Pattern Recognition Learning Applied to Stereovision Matching

Gonzalo Pajares, Jesús Manuel de la Cruz and José A. López

Dpto. Arquitectura de Computadores y Automática. Facultad de Ciencias Físicas.
Universidad Complutense. 28040 Madrid. SPAIN
pajares@eucmax.sim.ucm.es
jmcruz@dia.ucm.es

Abstract

This paper presents an approach to the local stereovision matching problem by developing a statistical pattern recognition learning strategy. We use 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 a maximum likelihood estimates method to get the best cluster center. A comparative analysis against a classical approach using the squared Euclidean distance (i.e. without learning) is illustrated.

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 presents an approach to the local stereopsis correspondence problem by developing a statistical learning strategy based on maximum likelihood estimates (Duda and Hart, 1973) where the matching problem is considered as a pattern classification problem.

Two sorts of techniques have been broadly used for stereo matching, (Ozanian, 1995, Pajares, 1995) area-based and feature-based: 1) Area-based stereo techniques use correlation 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, 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; Mousavi and Schalkoff, 1994) or the corresponding edges themselves (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. A test strategy is designed by varying environmental conditions (new images and different illumination). The experimental results prove that these conditions do not affect the behaviour of the system but the intrinsic and extrinsic factors are decisive.

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. Our experiment has been carried out in an artificial environment where the edge segments are abundant. This fact justifies that we choose edge-segments as features.

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 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 differences in gradient magnitude and gradient direction don't overpass the quantities of $\pm$ 20% and $\pm 45°$ 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 $\mathbf{x}_l$ and $\mathbf{x}_r$, 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 $\mathbf{x} = \{x_m, x_d, x_p, x_v\}$ from $\mathbf{x}_l$ and $\mathbf{x}_r$. The components of $\mathbf{x}$ 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 $\mathbf{x}$, null. Nevertheless, in any real system and due to the intrinsic and extrinsic factors, $\mathbf{x}$ differs at least slightly from the null vector.

In stereovision matching we are only concerned with the true matches and correspondence is based on a minimum distance criterion (e.g. Euclidean distance), which is the similarity constraint, between attributes of features. We have verified that their differences in attributes cluster in a cloud around a center (Pajares, 1995). We call to this cloud the class of true matches. This class is considered a hyper-sphere in $\Re^4$ with radius R. False matches are surrounding the class of true matches and they can be grouped in classes. Therefore, although there are N classes only one is the class associated to the true matches.

This paper is organized as follows: in section 2 the probability densities associated to each class is obtained; in section 3) the maximum likelihood estimates approach leads to a learning law; in section 4) we prove the validity of our method. Finally in section 5) the conclusions are presented

2. Finite Mixture of Multivariate Densities

Following Duda and Hart (1973), we start by assuming that we know the complete probability structure for the problem with the sole exception of the values of some parameters. To be more specific, we make the following assumptions: 1) The samples come from a set of N known classes w_j, where $j=1..N$; 2) The a priori probabilities $P(w_j)$ are also known; 3) The forms for the class-conditional probability densities $p(\mathbf{x}/w_j, \delta_j)$ are known, $j=1..N$. We suppose such densities as normal ones, with the parameter vector given by $\delta_j = (\mu_j, \Sigma_j)$, where μ and Σ are the mean vector and the covariance matrix, respectively and 4) All that is unknown are the values for the parameter vector μ_j, Σ_j.

Stimuli patterns are assumed to be obtained by selecting a class w_j with probability $P(w_j)$ and then selecting an $\mathbf{x}$ according to the probability law $p(\mathbf{x}/w_j, \delta_j)=N(\mathbf{x}/w_j, \mu_j, \Sigma)$. Thus, the probability density function for the samples is given by:

$$p(\mathbf{x} \mid \delta) = \sum_{j=1}^{N} p(\mathbf{x} \mid w_j, \delta_j) P(w_j) \qquad (1)$$

A density function of this form is called a mixture density. The conditional densities probabilities $p(\mathbf{x}/w_j, \delta_j)$ are the component densities, and the a priori probabilities $P(w_j)$ are the mixing parameters (Duda and Hart, 1973). Our basic goal will be to use samples drawn from this mixture density to estimate the unknown parameter vector δ.

