Using Support Vector Machines to enhance the performance of elastic graph matching for frontal face authentication

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

A novel method for enhancing the performance of elastic graph matching in frontal face authentication is proposed. The starting point is to weigh the local similarity values at the nodes of an elastic graph according to their discriminatory power. Powerful and well-established optimization techniques are used to derive the weights of the linear combination. More specifically, we propose a novel approach that reformulates Fisher's discriminant ratio to a quadratic optimization problem subject to a set of inequality constraints by combining statistical pattern recognition and Support Vector Machines. Both linear and nonlinear Support Vector Machines are then constructed to yield the optimal separating hyperplanes and the optimal polynomial decision surfaces, respectively. The method has been applied to frontal face authentication on the M2VTS database. Experimental results indicate that the performance of morphological elastic graph matching, is highly improved by using the proposed weighting technique.

Index terms - Face authentication, Elastic Graph Matching, Fisher's Discriminant ratio, Constrained least squares optimization, Support vector machines

I. Introduction

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. Machine recognition of faces yields problems that belong to the following categories whose objectives are briefly outlined:

- **Face recognition.** Given a test face and a set of reference faces in a database find the N most similar reference faces to the test face.
- **Face authentication.** Given a test face and a reference one, decide if the test face is identical to the reference face.

Face recognition has been studied more extensively than face authentication. The two problems are conceptually different. On the one hand, a face recognition system usually assists a human expert to determine the identity of a test face by computing all similarity scores between the test face and each human face stored in the system database and by ranking them. On the other hand, a face authentication system should decide itself if a test face is assigned to a *client* (i.e., one who claims his/her own identity) or to an *impostor* (i.e., one who pretends to be someone else). The evaluation criteria for face recognition and face authentication systems are different. The performance of face recognition systems is quantified in terms of the percentage of correctly identified faces within the N best matches [2]. The performance of face authentication systems is measured in terms of the *false rejection rate* (FRR) achieved at a fixed *false acceptance rate* (FAR) or vice versa. By varying FAR, the *receiver operating characteristic* (ROC) curve is obtained. If a scalar figure of merit is used to judge the performance of an authentication algorithm, then we usually choose the operating point having FAR=FRR, the so called *equal error rate* (EER). A third difference is in the requirements needed when face recognition/authentication systems are trained. Face authentication systems need more than one training images per person while face recognition systems are usually trained on sets having one frontal image per person.

A well-known approach to face recognition/authentication is the so-called dynamic link architecture, 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. A simplified implementation of dynamic link architecture, the so-called elastic graph matching (EGM), is often preferred for locating objects in a scene with a known reference [3,4,5]. It was found that elastic graph matching achieves a better performance than the eigenfaces [6,7], auto-association and classification neural

networks [8,9] due to its robustness to lighting, varying face position and facial expression variations. The research on elastic graph matching and its applications has been an active research topic since its invention. A different topology cost for a particular pair of nodes, that was based on the radius of the Apollonius sphere defined by the Euclidean distances between the nodes being matched, was proposed in [10]. Three major extensions to elastic graph matching that allowed for handling larger galleries, tolerated larger variations in pose and increased its matching accuracy were introduced in [11]. Procedures that increase the robustness of elastic graph matching in translations, deformations and changes in background were proposed in [12]. Recently, a variant of elastic graph matching based on multiscale dilation-erosion, the so-called *morphological elastic graph matching* (MEGM)[1], was proposed and tested for face authentication [13,14].

This paper addresses the derivation of optimal coefficients that weigh the local similarity values determined by the elastic graph matching procedure at each grid node. The starting point is the observation that the distance measure between two persons in elastic graph matching is simply the sum of local similarity values. However, this observation contradicts the fact that some facial features (e.g., the eyes, the nose, etc.) are more distinctive for one person than other features. Therefore, it is reasonable to assume that the response of local image descriptors attempting to capture the image properties in the region of these features should be weighted heavier than the other responses. To alleviate the just described weakness of the elastic graph matching, we propose to weigh the local similarity values at the grid nodes by a novel approach that combines statistical pattern recognition (i.e., discriminant analysis) [15,16] and Support Vector Machines [17,18,19]. Our approach reformulates Fisher's discriminant ratio to a quadratic optimization problem subject to a set of inequality constraints. Both linear and nonlinear Support Vector

[1]In this paper, the term morphological elastic graph matching is used as synonymous to morphological dynamic link architecture.

Machines are then constructed to yield the optimal separating hyperplanes and the optimal polynomial decision surfaces, respectively. The proposed method has been applied to frontal face authentication on the M2VTS database [20]. Experimental results indicate that the performance of morphological elastic graph matching, a variant of elastic graph matching, is highly improved by using the proposed weighting technique reaching an EER of 2.4 %. It is worth noting that the objective of our approach differs significantly from that of:

- *Linear Discriminant Analysis (LDA) to face recognition* [21,22,23], because there Fisher Discriminant Analysis is applied to the entire face image as a linear projection algorithm, i.e., a parametric projection pursuit, or
- *LDA for feature selection* [24,13], because there Fisher Discriminant Analysis is applied to feature vectors for selecting the most discriminating features, that is, to reduce the dimensionality of feature vectors by throwing away the local image descriptor responses that are not significant to the authentication task.

On the contrary, our objective is closely related to the Bayesian approach that yields the most reliable nodes for gender identification, beard and glass detection in bunch graphs [4,11] or the algorithm for learning the weights in discrimination functions using a priori constraints [25]. The latter algorithm uses the Simplex method to find the local minima of a multidimensional function in order to optimize the weights on the local similarity values.

The outline of the paper is as follows. A brief description of elastic graph matching and the problem treated throughout the paper is given in Section II. Local discriminatory power coefficients are derived in Section III through the formulation of the problem as a constrained Least Squares optimization problem. The extension of constrained Least Squares formulation to the construction of a class of Support Vector Machines (both linear

and nonlinear ones) is described in Section IV. The weights derived by the approaches described in Sections III and IV are incorporated to the morphological elastic graph matching and the combined scheme is tested for frontal face authentication. The performance of the weighted morphological elastic graph matching is assessed in Section V and conclusions are drawn in Section VI.

