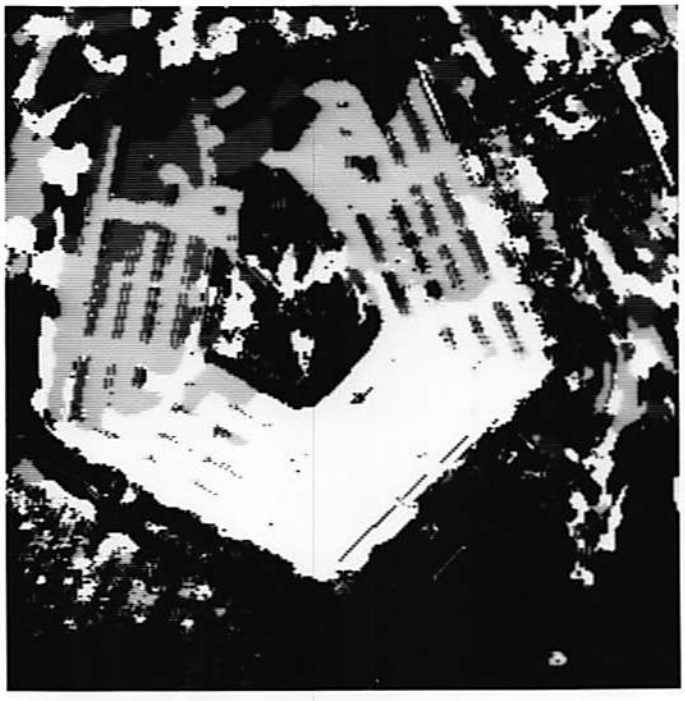


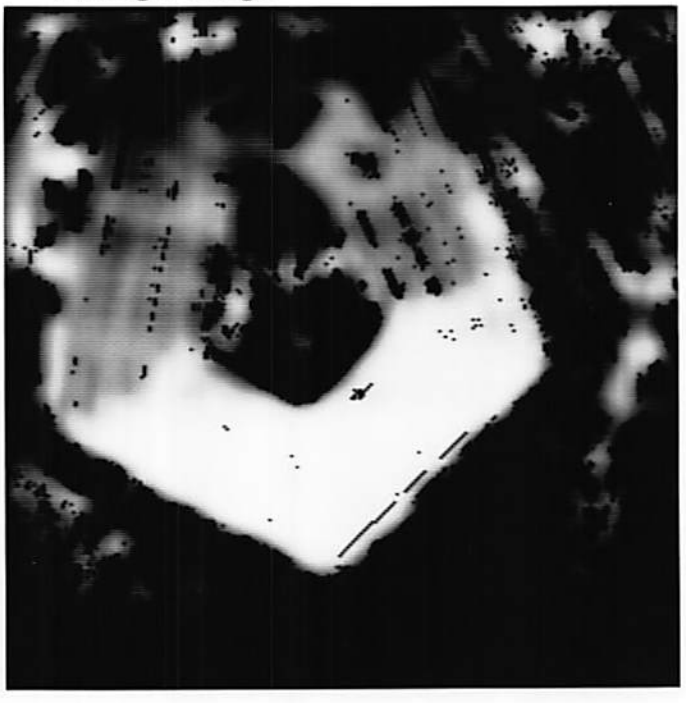
Figure 9: 3-D plots of the smoothed disparity map for the Renault part images.

(a) Left image.

(b) Right image.



(c) Disparity map.



(d) Smoothed disparity map.

Figure 10: The Pentagon part images. The disparity maps are represented by intensity images.

