Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 25 (2004) 1743–1758

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Neural adaptive stereo matching

Elisabetta Binaghi *, Ignazio Gallo, Giuseppe Marino, Mario Raspanti

Dipartimento di Informatica e Comunicazione, University of Insubria, Via Ravasi 2, 21100 Varese, Italy

Received 5 June 2003; received in revised form 1 July 2004
Available online 15 September 2004

Abstract

The present work investigates the potential of neural adaptive learning to solve the correspondence problem within a two-frame adaptive area matching approach. A novel method is proposed based on the use of the *zero mean normalized cross-correlation coefficient* integrated within a neural network model which uses a least-mean-square delta rule for training.

Two experiments were conducted for evaluating the neural model proposed. The first aimed to produce dense disparity maps based on the analysis of standard test images. The second experiment, conducted in the biomedical field, aimed to model 3D surfaces from a varied set of scanning electron microscope stereoscopic image pairs.

Keywords: Adaptive stereo matching; Cross-correlation; Neural learning; Scanning electron microscope

1. Introduction

Stereoscopic image analysis is a well-known technique to recover the third dimension (Faugeras, 1993). The accuracy of the overall reconstruction process depends on the accuracy with which the ''correspondence problem'' is solved. It concerns the matching of points or other kinds of primitives in two (or more) images such that the matched image points are the projections of the same point in the scene. The disparity map obtained from the matching stage is then used to compute the 3D positions of the scene points given the imaging geometry.

Stereo matching has been intensively studied (Barnard and Fischler, 1982; Dhond and Aggarwal, 1989) and is still a major research topic within the computer vision community to satisfy demands in new application domains such as virtual reality and virtual studio (McMillan and Bishop, 1995).

The large number of stereo matching methods can be roughly classified into area-based and

* Corresponding author. Tel.: +39 332 218941; fax: +39 332 218909.
E-mail address: elisabetta.binaghi@uninsubria.it (E. Binaghi).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.07.001

feature-based or a combination of the two. Other types of stereo matching methods such as diffusion-based, wavelet-based and phase-based techniques have also been developed (Sun, 2002).

Feature matching techniques proceed from human vision studies (Marr and Poggio, 1976) and are based on the application of edge detection operators to extract features such as segments or contours then matched in two or more views using heuristics or constraints. All the features are described by their signatures and the correspondences are established based on the selection of the best matching signatures. These techniques have the potential of precise positioning and reliable results (Shapiro and Haralick, 1981; Barnard and Thompson, 1980; Weng and Ahuja, 1989). The main drawback of the feature-based approach is that typically, irregularly distributed features are matched producing sparse disparity maps. If a dense disparity map is required an extra surface fitting stage is needed. The literature contains several detailed reviews of various approaches and comparisons of performances (Dhond and Aggarwal, 1989; Hsieh et al., 1992).

The requirement of dense output arises from modern applications such as view synthesis and image based rendering which require high quality and resolution.

Area based approaches have the advantage of directly generating dense disparity maps. Matching elements for area-based methods are the individual pixels over which the matching cost is evaluated; pixel-to-pixel correspondence is evaluated on image intensity function and similarity statistics.

According to the taxonomy proposed by Scharstein and Szelinsky, the dense stereo matching process can be divided into three tasks: *matching cost computation*, *aggregation of local evidence* and *computation of disparity values* (Scharstein and Szeliski, 2002). Many dense stereo matching methods have presented several different solutions to one or more of these tasks. The most common *matching costs* include *squared intensity differences* (SD) and *absolute intensity difference* (AD) (Cox et al., 1996; Scharstein and Szeliski, 2002).

The actual sequence of steps in the overall matching procedure depends on the matching algorithm and in particular, on its local or global nature. Local, window-based algorithms, implicitly assuming smoothness, perform the *aggregation* task by summing or averaging *matching cost* over a support region. Some local algorithms, those based on *normalised cross-correlation* (Chen and Medioni, 1999) and *rank methods* (Zabih and Woodfill, 1994) combine the first and second steps directly computing *cost* on a support region. Global algorithms make an explicit smoothness assumption and directly solve an optimisation problem usually formulated as an energy minimisation, skipping the *aggregation* step and directly computing *disparity values* (Scharstein and Szeliski, 2002).

Local methods usually compute final *disparity* adopting a local Winner Take All strategy which selects the pair with the best *matching cost* under assumption of uniqueness.

Generally speaking, area-based methods work well especially when surfaces vary in smoothness and images have an adequate visual texture. Serious difficulties may be encountered in regions with low texture, periodic structures and depth discontinuities. Occlusion is another crucial issue in generating high-quality stereo maps. Many approaches ignore the effects of occlusion, whereas others attempt to mitigate its effects in various ways (Scharstein and Szeliski, 1996).

One of the principal factors influencing the success of local area-based methods is the proper selection of window shape and size. The windows must be large enough to capture intensity variation for reliable matching but small enough to avoid the effects of projective distortions at the same time. Appropriate window selection should improve matching accuracy but requires an optimised balance between the above opposite criteria.

Various approaches have tackled this problem such as shiftable windows (Bobik and Intille, 1999), windows with adaptive size (Okutomi and Kanade, 1992), windows using image segmentation (Zhang and Kambhamettu, 2002), windows based on connected components (Boykov et al., 1997) and windows based on disparity space (Szeliski and Scharstein, 2002). The adaptive window addresses the problem of finding the appropriate window size typically enlarging/reducing the win-

dow size according to the examined area. The proposed techniques share the use of explicit constraints and criteria influencing and varying the dimension of the aggregation window contextually within the image.

The present work investigates the potential of neural adaptive learning (Pao, 1989; Rumelhart et al., 1986) to solve the correspondence problem within a two-frame adaptive area matching approach. A novel method is proposed based on the use of the *zero mean normalized cross-correlation coefficient* (ZNCC) (Gonzalez and Woods, 2002) integrated within a neural network model which uses the least-mean-square delta rule for training (Rumelhart et al., 1986). In this context, the neural learning task can be formulated as the search for the proper window shape and size for each support region.

Two experiments were conducted for evaluating the neural model proposed. The first aimed to produce dense disparity maps based on the analysis of test images from Scharstein and Szelinski's Test Data Set (Scharstein and Szeliski, 2002). For this purpose, our matching algorithm was structured according to the taxonomy mentioned above and integrated in the stand-alone C++ implementation framework made available on the Web at www.middlebury.edu/stereo. We followed the evaluation methodology proposed by Scharstein and Szelinski using test images which include data with a ground truth disparity map and evaluating performances based on suggested quality metrics.

The second experiment, conducted in the biomedical field, aimed to produce disparity maps from a varied set of scanning electron microscope (SEM) stereoscopic image pairs. Application requirements suggested reformulating the matching algorithm by inserting a pre-processing stage based on the use of Laplacian Pyramid filtering technique for the extraction of points of interest in the reference image from which to attempt the matching process and produce a disparity map.

