ELSEVIER

Pattern Recognition Letters 19 (1998) 319–330

Pattern Recognition Letters

Stereo matching based on the self-organizing feature-mapping algorithm

G. Pajares [a,*], J.M. Cruz [a], J. Aranda [b]

[a] *Dpto. Informática y Automática, Facultad de CC Físicas, Universidad Complutense, 28040 Madrid, Spain*
[b] *Dpto. Informática y Automática, Facultad de CC Físicas, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain*

Received 24 September 1996; revised 1 December 1997

Abstract

This paper presents an approach to the local stereo matching problem using edge segments as features with several attributes. We have verified that the differences in attributes for the true matches cluster in a cloud around a center. The correspondence is established on the basis of the minimum squared Mahalanobis distance between the difference of the attributes for a current pair of features and the cluster center (similarity constraint). We introduce a learning strategy based on the Self-Organizing feature-mapping method to get the best cluster center. A comparative analysis among methods without learning is illustrated.

Keywords: Self-organizing feature-mapping; Stereovision; Local matching; Learning; Training

1. Introduction

The key step in stereovision is image matching, namely, the process of identifying the corresponding points in two images that are generated by the same physical point in space. This paper is devoted solely to this problem. The stereo correspondence problem can be defined in terms of finding pairs of true matches that satisfy three competing constraints: similarity, smoothness and uniqueness (Marr and Poggio, 1979). The similarity constraint is associated to a local matching process where a minimum difference attribute criterion is applied. The results computed in the local process are later used by a global matching process where other constraints are imposed, for example, smoothness (Marr and Poggio, 1979), Minimum differential disparity (Medioni and Nevatia, 1985), Figural continuity (Pollard et al., 1981). A good choice of local matching strategy is the key for good results in the global matching process.

This paper presents an approach to the local stereopsis correspondence problem by developing a learning strategy based on the Self-Organizing Feature-Mapping (SOFM) algorithm (Kohonen, 1989; Kosko, 1992; Martin-Smith et al., 1993; Haykin, 1994; Sonka et al., 1995; Flanagan and Hasler, 1995).

Two sorts of techniques have been broadly used for stereo matching, (Dhond and Aggarwal, 1989; Ozanian, 1995; Pajares, 1995) area-based and feature-based. (1) Area-based stereo techniques use cor-

* Corresponding author. E-mail: pajares@eucmax.sim.ucm.es.

0167-8655/98/$19.00
PII S0167-8655(98)00003-8

relation between brightness (intensity) patterns in the local neighbourhood of a pixel in one image and brightness patterns in the local neighbourhood in the other image (Fua, 1993), where the number of pairs of features to be considered becomes high. (2) Feature-based methods use sets of pixels with similar attributes, normally either pixels belonging to edges (Kim and Aggarwal, 1987; Marr and Poggio, 1979; Mousavi and Schalkoff, 1994; Pollard et al., 1981) or the corresponding edges themselves (Ayache and Faverjon, 1987; Cruz et al., 1995a,b; Hoff and Ahuja, 1989; Medioni and Nevatia, 1985; Pajares, 1995). As shown in (Ozanian, 1995), these last methods lead to a sparse depth map only, leaving the rest of the surface to be reconstructed by interpolation; but they are faster than area-based methods, because there are much fewer points (features) to be considered.

There are intrinsic and extrinsic factors affecting the stereovision matching system: (a) *extrinsic*, in a practical stereo vision system, the left and right images are obtained at different positions/angles; (b) *intrinsic*, the stereovision system is equipped with two different physical cameras (i.e., with different components), which are always placed at the same relative position (left and right). A systematic noise appears for each camera.

Due to the above-mentioned factors, the corresponding features in the two images may display different attribute values. This may lead to incorrect matches. Thus, it is very important to find features in both images which are unique or independent of possible variation in the images (Wuescher and Boyer, 1991). Our experiment has been carried out in an artificial environment where the edge segments are abundant. Such features have been studied in terms of reliability (Breuel, 1996) and robustness (Wuescher and Boyer, 1991) and, as mentioned before, have also been used in previous stereovision matching works. This fact justifies our choice of features, although they may be too localised. Four average attribute values (module and direction gradient, variance and Laplacian) are computed for each edge-segment as shown later.

The extrinsic factors have been broadly considered in the literature. This paper deals with both kinds of factors but it is mainly concerned with the intrinsic factors since we have verified their significance and as a result, a research line based in learning strategies has been opened to solve the stereovision matching problem (Cruz et al., 1995a,b; Pajares, 1995). In (Cruz et al., 1995a; Pajares, 1995), for each pair of features the difference in attributes is computed and a Gaussian Probability Density Function (PDF) is associated with all differences in attributes for all pairs of features classified as true matches. The mean vector and covariance matrix needed by the PDF are estimated following a maximum likelihood method, which leads to a learning law. Afterwards, given a pair of features, a probability of matching is computed based on the PDF. In (Cruz et al., 1995b; Pajares, 1995), the perceptron criterion function is used to establish the difference between the attribute vectors for left and right images, then the synaptic weights of the perceptron are updated through a learning law (this learning law is different from that given in (Cruz et al., 1995b; Pajares, 1995)). Following this, for each pair of features the difference of attributes combined with the synaptic weight vector determines if the pair is a true or false match.

In stereovision matching we are only concerned with the true matches and correspondence is based on a minimum distance criterion (similarity constraint) between attributes of features. We have verified that their differences in attributes cluster in a cloud around a center (Pajares, 1995). Hereinafter, we will associate the terms cloud and cluster, without distinction, with the grouping of the differences in attributes for the true matches. This cluster is surrounded by differences in attributes corresponding to false matches. Hence, we attempt to design and optimize our stereo matching system at the same time as we provide a robust method for any stereo matching system. Our goal is to learn the best cluster center without target prototypes. Therefore this is an unsupervised learning approach. Unsupervised learning is also used by Cruz et al. (1995a) and Pajares (1995). The SOFM technique is chosen due to its self-organizing capability. The variability in attribute values in the two images suggests that a better representative cluster center can be obtained considering differences in attributes close to the cloud although they do not belong to the cloud. The SOFM also embodies this possibility. Moreover, the convergence of weights in the learning process, in terms of training patterns, is faster in the SOFM than in the

method of Cruz et al. (1995a) and Pajares (1995), due to the conjunction in the learning rate of two functions: (1) a neighbourhood function associated with the dispersion of the vectors in the cluster through the Mahalanobis distance (Duda and Hart, 1973; Maravall, 1993), and (2) a decreasing function related to the number of training vectors.

