FACE AUTHENTICATION BY USING ELASTIC GRAPH MATCHING AND SUPPORT VECTOR MACHINES

Anastasios Tefas, Constantine Kotropoulos and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 540 06, GREECE, {tefas,costas,pitas}@zeus.csd.auth.gr

ABSTRACT

In this paper, a novel method for enhancing the performance of elastic graph matching in face authentication is proposed. The starting point is to weigh the local matching errors at the nodes of an elastic graph according to their discriminatory power. We propose a novel approach to discriminant analysis that re-formulates Fisher's Linear Discriminant ratio to a quadratic optimization problem subject to inequality constraints by combining statistical pattern recognition and support vector machines. The method is applied to frontal face authentication on the M2VTS database.

1. INTRODUCTION

Automated face recognition has exhibited a tremendous growth for more than two decades. Many techniques for face recognition have been developed whose principles span several disciplines, such as image processing, pattern recognition, computer vision and neural networks [1]. The increasing interest in face recognition is mainly driven by application demands, such as nonintrusive identification and verification for credit cards and automatic teller machine transactions, nonintrusive access-control to buildings, identification for law enforcement, etc.

A well-known approach to face recognition and authentication is the so-called dynamic link architecture (DLA), a general object recognition technique, that represents an object by projecting its image onto a rectangular elastic grid where a Gabor wavelet bank response is measured at each node [2]. Recently, a variant of dynamic link architecture based on multiscale dilation-erosion, the so-called *morphological dynamic link architecture* (MDLA), was proposed and tested for face authentication [3].

This paper addresses the derivation of optimal coefficients that weigh the local matching errors determined at each grid node by the elastic graph matching procedure. We propose to weigh the local matching errors at the grid nodes by a novel approach that combines statistical pattern recognition (i.e., discriminant analysis) [4] and Support Vector Machines [5]. Our approach re-formulates Fisher's Linear Discriminant ratio to a quadratic optimization problem subject to inequality constraints. Linear Support Vector Machines are then constructed to yield the optimal separating hyperplanes.

2. PROBLEM STATEMENT

A widely known face recognition algorithm is the elastic graph matching [2]. The method is based on the analysis of a facial image region and its representation by a set of local descriptors extracted at the nodes of a sparse grid (i.e., a feature vector):

$$\mathbf{j}(\mathbf{x}) = \left(\hat{f}_1(\mathbf{x}), \ldots, \hat{f}_M(\mathbf{x})\right) \tag{1}$$

where $\hat{f}_i(\mathbf{x})$ denotes the output of a local operator applied to image f at the i-th scale or at the i-th pair (scale, orientation), $\mathbf{x}$ defines the pixel coordinates and M is feature vector dimensionality. The grid nodes are either evenly distributed over a rectangular image region or they are placed on certain facial features (e.g., nose, eyes, etc.) called fiducial points. In both cases a face/facial feature detection algorithm is needed.

Let the superscripts t and r denote a test and a reference person (or grid), respectively. The L_2 norm between the feature vectors at the l-th grid node is used as a (signal) similarity measure, i.e., $C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r)) = \|\mathbf{j}(\mathbf{x}_l^t) - \mathbf{j}(\mathbf{x}_l^r)\|$. The objective in elastic graph matching is to find the set of test grid node coordinates $\{\mathbf{x}_l^t,\ l \in \mathcal{V}\}$ that minimizes the cost function:

$$D(t,r) = \sum_{l \in \mathcal{V}} C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r))$$
$$\text{subject to } \mathbf{x}_l^t = \mathbf{x}_l^r + \mathbf{s} + \boldsymbol{\delta}_l, \quad \|\boldsymbol{\delta}_l\| \leq \delta_{\max} \tag{2}$$

where $\mathbf{s}$ denotes a global translation of the graph, $\boldsymbol{\delta}_l$ is a local perturbation and $\delta_{\max}$ controls the rigidity/plasticity of the graph.