2. Neural adaptive image matching

Table 1 lists the solutions adopted by our adaptive neural model based on ZNCC (AZNCC) in accordance with the taxonomy proposed by Scharstein and Szeliski (2002).

Table 1
Solutions adopted by AZNCC

Matching cost	Zero mean normalised cross-correlation coefficient
Aggregation	Rectangular adaptive window
Optimisation	WTA

Several matching costs have been proposed varying in performance and computational costs (Sun, 2002). ZNCC was used in our model showing useful properties such as (Gonzalez and Woods, 2002):

- optimal signal-to-noise ratio estimation,
- insensitivity to image intensity variations, due to normalisation with respect to mean and standard deviation.

The overall proposed algorithm is local, actually combining matching cost and aggregation steps which act on the support region.

Setting the appropriate window size is a critical task; problems may arise both with small windows that do not cover enough intensity variation in textureless areas and with large windows near discontinuities and occluded regions (Lotti and Giraudon, 1994).

To overcome this problem adaptive techniques were investigated. Okutomi and Kanade (1992) proposed a refinement of an initial disparity map by adapting the local window to local variation of intensity and disparity. Lotti and Giraudon (1994) proposed a direct method based on the use of correlation windows with dimensions constrained by an edge map extracted from the image.

Adaptation criteria in both cases are statically formulated in the light of some explicit assumptions. In this work we address the matching problem proposing an adaptive supervised correlation model based on correlation coefficient measure, able to learn the appropriate window shape and size for each search in the target image from a supervised set of examples, automatically extracted from the reference image.

2.1. Neural representation of correlation coefficient for stereo matching

Let $w(x,y)$ be a region (window) of the reference image with dimensions JXK and f the matching image. The correlation coefficient can be written as follows:

$$c(x_0,y_0)=\frac{\sum_{s=-a}^{a}\sum_{t=-b}^{b}[f(x_0+s,y_0+t)-\bar{f}]\cdot[w(s,t)-\bar{w}]}{\sqrt{\sum_{s=-a}^{a}\sum_{t=-b}^{b}[f(x_0+s,y_0+t)-\bar{f}]^2\sum_{s=-a}^{a}\sum_{t=-b}^{b}[w(s,t)-\bar{w}]^2}} \tag{1}$$

with $J = 2a + 1$ and $K = 2b + 1$.

It expresses the matching of window w within the image f. Summations are taken where f and w overlap. For one value (x_0, y_0) inside f, the application of Eq. (1) yields one value of c. The window w is centred in x_0, y_0; as x_0, y_0 are varied, w moves around the image area consistently. $\bar{f}$ is the average value of f in the region coincident with the current location of w and $\bar{w}$ is the average value of the pixels in w.

The function c returns values ranging from -1 for situations in which window w and sub-image f are not similar at all, to 1 for situations in which they are identical. Eq. (1) can be rewritten in the following form to fulfil the analytical requirements imposed by our model:

$$\begin{aligned}c(x_0,y_0)=&\frac{[f(x_0-a,y_0-b)-\bar{f}]}{\sqrt{\sum_{s=-a}^{a}\sum_{t=-b}^{b}[f(x_0+s,y_0+t)-\bar{f}]^2}}\\&\times\frac{[w(-a,-b)-\bar{w}]}{\sqrt{\sum_{s=-a}^{a}\sum_{s=-b}^{b}[w(s,t)-\bar{w}]^2}}+\cdots+\cdots\\&+\frac{[f(x_0+a,y_0+b)-\bar{f}]}{\sqrt{\sum_{s=-a}^{a}\sum_{t=-b}^{b}[f(x_0+s,y_0+t)-\bar{f}]^2}}\\&\times\frac{[w(a,b)-\bar{w}]}{\sqrt{\sum_{s=-a}^{a}\sum_{s=-b}^{b}[w(s,t)-\bar{w}]^2}}\\=&\sum_{s=-a}^{a}\sum_{t=-b}^{b}F(x_0+s,y_0+t)\cdot W(s,t)\end{aligned} \tag{2}$$

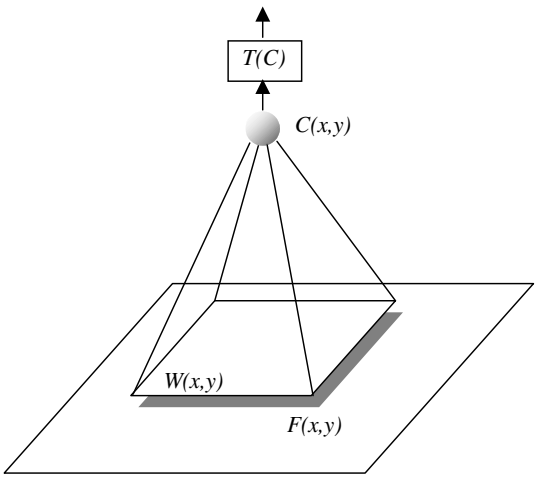


Fig. 1. Neural representation of correlation measure.

Proceeding from this formulation we represent correlation within a neural structure as illustrated in Fig. 1.

The neuron activation function $C(x,y)$ is given by the following formula:

$$C(x_0,y_0)=\sum_{s=-a}^{a}\sum_{t=-b}^{b}F(x_0+s,y_0+t)\cdot W(s,t)\cdot G(s,t) \tag{3}$$

where $G(s,t) = G(s)\cdot G(t)$ is the product of two functions each obtained as the difference between two sigmoid functions (Fig. 2).

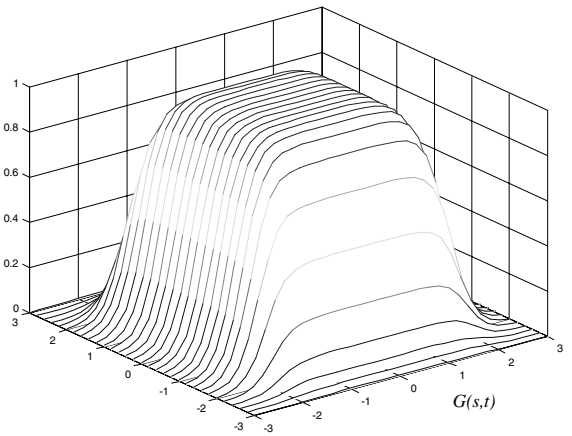


Fig. 2. Function $G(s,t)$, component of the activation function obtained as difference of two sigmoid functions.

In the formula, exemplifying for one dimension we have:

$$G(s) = L(s,a,-c_x) - L(s,a,c_x);$$
$$L(s,a,c_x) = \frac{1}{1+\mathrm{e}^{-a(s-c_x)}} \tag{4}$$

with a parameter which controls the shape of the sigmoid function, c_x is offset by the function L. The neural model makes use of a *Gaussian transfer function* $T(x)$ having the following formula

$$T(x) = \mathrm{e}^{\frac{-(x-d)^2}{2\sigma^2}} \tag{5}$$

with d and σ the centre and standard deviation, respectively; d is assumed constant and equal to 1. The function $T(x)$ maps values of the neuron activation function in the 0–1 range. This function was chosen to control the output of the neural structure and to accelerate the convergence maintaining significance with similarity correlation measures; for $\sigma \to 0$ the behaviour of the neuron output increasingly becomes crisp in [0,1] range (see Fig. 3).