This paper is organized as follows. In Section 2 the local stereo matching system is designed with three basic modules performing the following three operations: (1) extraction of features and attributes, (2) training, using the SOFM and (3) matching for the current stereo-pairs. In Section 3, to show the effectiveness of the learning process, a test strategy is designed, and a comparative analysis is performed against classical methods without learning and other recent works using learning processes. Finally, in Section 4, the conclusions are presented.

2. Local stereo correspondence

Our local stereo matching system is equipped with a parallel optical axis geometry and designed with three basic modules: (1) image analysis, (2) training, and (3) current stereo matching. The function of the image analysis module is to extract information (features and their properties or attributes) from the scene and to make this information available to the training module or to the current stereo matching module. The image analysis is also responsible for performing an initial selection of pairs of features, supplying, to either of the other two systems, only those pairs that verify two conditions: (1) their absolute value of the difference in the direction of the gradient is below a specific threshold, fixed to 30° (the direction of the gradient is an attribute that will be defined later) and (2) their overlap rate, percentage of coincidence when one segment slides over another one following an epipolar line (Medioni and Nevatia, 1985), surpasses a certain value, fixed to 70%.

The system works in two mutually exclusive modes: OFF-LINE or training process and ON-LINE or decision matching process. In both modes the image analysis extracts features and attributes. In the OFF-LINE mode the system updates, through the corresponding training process, the cluster center vector and the covariance matrix associated with the cloud of the differences in attributes. During the ON-LINE process the system uses the updated cluster center vector and the covariance matrix obtained during the last OFF-LINE process, and computes the Mahalanobis distance between the cluster center vector and the attribute difference vector of a given incoming pair of features to decide if it is a true or a false match.

2.1. Feature and attribute extraction

As stated in the Introduction, due to the intrinsic and extrinsic factors, the corresponding features in both images may display different values. This may produce incorrect matches. Hence, we have chosen edge-segments as features due to the following reasons: (a) they have high reliability (Breuel, 1996) and robustness (Wuescher and Boyer, 1991), (b) they are abundant in the environment where the experiments have been carried out, and (c) they have been successfully used in previous stereovision matching works (Ayache and Faverjon, 1987; Cruz et al., 1995a,b; Hoff and Ahuja, 1989; Medioni and Nevatia, 1985; Pajares, 1995).

The contour edges in both images are extracted using the Laplacian of Gaussian filter in accordance with the zero-crossing criterion (Huertas and Medioni, 1986). For each zero-crossing in a given image, its gradient vector (magnitude and direction) as in (Leu and Yau, 1991), Laplacian as in (Lew et al., 1994) and variance as in (Krotkov, 1989), are computed from the gray levels of a central pixel and its eight immediate neighbours. To find the gradient magnitude of the central pixel, we compare the gray level differences from the four pairs of opposite pixels in the 8-neighbourhood; the largest difference is taken as the gradient magnitude. The gradient direction of the central pixel is the direction out of the eight principal directions whose opposite pixels yield the largest gray level difference and also points in the direction which the pixel gray level is increasing. A chain-code with 8 principal directions allows the normalization of the gradient direction. Once the zero-crossings are detected we use the following two algorithms for extracting the edge-segments or features: (a) Tanaka and Kak (1990), adjacent zero-crossings are connected if their corresponding differ-

ences in gradient magnitude and gradient direction do not overpass the quantities of $\pm 20\%$ and $\pm 45^{\circ}$ respectively, (b) Nevatia and Babu (1980), each detected contour according to the preceding algorithm is approximated by a series of piecewise linear line segments. Hence, we have built edge-segments made up of a certain number of zero-crossings. As stated before, for each zero-crossing we have computed four attributes (magnitude and direction gradient, Laplacian and variance). We consider the four attributes for all zero-crossings belonging to a given edge-segment and for each attribute an average value is finally obtained. All average attribute values are scaled, so that they fall within the same range. These four averaged values are the associated attributes to the given edge-segment. Moreover, each edge-segment is identified with initial and final pixel coordinates, its length and its label.

Therefore, given a stereo-pair of edge-segments, where an edge-segment comes from the left image and the other from the right image, we have four associated attributes for each edge-segment (i.e., two groups of four attributes). With the two groups of attributes we make up two 4-dimensional vectors $\boldsymbol{x}_l$ and $\boldsymbol{x}_{\mathrm{r}}$, where their four components are the four averaged attribute values of each edge-segment. The sub-indices ''l'' and ''r'' are denoting edge-segments belonging to the left and right images respectively. Now, for the given stereo-pair of edge-segments, we obtain a 4-dimensional difference vector of attributes $\boldsymbol{x} = \{x_{\mathrm{m}}, x_{\mathrm{d}}, x_{\mathrm{p}}, x_{\mathrm{v}}\}$ from $\boldsymbol{x}_l$ and $\boldsymbol{x}_{\mathrm{r}}$. The components of $\boldsymbol{x}$ are the corresponding differences for module and direction gradient, Laplacian and variance, respectively. We must consider that an ideal true match has its representative difference vector, $\boldsymbol{x}$, null. Nevertheless, in any real system and due to the intrinsic and extrinsic factors, $\boldsymbol{x}$ differs at least slightly from the null vector.

2.2. Training process: SOFM applied to stereovision

2.2.1. Brief description of the SOFM

The SOFM is a particular case of competitive learning neural networks developed by Kohonen (1989). Basically, given a set of neurons, they can learn competitively if they have common input connections and they learn stimuli patterns selectively, in such a way that this selectivity depends only on the specialization that neurons themselves develop during the training phase. The simplest scheme of competitive learning consists of the iterative execution of the following steps: (1) present a stimulus vector, $\boldsymbol{x}$, (2) compute the winning neuron, (3) change its synaptic weight vector and those of the neurons belonging to a certain neighbourhood, by small quantities towards the applied input stimulus vector.