Once δ is known, we can break the mixture down into its components. Then, when a difference vector x associated with a pair of features is presented to the system as an input, the pair will be classified as belonging to the class of true matches or to a class of false matches. An unsupervised learning approach is the key step to computing the unknown parameter vector δ. The unsupervised learning process is solved by the Maximum Likelihood approach (Duda and Hart, 1973).

3. Maximum Likelihood Estimates

Suppose now, that we are given a set $X = \{x_1,..., x_n\}$ of n unlabeled samples drawn independently from the mixture density (1). The likelihood of the observed samples is by definition the joint density (Edwards, 1972):

$$p(X|\delta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\delta) \qquad (2)$$

The maximum likelihood estimate is that value of $\delta = (\mu, \Sigma)$ that maximizes $p(X/\delta)$. See Duda and Hart (1973) for an exhaustive treatment. Through a *Stochastic Gradient Descent Solution* and after an amount of manipulation the expressions for μ_j and Σ_j can be obtained from the following recursive expression (Traven, 1991)

$$\theta_{k+1} = \theta_k + \eta_{k+1}(\theta(\mathbf{x}_{k+1}) - \theta_k) \qquad (3)$$

where θ means either μ_j or Σ_j. This is a learning law where the new value is obtained by adding to the old value a fraction of the difference between the sample and the old value. The learning rate η_{k+1} is computed as follows,

$$\eta_{k+1} = \frac{p(w/\mathbf{x}_{k+1})}{\sum_{i=1}^{k+1} p(w/\mathbf{x}_i)} \qquad (4)$$

In our approach we replace the learning rate given in (4) by the following,

$$\eta_{k+1} = \beta \frac{2l_c}{(l_l + l_r)} \qquad (5)$$

where l_c is the common overlap length, and l_l, l_r are the corresponding lengths for left and right edge-segments under matching. The factor $2l_c/(l_l + l_r)$ defines the overlap rate. The "overlapping" is a concept introduced by Medioni and Nevatia (1985) as follows: "*two segments overlap if by sliding one of them in a direction parallel to the epipolar line, i.e. to a horizontal line, they would intersect*"; $\beta = (1+k)^{-1}$ and permits to the learning rate to shrink with the sample k. This kind of coefficient is commonly used in Self-Organizing Feature Maps (Kohonen, 1995). Finally the Probability Density Function for each class w_j is given by equation (6), so given a $\mathbf{x}$ input vector we can compute its associated probability to the class w_j through (6)

$$p(\mathbf{x}/w_j,\delta_j)=\frac{1}{4\pi^2|\Sigma_j|^{0.5}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_j)^t\Sigma_j^{-1}(\mathbf{x}-\mu_j)\right] \quad (6)$$

From the above we can infer that the probability is large when the embedded squared Mahalanobis distance $d_{Mj}=(\mathbf{x}-\mu_j)^t\Sigma_j^{-1}(\mathbf{x}-\mu_j)$ is small. Hence we use the Mahalanobis distance as the criterion to classify the $\mathbf{x}$ input vector as a true or false match.

As we are solely concerned with the true matches, we only compute the parameter vector associated to the class of true matches, i.e. $\delta_T = (\mu_T,\Sigma_T)$ and use the following criterion: "given a input vector $\mathbf{x}$, coming from a stereo image pair, we compute the squared Mahalanobis distance d_{MT}, for such vector using the last updated parameter vector δ_T; if d_{MT} is less than the radius R of the cloud (class) then the input vector $\mathbf{x}$ is classified as a true match and it is used to update δ_T when the corresponding stereo image pair is fully processed (the parameter vector δ_T is updated according to equation (3)), otherwise do nothing because the input vector is classified as a false match". R is the radius of the hyper-sphere in $\Re^4$ associated to the cloud where true matches cluster and it is fixed to 10 in our approach.