2.2. Neural learning

The first goal of neural learning can be formulated as the search for the most adequate correlation window shape and size; it is achieved by applying the delta rule training algorithm (Pao, 1989; Rumelhart et al., 1986; Bishop, 1995) to parameters c_x and c_y in the function $G(s,t)$ as exemplified below:

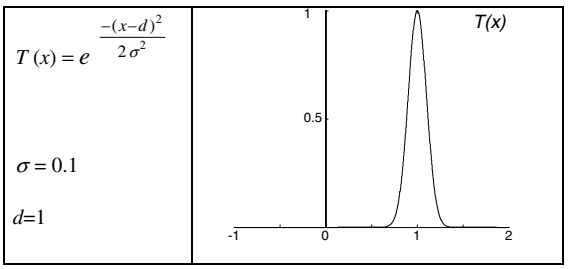


Fig. 3. Neuron transfer function $T(x)$; d and σ centre and standard deviation, respectively; the function is exemplified for $d = 1$ and $\sigma = 0.1$.

$$c_x^{\text{new}} = c_x^{\text{old}} - \eta\frac{\partial E}{\partial c_x} \quad \text{with} \quad \frac{\partial E}{\partial c_x} = \frac{\partial E}{\partial O}\frac{\partial O}{\partial C}\frac{\partial C}{\partial c_x}$$
$$\frac{\partial E}{\partial O} = (O-D)$$
$$\frac{\partial O}{\partial C} = \frac{\partial T(C)}{\partial C} = T(C)\frac{d-C}{\sigma^2}$$
$$\frac{\partial C}{\partial c_x} = \sum_s\sum_t\left[F(x_0+s,y_0+t)\cdot W(s,t)\cdot\frac{\partial G(s,t)}{\partial c_x}\right]$$
$$\frac{\partial G(s,t)}{\partial c_x} = \frac{\partial G(s)}{\partial c_x}G(t)$$
$$\frac{\partial G(s)}{\partial c_x} = a\cdot[L(s,a,-c_x)\cdot(1-L(s,a,-c_x)) + L(s,a,c_x)\cdot(1-L(s,a,c_x))] \tag{6}$$

$$c_y^{\text{new}} = c_y^{\text{old}} - \eta\frac{\partial E}{\partial c_y} \quad \text{with} \quad \frac{\partial E}{\partial c_y} = \frac{\partial E}{\partial O}\frac{\partial O}{\partial C}\frac{\partial C}{\partial c_y}$$
$$\frac{\partial C}{\partial c_y} = \sum_s\sum_t\left[F(x_0+s,y_0+t)\cdot W(s,t)\cdot\frac{\partial G(s,t)}{\partial c_y}\right]$$
$$\frac{\partial G(s,t)}{\partial c_y} = G(s)\frac{\partial G(t)}{\partial c_y}$$
$$\frac{\partial G(t)}{\partial c_y} = a\cdot[L(t,a,-c_y)\cdot(1-L(s,a,-c_y)) + L(s,a,c_y)\cdot(1-L(s,a,c_y))] \tag{7}$$

where D and O indicate the desired and obtained outputs, respectively; c_x and c_y offset by the function L in x and y dimensions; $E = 1/2(O-D)^2$ criterion function representing the total squared error between obtained and desired outputs.

The second learning goal consists in the search for appropriate output function achieved by applying the learning rule to σ parameter in the transfer function $T(C)$, as exemplified below.

$$\sigma^{\text{new}} = \sigma^{\text{old}} - \eta\frac{\partial E}{\partial \sigma} \quad \text{with} \quad \frac{\partial E}{\partial d} = \frac{\partial E}{\partial O}\frac{\partial O}{\partial d}$$
$$\frac{\partial E}{\partial O} = (O-D) \tag{8}$$
$$\frac{\partial O}{\partial \sigma} = \frac{\partial T(C)}{\partial \sigma} = T(C)\frac{(C-d)^2}{\sigma^3}$$

2.3. Supervised training

The neural correlation model must be trained with a supervised learning procedure before computing the matching cost for each pixel. In general

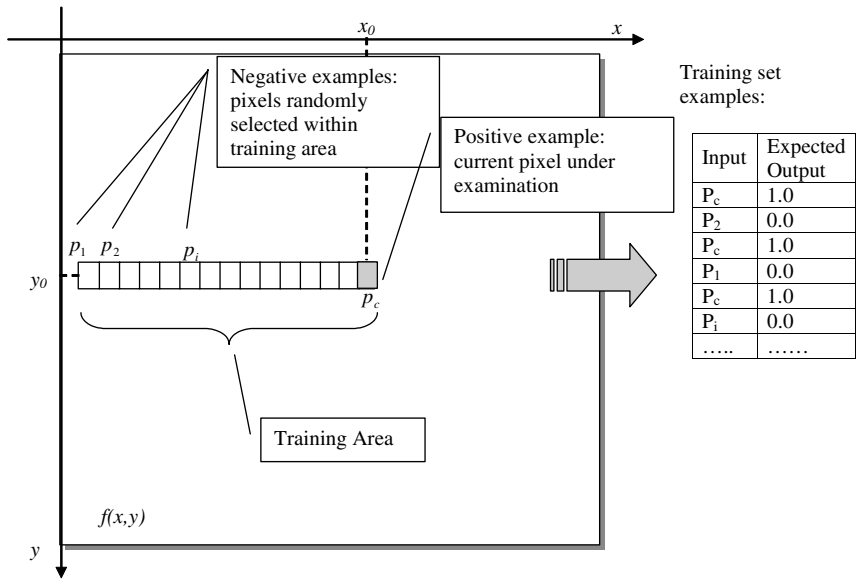


Fig. 4. Exemplification of training set construction.

neural supervised learning is based on the presentation to the network of a set of training examples having the structure [<features>; <expected value>].

In our context, the *feature part* of a given training example is the set of $F(x,y)$ and $W(s,t)$ values (see formula (3) and Fig. 1 in Section 2.1); the *expected value* is the degree of correlation. Fig. 4 exemplifies the training set construction. Initially, the template values are extracted positioning $W(x,y)$ on the reference image centered on the pixel concerned p_c. When F also is centered on p_c the training procedure associates an expected correlation value equal to 1.0 (positive example). When F is centered on pixels p_i with $i \neq c$, randomly selected within a given *training area* in the reference image, the training procedure associates a correlation value equal to 0 (negative example). The training area is defined as the region, within the reference image, positioned and dimensioned as the search area in the matching image. To give equal occurrence of positive and negative examples in training, the overall training set is formed including for each negative example the positive one.

The complete training set is employed by the traditional delta rule algorithm to train the neural correlation model. The underlying assumption of this training strategy is that conditions for correspondence, learned on the reference image, can be successfully applied by the trained network when acting on the matching image.