In the Kohonen network the winning neuron is, by definition, the neuron that has the synaptic weight vector closest to the input stimulus, where ''closest'' can be defined in different ways. Typically, it is based on a distance, $d(\boldsymbol{x},\boldsymbol{m})$. Moreover, the weight modification not only affects the winning neuron but also those neurons belonging to a certain neighbourhood.

2.2.2. The SOFM in stereovision matching

During our experiments (Cruz et al., 1995a; Pajares, 1995), we have verified that the differences in attributes for the true matches cluster in a cloud around a center (which differs from the null vector) with an associated Probability Density Function (PDF). Without loss of generality the PDF is considered a Gaussian one, with two unknown parameters: (a) the mean difference vector $\boldsymbol{m}$, that will be learnt and which is the center of the cloud, and (b) the covariance matrix C, that is to be estimated as seen later.

We consider the pattern space $\mathbb{R}^4$ of the differences in attributes for true and false matches quantized into r regions of quantization or neurons, each one with an associated prototype $\boldsymbol{m}_{\mathrm{r}}$ or synaptic weight vector. Fig. 1 illustrates the quantization of the space $\mathbb{R}^4$. This quantization could be considered as a Voronoi tessellation bordered by hypersurfaces in $\mathbb{R}^4$ so that each partition contains a reference vector (synaptic weight vector) that is the ''nearest neighbour'' to any vector within the same partition (Patterson, 1996; Kohonen, 1995).

The cloud where true matches cluster is a neuron, with its cluster center $\boldsymbol{m}$ as the synaptic weight vector. Without loss of generality, this neuron is considered a hyper-sphere in $\mathbb{R}^4$ with radius R. The remainder of neurons in $\mathbb{R}^4$ are associated with false matches and these neurons, different from the central one, are similar in size to the central one. In stereovi-

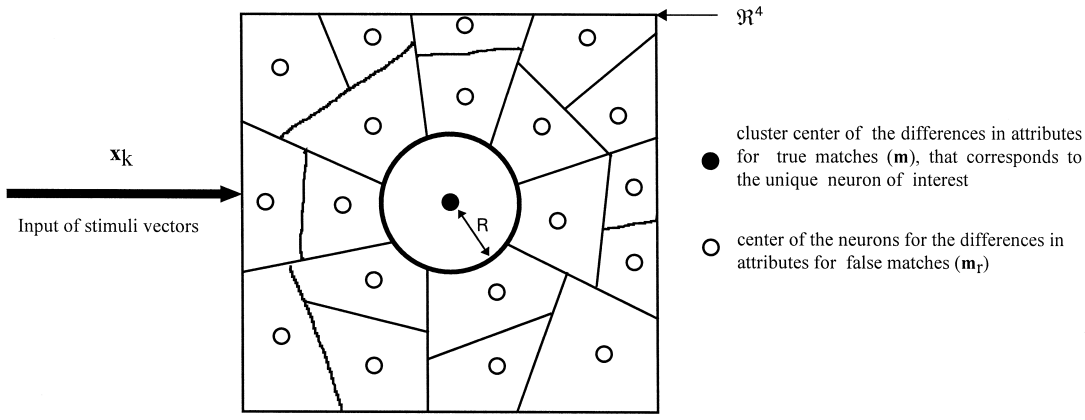

Fig. 1. Partition of pattern space into r regions of quantization or neurons. The internal hyper-sphere of radius R corresponds to the cluster of the differences in attributes.

sion matching we are only concerned in the detection of true matches. This is because, if a given pair of features is not a true match, it is obviously a false one (i.e., only two possible classes are used). According to the SOFM if this neuron wins, its synaptic weight vector is updated and also the corresponding synaptic weight vectors of the neurons belonging to a certain neighbourhood. Each neuron and all neurons surrounding it form the neighbourhood required by the SOFM. Therefore, the cluster center ($\boldsymbol{m}$) is updated in the following two cases: (1) when the central neuron results in being the winner and (2) when the central neuron belongs to the neighbourhood of any other neuron (associated to false matches) which result in being the winner. We have verified in our experiments, that in the second case, the movement of $\boldsymbol{m}$ is insignificant compared to the first one. Also, the movement of any $\boldsymbol{m}_r$ is insignificant when it is updated due to the activation of a neuron different from itself (including the case in which the winning neuron is the central one). Whereby and taking into account that we are only concerned with the detection of true matches, we have decided that the learning in our model is performed only by the central neuron if it results in being the winner. Due to the verified specific behaviour of our stereovision matching system, the neighbourhood is reduced to include the winning neuron only and no learning takes place among the losers, hence, the SOFM performs vector quantization (Patterson, 1996).

Now the goal is to compute the best representative cluster center $\boldsymbol{m}$, through the corresponding OFF-LINE learning process. The training is carried out with a set of n stimuli vectors $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, each stimulus vector is the difference vector of the attributes for a pair of features (see Section 2.1). At each iteration k a stimulus vector $\boldsymbol{x}_k$ is supplied to the system, so k ranges from 1 to n. When the winning neuron is the central one the synaptic weight vector $\boldsymbol{m}$, is moved in the direction of input stimulus $\boldsymbol{x}_k$. The adaptive update rule is given by

$$\boldsymbol{m}(k+1) = \begin{cases} \boldsymbol{m}(k) + c(k)h(d)[\boldsymbol{x}_k - \boldsymbol{m}(k)], & \text{if } d_M(\boldsymbol{x}_k, \boldsymbol{m}(k)) \leqslant R, \\ \boldsymbol{m}(k), & \text{otherwise,} \end{cases} \tag{1}$$