4. Experimental Validation

The objective is to prove the validity and performance of the method by varying environmental conditions in two ways: by using new images with different features (different objects) and by changing the illumination.
All images come from the same indoor environment, this is the robotics laboratory with north orientation. We have taken 30 stereo images with different external illumination conditions. Such stereo images provide 1647 pattern samples or pairs of edge segments. The 30 stereo images are grouped in four sets: 1) 8 stereo images captured with natural illumination in the morning (**SI1**) 2) 6 stereo images captured with natural illumination in the afternoon (**SI2**), 3) 6 stereo images captured with artificial illumination (**SI3**) and 4) 10 stereo images captured with both artificial and natural illumination (**SI4**). A representative pair of stereo images is shown for SI1 and SI2 in Figures 1(a-d) and 2(a-d) respectively.

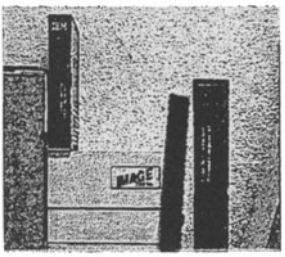

Figure 1a. SI1: Original Left Stereo Image

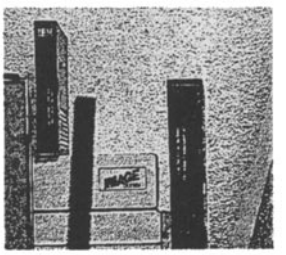

Figure 1b. SI1: Original Right Stereo Image

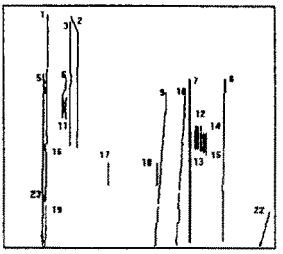

Figure 1c. SI1: Labeled Segments Left Image

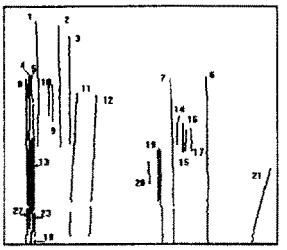

Figure 1d. SI1: Labeled Segments Right Image

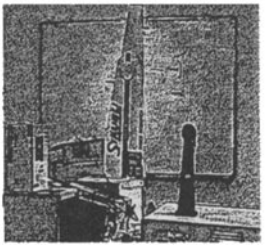

Figure 2a. SI3: Original Left Stereo Image

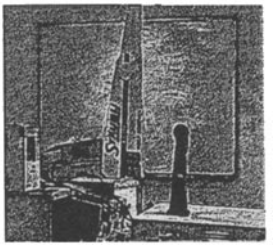

Figure 2b. SI3: Original Right Stereo Image

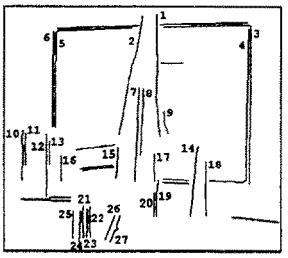

Figure 2c. SI3: Labeled Segments Left Image

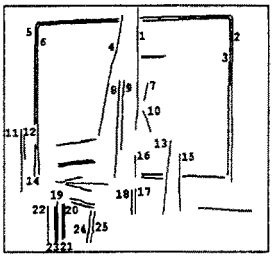

Figure 2d. SI3: Labeled Segments Right Image

Initially, the cluster center associated to the class of true matches μ_T is set to the null vector and the covariance matrix Σ_T, is the identity one. After each stereo image pair is processed the cluster center and the covariance matrix are both updated according to equation (3) by using the pairs of edge segments classified by the system as true matches.

Table 1 displays in second row the number of pattern samples corresponding to each set of stereo image pairs considered (first row), the third row shows the updated cluster center μ_T after the number of pattern samples in each column have been processed.

Table 1 Number of pattern samples used and cluster center computed for each set of Stereo Image pairs

	SI1	SI2	SI3	SI4
pattern samples	459	355	311	522
μ_T	{0.15, -0.01, 0.39, 0.37}	{0.20, -0.04, 0.35, 0.44}	{0.32, -0.09, 0.58, 0.80}	{0.44, -0.18, 0.79, 0.95}

The changes in the covariance matrix throughout the 4 sets are not statistically significant, in which case it suffices to give Σ_T for SI4 by example.