We base the termination of the learning process on the following conditions (combined in AND):

- reaching the maximum number of epochs,
- reaching the minimum σ value,
- reaching the minimum relative learning error; the relative learning error is computed according to the following formula

  ((mse_old − mse_current)/mse_old)

  where mse_old is the mean squared error computed by the network in a previous learning cycle and mse_current the mean squared error computed by the network in the current cycle.

Parameters were experimentally assessed; Section 3.2 illustrates experiments aimed to find the optimal setting of maximum number of epochs parameter.

Suggested values for the other parameters are:

- minimum σ value = 0.1,
- minimum relative learning error = 0.001.

When trained, the correlation model is applied to pixels of the matching image situated within the search area to compute disparity values.

3. Experimental evaluation based on Scharstein and Szeliski test dataset

The experiments illustrated in this section addressed the following questions:

- how did the adaptive neural model compare with other correspondence matching approaches? and
- how did the performance of the adaptive neural model depend upon their main parameters?

The overall experimental activity was supported by tools and test data available within the implementation framework proposed by Scharstein and Szeliski (2002) in their paper and made available on the Web at www.middlebury.edu/stereo.

We included our stereo correspondence algorithm in this framework, and applied it to the test data available. Performances of the algorithm were evaluated based on the available evaluation module which allows comparison with other implemented algorithms.

We selected as quality evaluation measure the percentage of bad matching pixels expressed by the following formula:

$$B = \frac{1}{N} \sum_{(x,y)} (|d_C(x,y) - d_T(x,y)| > \delta_d) \tag{9}$$

where $d_C(x,y)$ is the computed disparity map, $d_T(x,y)$ is the ground truth disparity map and δ_d is a disparity error tolerance set to the suggested value equal to 1.

This measure is intended computed over the whole image and on three different kinds of regions in the whole image:

- textureless regions: regions where the squared horizontal intensity gradient averaged over a square window of a given size (suggested value 3) is below a given threshold (suggested value 4.0);
- occluded regions: regions that are occluded in the matching image;
- depth discontinuity regions: pixels whose neighboring disparity differs by more than a given threshold (suggested value 2.0), dilated by a window of a given width (suggested value 9).

These regions are computed by pre-processing reference images and ground truth disparity maps yielding binary segmentation.

We have selected two data sets.

- the monochromatic "Map" constituted by a pair of images and corresponding disparity map,
- the "Sawtooth" constituted by a 9 frame stereo sequence and corresponding disparity map.

In all our experiments, we used a stereo pair of images as input even when more images were available in the data set considered.

Readers interested in additional explanations concerning the data sets, their segmentation and the metrics used are referred to (Scharstein and Szeliski, 2002).

Fig. 5 show the two reference images together with the ground-truth disparities of the selected data.

3.1. Matching cost

In this section we describe the experiment conducted to evaluate the performances of the individual matching cost component of our algorithm without considering adaptive learning solutions. The evaluation was based on the monochromatic Map image pair.

To this purpose the matching cost ZNCC was implemented and integrated within the above mentioned implementation framework. To compare performances of ZNCC with those of other standard matching costs, we implemented the correlation-based matching cost NCC, and used the available AD and SD standard matching cost. All the matching costs were implemented within a local algorithm with fixed aggregation window and WTA optimization strategy.

Fig. 6 shows the results obtained aggregating the matching costs with 5×5, 9×9, 13×13

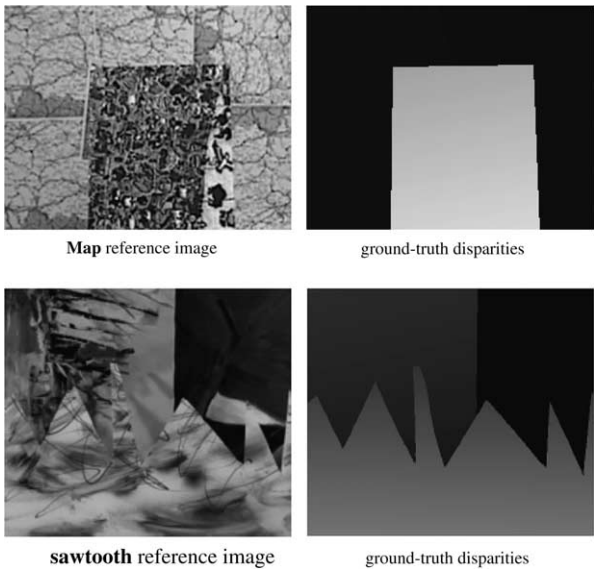


Fig. 5. Stereo images with ground truth used in this study: the monochromatic "Map" and the "Sawtooth" color image. The figure shows the reference images and the ground truth disparities.

windows (the aggregation with AD and SD is named SAD and SSD, respectively). All the matching costs considered show comparable performances in the whole image and there is little difference among them in specific regions. In more detail, the performances of ZNCC are strongly influenced by the window dimension in textureless areas. A large window dimension can help for ZNCC and SAD: when 9×9 and 13×13 windows are used, they prevail on the other matching costs. In occluded regions, the ZNCC shows the best behavior independently from the window dimension. SAD and SSD prevail on both ZNCC and NCC correlation measures near discontinuities.

3.2. Sensitivity analysis

In this experiment we attempted to demonstrate how performances of the global adaptive algorithm AZNCC depend upon two main parameters: maximum number of epochs which controls the number of iterative presentation of examples during learning and maximum window size which limits the automatic enlarging of the correlation window.

Fig. 7 shows plots of the performances obtained in occluded, non-occluded, textureless and discontinuity regions of the data set Map and Sawtooth. Plots are as function of maximum number of epochs. Overall there are few differences varying the parameter value. In discontinuity regions performances slightly increase when the maximum number of epochs decrease. Inversely, in textureless regions of Sawtooth image, performances decrease when the maximum number of epochs decrease. In the other regions performances are nearly constant.

It is important to note that a relatively small number of epochs can be used to train the network in this application; experiments were conducted demonstrating that the increase in the number of epochs over the value 80, implies a decrease of performances due to overfitting conditions.

Fig. 8 shows plots of the performances of AZNCC as a function of maximum window size parameter for Map and Sawtooth images. Only in regions with discontinuities of both the images, does the AZNCC show performances influenced by the maximum window size parameter which varies from 31 to 9. In the other regions performances show few differences.