where $\boldsymbol{m}(k)$ is the corresponding value for the synaptic weight vector or cluster center before the stimulus $\boldsymbol{x}_k$ is processed, and the vector $\boldsymbol{m}(k+1)$ is the corresponding synaptic updated weight vector after $\boldsymbol{x}_k$ is processed; $d_M(\boldsymbol{x}_k, \boldsymbol{m}(k)) = (\boldsymbol{x}_k - \boldsymbol{m}(k))^T C^{-1} (\boldsymbol{x}_k - \boldsymbol{m}(k))$ is the squared Mahalanobis distance between the current stimulus vector and $\boldsymbol{m}(k)$ (Duda and Hart, 1973; Maravall, 1993). It is the metric used to determine whether the central neuron wins or does not during the competition process. The covariance matrix C, used in the computation of such distance, is the one obtained during the last OFF-LINE process, since the one corresponding to the current OFF-LINE process is not available yet (C measures the dispersion of the

stimuli vectors in the cluster); $c(k)$ is either constant or defines a decreasing sequence of positive numbers on the interval (0,1]. Numerous simulations have shown that the best results are obtained if it is selected fairly wide in the beginning and then permitted to shrink with iteration k, and fulfills $c(k) \to 0$, $k \to \infty$ (Kohonen, 1989; Kosko, 1992; Martin-Smith et al., 1993; Haykin, 1994). The function $h(d)$ is a neighbourhood function (Martin-Smith et al., 1993; Flanagan and Hasler, 1995) such that $h: \mathbb{R}^+ \to \mathbb{R}^+$. Finally, R is the radius of the hyper-sphere. Experimental tests have shown that $R = 10$ is a satisfying value for the radius. In practice, we have chosen the following expressions to define both $c(k)$ and $h(d)$:

$$c(k) = \frac{1}{a+k}, \quad h(d) = \frac{1}{b + d_{\mathrm{M}}(x_k, \boldsymbol{m}(k))}, \quad (2)$$

where a and b are both integer constants, that in our work, after experimentation, have been fixed to 20 and 1, respectively. Both $c(k)$ and $h(d)$ control the movement in updating $\boldsymbol{m}$. So, $c(k)$ decreases as the number of samples increases according to the following law: ''*the learning rate decreases as learning progress*''; the neighbourhood function $h(d)$ takes into account how close the stimulus vector is to the synaptic weight vector (cluster center) $\boldsymbol{m}$, so $h(d)$ increases as the distance value decreases; the expressions $c(k)$ and $h(d)$ define the learning rate when they are taken jointly. This definition of the learning rate introduces an important improvement with regard to previous research works involving learning (Cruz et al., 1995a; Pajares, 1995). In these references, the learning rate is computed as follows:

$$\eta_k = \frac{p(\boldsymbol{x}_k)}{\sum_{i=1}^{k} p(\boldsymbol{x}_i)}, \quad (3)$$

where $p(\boldsymbol{x}_i)$ is the matching probability of the pattern stimulus $\boldsymbol{x}_i$ associated to the central neuron, and is computed as a decreasing exponential function of the Mahalanobis distance (defined by the Gaussian PDF). The denominator in Eq. (3) grows with the number of iterations k. For values of k smaller than a given threshold T (T is about 100 in our experiments), the number of training patterns is not still significant and the values of the learning rate η_k are still high. This leads to undesired high variabilities of $\boldsymbol{m}$ at this initial training phase. We avoid this by introducing the constant value a in $c(k)$. However, when k grows and the Mahalanobis distance is smaller than R, the $c(k)h(d)$ learning rate in Eq. (1) takes greater values than η_k. This is a training phase, where the number of training patterns can be considered significant, and the convergence of the SOFM is faster than the convergence in (Cruz et al., 1995a; Pajares, 1995). So, we need a smaller number of stimuli vectors to get a similar $\boldsymbol{m}$ to the one computed in (Cruz et al., 1995a; Pajares, 1995) and therefore, the computational cost is reduced.

Finally, we must estimate the covariance matrix C in order to get the second parameter of the PDF describing the cluster as stated before. This process is carried out according to a maximum likelihood method (Cruz et al., 1995a; Pajares, 1995), where as before the same set of n stimuli vectors is supplied (i.e., processed):

$$C(k+1) = C(k) + c(k)\,h(d) \times \left[(x_k - \boldsymbol{m}(k))^{\mathrm{T}} (x_k - \boldsymbol{m}(k)) - C(k) \right], \quad (4)$$

where it is important to point out that $\boldsymbol{m}(k)$ is the value computed through Eq. (1) during the current OFF-LINE process; $c(k)$ and $h(d)$ are, as before, given by Eq. (2), but using $\boldsymbol{m}(k)$ of the current OFF-LINE process for computing $h(d)$. $C(k)$ and $C(k+1)$ are the covariance matrices before and after the stimulus $\boldsymbol{x}_k$ is processed, respectively. Finally, ''T'' denotes transpose.

From the above considerations we can infer that the learning rules governing the described training process are given by Eqs. (1), (2) and (4).

2.3. The current stereo matching process

This is an ON-LINE process in which a pair of new stereo images are to be matched. The image analysis system extracts pairs of features and supplies their corresponding four-dimensional difference vectors of the attributes $\boldsymbol{x}$, to the stereo matching system. For each $\boldsymbol{x}$ received, the system computes the squared Mahalanobis distance $d_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{m})$, where the involved cluster center $\boldsymbol{m}$ and the covariance matrix C are the ones updated during the last OFF-LINE training process according to Eqs. (1), (2) and

(4). The incoming pair of features is classified as a true match if $d_M(\boldsymbol{x},\boldsymbol{m})$ is less than R, otherwise it is a false match. R is the radius of the hyper-sphere as defined in Section 2.2.

3. Experimental validation, comparative analysis and performance evaluation

To assess the validity and performance of our method, we design a test strategy with the following two goals:

1. To show the validity of the learning process, i.e., to show how results are improved as the learning process grows.
2. To show the effectiveness of our method as compared to classical local stereo matching techniques as those proposed by Kim and Aggarwal (1987) (KA) and Medioni and Nevatia (1985) (MN), and other more recent learning local methods as those proposed by Cruz et al. (1995a,b) and Pajares (1995).