II. Problem Statement

A widely known face recognition algorithm is the elastic graph matching [3,11]. 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. 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, the elastic graph matching algorithm consists of the following steps:

Step 1: Build an information pyramid by using scale-space image analysis techniques. For example, the responses of a set of 2D Gabor filters tuned to different orientations and scales [3,4,5,11,24] or the output of multiscale morphological dilation-erosion at several scales can be employed to form a local descriptor (i.e., a feature vector) [13]:

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

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.

Step 2: Detect the face or the fiducial points on the reference image and place the grid nodes over the facial image region or the fiducial points, respectively. Many face detection algorithms were proposed in the literature. In this paper, we mainly resort to a variant of the approach proposed by Yang and Huang [26] that is based on multiresolution images

(the so-called mosaic images) [27]. The method that represents the face as a labeled graph whose nodes are placed at different fiducial points is more difficult to be applied automatically, since the detection module has to find the precise coordinates of facial features. The detection of a rectangular facial region that encloses the face is generally an easier task.

Step 3: Translate and deform an elastic graph comprised of local descriptors measured at variable pixel coordinates on the test image so that a cost function is minimized. The cost function is based on both the norm of the difference between the reference and test local descriptors and the distortion between the reference grid and the variable test graph. 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., $C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r)) = \|\mathbf{j}(\mathbf{x}_l^t) - \mathbf{j}(\mathbf{x}_l^r)\|$. Let us define by $\mathcal{V}$ the set of grid nodes that are considered as graph vertices. Let also $\mathcal{N}(l)$ denote the four-connected neighborhood of vertex l. 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:

$$C(\{\mathbf{x}_l^t\}) = \sum_{l \in \mathcal{V}} \left\{ C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r)) + \lambda \sum_{\xi \in \mathcal{N}(l)} C_e(l, \xi) \right\} \tag{2}$$

where $C_e(l, \xi)$ penalizes grid deformations, i.e., $C_e(l, \xi) = \|(\mathbf{x}_l^t - \mathbf{x}_l^r) - (\mathbf{x}_\xi^t - \mathbf{x}_\xi^r)\|$, $\xi \in \mathcal{N}(l)$. Obviously, the cost function (2) defines a distance measure between two persons. One may interpret the optimization of (2) as a simulated annealing with additional penalties imposed by the grid deformations. Accordingly, (2) can be simplified to

$$D(t, r) = \sum_{l \in \mathcal{V}} C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r)) \ \text{ subject to } \ \mathbf{x}_l^t = \mathbf{x}_l^r + \mathbf{s} + \boldsymbol{\delta}_l, \ \ \|\boldsymbol{\delta}_l\| \leq \delta_{\max} \tag{3}$$

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

In this paper, we rely on the morphological elastic graph matching [13,14]. However, the developed methods are applicable to any elastic graph matching algorithm, e.g., [3,4,5,11]. Figure 1 demonstrates the image analysis step employed in morphological elastic graph matching. The grids formed in the matching procedure of a test person with himself and with another person, for a pair of persons extracted from M2VTS database, are depicted in Figure 2. It is worth noting that each of the aforementioned steps affects significantly the performance of elastic graph matching in both the computational complexity and the authentication efficiency.

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

$$\mathbf{c}_t = \begin{bmatrix} C_v(\mathbf{j}(\mathbf{x}_1^t), \mathbf{j}(\mathbf{x}_1^r)) \\ C_v(\mathbf{j}(\mathbf{x}_2^t), \mathbf{j}(\mathbf{x}_2^r)) \\ \vdots \\ C_v(\mathbf{j}(\mathbf{x}_L^t), \mathbf{j}(\mathbf{x}_L^r)) \end{bmatrix} \tag{4}$$

where L is the cardinality of $\mathcal{V}$. Hereafter, $\mathbf{c}_t$ is referred as the *similarity vector* between the test person t and the reference person r. Using matrix notation, (3) is rewritten as

$$D(t,r) = \mathbf{1}^T \mathbf{c}_t, \tag{5}$$

where $\mathbf{1}$ is an $L \times 1$ vector of ones. That is, the classical elastic graph matching treats uniformly all local similarity values $C_v(\mathbf{j}(\mathbf{x}_l^t), \mathbf{j}(\mathbf{x}_l^r))$. Each of them can be considered as the noisy output of the local similarity detector due to either localization errors inherent in the face detection module or errors attributed to the limited number of iterations allowed in the simulated annealing during the elastic graph matching procedure. Moreover, it is well known that some facial features (e.g., the eyes, the nose, etc.) are more crucial in the authentication procedure of a certain person than other features. Therefore, it sounds reasonable to weigh the local similarity vector $\mathbf{c}_t$ by a coefficient vector $\mathbf{w}_r$ that

quantifies its discriminatory power at each grid node, i.e.:

$$D'(t,r) = \mathbf{w}_r^T \mathbf{c}_t. \tag{6}$$

Let us denote by $\mathcal{S}_C$ the set that includes all similarity vectors between two images of the same reference person. Let also $\mathcal{S}_I$ denote the set that includes all similarity vectors between the reference person and any other person in the training set. Throughout the paper we study a two-class problem, namely, to separate efficiently all similarity vectors that are attributed to a client (i.e., the reference person r) from the similarity vectors that belong to anybody else (i.e., the class of $\mathbf{c}_t \in \mathcal{S}_I$, i.e., the set of impostors for client r). That is, we demand the weighting coefficient of the l-th grid node $w_r(l)$ for reference person r to quantify how well the within-class similarity vectors are separated from the similarity vectors that belong to the class of his/her impostors. The new distance measure $D'(t,r)$ corresponds to a projection of similarity vector $\mathbf{c}_t$ onto a line in the direction of $\mathbf{w}_r$. A constrained Least Squares formulation of this binary classification problem is treated in Section III and a construction of a novel class of Support Vector Machines is proposed in Section IV. Accordingly, the weighting schemes proposed in this paper can be considered as post-processing algorithms that are applied after elastic graph matching aiming at improving its verification capability.