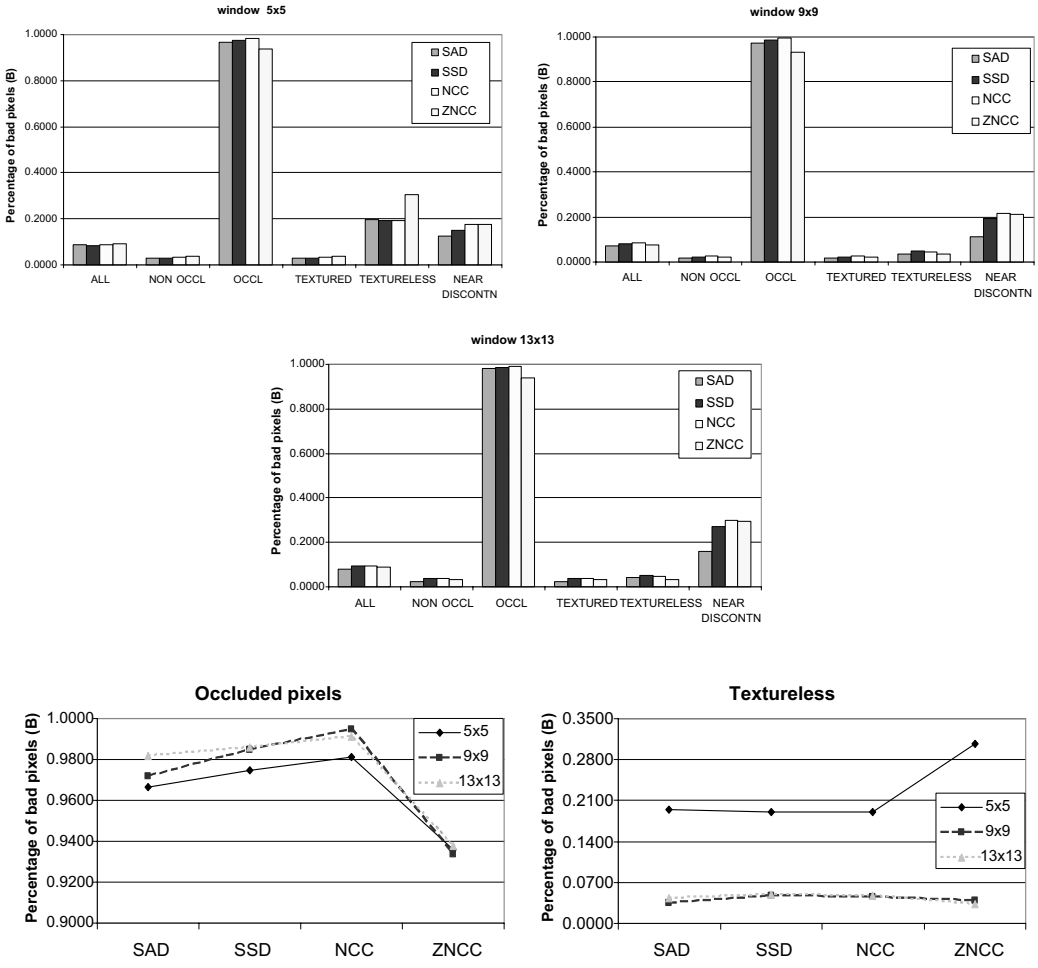
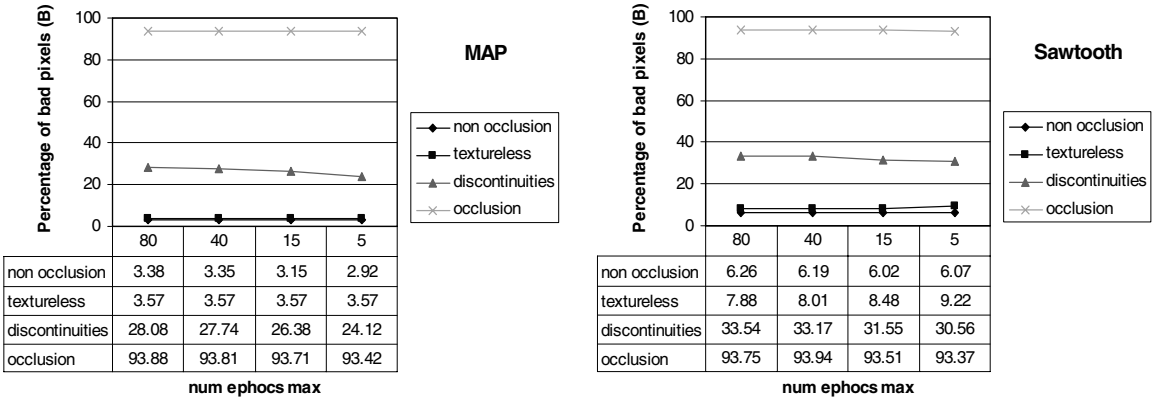


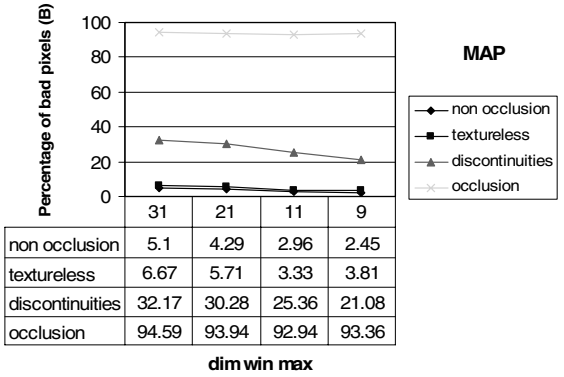
Fig. 6. Performances of different matching costs aggregated with 5×5, 9×9 and 13×13 window and applied on MAP image.



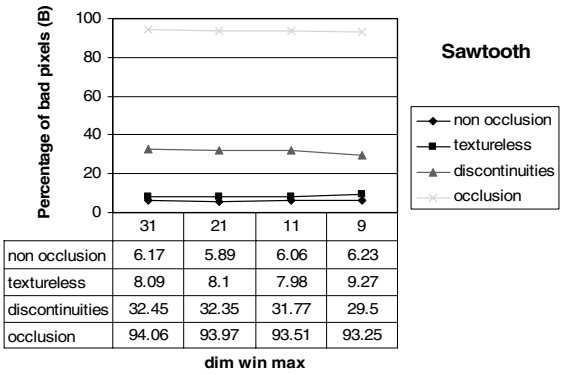
MAP	80	40	15	5
non occlusion	3.38	3.35	3.15	2.92
textureless	3.57	3.57	3.57	3.57
discontinuities	28.08	27.74	26.38	24.12
occlusion	93.88	93.81	93.71	93.42

Sawtooth	80	40	15	5
non occlusion	6.26	6.19	6.02	6.07
textureless	7.88	8.01	8.48	9.22
discontinuities	33.54	33.17	31.55	30.56
occlusion	93.75	93.94	93.51	93.37

Fig. 7. Plots of the performances of AZNCC as a function of maximum number of epochs.



	31	21	11	9
non occlusion	5.1	4.29	2.96	2.45
textureless	6.67	5.71	3.33	3.81
discontinuities	32.17	30.28	25.36	21.08
occlusion	94.59	93.94	92.94	93.36



	31	21	11	9
non occlusion	6.17	5.89	6.06	6.23
textureless	8.09	8.1	7.98	9.27
discontinuities	32.45	32.35	31.77	29.5
occlusion	94.06	93.97	93.51	93.25

Fig. 8. Plots of performances as a function of the maximum window size parameter.