Let $\mathbf{c}_t \in I\!R^L$ be a column vector comprised by the matching errors between a test person t and a reference person r at all grid nodes, i.e.:

$$\mathbf{c}_t = \begin{bmatrix} C_v(\mathbf{j}(\mathbf{x}_1^t), \mathbf{j}(\mathbf{x}_1^r)) \\ C_v(\mathbf{j}(\mathbf{x}_2^t), \mathbf{j}(\mathbf{x}_2^r)) \\ \vdots \\ C_v(\mathbf{j}(\mathbf{x}_L^t), \mathbf{j}(\mathbf{x}_L^r)) \end{bmatrix} \tag{3}$$

where L is the cardinality of $\mathcal{V}$. Hereafter, $\mathbf{c}_t$ is referred as the matching vector between the test person t and the reference person r. Using matrix notation, (2) is rewritten as

$$D(t,r) = \mathbf{1}^T \mathbf{c}_t, \tag{4}$$

where $\mathbf{1}$ is an $L \times 1$ vector of ones. That is, the classical elastic graph matching treats uniformly all local matching errors $C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r))$. We would like to weigh the local matching errors, i.e., to compute a weighted distance measure:

$$D'(t,r) = \mathbf{w}_r^T \mathbf{c}_t \tag{5}$$

where $\mathbf{w}_r$ is an appropriate vector of coefficients. Let us denote by $\mathcal{S}_r$ the class of matching vectors that belong to the reference person. Let also $\mathcal{S}$ denote the set of matching errors of the training set. Throughout the paper we study a two-class problem, namely, to separate efficiently all matching vectors that are attributed to a client (i.e., the reference person r) from the matching vectors that belong to anybody else (i.e., the class of $\mathbf{c}_t \in (\mathcal{S} - \mathcal{S}_r)$, which constitutes the set of impostors for client r).

3. CONSTRAINED LEAST SQUARES OPTIMIZATION

Let $\hat{\mathbf{m}}_C$ and $\hat{\mathbf{m}}_I$ denote the class sample mean of the matching vectors $\mathbf{c}_t$ that correspond to client claims, and of those that correspond to impostor claims related to the reference person r, respectively. Let also N_C and N_I be the corresponding numbers of matching vectors that belong to these two classes. Obviously, the total number of matching vectors N is equal to $N_C + N_I$. Let $\mathbf{S}_W$ and $\mathbf{S}_B$ be within-class and between-class scatter matrices, respectively.

Let us suppose that we would like to linearly transform the matching vector (e.g., to apply (5)). Four feature selection criteria are studied in detail in [4]. The most known criterion is to choose $\mathbf{w}_r$ so that the ratio of the trace of the between-class scatter matrix and the trace of the within-class scatter matrix of the transformed matching vectors is maximized. Since, in our case, the transformed matching vector is merely the scalar $\mathbf{w}_r^T \mathbf{c}_t$, the optimization criterion is simplified to the ratio of between-class and within-class variances, i.e.:

$$J(\mathbf{w}_r) = \frac{\mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_r}{\mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r}. \tag{6}$$

This is the so-called Fisher's discriminant ratio. The coefficient vector $\mathbf{w}_{r,o}$ that maximizes (6) is given by:

$$\mathbf{w}_{r,o} = \mathbf{S}_W^{-1} (\mathbf{m}_I - \mathbf{m}_C) \tag{7}$$

and yields Fisher's linear discriminant $\mathbf{w}_{r,o}^T \mathbf{c}_t$. It is straightforward to prove that the minimization of:

$$J'(\mathbf{w}_r) = \mathbf{w}_r^T (\mathbf{S}_W + \mathbf{S}_B) \mathbf{w}_r \tag{8}$$

subject to the equality constraint:

$$\mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_r = \zeta = \text{const}, \quad \zeta > 0 \tag{9}$$

yields the coefficient vector:

$$\mathbf{w}_r' = \kappa \, \mathbf{S}_W^{-1} (\mathbf{m}_I - \mathbf{m}_C) \tag{10}$$

where κ is a proporptionality constant given by:

$$\kappa = \sqrt{\frac{\zeta}{\hat{P}_C \hat{P}_I}} \frac{1}{(\hat{\mathbf{m}}_I - \hat{\mathbf{m}}_C)^T \mathbf{S}_W^{-1} (\hat{\mathbf{m}}_I - \hat{\mathbf{m}}_C)}. \tag{11}$$