III. Constrained Least Squares Optimization

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

matrices are defined by:

$$\mathbf{S}_W = \hat{P}_C \frac{1}{N_C} \sum_{\mathbf{c}_t \in \mathcal{S}_C} (\mathbf{c}_t - \hat{\mathbf{m}}_C)(\mathbf{c}_t - \hat{\mathbf{m}}_C)^T + \hat{P}_I \frac{1}{N_I} \sum_{\mathbf{c}_t \in \mathcal{S}_I} (\mathbf{c}_t - \hat{\mathbf{m}}_I)(\mathbf{c}_t - \hat{\mathbf{m}}_I)^T \quad (7)$$

$$\mathbf{S}_B = \hat{P}_C (\hat{\mathbf{m}}_C - \hat{\mathbf{m}})(\hat{\mathbf{m}}_C - \hat{\mathbf{m}})^T + \hat{P}_I (\hat{\mathbf{m}}_I - \hat{\mathbf{m}})(\hat{\mathbf{m}}_I - \hat{\mathbf{m}})^T$$

$$= \hat{P}_C \hat{P}_I (\hat{\mathbf{m}}_I - \hat{\mathbf{m}}_C)(\hat{\mathbf{m}}_I - \hat{\mathbf{m}}_C)^T, \quad (8)$$

respectively, where $\hat{P}_C$ and $\hat{P}_I$ are estimates of the a priori class probabilities. Let us suppose that we would like to linearly transform the similarity vector (e.g., (6)). Four feature selection criteria are studied in detail in [15]. 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 similarity vectors is maximized. Since in our case the transformed similarity vector is merely the scalar $\mathbf{w}_r^T \mathbf{c}_t$ (i.e., the weighted distance measure), the optimization criterion is simplified to the ratio of between-class and within-class variances, i.e.:

$$J(\mathbf{w}_r) = \frac{\mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_r}{\mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r}. \quad (9)$$

This is the so-called *Fisher's discriminant ratio*. The coefficient vector $\mathbf{w}_{r,o}$ that maximizes (9) is given by:

$$\mathbf{w}_{r,o} = \mathbf{S}_W^{-1} (\mathbf{m}_I - \mathbf{m}_C) \quad (10)$$

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

$$J'(\mathbf{w}_r) = \mathbf{w}_r^T (\mathbf{S}_W + \mathbf{S}_B) \mathbf{w}_r \quad (11)$$

subject to the equality constraint:

$$\mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_r = \zeta = \text{const}, \quad \zeta > 0 \quad (12)$$

yields the coefficient vector:

$$\mathbf{w}_r' = \kappa \, \mathbf{S}_W^{-1} (\mathbf{m}_I - \mathbf{m}_C) \quad (13)$$

where κ is a proportionality constant given by:

$$\kappa = \sqrt{\frac{\zeta}{\hat{P}_C \hat{P}_I}} \frac{1}{(\hat{\mathbf{m}}_I - \hat{\mathbf{m}}_C)^T \mathbf{S}_W^{-1} (\hat{\mathbf{m}}_I - \hat{\mathbf{m}}_C)}. \tag{14}$$

It is seen that the coefficient vector given by (13), which is optimal with respect to the criterion (11)-(12), is still in the direction of the coefficient vector that minimizes Fisher's discriminant ratio. The nice property of the optimality criterion (11) is that it rewrites Fisher's discriminant ratio as a quadratic optimization criterion subject to an equality constraint (e.g., a constraint least-squares 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 (12) seems to be too restrictive. We shall modify the objective and the constraint functions as follows:

$$\text{minimize} \qquad \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r \tag{15}$$

$$\text{subject to} \qquad \mathbf{w}_r^T (\mathbf{m}_I - \mathbf{m}_C) \geq \mathbf{1}^T (\mathbf{m}_I - \mathbf{m}_C). \tag{16}$$

The new optimization problem minimizes the within-class variance while the difference between class centers (i.e., the average distance measure over client claims $E\{D'(t,r) \mid \mathbf{c}_t \in \mathcal{S}_C\}$ and the average distance measure over impostor claims $E\{D'(t,r) \mid \mathbf{c}_t \in \mathcal{S}_I\}$) is not reduced after linear weighting, i.e.:

$$E\{D'(t,r) \mid \mathbf{c}_t \in \mathcal{S}_I\} - E\{D'(t,r) \mid \mathbf{c}_t \in \mathcal{S}_C\} \geq E\{D(t,r) \mid \mathbf{c}_t \in \mathcal{S}_I\} - E\{D(t,r) \mid \mathbf{c}_t \in \mathcal{S}_C\}, \tag{17}$$

where $D(t,r)$ and $D'(t,r)$ are the scalars given by (5) and (6), respectively. It can be seen that the objective function (15) is in par with the aims of Fisher's discriminant ratio. Moreover, (16) is directly related to the problem of face authentication using elastic graph matching. The inequality constraint (16) can be rewritten as:

$$\sum_{t=1}^{N} k_t (\mathbf{w}_r^T - \mathbf{1}^T) \mathbf{c}_t \geq 0 \tag{18}$$

where

$$k_t = \begin{cases} -N_I, & \mathbf{c}_t \in \mathcal{S}_C \\ N_C, & \mathbf{c}_t \in \mathcal{S}_I. \end{cases} \tag{19}$$

The inequality constraint (18) can be combined with the quadratic objective function (15) to yield a linearly constrained least squares problem that can be solved by constrained quadratic optimization methods [28]. The Lagrangian function to be minimized is:

$$L_p(\mathbf{w}_r, \alpha) = \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r - \alpha \sum_{t=1}^{N} k_t (\mathbf{w}_r^T - \mathbf{1}^T)\mathbf{c}_t \tag{20}$$

where α is the Lagrange multiplier. To find the stationary point ($\mathbf{w}_{r,o}$, α_o) of (20), we solve the set of equations:

$$\nabla_{\mathbf{w}_r} L_p(\mathbf{w}_{r,o}, \alpha_o) = \mathbf{0} \tag{21}$$

$$\frac{\partial}{\partial \alpha} L_p(\mathbf{w}_{r,o}, \alpha_o) = 0. \tag{22}$$