3.1. Design of a test strategy

The objective is to prove the validity and generalization of the method by varying environmental conditions in two ways: by using new images with different features (different objects) and by changing the illumination. With this aim in mind, and the two goals pointed out before, 8 pairs of stereo-images captured with natural illumination are used as initial samples. Figs. 2–4 show three representative left images. Furthermore, three sets of stereo-images, which are different from each other, are used and will constitute the inputs for the test: SP1, SP2 and SP3 with 6, 6 and 10 stereo-images. A representative stereo pair is shown for each SP set in Figs. 5–7, respectively. The first set of stereo images (SP1) has been captured with natural illumination (as the initial stereo-images samples) and the remaining two sets with artificial illumination.

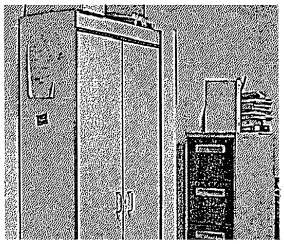

Fig. 3. Left original training image (furniture).

We assume that this is the first time that an OFF-LINE training process is carried out by the system, i.e., the parameter vector $(\boldsymbol{m},\boldsymbol{C})$ is set initially to $(\boldsymbol{0},\boldsymbol{I})$, because at this moment we have no knowledge of the behaviour of the system and it is considered as an ideal system where the differences in attributes values are null and no dispersion of the patterns in the cluster is considered.

The test process involves the following steps in which the parameter vector changes:

Step 0. The system performs an OFF-LINE training process using the stimuli vectors provided by the 8 pairs of initial stereo-images. As explained in Section 2.2.2, the stimuli vectors are the difference vectors of the attributes.

Step 1. The system processes two sets of stereo-images SP1 and SP3 during an ON-LINE process. Only the stimuli vectors coming from set SP1 are used for a new OFF-LINE training process because the stimuli vectors from set SP3 will again

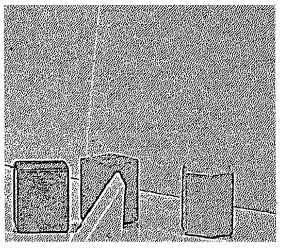

Fig. 2. Left original training image (blocks).

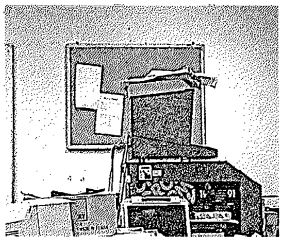

Fig. 4. Left original training image (computers).

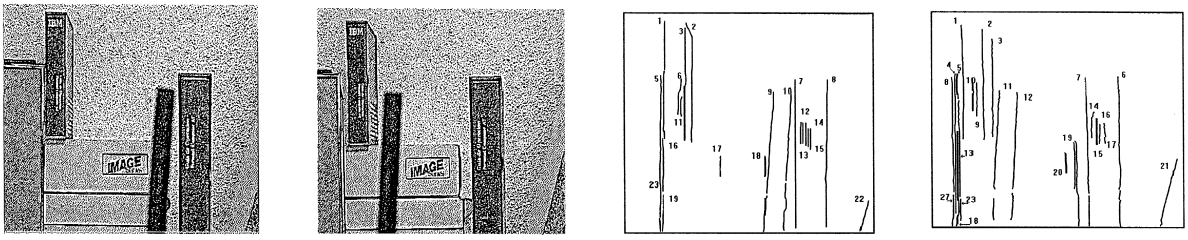

Fig. 5. SP1. (a) Original left stereo image. (b) Original right stereo image. (c) Labeled segments left image. (d) Labeled segments right image.

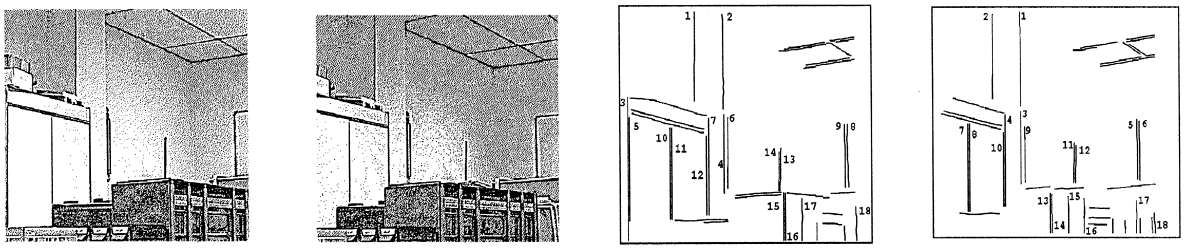

Fig. 6. SP2. (a) Original left stereo image. (b) Original right stereo image. (c) Labeled segments left image. (d) Labeled segments right image.

be ON-LINE processed later in Step 3, so that no interferences derived from its own processing arise at this step.

Step 2. The system processes set SP2 in a new ON-LINE process. The processing conditions are similar to those of set SP3 in Step 1, however, here the stimuli vectors are incorporated into a new OFF-LINE training process.

Step 3. The system once again performs an ON-LINE process with set SP3. At this point, the system is already familiar with the environment of this set, because it was trained with stimuli vectors from SP2 during Step 2 with similar illumination. Here it is intended to show better results than those obtained in Step 1 for set SP3.