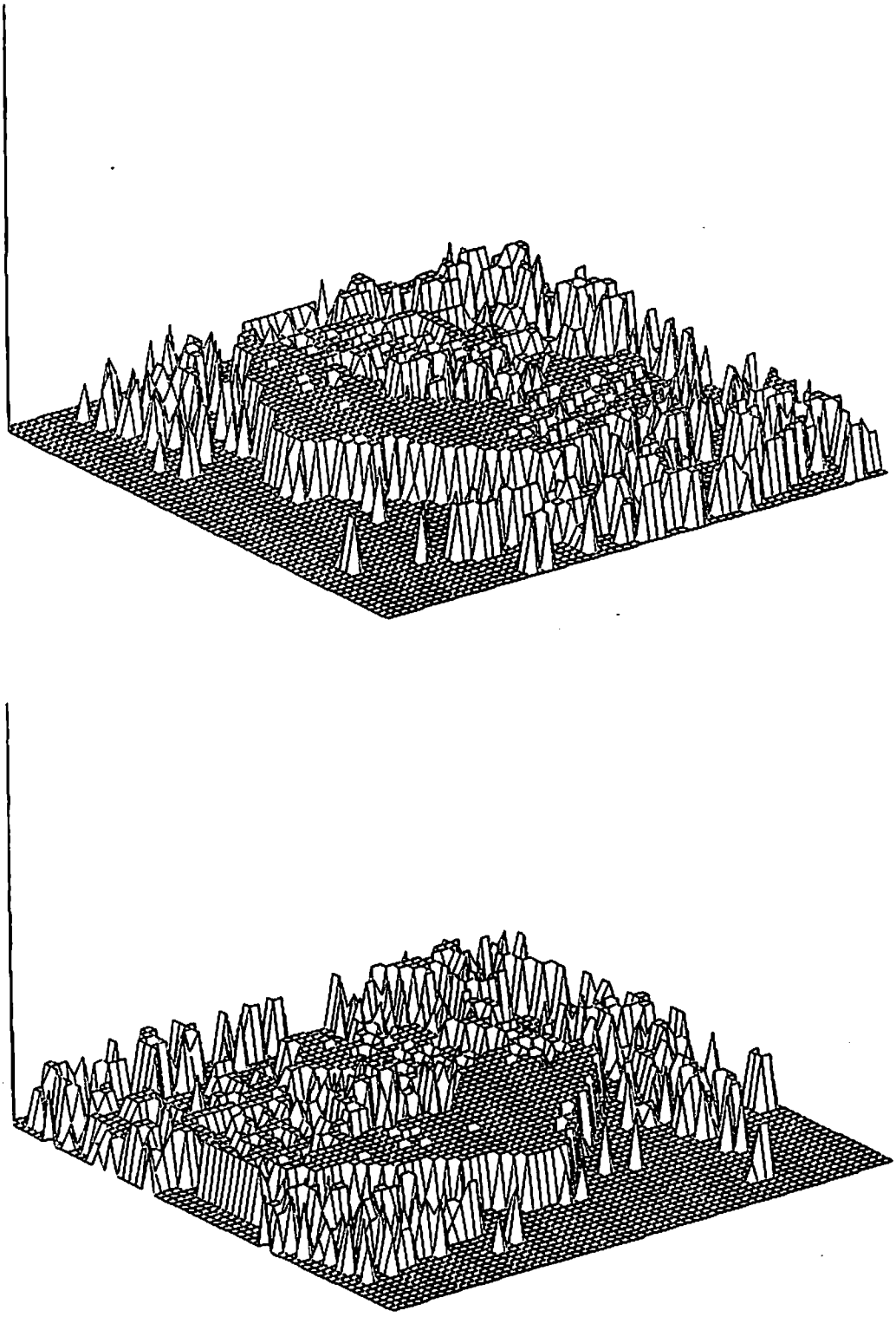


Figure 11: 3-D plots of the disparity map for the Pentagon images.

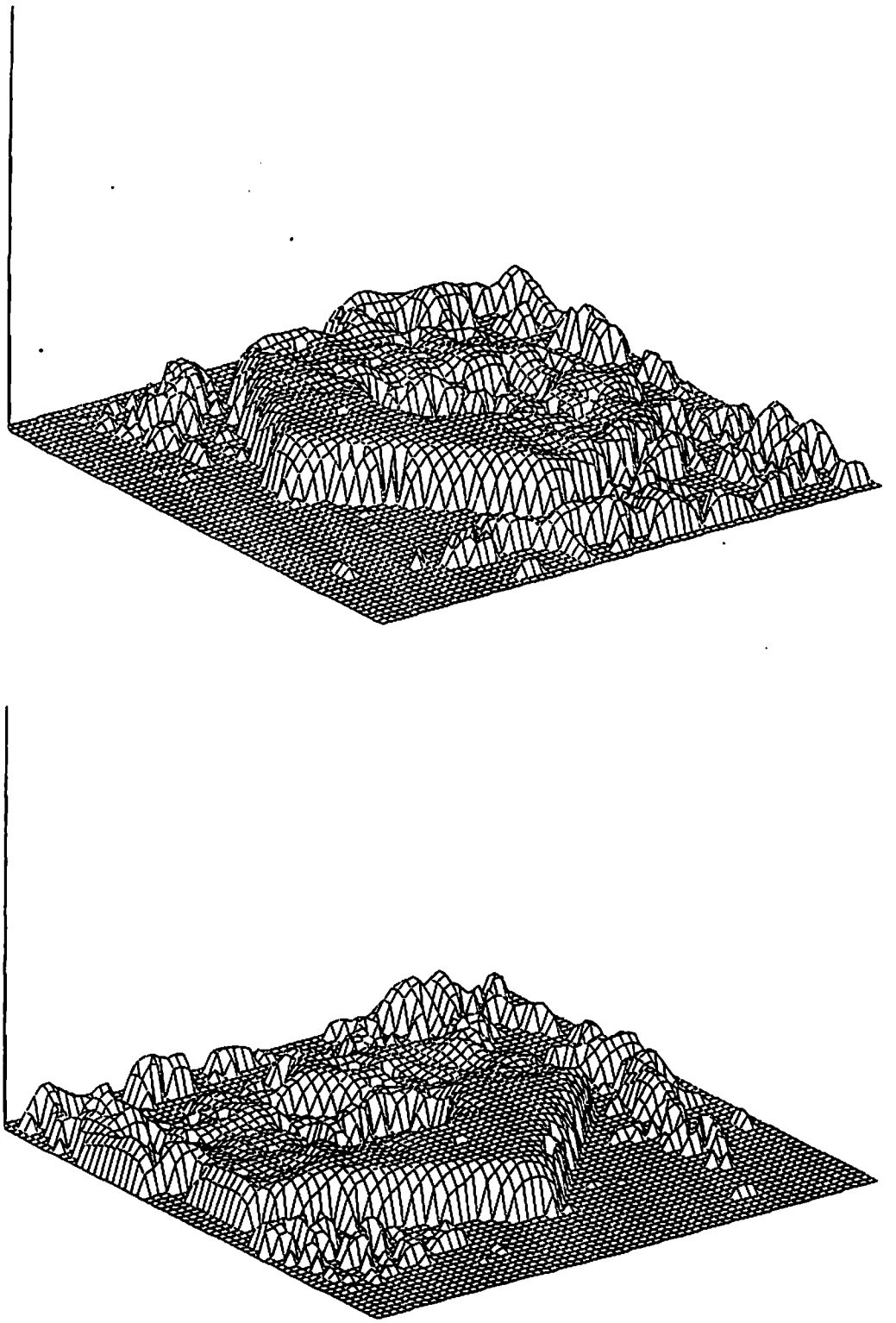


Figure 12: 3-D plots of the smoothed disparity map for the Pentagon images.