The first-order necessary conditions or *Kuhn-Tucker (KT) conditions* [28] imply that, if $\mathbf{w}_{r,o}$ is a local minimum of the problem (15) and (18), it should satisfy (21), under the regularity assumption that the intersection of the set of feasible directions with the set of descent directions coincides with the intersection of the set of feasible directions for linearized constraints with the set of descent directions, i.e.:

$$\mathbf{w}_{r,o} = \frac{1}{2}\alpha_o \mathbf{S}_W^{-1} \sum_{t=1}^{N} k_t \mathbf{c}_t \tag{23}$$

$$\text{subject to} \qquad \alpha_o \geq 0 \tag{24}$$

$$\alpha_o \sum_{t=1}^{N} k_t (\mathbf{w}_{r,o}^T - \mathbf{1}^T)\mathbf{c}_t = 0 \tag{25}$$

where (25), also known as *complementary condition*, states that both the Lagrange multiplier α_o and the constraint $\sum_{t=1}^{N} k_t (\mathbf{w}_{r,o}^T - \mathbf{1}^T)\mathbf{c}_t$ cannot be nonzero. The stationary solution α_o of (22) is found by solving the *Wolfe dual problem* [28], i.e.:

$$\text{maximize} \quad L_p(\mathbf{w}_r, \alpha) \quad \text{subject to (23) and } \alpha \geq 0. \tag{26}$$

By substituting (23) into (20), we obtain the Wolfe dual objective function:

$$\mathcal{W}(\alpha) = \alpha \sum_{t=1}^{N} k_t \mathbf{1}^T \mathbf{c}_t - \frac{1}{4}\alpha^2 \sum_{t=1}^{N}\sum_{j=1}^{N} \underbrace{k_t k_j \mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}_j}_{\mathbf{H}_{tj}} \tag{27}$$

which is maximized for α_o given by:

$$\alpha_o = \frac{2\sum_{t=1}^{N} k_t \mathbf{1}^T \mathbf{c}_t}{\mathbf{1}^T \mathbf{H} \mathbf{1}} \tag{28}$$

provided that $\mathbf{1}^T\mathbf{H}\mathbf{1} > 0$. The numerator in (28) 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). By substituting α_o given by (28) into (23), we obtain the optimal coefficient vector for the criterion (15) and (18), i.e.:

$$\mathbf{w}_{r,o} = \left[\frac{\sum_{t=1}^{N} k_t \mathbf{1}^T \mathbf{c}_t}{\mathbf{1}^T \mathbf{H} \mathbf{1}}\right] \mathbf{S}_W^{-1} \sum_{t=1}^{N} k_t \mathbf{c}_t. \tag{29}$$

It is obvious that, except the scaling factor given by the term inside brackets, the direction of $\mathbf{w}_{r,o}$ given by (29) coincides with that of (10) which maximizes Fisher's discriminant ratio as well as with that of (13) which maximizes the criterion (11) and (12).

IV. Support Vector Machine formulation

Support Vector Machines (SVMs) is a state-of-the-art pattern recognition technique whose foundations stem from statistical learning theory [17,29]. However, the scope of SVMs is beyond pattern recognition, because they can handle also another two learning problems, i.e., regression estimation and density estimation. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, i.e., the so-called *structural risk minimization* principle. There is a close relationship between SVMs and regularization networks [30] and the Radial Basis Function classifiers [31]. Among the many tutorials on SVMs that can be found in literature, we refer to [19,32].

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) for both the separable and non-separable case. The optimal separating decision surface in dot product spaces by mapping the similarity vectors to a high-dimensional space where an optimal hyperplane is constructed as well.

A. The Separable Case

Suppose the training data:

$$(\mathbf{c}_1, y_1), \ldots, (\mathbf{c}_N, y_N), \quad \mathbf{c}_t \in I\!R^L, \quad y_t = \begin{cases} 1 & \text{if } \mathbf{c}_t \in \mathcal{S}_I \\ -1 & \text{if } \mathbf{c}_t \in \mathcal{S}_C \end{cases} \tag{30}$$

can be separated by a hyperplane:

$$g_{\mathbf{w}_r,b}(\mathbf{c}_t) = \mathbf{w}_r^T \, \mathbf{c}_t - b = 0 \tag{31}$$

with the property:

$$\begin{aligned} g_{\mathbf{w}_r,b}(\mathbf{c}_t) &\geq 1 \qquad \text{if } y_t = 1 \\ g_{\mathbf{w}_r,b}(\mathbf{c}_t) &\leq -1 \qquad \text{if } y_t = -1 \end{aligned} \tag{32}$$

where b is a bias term. In compact notation, the set of inequalities (32) can be rewritten as:

$$y_t(\mathbf{w}_r^T \mathbf{c}_t - b) - 1 \geq 0 \quad t = 1, \ldots, N. \tag{33}$$

For our purposes, let us define the distance $v(\mathbf{w}_r, b; \mathbf{c}_t)$ of a similarity vector $\mathbf{c}_t$ from the hyperplane (31) as:

$$v(\mathbf{w}_r, b; \mathbf{c}_t) = \frac{\mid \mathbf{w}_r^T \, \mathbf{c}_t - b \mid}{\|\mathbf{w}_r\|_{\mathbf{S}_W}} = \frac{\mid \mathbf{w}_r^T \, \mathbf{c}_t - b \mid}{(\mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r)^{1/2}} \tag{34}$$

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

$$\varrho(\mathbf{w}_r) = \min_{\mathbf{c}_t \in \mathcal{S}_I} v(\mathbf{w}_r, b; \mathbf{c}_t) + \min_{\mathbf{c}_t \in \mathcal{S}_C} v(\mathbf{w}_r, b; \mathbf{c}_t) = \frac{2}{(\mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r)^{1/2}}. \tag{35}$$