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

As mentioned before, the squared Mahalanobis distance d_{MT} is the metric used in our approach to select a pair of edge segments as a true or false match. To show the validity of our approach we choose the squared Euclidean distance d_{ET} as the similarity criterion. So, we compare the squared Mahalanobis distance against the squared Euclidean one. Figure 3 shows the percentage of successes obtained by using d_{MT} and d_{ET} for each set of stereo image pairs.

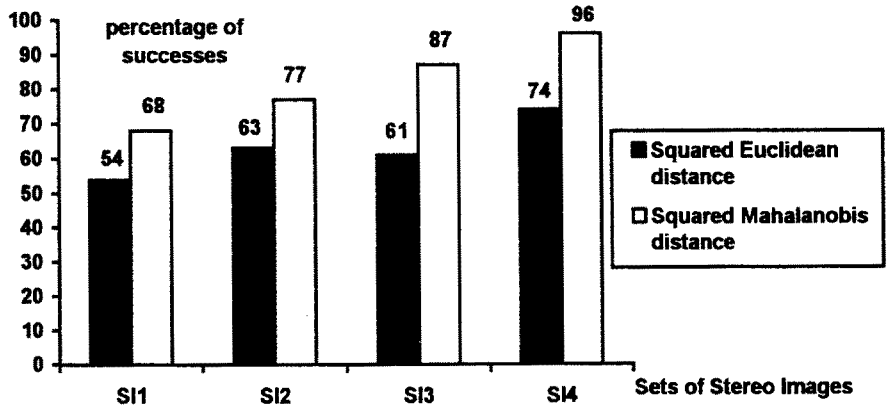

Figure 3. Average percentage of successes for each set of stereo images

5. Conclusion

From Figure 3 we can see that the best results in percentage of successes are always obtained by using the squared Mahalanobis distance. Also the percentage of successes increases as the number of pattern samples increases. Therefore, the stereovision matching is improved by using a pattern classification approach based on maximum likelihood estimates. This is because a learning process is involved. Nevertheless such behaviour is not affected by the nature of the different objects nor by illumination conditions.

References

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

Edwards, A.W.F. (1972). *Likelihood*. Cambridge: University Press.

Huertas, A. and G. Medioni (1986). Detection of Intensity Changes with Subpixel Accuracy Using Laplacian-Gaussian Masks. *IEEE Trans. Pattern Analysis Mach. Intell.* 8(5), 651-664.

Kim, D.H. and J.K. Aggarwal (1987) Positioning Three-Dimensional Objects using Stereo Images," *IEEE Journal of Robotics and Automation*. 3, 361-373.

Kohonen T. (1995). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Krotkov, E.P. (1989). *Active Computer Vision by Cooperative Focus and Stereo*, Springer-Verlag, New York.

Leu, J.G. and H.L. Yau (1991). Detecting the Dislocations in Metal Crystals from Microscopic Images, *Pattern Recognition*, 24(1), 41-56.

Medioni, G & Nevatia, R. (1985). Segment Based Stereo Matching. *Computer Vision, Graphics, and Image Processing*, **31**, 2-18.

Mousavi, M.S. and R.J. Schalkoff (1994). ANN Implementation of stereo vision using a multi-layer feedback architecture. *IEEE Trans. Sys. Man Cybern.* 24(8), 1220-1238.

Nevatia, R. and K.R. Babu (1980). Linear Feature Extraction and Description. *Computer Vision Graphics Image Processing* 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*, PhD thesis, Dpto. Informática y Automática, Facultad Ciencias UNED: Madrid.

Tanaka, S. and A.C. Kak (1990). A Rule-Based Approach to Binocular stereopsis. In: R.C. Jain and A.K. Jain, Eds., *Analysis and interpretation of range images*, Springer-Verlag, Berlin, 33-139.

Traven, H.G.C. (1991). A neural Network Approach to Statistical Pattern Classification by Semiparametric Estimation of Probability Density Functions. *IEEE Transactions Neural Networks*, **2**, 366-377.