According to point (2) of Section 2.2.2, we have a unique neuron associated to the cloud where the differences in attributes for the true matches cluster around a center $\boldsymbol{m}$. The cloud was considered a hyper-sphere with radius R. With this approach the size of the cloud is constant but the cloud position varies as the synaptic weight vector or cluster center $\boldsymbol{m}$ changes after each OFF-LINE training process (i.e., during the four test steps). The computed results of the cluster center vector $\boldsymbol{m}$ during Steps 0–3

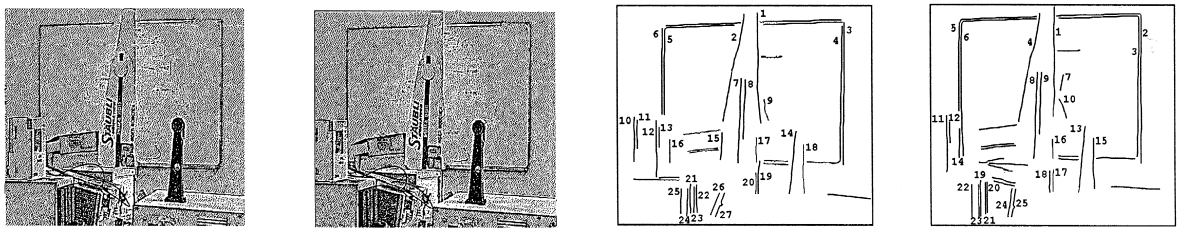

Fig. 7. SP3. (a) Original left stereo image. (b) Original right stereo image. (c) Labeled segments left image. (d) Labeled segments right image.

are respectively: {0.161, −0.014, 0.402, 0.393}, {0.201, −0.014, 0.402, 0.393}, {0.311, −0.968, 0.555, 0.735}, {0.415, −0.121, 0.720, 0.896}. At first, we are unable to fix the range of values for each element of the $\boldsymbol{m}$ vector, because these values depend on the differences in attribute values; such differences are due to intrinsic and extrinsic factors and we have no control over such factors. Nevertheless, for illustrative purposes, we have verified during our experiments (Cruz et al., 1995a; Pajares, 1995) that the values for the elements of $\boldsymbol{m}$ are restricted to the range $[-1.5..1.5]$.

The changes in the covariance matrix C throughout the 4 steps are not statistically significant, in which case it suffices to give C for Step 0.

$$C = \begin{bmatrix} 0.990 & 0.011 & 0.022 & 0.015 \\ 0.011 & 0.966 & -0.009 & -0.016 \\ 0.022 & -0.009 & 1.202 & -0.025 \\ 0.015 & -0.016 & -0.025 & 1.178 \end{bmatrix}.$$

To compare the effectiveness of the method, the KA and MN local matching techniques are selected. These methods work in the following way: (a) in KA, given two potential pixels for matching, a probability is computed through two weighting functions. One is based on the directional difference according to 16 fixed patterns, and the other is based on the difference in the gradients of gray-level intensity, (b) in the MN method, the local stereo-correspondence is established between edge segments by defining a boolean function indicating whether two segments are potential matches if they overlap (two segments overlap if by sliding one of them in a horizontal direction, they intersect) and have similar contrast and orientation. In short the KA and MN methods measure differences between attribute values and for comparison purposes, they can be replaced by the Euclidean distance, as it computes the same measurement. Thus, a comparison can be established with the Mahalanobis distance proposed in our method.

3.2. Comparative analysis

Table 1 records all computed results for the stereo-pair representative of the set SP1. There are four columns: the first one indicates the order number (on) for each pair of features appearing in the second column (pair), where the ''∗'' symbol denotes a true match as tested by a human expert; in the third and fourth columns, the Euclidean $d_E(\boldsymbol{x},\boldsymbol{0})$ and the Mahalanobis distances $d_M(\boldsymbol{x},\boldsymbol{m})$ are computed, respectively.

Table 1
Matching results from the stereo-pair representative of the set SP1, on: order number for the 39 pairs of considered features; pair: pairs of labeled features (l,r) from left and right images; ''∗'' symbol means a true match; $d_E(\boldsymbol{x},\boldsymbol{0})$, $d_M(\boldsymbol{x},\boldsymbol{m})$: computed results for the Euclidean distance (no learning) and Mahalanobis distance (with learning), respectively

On	Pair	$d_E(\boldsymbol{x},\boldsymbol{0})$	$d_M(\boldsymbol{x},\boldsymbol{m})$
1	∗(1, 1)	9.81	10.22
2	(2, 2)	7.92	8.15
3	∗(2, 3)	8.10	6.22
4	(2, 6)	18.14	11.12
5	(2, 8)	23.21	66.13
6	(2, 14)	15.17	11.25
7	∗(3, 2)	8.64	4.17
8	(3, 3)	7.21	7.97
9	(3, 8)	60.17	85.21
10	(3, 9)	7.23	9.22
11	(3, 10)	7.61	10.61
12	(3, 14)	8.92	7.88
13	∗(7, 7)	8.22	2.97
14	(7, 19)	8.32	7.15
15	(8, 2)	17.22	6.61
16	(8, 3)	16.32	7.82
17	∗(8, 6)	8.51	1.12
18	(8, 8)	17.21	32.31
19	(8, 9)	10.12	6.11
20	(8, 12)	8.14	14.10
21	∗(9, 11)	5.82	1.97
22	(10, 6)	9.05	7.22
23	∗(10, 12)	9.15	2.55
24	(11, 2)	7.55	5.85
25	(11, 3)	7.41	7.15
26	∗(11, 9)	7.36	1.15
27	(11, 10)	8.65	6.12
28	∗(13, 14)	5.45	4.15
29	(13, 16)	5.32	6.66
30	(14, 14)	7.01	5.27
31	∗(14, 16)	7.21	3.91
32	(15, 15)	4.90	5.82
33	∗(15, 17)	2.89	1.98
34	(16, 6)	12.12	20.25
35	(16, 12)	11.10	19.76
36	(16, 20)	10.21	9.10
37	(17, 18)	15.18	25.14
38	∗(18, 20)	4.11	3.71
39	∗(23, 27)	7.15	1.36

Of all the possible combinations of pairs of matches formed by segments of left and right images, only 39 of them are considered, as the remainder do not meet the initial restriction, which states that the value of the difference in the direction of the gradient must be less than $\pm 45°$ and the overlap rate (percentage of overlapping coincidence) greater than 75%. These matches are directly classified as False by the system and omitted from the results' table. The choice of such thresholds is supported by the parallel optical axis geometry with the given flexibility in order to avoid errors during previous stages. Of the 39 pairs considered, there are unambiguous and ambiguous ones, depending on whether a given left image segment corresponds to one and only one, or several right image segments, respectively. In any case, the decision about the correct match is made by choosing the result of the smaller value for each one of the methods (in the unambiguous case, there is only one) as long as it does not surpass a fixed threshold, which coincides with the radius R of the hyper-sphere (see Sections 2.2 and 2.3). R is set to 10 in this paper.