That is, the optimal hyperplane separates the data so that the within-class variance is minimized, i.e., the objective function (15). The optimization is subject to the constraint functions (33). By comparing (16) with (33), 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 [17]:

$$\text{minimize } J_{\text{SVM}}(\mathbf{w}_r) = \mathbf{w}_r^T \mathbf{w}_r \text{ subject to (33).} \tag{36}$$

The solution of the minimization of (15) subject to the inequalities (33) is given by the saddle point of the Lagrangian:

$$L(\mathbf{w}_r, b, \boldsymbol{\alpha}) = \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r - \sum_{t=1}^{N} \alpha_t \{ y_t (\mathbf{w}_r^T \mathbf{c}_t - b) - 1 \} \tag{37}$$

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 [28] imply that:

$$\begin{aligned}
&\nabla_{\mathbf{w}_r} L(\mathbf{w}_{r,o}, b_o, \boldsymbol{\alpha}_o) = \mathbf{0} \Leftrightarrow \mathbf{w}_{r,o} = \frac{1}{2} \mathbf{S}_W^{-1} \sum_{t=1}^{N} \alpha_{t,o} y_t \mathbf{c}_t \\
&\frac{\partial}{\partial b} L(\mathbf{w}_{r,o}, b_o, \boldsymbol{\alpha}_o) = 0 \Leftrightarrow \sum_{t=1}^{N} \alpha_{t,o} y_t = 0 \\
&y_t \left(\mathbf{w}_{r,o}^T \mathbf{c}_t - b_o \right) - 1 \geq 0 \quad t = 1, \ldots, N \\
&\alpha_{t,o} \geq 0 \quad t = 1, \ldots, N \\
&\alpha_{t,o} \left\{ y_t \left(\mathbf{w}_{r,o}^T \mathbf{c}_t - b_o \right) - 1 \right\} = 0 \quad t = 1, \ldots, N.
\end{aligned} \tag{38}$$

From the conditions (38), one can see that the weighting vector we search for, is a linear combination of the similarity vectors in the training set multiplied by the inverse matrix of $\mathbf{S}_W$. Moreover, it is the linear combination of the similarity vectors having nonzero Lagrange multipliers α_t. These similarity vectors are the *support vectors* [17,32] in the

problem under study, provided that the regularity assumption analyzed in the previous section for the intersection of the set of feasible directions with the set of descent directions holds. Putting the expression for $\mathbf{w}_{r,o}$ into the Lagrangian (37) and taking into account the KT conditions, we obtain the Wolf dual functional:

$$\mathcal{W}(\boldsymbol{\alpha}) = \sum_{t=1}^{N} \alpha_t - \frac{1}{4} \sum_{t=1}^{N} \sum_{j=1}^{N} \alpha_t \alpha_j \underbrace{y_t y_j \left(\mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}_j \right)}_{\mathbf{H}_{tj}}. \tag{39}$$

where $\mathbf{H}_{tj}$ is the tj-th element of the Hessian matrix $\mathbf{H}$. The maximization of (39) in the non-negative quadrant of α_t, i.e.:

$$\alpha_t \geq 0 \quad t = 1, \ldots, N \tag{40}$$

under the constraint:

$$\sum_{t=1}^{N} \alpha_t y_t = 0 \tag{41}$$

is equivalent to the optimization problem:

$$\text{minimize} \quad \frac{1}{4} \boldsymbol{\alpha}_o^T \, \mathbf{H} \, \boldsymbol{\alpha}_o - \mathbf{1}^T \, \boldsymbol{\alpha}_o \;\; \text{subject to (40) and (41).} \tag{42}$$

The optimization problem (42) can be solved by using any optimization software package (e.g. [33]). For a review of optimization algorithms the interested reader is referred to [32]. Having found the non-zero Lagrange multipliers $\alpha_{t,o}$, the optimal separating hyperplane is given by:

$$g(\mathbf{c}) = \text{sgn} \left(\frac{1}{2} \sum_{\alpha_{t,o} > 0} y_t \alpha_{t,o} (\mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}) - b_o \right) \tag{43}$$

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 (6).

B. The Non-Separable Case

When the similarity values are not linearly separable, we would like to relax the constraints (32) by introducing non-negative slack variables ξ_t, $t = 1, \ldots, N$ [17], such that:

$$\mathbf{w}_r^T \mathbf{c}_t \geq b + 1 - \xi_t \qquad \text{if } y_t = 1 \tag{44}$$

$$\mathbf{w}_r^T \mathbf{c}_t \leq b - 1 + \xi_t \qquad \text{if } y_t = -1 \tag{45}$$

$$\xi_t \geq 0, \qquad t = 1, \ldots, N. \tag{46}$$

For $\xi_t \geq 0$, the above constraints can be rewritten in a compact notation as:

$$y_t(\mathbf{w}_r^T \mathbf{c}_t - b) + \xi_t - 1 \geq 0 \quad t = 1, \ldots, N. \tag{47}$$

The so-called generalized optimal hyperplane is determined by the vector $\mathbf{w}_{r,o}$ that minimizes the functional:

$$J(\mathbf{w}_r, b, \boldsymbol{\xi}) = \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r + Q(\sum_{t=1}^{N} \xi_t)^\sigma, \quad \sigma > 0 \tag{48}$$

where Q is a given value, that defines the cost of constraint violations, subject to:

$$\xi_t \geq 0 \quad t = 1, \ldots, N. \tag{49}$$

The larger the Q is, the higher penalty to the errors is assigned. The minimization of (48) subject to (47) and (49) is a convex programming problem for any integer σ. For $\sigma = 1, 2$, it is a quadratic programming problem. Moreover, the choice $\sigma = 1$ has the advantage that neither ξ_t nor their Lagrange multipliers appear in the Wolfe dual problem [19]. The solution of the minimization of (48) subject to the inequalities (47) and (49) can be obtained by applying the procedure described in Section IV-B. It can be shown that the Lagrange multipliers α_t, $t = 1, \ldots, N$ maximize