It is seen that the coefficient vector given by (10), which is optimal with respect to the criterion (8)-(9), is still in the direction of the coefficient vector that minimizes Fisher's discriminant ratio. The nice property of the optimality criterion (8) is that it rewrites Fisher's discriminant ratio as a quadratic optimization criterion subject to an equality constraint (e.g., a constraint least-squares (CLS) criterion), thus enabling the use of Lagrange multipliers which is a more straightforward optimization procedure than the solution of a generalized eigenvalue problem. However, the equality constraint (9) seems to be too restrictive. We shall modify the objective and the constraint functions as follows:

$$\text{minimize} \qquad \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r \tag{12}$$

$$\text{subject to} \qquad \mathbf{w}_r^T (\mathbf{m}_I - \mathbf{m}_C) \geq \mathbf{1}^T (\mathbf{m}_I - \mathbf{m}_C). \tag{13}$$

The new criterion minimizes the within-class variance, while the difference between the class centers (i.e., the average distance measure over client claims $E\{D'(t,r) \mid \mathbf{c}_t \in \mathcal{S}_r\}$ and the average distance measure over impostor claims $E\{D'(t,r) \mid \mathbf{c}_t \in (\mathcal{S} - \mathcal{S}_r)\}$) is not reduced after linear weighting. Therefore, the interpretation of (12)-(13) agrees with that of FLD ratio. It is straightforward to show that the inequality constraint (13) can be rewritten as

$$\sum_{t=1}^{N} k_t (\mathbf{w}_r^T - \mathbf{1}^T) \mathbf{c}_t \geq 0, \quad k_t = \begin{cases} -N_I, & \mathbf{c}_t \in \mathcal{S}_r \\ N_C, & \mathbf{c}_t \in (\mathcal{S} - \mathcal{S}_r). \end{cases} \tag{14}$$

The inequality constraint (14) can be combined with the quadratic objective function (12) to yield a linearly constrained least squares problem that can be solved by constrained quadratic optimization methods [6]. The Lagrangian function to be minimized is:

$$L_p(\mathbf{w}_r, \alpha) = \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r - \alpha \sum_{t=1}^{N} k_t (\mathbf{w}_r^T - \mathbf{1}^T) \mathbf{c}_t \tag{15}$$

where α is the Lagrange multiplier. The stationary point $(\mathbf{w}_{r,o}, \alpha_o)$ of (15) can be easily calculated by maximizing the Wolfe dual objective function. The latter function is maximized for α_o given by:

$$\alpha_o = \frac{2 \sum_{t=1}^{N} k_t \mathbf{1}^T \mathbf{c}_t}{\sum_{t=1}^{N} \sum_{j=1}^{N} k_t k_j \mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}_j} \tag{16}$$

provided that the denominator in (16) is positive. The numerator in (16) is always non-negative by construction (i.e., the average distance measure over client claims is always less than the average distance measure over impostor claims). The optimal coefficient vector for the criterion (12,14) is given by:

$$\mathbf{w}_{r,o} = \frac{1}{2} \alpha_o \mathbf{S}_W^{-1} \sum_{t=1}^{N} k_t \mathbf{c}_t. \tag{17}$$

It is obvious that, except for the scaling factor $\frac{1}{2}\alpha_o$, the direction of $\mathbf{w}_{r,o}$ given by (17) coincides with that of (7), which maximizes Fisher's discriminant ratio, as well as with that of (10), which maximizes the objective criterion (8) and (9).