3.3. Adaptive correlation versus fixed window correlation

This experiment was aimed to identify and evaluate the contribute of neural adaptive learning in the global matching algorithm. To this purpose we compared AZNCC with a local fixed window algorithm which uses the same matching cost, ZNCC and the same optimisation strategy based on WTA. The evaluation was based on a monochromatic MAP pair of images.

AZNCC was used selecting the best parameter values, determined by the experiments described in the section above:

maximum number of epochs = 10,
maximum window size = 11.

The local fixed window algorithm aggregated with 5×5, 10×10 and 13×13 window.

Fig. 9 shows the disparity map and the performances obtained. In all the regions considered, with the exception of regions with discontinuity, the adaptive algorithm shows performances comparable with those obtained by the best fixed window algorithm. In regions with discontinuities, the fixed window algorithm with window dimension 5×5 slightly prevails.

To evaluate the computational burden of the adaptive algorithm, we evaluated the CPU time on a Windows platform with a 300 MHz processor, and 384 Mb RAM. The training phase for all the pixels in the image took about 97.511 s. The whole procedure including optimisation took 410.42 s.

This can be considered an acceptable performance; a small number of epochs is sufficient in the learning stage for training the network.

3.4. Overall comparison

We compared the results of our neural model with results provided by Scharstein and Szeliski (2002) in their paper. The authors developed an overall comparison analysis of different stereo methods using the data sets available on the web site. Among the algorithms considered, we have selected the following six (three among those with better performances and three among those with worse performances):

- SSD—21×21 shiftable window SSD;
- Max-flow/min-cut algorithm (one of the first method to formulate matching as a graph flow problem);
- Genetic algorithm (a global optimisation technique operating on quadtrees);
- Fast correlation algorithm (an efficient implementation of correlation based matching with consistency and uniqueness validation);
- Discontinuity-preserving regularisation (a multiview technique for virtual view generation);
- Maximum surface technique (a fast stereo algorithm using rectangular subregions).

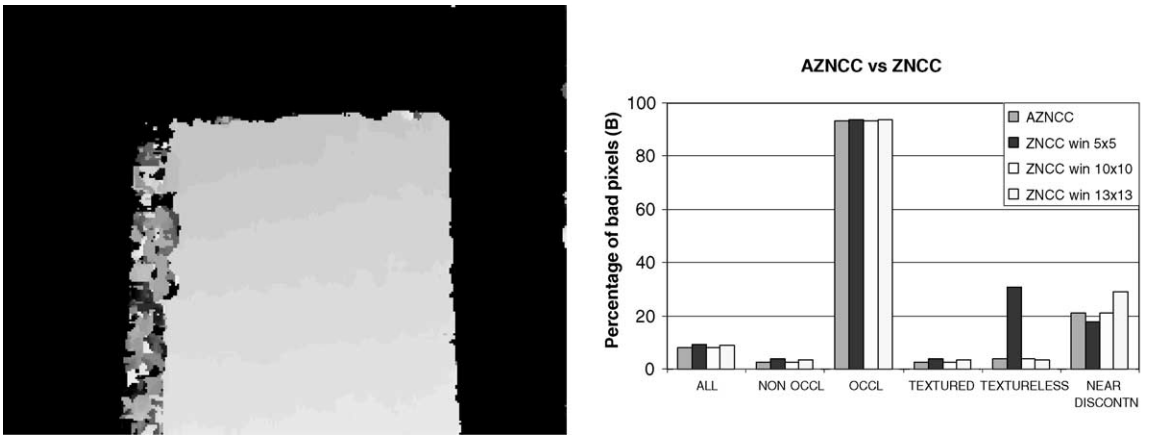


Fig. 9. Disparity map and plots of performances obtained by AZNCC and local algorithms implementing the matching cost ZNCC and 5×5, 10×10 and 13×13 aggregation windows.

Readers interested in additional explanations concerning the algorithms considered and their parameter settings, are referred to (Scharstein and Szeliski, 2002).

Table 2 lists the performances obtained for the Map image in non-occluded and discontinuity regions.

Table 2—comparative performances of stereo algorithms. The algorithms are listed roughly in decreasing order of overall performance. The metrics used in the comparison is that proposed by Scharstein and Szelisky for the Map image.

For all the algorithms considered, a complete analysis of the comparison results involves an in-depth investigation of factors influencing performances such as parameter setting, capability of the individual algorithm components in relation to stereo image characteristics. This is beyond the scope of the present work. We conclude that our approach shows competitive performances, but improvements are required to exploit the potential of the solutions adopted. The neural adaptive model processes textureless, occluded and discontinuity regions with different performances. In particular, the combined use of ZNCC and adaptive window size reduction yields satisfactory results when dealing with textureless areas, as shown in Figs. 6 and 9. On the contrary, solutions for occlusion and discontinuity are ineffective as confirmed by the experimental findings shown in Table 2.

Table 2
Comparative performances of stereo algorithms

	B measurements in non-occluded measures	B measurements in discontinuity regions
SSD	0.66	9.35
Genetic	1.04	10.91
Max flow	3.13	15.98
AZNCC	2.45	21.08
Discont-preserving	2.36	33.01
Fast correlation	8.42	12.68
Max surf.	4.17	27.88

4. 3D surface reconstruction from SEM images

The proposed image matching strategy was applied to stereoscopic SEM image analysis for 3D automatic surface reconstruction.

4.1. Imaging geometry

A SEM generates an image by a finely focused electron beam, sequentially raster-scanned across the specimen by conventional electromagnetic coils. The interaction of the electron beam with the specimen surface originates a wide range of signals (secondary electrons, back-scattered electrons, Auger electrons, X-rays, etc.), reflecting different features of the specimen, which are individually collected by dedicated sensors. An image of the

specimen is then reconstructed on a separated cathode-ray screen in synchronism with the scanning beam. As is evident, the optical system of the SEM is far simpler in principle than that of an optical microscope, since lenses and apertures are only used in the generation and focusing of the electron beam while the signal emerging from the specimen is just collected, with no further processing.

The image geometry, which is therefore entirely determined by the focusing and scanning parameters, is a central (perspective) projection where the specimen is imaged as seen from a nodal point of the electro-optical system corresponding to the centre of the final diaphragm. The *projection length* is defined as the distance between this point and the specimen surface. Different magnifications are obtained by just varying the scan width, whilst the projection length is usually kept constant. At adequately high magnifications, the scan width becomes negligible with respect to the projection length, and the imaging geometry can be reasonably approximated to a parallel (orthoscopic) projection.

In the Philips XL-30 FEG SEM used in the present study the specimen was mounted on a motorized micrometric X–Y–Z stage. It can be freely rotated around the optical axis of the system and manually tilted around an orthogonal axis, corresponding to the X axis of the image (the "fast" scanning axis). The tilt angle can be directly read on a large, analog readout.