$$\mathcal{W}(\alpha) = \sum_{t=1}^{N} \alpha_t - \frac{1}{4} \sum_{t=1}^{N} \sum_{j=1}^{N} \alpha_t \alpha_j \underbrace{y_t y_j \left(\mathbf{c}_t \mathbf{S}_W^{-1} \mathbf{c}_j \right)}_{\mathbf{H}_{tj}} = \mathbf{1}^T \boldsymbol{\alpha}_o - \frac{1}{4} \boldsymbol{\alpha}_o^T \mathbf{H} \boldsymbol{\alpha}_o \tag{50}$$

subject to:

$$0 \leq \alpha_t \leq Q \quad t = 1, \ldots, N \tag{51}$$

$$\sum_{t=1}^{N} \alpha_t y_t = 0 \tag{52}$$

The comparison of (50)-(52) and (42) reveals that the objective function (50) and the equality constraint (52) remain unchanged, while the Lagrange multipliers are now upper-bounded by Q. As in the separable case, only some of the Lagrange multipliers α_t are non-zero. These multipliers are used to determine the support vectors. Having determined the support vectors, $\mathbf{w}_{r,o}$ is found by the first equation in (38) and the weighted distance measure is computed by (6). The equations derived for the optimal separating hyperplane and the bias term in the separable case are valid for the non-separable case as well.

C. Nonlinear Support Vector Machines

Thus far, we have described the case of linear decision surfaces. By examining the training procedure (50)-(52), one may notice that the similarity vectors $\mathbf{c}_t$ appear in quadratic forms $\mathbf{c}_t^T \mathbf{S}_W^{-1} \mathbf{c}_j$. The just described quadratic form can be expressed by an inner product of the form $(\mathbf{S}_W^{-1/2}\mathbf{c}_t)^T(\mathbf{S}_W^{-1/2}\mathbf{c}_j)$, because $\mathbf{S}_W^{-1}$ is a positive definite matrix. To allow for a more complex decision surface, the rotated similarity vectors $\mathbf{S}_W^{-1/2}\mathbf{c}_t$, $t = 1, \ldots, N$ are nonlinearly transformed into a high-dimensional feature space by a map $\Phi : I\!R^L \mapsto \mathcal{H}$ and then linear separation is done in a Hilbert space $\mathcal{H}$. It is obvious that the training procedure in $\mathcal{H}$ would depend only on inner products of the form $< \Phi(\mathbf{S}_W^{-1/2}\mathbf{c}_t), \Phi(\mathbf{S}_W^{-1/2}\mathbf{c}_j) >$. If the inner product in space $\mathcal{H}$ had an equivalent kernel in the input space $I\!R^L$, i.e.:

$$< \Phi(\mathbf{S}_W^{-1/2}\mathbf{c}_t), \Phi(\mathbf{S}_W^{-1/2}\mathbf{c}_j) >= K(\mathbf{S}_W^{-1/2}\mathbf{c}_t, \mathbf{S}_W^{-1/2}\mathbf{c}_j) \tag{53}$$

the inner product would not need to be evaluated in the feature space, thus avoiding the curse of dimensionality problem. In order (53) to hold, the kernel function has to be a positive definite function that satisfies *Mercer's condition* [17]. Functions that are usually employed in constructing nonlinear SVMs are tabulated in Table I. The polynomial kernel

defined by:

$$K(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_t, \mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_j) = \left(\mathbf{c}_t^T \mathbf{S}_W^{-1}\mathbf{c}_j + 1\right)^p \tag{54}$$

was used in the experiments reported in the next section.

Nonlinear SVMs yield a higher computational cost than linear SVMs during the test phase. Indeed, in nonlinear SVMs the distance between the reference person r and the test person τ is given by:

$$D(\tau, r) = \frac{1}{2}\sum_{t=1}^{N_s} \alpha_t y_t K(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_t, \mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_\tau), \tag{55}$$

where N_s denotes the number of support vectors extracted in the training phase, instead of the much simpler distance computed by the linear SVM, i.e.:

$$D(\tau, r) = \frac{1}{2}\sum_{t=1}^{N_s} \alpha_t y_t \mathbf{c}_t^T \mathbf{S}_W^{-1}\mathbf{c}_\tau = \mathbf{w}_r^T \mathbf{c}_\tau. \tag{56}$$

In the latter case, the inner product between the optimal weighting vector, found in the training phase, and the test similarity vector suffices. This is not the case in (55), where a sum of N_s terms has to be computed. Thus, the test phase of nonlinear SVMs is N_s times slower than that of the linear SVMs.

V. Experimental results

The optimal coefficient vectors derived by the procedures described in Sections III and IV have been used to weigh the raw similarity vectors $\mathbf{c}$ that are provided by the morphological elastic graph matching [13,34] applied to frontal face authentication. Let us call the combination of the morphological elastic graph matching and either the constrained Least Squares or the SVM weighting approach *weighted MEGM*. The weighted MEGM has been tested on the M2VTS database [20]. The database contains 37 persons' video data, which include speech consisting of uttering digits and color image sequences of rotated heads. Four recordings of the 37 persons have been collected at different time

instants. In our experiments, the sequences of rotated heads were considered by using only the luminance information at a resolution of 286 $\times$ 350 pixels. From each image sequence, one frontal image was chosen based on symmetry considerations for the experimental protocol defined subsequently. Let the term *shot* denote the collection of one frontal face image per person (i.e., 37 frontal face images in total).

Four experimental sessions were implemented by employing the "leave-one-out" and "rotation" estimates. In each session, one shot was left out to be used as test set. To implement test impostor claims, we rotated over the 37 person identities by considering the frontal face image of each person in the test set as impostor. By excluding any frontal face image of the test impostor from the remaining three shots, a training set consisted of 36 clients was built. The test impostor pretended to be one of the 36 clients and this attempt was repeated for all client identities. As a result, 36 impostor claims were produced. In a similar manner, 36 test client claims were tested by employing the clients' frontal faces from the shot that was left out, and those of the training set. Let BP, BS, CC, ..., XM be the identity codes of the persons included in the database. Figure 3 depicts the experimental protocol when person BP is considered to be an impostor and shot 4 is employed as test set. It can be seen that the training set is built of 3 out of the 4 available shots each one consisted of 36 out of the 37 available persons. The comparisons shown for person BP are repeated for all other persons in the database. Obviously, similar comparisons are made by rotating among the available shots.