4. SUPPORT VECTOR MACHINE FORMULATION

Support Vector Machines (SVMs) is a state-of-the-art pattern recognition technique whose foundations are stemming from statistical learning theory [5]. SVM is a learning machine capable of implementing a set of functions that approximate best the supervisor's response with an expected risk bounded by the sum of the empirical risk and the Vapnik-Chervonenkis (VC) confidence, a bound on the generalization ability of the learning machine, that depends on the so-called VC dimension of the set of functions implemented by the machine. Motivated by the fact that SVM training algorithm consists of a quadratic programming problem, we shall reformulate the criterion of minimizing the within-class variance so that it can be solved by constructing the optimal separating hyperplane (linear SVM).

Suppose the training data:

$$(\mathbf{c}_1, y_1), \ldots, (\mathbf{c}_N, y_N), \quad \mathbf{c}_t \in I\!R^L, \qquad (18)$$
$$y_t = \begin{cases} 1 & \text{if } \mathbf{c}_t \in (\mathcal{S} - \mathcal{S}_r) \\ -1 & \text{if } \mathbf{c}_t \in \mathcal{S}_r \end{cases}$$

can be separated by a hyperplane:

$$g_{\mathbf{w}_r,b}(\mathbf{c}_t) = \mathbf{w}_r^T\, \mathbf{c}_t - b = 0 \qquad (19)$$

with the property:

$$\begin{aligned} g_{\mathbf{w}_r,b}(\mathbf{c}_t) &\geq 1 \qquad \text{if } y_t = 1 \\ g_{\mathbf{w}_r,b}(\mathbf{c}_t) &\leq -1 \qquad \text{if } y_t = -1 \end{aligned} \qquad (20)$$

where b is a bias term. In compact notation, the set of inequalities (20) can be rewritten as:

$$y_t(\mathbf{w}_r^T \mathbf{c}_t - b) - 1 \geq 0 \quad t = 1, \ldots, N. \qquad (21)$$

Let us define the distance $v(\mathbf{w}_r, b; \mathbf{c}_t)$ of a matching vector $\mathbf{c}_t$ from the hyperplane (19) as:

$$v(\mathbf{w}_r, b; \mathbf{c}_t) = \frac{\mid \mathbf{w}_r^T\, \mathbf{c}_t - b \mid}{\|\mathbf{w}_r\|_{\mathbf{S}_W}} = \frac{\mid \mathbf{w}_r^T\, \mathbf{c}_t - b \mid}{(\mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r)^{1/2}} \qquad (22)$$

where the norm of the coefficient vector $\mathbf{w}_r$ is measured with respect to the within-scatter matrix $\mathbf{S}_W$. In our case, the optimal hyperplane is given by maximizing the margin:

$$\varrho(\mathbf{w}_r, b) = \min_{\mathbf{c}_t \in (\mathcal{S} - \mathcal{S}_r)} v(\mathbf{w}_r, b; \mathbf{c}_t) + \qquad (23)$$
$$+ \min_{\mathbf{c}_t \in \mathcal{S}_r} v(\mathbf{w}_r, b; \mathbf{c}_t) = \frac{2}{(\mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r)^{1/2}}.$$

Equivalently, the optimal hyperplane separates the data so that the within-class variance, i.e., the objective function (12), is minimized. The optimization is subject to the constraint functions (21). By comparing (13) with (21), we observe that more than one inequality constraints are now imposed that demand the distance measures $D'(t, r)$ related to impostor claims to be linearly separable from the distance measures $D'(t, r)$ related to client claims on the training set. For completeness, we mention that the standard SVM would solve the problem [5]:

$$\text{minimize } J_{\text{SVM}}(\mathbf{w}_r) = \mathbf{w}_r^T\, \mathbf{w}_r \text{ subject to (21).} \qquad (24)$$