The electron beam is a coherent beam generated by a Schottky field-effect electron source; the scan coils are driven by computer-controlled and linearized DACs, and the projection length between the nodal point and the specimen surface, also called *working distance*, can be determined with a typical accuracy of 1%.

4.2. Materials used and application-specific assumptions

The present study was explicitly aimed at the reconstruction of *surfaces*, rather than *volumes*. The technical approach used is indeed appropriate for the reconstruction of height maps, while different, tomographic techniques would be required for true 3D reconstruction.

All specimens were imaged from a working distance of 10.0 mm, which is the distance for which the electro-optical system is optimized, while the tilt angle was kept at 5 degrees to limit the reciprocal occlusion of tall structures protruding from the specimen surface.

All pictures were digitally acquired *on-board* either as 720×484 or 1440×968 pixel TIFF files. The only constraints hard-wired in the procedure were that all images were expected to be 8bpp grayscale, and that the tilt axis had to be parallel to the X axis and to intercept the Y axis at mid-height of the image.

4.3. Extracting points of interest

For this application, an important phase preliminary to the image matching procedure consisted in the search for points of interest within the reference image, from which to attempt the matching process. First of all the Laplacian Pyramid filtering technique (Burt and Adelson, 1983) was applied to the reference image; it enhanced salient image points by subtracting a sampled low-pass filtered copy of the image from the image itself. The filtered image was then partitioned uniformly in windows of a given dimension (suggested window dimension 7×7) and pixels corresponding to local maxima within each window were selected as points of interest.

4.4. Application results

The adaptive model was tested on several image pairs, mostly taken—at different magnifications—from specimens of animal extracellular matrix: fragments of human aortic wall and colon, and of the ciliary body of the eye. For these samples, the reconstruction process was evaluated qualitatively by experts. It was plainly obtained by visual comparison with the original specimen.

The results obtained are shown in Figs. 10–12 where the source image, the digital elevation model and the 3D surface reconstruction based on Delaunay triangulation technique, are illustrated for each case.

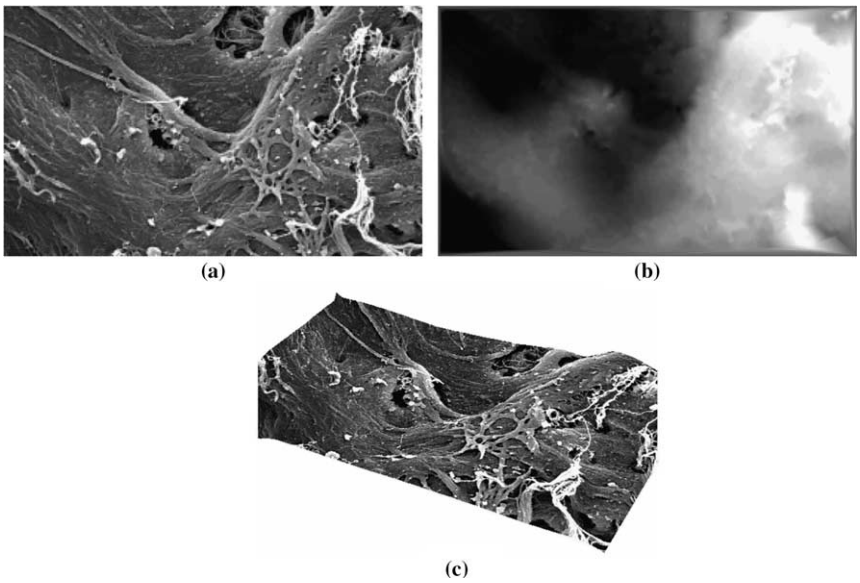


Fig. 10. Fragment of human aortic wall: (a) source image, (b) digital elevation model, and (c) 3D surface reconstruction.

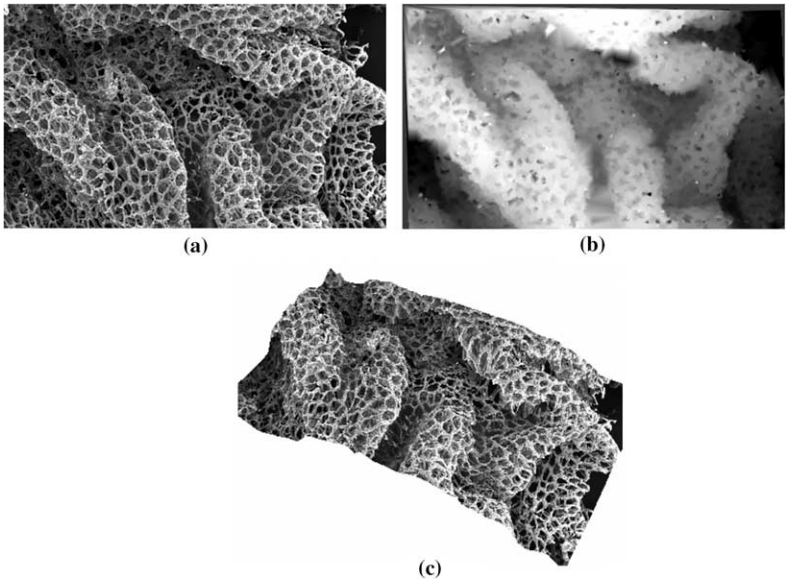


Fig. 11. Fragment of human colon wall: (a) source image, (b) digital elevation model, and (c) 3D surface reconstruction.

To compare the accuracy of the proposed image matching procedure when dealing with SEM imagery, specific preparations of inorganic crystals (NaCl) were imaged (see Fig. 13a). The regular and known structure of the image patterns facilitates the qualitative evaluation of results.

We proceeded in the image matching task by applying the adaptive correlation model to search

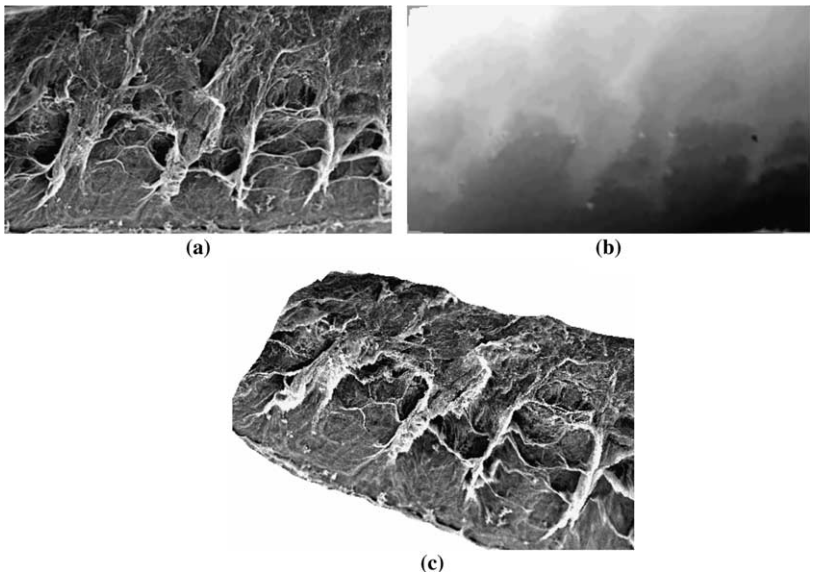


Fig. 12. Fragment of ciliary body of the eye: (a) source image, (b) digital elevation model, and (c) 3D surface reconstruction.