Next, we describe the training procedure. The training procedure is analogous to the test procedure described above. It is applied to the training set of the 36 clients. For each client we had three frontal face images at our disposal. We implemented six training client claims by considering all permutations of the three frontal images of the same person taken two at a time. It is evident that each training client claim yields a raw

similarity vector. Moreover, we implemented $35 \times 6 = 210$ training impostor claims when each of the other 35 persons attempted to access the system with the identity of the person under consideration. That is, another six raw similarity vectors corresponding to all pairwise comparisons between the frontal images of any two different persons taken from different shots were computed. Accordingly, 6 intra-class similarity vectors and 210 inter-class similarity vectors were computed for each of the 36 trained classes.

Morphological elastic graph matching was used to yield all the raw similarity vectors required. In each client or impostor claim during either the training or the test procedure, the reference grids built for each person (i.e., class) were matched and adapted to the feature vectors computed at every image pixel of the frontal face image of a test person who could be either a client or an impostor using MEGM.

Let us assume that person BP using his frontal face image from shot 4, denoted by BP_4, pretends to be person BS during the test procedure. To test such a claim, we develop a training procedure aiming at the derivation of:

1. the optimal weighting vector to weigh the raw similarity vectors so that the within-class variance of the weighted distance measures $D'(t, r)$ is minimized, and

2. the threshold on the weighted distance measures that should ideally enable the distinction between the distance measures that correspond to client claims within the trained class under study, and the distance measures that correspond to impostor claims for impostors that belong to any other class.

In the example being discussed, the training procedure determines the following 36 pairs of weighting vectors and thresholds: $(\mathbf{w}_{BS}(4, BP), T_{BS}(4, BP))$, $(\mathbf{w}_{CC}(4, BP), T_{CC}(4, BP))$, ..., $(\mathbf{w}_{XM}(4, BP), T_{XM}(4, BP))$. The weighting vector $\mathbf{w}_{BS}(4, BP)$ and the threshold $T_{BS}(4, BP)$ are derived so that they discriminate the distance measures between frontal face images of person BS that originate from shots 1, 2, and 3 against all other possible

distance measures between frontal face images of the remaining 35 persons (e.g., $CC - XM$) and person BS taken from different shots having excluded any frontal face image of person BP.

The aforementioned optimal weighting vectors were derived by the procedures of Section III and IV. Obviously, the 6 intra-class similarity vectors were very few for applying discriminating techniques. In order to increase the number of intra-class similarity vectors, additional client images were extracted from the database and included in the training set. Furthermore, the number of frontal face images for each client was increased by augmenting one's original frontal images with others produced by adding slightly Gaussian noise to the original images, as has been proposed in [23]. Overall, a number of 60-100 intra-class similarity vectors were produced for each class by means of 20-30 additional images per client (i.e., per class). The additional client images prevent any overfitting during the training caused by the lack of data, especially in the case of non-linear SVMs. It is worth noting that the additional similarity vectors were used only for the derivation of the optimal separating hyperplane.

Next, any weighted distance measure was compared against an appropriate threshold. The thresholds were computed as follows. The minimum of the six inter-class distance measures was found when each of the 35 training impostors , i.e., when $CC, \ldots, XM$ pretends to be BS during the training procedure. The vector of 35 minimum distance measures is ordered in ascending order according to their magnitude. Let $D'_{(Q)}$ denote the Q-th order statistic in the vector of these 35 inter-class distance measures. Obviously, $D'_{(1)}$ is the minimum impostor distance measure. The threshold is chosen as follows:

$$T_{BS}(4, BP) = D'_{(Q)}, \quad Q = 1, 2, \ldots \tag{57}$$

where we explicitly note that the frontal face image of person BP from shot 4 was excluded.

Let us now explain how test claims are assessed. To be specific, we consider the case of person BP being an impostor and persons $BS, \ldots, XM$ being clients. We assume that person BP uses his frontal face image BP_4 to pretend to be person BS. First, the raw similarity vectors (4) produced by MEGM were derived using the three reference grids for person BS that were available in the training set. Next, the weighted distance measures (6) were computed by using $\mathbf{w}_{BS}(4, BP)$ derived during the training procedure, i.e., $D'(BP_4, BS_1), D'(BP_4, BS_2)$ and $D'(BP_4, BS_3)$ and their minimum was found:

$$D'(BP_4, \{BS\}) = \min\{D'(BP_4, BS_1), D'(BP_4, BS_2), D'(BP_4, BS_3)\}. \tag{58}$$

The same steps were repeated for any of the 36 clients, i.e., $BS, \ldots, XM$. A false acceptance occurred when:

$$D'(BP_4, \mathcal{X}) \leq T_{\mathcal{X}}(4, BP) \quad \mathcal{X} = BS, \ldots, XM \tag{59}$$

while a false rejection occurred when:

$$D'(\mathcal{X}_4, \{\mathcal{X}\}) > T_{\mathcal{X}}(4, BP) \quad \mathcal{X} = BS, \ldots, XM. \tag{60}$$

By repeating the procedure four times, $4 \times 37 \times 36 = 5328$ client claims and 5328 impostor claims were realized in total.

For a particular choice of parameter Q, a collection of thresholds was determined that defined an *operating state* of the test procedure. For each operating state, a false acceptance rate (FAR) and a false rejection rate (FRR) were computed. By varying the parameter Q, the *Receiver Operating Characteristic* (ROC) was created.