The solution of the optimization problem under study is given by the saddle point of the Lagrangian:

$$L(\mathbf{w}_r, b, \boldsymbol{\alpha}) = \mathbf{w}_r^T \mathbf{S}_W\, \mathbf{w}_r - \sum_{t=1}^{N} \alpha_t \{y_t(\mathbf{w}_r^T \mathbf{c}_t - b) - 1\} \qquad (25)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T$ is the vector of Lagrange multipliers. The Lagrangian has to be minimized with respect to $\mathbf{w}_r$ and b and maximized with respect to $\alpha_t > 0$. The Kuhn-Tucker (KT) conditions [6] imply that:

$$\begin{aligned} &\nabla_{\mathbf{w}_r} L(\mathbf{w}_{r,o}, b_o, \boldsymbol{\alpha}_o) = \mathbf{0} \Leftrightarrow \mathbf{w}_{r,o} = \frac{1}{2}\mathbf{S}_W^{-1} \sum_{t=1}^{N} \alpha_{t,o} y_t \mathbf{c}_t \\ &\frac{\partial}{\partial b} L(\mathbf{w}_{r,o}, b_o, \boldsymbol{\alpha}_o) = 0 \Leftrightarrow \sum_{t=1}^{N} \alpha_{t,o} y_t = 0 \\ &y_t\, (\mathbf{w}_{r,o}^T \mathbf{c}_t - b_o) - 1 \geq 0 \quad t = 1, \ldots, N \\ &\alpha_{t,o} \geq 0 \quad t = 1, \ldots, N \\ &\alpha_{t,o}\, \{y_t\, (\mathbf{w}_{r,o}^T \mathbf{c}_t - b_o) - 1\} = 0 \quad t = 1, \ldots, N. \end{aligned} \qquad (26)$$

From the conditions (26), one can see that the weighting vector we search for is the linear combination of the matching vectors having nonzero Lagrange multipliers α_t. These matching vectors are the *support vectors* [5]. Putting the expression for $\mathbf{w}_{r,o}$ into the Lagrangian (25) and taking into account the KT conditions, we obtain the Wolf dual functional:

$$\mathcal{W}(\boldsymbol{\alpha}) = \sum_{t=1}^{N} \alpha_t - \frac{1}{4} \sum_{t=1}^{N} \sum_{j=1}^{N} \alpha_t \alpha_j \underbrace{y_t y_j \left(\mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}_j\right)}_{\mathbf{H}_{tj}} \qquad (27)$$

where $\mathbf{H}_{tj}$ is the ij-th element of the Hessian matrix $\mathbf{H}$. The maximization of (27) in the non-negative quadrant of α_t, i.e.:

$$\alpha_t \geq 0 \quad t = 1, \ldots, N \qquad (28)$$

under the constraint:

$$\sum_{t=1}^{N} \alpha_t y_t = 0 \qquad (29)$$

is equivalent to the optimization problem:

$$\text{minimize } \frac{1}{4} \boldsymbol{\alpha}_o^T\, \mathbf{H}\, \boldsymbol{\alpha}_o - \mathbf{1}^T\, \boldsymbol{\alpha}_o \text{ subject to (28) and (29).} \qquad (30)$$

Having found the non-zero Lagrange multipliers $\alpha_{t,o}$, the optimal separating hyperplane is given by:

$$g(\mathbf{c}) = \text{sgn}\left(\frac{1}{2} \sum_{\alpha_{t,o} > 0} y_t \alpha_{t,o} (\mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}) - b_o\right) \qquad (31)$$

where $b_o = \frac{1}{2} \mathbf{w}_{r,o}^T (\mathbf{c}_p + \mathbf{c}_q)$ for any pair of support vectors $\mathbf{c}_p$ and $\mathbf{c}_q$, such that $y_p = 1$ and $y_q = -1$. The weighted distance measure is given by (5). The extension of the proposed method to deal with matching errors that are not linearly separable as well as with nonlinear decision surfaces can be done following a similar approach [7].

5. EXPERIMENTAL RESULTS

The optimal coefficient vectors derived by the procedures described in Sections 3 and 4 have been used to weigh the raw matching vectors $\mathbf{c}$ that are provided by the morphological dynamic link architecture [3], a variant of elastic graph matching, applied to frontal face authentication. Let us call the combination of the CLS/SVM weighting approach and the morphological dynamic link architecture weighted MDLA. The weighted MDLA has been tested on the database of the European research project *Multi-Modal Verification for Telecommunication Services* (M2VTS). The database contains 37 persons' video data, which include speech consisting of uttering digits and image sequences of rotated heads. Four recordings (i.e., shots) of the 37 persons have been collected. Four experimental sessions have been implemented by employing the "leave-one-out" principle. Each experimental session consists of a training and a test procedure that are applied to their training set and test set, respectively. To apply the proposed methods additional client images are extracted from the database in order to have a larger set of intra-class distances for each client class. Moreover, additional client images are extracted in order to prevent overfitting during the training caused by the lack of data.