Table 2 shows results for the stereo-pair representative of set SP3 with the same symbols and criteria as those explained in Table 1, although two values, identified by numbers 1 and 3, are obtained for the Mahalanobis distance according to the processing for this stereo-pair in Steps 1 and 3, respectively. An overview of the results in Table 2, allows us to check that, in general, for true/false matches the minimum/maximum values for the Mahalanobis distance are obtained during Step 3 as compared with those obtained in Step 1. These results are the best ones and correspond to a phase of increased learning.

From results in Tables 1 and 2 and results for the stereo-pair representative of set SP2 (these last are omitted because they are similar to those of the stereo-pair representative of set SP3 in Step 1) we build Table 3 in which the final results are summarized. It displays the number of successes and failures for the stereo-pairs representing sets SP1, SP2 and SP3 according to the decision process explained above. Also, it shows a coefficient μ, which provides a decision margin when ambiguities arise. Such a coefficient is obtained as follows: (a) for each ambiguity case, we select two pairs of matches, one is the true match ($*$) and the other one the match with the closest distance value to the true match, (b) with the two selected pairs of matches we compute the difference between their corresponding distance values, (c) finally, the coefficient μ is the average value for all ambiguity cases. Hence, a minimum value (most negative value) for μ indicates that decisions can be taken with a higher degree of confidence. The processing of set SP3 in Steps 1 and

Table 2
Results from stereo-pair representative of SP3; on: order number for the 35 pairs of features; pair: pairs of labeled features (l,r) from left and right images, respectively, where "$*$" means a true match; $d_M(\boldsymbol{x},\boldsymbol{m})$, $d_E(\boldsymbol{x},\boldsymbol{0})$: computed results for the Mahalanobis distance (learning) and Euclidean distance (without learning), respectively, where (1) and (3) mean results computed according to test strategies in Steps 1 and 3, respectively

On	Pair	$d_E(\boldsymbol{x},\boldsymbol{0})$	$d_{M1}(\boldsymbol{x},\boldsymbol{m})$	$d_{M3}(\boldsymbol{x},\boldsymbol{m})$
1	$*(1,1)$	2.50	1.71	1.46
2	$*(2,4)$	3.08	3.01	2.82
3	$*(3,2)$	1.58	1.36	0.96
4	$(3,6)$	3.40	4.01	4.26
5	$*(4,3)$	2.04	2.05	1.70
6	$(4,5)$	11.14	12.11	12.50
7	$(5,1)$	80.12	91.00	94.60
8	$(5,2)$	6.76	12.18	13.11
9	$*(5,6)$	3.32	3.21	3.20
10	$(6,3)$	12.70	13.36	13.68
11	$*(6,5)$	4.01	3.25	3.11
12	$*(7,8)$	2.30	2.35	1.66
13	$(8,9)$	38.26	55.23	57.61
14	$(10,3)$	13.47	16.41	17.08
15	$(11,2)$	79.63	79.42	79.41
16	$(11,6)$	80.06	93.91	94.81
17	$*(12,11)$	0.51	0.51	0.48
18	$(12,15)$	7.45	10.29	10.60
19	$*(13,12)$	28.55	23.12	18.10
20	$*(14,13)$	3.50	2.69	2.55
21	$*(17,16)$	8.86	6.23	6.12
22	$(18,11)$	6.11	7.52	5.42
23	$*(18,15)$	7.89	7.76	4.96
24	$*(21,19)$	4.32	4.32	4.01
25	$(22,19)$	2.06	4.68	5.11
26	$*(22,20)$	2.55	2.02	1.67
27	$*(23,21)$	4.08	4.12	3.75
28	$(23,23)$	12.65	15.44	15.96
29	$*(24,22)$	15.78	11.12	8.22
30	$(25,21)$	1.19	3.32	3.61
31	$*(25,23)$	2.03	2.04	1.83
32	$(26,21)$	3.32	3.99	4.51
33	$(26,23)$	5.44	5.96	6.25
34	$*(26,24)$	2.35	2.15	2.08
35	$*(27,25)$	3.91	3.62	2.98

Table 3
Matching results for the stereo-pairs representing its SP1, SP2 and SP3 and decision margin (μ)

	$SP1_{NL}$	$SP1_{L}$	$SP2_{NL}$	$SP2_{L}$	$SP3_{NL}$	$SP3_{L}$ (1)	$SP3_{L}$ (3)
Successes	9	12	13	16	17	19	21
Failures	6	3	6	3	5	3	1
μ	−0.16	−3.29	−1.53	−4.18	−3.67	−5.48	−6.68

3 is denoted as (1) and (3), respectively; the subindex NL means ''No Learning'' (identified with KA and MN classical methods) and L means ''Learning'' (identified with our proposed SOFM local stereovision matching method).

Analyzing all results, the following conclusions may be inferred:

1. The learning process improves the matching results. As training progresses the results are better.
2. The absolute value of the decision margin increases with the training (i.e., better decisions are made with a greater learning).
3. Our stereovision local matching approach produces better results than the KA and MN classical local stereovision matching techniques.
4. Due to the definition of the learning rates, Eqs. (2) and (3), we have verified that the approach developed in this paper requires less training than other local stereovision matching techniques using also learning (Cruz et al., 1995a; Pajares, 1995) (CP) to get a similar number of successes. Indeed, Table 4 shows the number of training patterns to get a similar $\boldsymbol{m}$ at each step for the SOFM and CP methods. We can point out that with a similar $\boldsymbol{m}$ we obtain also similar percentages of successes over the same dataset described in Section 3.1. The total results show a reduction of a 12% in SOFM over CP considering the four steps. We can point out that the difference in percentage increases as the training grows. The extra number of training patterns processed by CP were supplied through additional stereo images, exactly 1, 1, 2 and 2 in the four steps, respectively.
5. As stated in Section 2.2.2 the processing complexity of the original SOFM has been reduced and now it can be considered similar to that in CP. Therefore, no additional processing time is added by the SOFM as compared to CP.