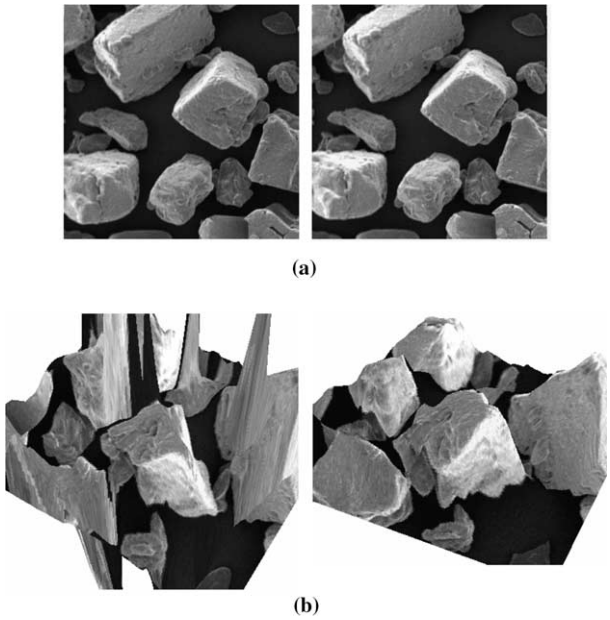


Fig. 13. (a) Reference and target images of specific preparations of inorganic crystals (NaCl). (b) Surface reconstructions obtained with conventional correlation coefficient and adaptive neural correlation model.

for corresponding points. Results were compared with those obtained with a conventional correlation coefficient operator which makes use of a fixed 9×9 correlation window. The window

dimension in the conventional procedure was selected optimising the performances during a trial and error procedure.

The results shown in Fig. 13b confirm that the adaptive method performs better.

5. Conclusions

Our objective in this study was to investigate the potentialities of neural learning in the field of stereo matching. A new approach based on the use of cross-correlation and neural techniques adapting local windows in shape and size, has been illustrated and experimentally evaluated.

Firstly the strategy was tested on standard data sets available on the Web. As seen in this experimental context the individual components of the matching algorithm positively contribute to solving the global matching problem and the allied use of correlation and adaptive techniques benefits the matching in general. However, global comparative evaluation highlighted some limitations. Better performances, in fact, were obtained by other methods both in non-occluded and discontinuity regions.

Satisfactory results were obtained in the second experimental evaluation aimed to reconstruct 3D surfaces from SEM stereo image pairs.

Overall the computational complexity implied by the neural learning stage was limited by the low number of epochs required for training.

At present, in the light of the results obtained in both the experimental contexts, we may confirm that neural learning has good potential in the field of stereo matching. We consider this study preliminary to further investigation involving both methodological and experimental issues. For example, the present solution can easily be reinforced implementing a varied set of matching costs within the neural structure and selecting the best solution for each case and providing specifying solutions for discontinuity such as adaptive variation of window shape and size based on edge detection measures. Further experiments are planned to substantiate these new ideas.

References

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Bobik, A.F., Intille, S.S., 1999. Large occlusion stereo. Int. J. Comput. Vis. 33, 181–200.

Boykov, Y., Veksler, O., Zabih, R., 1997. Disparity component matching for visual correspondence. In: Proc. of International Conference on Computer Vision and Pattern Recognition, pp. 470–475.

Burt, P.J., Adelson, E.H., 1983. The Laplacian Pyramid as a compact image code. IEEE Trans. Commun. 31, 532–540.

Barnard, S.T., Fischler, M.A., 1982. Computational stereo. Comput. Surv. 14, 553–572.

Barnard, T., Thompson, W.B., 1980. Disparity analysis of images. IEEE Trans. PAMI 2, 333–340.

Chen, Q., Medioni, G., 1999. A volumetric stereo matching method: Application to image-based modelling. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 29–34.

Cox, J.I., Higonani, S.L., Rao, S.P., Maggs, B.M., 1996. A maximum likelihoods stereo algorithm. Comput. Vis. Image Understanding 63, 542–567.

Dhond, U.R., Aggarwal, J.K., 1989. Structure from stereo—a review. IEEE Trans. Syst., Man, Cyber. 19, 1489–1510.

Faugeras, O.D., 1993. Three-Dimensional Computer Vision: A Geometric Viewpoint. Cambridge, MIT Press, MA.

Gonzalez, R.C., Woods, R.E., 2002. Digital Image Processing. Prentice Hall, Upper Saddle River, New Jersey.

Hsieh, Y.C., McKeown, D., Perlant, F.P., 1992. Performance evaluation of scene registration and stereo matching for cartographic feature extraction. IEEE TPAMI 14, 214–238.

Lotti, G.L., Giraudon, G., 1994. Correlation algorithm with adaptive window for aerial image in stereo vision. In: Proc. of Image and Signal Processing for Remote Sensing, Europto Symposium.

McMillan, L., Bishop, G., 1995. Plenoptic modelling: An image-based rendering system. Comput. Graphics (SIGGRAPH'95), 39–46.

Marr, D., Poggio, T., 1976. Cooperative computation of stereo disparity. Science 194, 283–287.

Okutomi, M., Kanade, T., 1992. A locally adaptive window for signal matching. Int. J. Comput. Vis. 7, 143–162.

Pao, Y.H., 1989. Adaptive Pattern Recognition and Neural Networks. Addison Wesley, MA.

Rumelhart, H., Hinton, G.E., Williams, R.J., 1986. Learning internal representation by error propagation. In: Rumelhart, H., McClelland, J.L. (Eds.), Parallel Distributed Process. MIT Press, Cambridge, MA, pp. 318–362.

Shapiro, L.G., Haralick, R.M., 1981. Structural description and inexact matching. IEEE Trans. Pami 3, 504–519.

Scharstein, D., Szeliski, R., 1996. Stereo matching with non-linear diffusion. In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. 47, 7–42.

Szeliski, R., Scharstein, D., 2002. Symmetric sub-pixel stereo matching. In: Proc. of 7th European Conference on Computer Vision, pp. 525–540.

Sun, C., 2002. Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. Int. J. Comput. Vis. 47 (1/2/3), 99–117.

Zabih, R., Woodfill, J., 1994. Non-parametric local transform for computing visual correspondence. In: ECCV, Vol. II, pp. 151–158.

Zhang, Y., Kambhamettu, C., 2002. Stereo matching with segmentation-based cooperation. In: Proc. of 7th European Conference on Computer Vision, pp. 556–571.

Weng, J., Ahuja, N., 1989. Motion and structure from two perspective views: Algorithms, error analysis and error estimation. IEEE Trans. PAMI 11, 451–476.