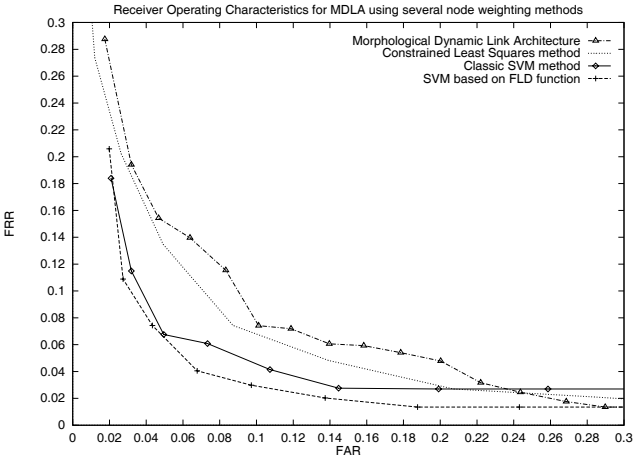

Figure 1: Receiver Operating Characteristics for MDLA for several discriminatory power coefficients.

For comparison reasons we have also weighted the raw matching vectors by the coefficient vector determined by the standard SVM algorithm for pattern recognition (24). By using the constrained least squares solution described in Section 3, we achieved an *Equal Error Rate* (EER) of 8.2%. That is, a drop of 1% from the original MDLA. Further improvements (i.e., an EER equal to 6.4 %) were obtained when the coefficient vector derived by the standard SVM was used to weigh the raw matching vectors. The best authentication performance was obtained when the proposed linear support vector machine that minimizes (25) was applied. In this case, we achieved an EER of 5.6%.

In Table 1, a performance comparison between several face authentication algorithms developed within the M2VTS research project is reported. It is clearly seen that the weighted MDLA algorithm attains the best performance. It is worth mentioning that all methods were tested on the same database according to the same protocol. The *Receiver Operating Characteristics* (ROC) curves of MDLA for each weighting algorithm are depicted in Figure 1. In the same Figure, the ROC curve for the original MDLA is also plotted for comparison reasons. We can see that the area under the ROC for the proposed methods is much smaller than the initial one.

Table 1: Comparison of equal error rates for several authentication techniques in the M2VTS database.

Authentication Technique	EER (%)
MDLA with discriminating grids	**5.6**
MDLA	9.2
Gray level frontal face matching [8]	8.5
Discriminant GDLA [9]	6.0-9.2
GDLA [9]	10.8-14.4

6. CONCLUSIONS

Novel methods for incorporating discriminant analysis into the elastic graph matching algorithm have been proposed. They are based on statistical learning theory. Starting from Fisher's discriminant ratio, a constrained least squares optimization problem was set up and solved. The constrained least squares problem was further extended to a problem that can solved by the construction of a Support Vector Machine.

7. REFERENCES

[1] R. Chellapa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[2] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R.P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. on Computers*, vol. 42, no. 3, pp. 300–311, March 1993.

[3] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using morphological elastic graph matching," *IEEE Transactions on Image Processing*, accepted for publication August 1999.

[4] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall International, London, 1982.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 1995.

[6] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley, New York, 1987.

[7] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance elastic graph matching for face authentication," submitted June 1999.

[8] S. Pigeon and L. Vandendorpe, "Image-based multimodal face authentication," *Signal Processing*, vol. 69, pp. 59–79, August 1998.

[9] B. Duc, S. Fischer, and J. Bigün, "Face authentication with gabor information on deformable graphs," *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 504–516, 1999.