Table 4
Number of training patterns used in the different STEPs to achieve a similar cluster center $\boldsymbol{m}$ for the SOFM and CP methods

	Step 0	Step 1	Step 2	Step 3	Total
SOFM	482	362	283	325	1452
CP	506	394	346	403	1649

4. Concluding remarks

Our local stereovision matching method improves results as the training progresses, and shows a greater effectiveness as compared to other recent learning strategies. Also, it has been found to compare favorably with classical local stereo matching methods, where no learning is involved. This last fact is justified because the cluster center vector moves away from the null vector as training progresses. Such behaviour is not affected by the nature of the different objects nor by illumination conditions, but the intrinsic factors are decisive. The mismatches could be solved by applying global matching constraints.

Acknowledgements

The authors wish to acknowledge Professor Dr S. Dormido, Head of Department of Informática y Automática, Facultad de CC Físicas, UNED, Madrid, for his support and encouragement. Part of this work has been performed under project CICYT TAP94-0832-C02-01. We would like to thank the reviewers for the valuable comments which helped us to improve the paper.

References

Ayache, N., Faverjon, B., 1987. Efficient registration of stereo images by matching graph descriptions of edge segments. Internat. J. Comput. Vision 1, 107–131.

Breuel, T.M., 1996. Finding lines under bounded error. Pattern Recognition 29 (1), 167–178.

Cruz, J.M., Pajares, G., Aranda, J., 1995a. A neural network approach to the stereovision correspondence problem by unsupervised learning. Neural Networks 8 (5), 805–813.

Cruz, J.M., Pajares, G., Aranda, J., Vindel, J.L.F., 1995b. Stereo matching technique based on the perceptron criterion function. Pattern Recognition Letters 16, 933–944.

Dhond, A.R., Aggarwal, J.K., 1989. Structure from stereo – A review. IEEE Trans. Systems Man Cybernet. 19, 1489–1510.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Flanagan, J.A., Hasler, M., 1995. Self-organising artificial neural networks. In: Mira, J., Sandoval, F. (Eds.), From Natural to Artificial Neural Computation. Springer, Berlin.

Fua, P., 1993. A parallel algorithm that produces dense depth maps and preserves image features. Machine Vision Appl. 6, 35–49.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Macmillan College Publishing, New York.

Hoff, W., Ahuja, N., 1989. Surface from stereo: Integrating feature matching, disparity estimation and contour detection. IEEE Trans. Pattern Anal. Machine Intell. 11, 121–136.

Huertas, A., Medioni, G., 1986. Detection of intensity changes with subpixel accuracy using Laplacian–Gaussian masks. IEEE Trans. Pattern Anal. Machine Intell. 8 (5), 651–664.

Kim, D.H., Aggarwal, J.K., 1987. Positioning three-dimensional objects using stereo images. IEEE J. Robotics and Automation 3, 361–373.

Kohonen, T., 1989. Self-Organization and Associative Memory. Springer, New York.

Kohonen, T., 1995. Self-Organizing Maps. Springer, Berlin.

Kosko, B., 1992. Neural Networks and Fuzzy Systems. Prentice-Hall, Englewood Cliffs, NJ.

Krotkov, E.P., 1989. Active Computer Vision by Cooperative Focus and Stereo. Springer, New York.

Leu, J.G., Yau, H.L., 1991. Detecting the dislocations in metal crystals from microscopic images. Pattern Recognition 24 (1), 41–56.

Lew, M.S., Huang, T.S., Wong, K., 1994. Learning and feature selection in stereo matching. IEEE Trans. Pattern Anal. Machine Intell. 16 (9), 869–881.

Maravall, D., 1993. Reconocimiento de Formas y Vision Artificial. RA-MA, Madrid.

Marr, D., Poggio, T., 1979. A computational theory of human stereovision. Proc. Roy. Soc. London Ser. B 207, 301–328.

Martin-Smith, P., Pelayo, F.J., Diaz, A., Ortega, J., Prieto, A., 1993. A learning algorithm to obtain self-organizing maps using fixed neighbourhood Kohonen Networks. In: Mira, J., Cabestany, J., Prieto, A. (Eds.), New Trends in Neural Computation. Springer, Berlin.

Medioni, G., Nevatia, R., 1985. Segment based stereo matching. Comput. Vision Graphics Image Process. 31, 2–18.

Mousavi, M.S., Schalkoff, R.J., 1994. ANN implementation of stereo vision using a multi-layer feedback architecture. IEEE Trans. Systems Man Cybernet. 24 (8), 1220–1238.

Nevatia, R., Babu, K.R., 1980. Linear feature extraction and description. Comput. Vision Graphics Image Process. 13, 257–269.

Ozanian, T., 1995. Approaches for stereo matching – A review. Modeling Identification Control 16 (2), 65–94.

Pajares, G., 1995. Estrategia de Solucion al Problema de la Correspondencia en Vision Estereoscópica por la Jerarquía Metodológica y la Integración de Criterios. Ph.D. Thesis, Dpto. Informática y Automática, Facultad Ciencias UNED, Madrid.

Patterson, D.W., 1996. Artificial Neural Networks. Prentice-Hall, Singapore.

Pollard, S.B., Mayhew, J.E.W., Frisby, J.P., 1981. PMF: A stereo correspondence algorithm using a disparity gradient limit. Perception 14, 449–470.

Sonka, M., Hlavac, V., Boyle, R., 1995. Image Processing, Analysis and Machine Vision. Chapman-Hall, London.

Tanaka, S., Kak, A.C., 1990. A rule-based approach to binocular stereopsis. In: Jain, R.C., Jain, A.K. (Eds.), Analysis and Interpretation of Range Images. Springer, Berlin, pp. 33–139.

Wuescher, D.M., Boyer, K.L., 1991. Robust contour decomposition using a constraint curvature criterion. IEEE Trans. Pattern Anal. Machine Intell. 13 (1), 41–51.