The EER, i.e., the operating point in the ROC having FAR=FRR, of the MEGM without weighting coefficients according to the described experimental protocol was found to be 9.2% [14]. By using the constrained least squares solution described in Section III, we achieved an EER of 8.2%. Further improvements (i.e., an EER equal to 6.4 %) were

obtained, when the coefficient vector derived by the standard SVM (36) extended to handle the nonseparable case was used to weigh the raw similarity vectors. A better authentication performance was obtained when the proposed variant of linear support vector machine that minimizes (48) was applied. In this case, we achieved an EER of 5.6%. The weighting coefficients derived by using the proposed variant of SVMs that minimizes (48) for several persons in the database are shown pictorially in Figure 4.

We also applied the standard and the proposed variant of nonlinear SVMs to weigh the raw similarity values. The polynomial kernel (54) was used for $p = 4$. The use of non-linear decision surfaces yields a higher computational complexity than using separating hyperplanes in the test phase, as is explained in Section IV-C. However, the performance of MEGM was considerably improved by reaching an EER of 4.5% for the standard SVMs, and 2.4% for the proposed non-linear SVMs. The EERs achieved by employing the weighting methods described in the paper are summarized in Table II. The ROC curve of MEGM for each weighting algorithm is plotted in Figure 5. In the same figure, the ROC curve of the original MEGM is also depicted for comparison purposes. We can see that the area under the ROC curve for the proposed methods is much smaller than the initial one. Furthermore, the minimum and the maximum number of support vectors found by considering all persons in the training sets that are constructed according to the experimental protocol is given in Table III for the standard SVMs and the proposed variant. It is obvious that the number of support vectors does not change significantly by using the proposed methods. In any case the number of support vectors is between 10% and 20% of the trained vectors.

A performance comparison between several face authentication algorithms developed within the European Union research project *Multi-modal Verification for Telecommunication Services* (M2VTS) is reported in Table IV. It can be seen that the weighted

MEGM algorithm by the proposed nonlinear SVM variant attains the best performance. It is worth noting that all methods were tested on the same database according to the protocol described in this Section.

VI. 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. The experimental results indicated the success of the proposed methods in frontal face authentication. A very low EER of 2.4% was obtained when the weighting coefficients determined by the nonlinear SVM are used to weigh the raw similarity vectors computed by the morphological elastic graph matching. The selection of an appropriate kernel in nonlinear SVM construction is subject of ongoing search.

References

[1] R. Chellappa, 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] P. J. Phillips, "Matching pursuit filters applied to face identification," *IEEE Trans. on Image Processing*, vol. 7, no. 8, pp. 1150–1164, August 1998.

[3] 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.

[4] L. Wiskott, "Phantom faces for face analysis," *Pattern Recognition*, vol. 30, no. 6, pp. 837–846, 1997.

[5] J. Zhang, Y. Yan, and M. Lades, "Face recognition: Eigenface, elastic matching and neural nets," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1423–1435, September 1997.

[6] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of faces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, January 1990.

[7] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[8] G.W. Cottrell and M. Fleming, "Face recognition using unsupervised feature extraction," in *Int. Neural Network Conf.*, Paris, July 1990, vol. 1, pp. 322–325.

[9] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, vol. 59, pp. 291–294, 1988.

[10] B.S. Manjunath, R. Chellappa, and C. v.d. Malsburg, "A feature based approach to face recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR-92)*, 1992, pp. 373–378.

[11] L. Wiskott, J.-M. Fellous, N. Krüger, and C. v.d. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, July 1997.

[12] R.P. Würtz, "Object recognition robust under translations, deformations, and changes in background," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 769–775, July 1997.

[13] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using morphological elastic graph matching," *IEEE Trans. on Image Processing*, vol. 4, no. 9, 2000, to appear.

[14] C. Kotropoulos, A. Tefas, and I. Pitas, "Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions," *Pattern Recognition*, 2000, to appear.

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

[16] R.J. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*, John Wiley and Sons, New York, 1992.

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

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, pp. 1–43, 1998.

[20] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," *Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication (J. Bigün, G. Chollet, and G. Borgefors, Eds.)*, vol. 1206, pp. 403–409, 1997.

[21] D.L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, August 1996.

[22] P.N. Belhumer, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 1997.

[23] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human faces,," Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication (J. Bigün, G. Chollet, and

G. Borgefors, Eds.), 1997.

[24] 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, April 1999.

[25] N. Krüger, "An algorithm for the learning of weights in discrimination functions using a priori constraints," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 764–768, July 1997.

[26] G. Yang and T.S. Huang, "Human face detection in a complex background," *Pattern Recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[27] C. Kotropoulos and I. Pitas, "Rule-based face detection in frontal views," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP-97)*, Munich, Germany, April 1997, vol. IV, pp. 2537–2540.

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

[29] V. Vapnik, *Statistical Learning Theory*, J. Wiley, New York, 1998.

[30] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, 1999, to appear.

[31] B. Schoelkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2758–2765, November 1997.

[32] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: Training and applications," Technical Memo AIM-1602, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Feb. 28, 1997.

[33] S. Gunn, "Support vector machines for classification and regression," Technical Report MP-TR-98-05, Image Speech and Intelligent Systems Group, University of Southampton, 1998.

[34] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using discriminating grids with morphological feature vectors," *IEEE Trans. on Multimedia*, vol. 2, no. 1, pp. 14–26, March 2000.

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

TABLE I

SOME KERNEL FUNCTIONS AND THE TYPE OF DECISION SURFACE THEY DEFINE.

Type of Classifier	Kernel Function
Linear	$K(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_t, \mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_j) = \mathbf{c}_t^T\mathbf{S}_W^{-1}\mathbf{c}_j$
Polynomial of degree p	$K(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_t, \mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_j) = \left(\mathbf{c}_t^T\mathbf{S}_W^{-1}\mathbf{c}_j + 1\right)^p$
Gaussian RBF	$K(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_t, \mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_j) = \exp\left(-(\mathbf{c}_t - \mathbf{c}_j)^T\mathbf{S}_W^{-1}(\mathbf{c}_t - \mathbf{c}_j)\right)$
Multi Layer Perceptron	$K(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_t, \mathbf{S}_W^{-\frac{1}{2}}\mathbf{c}_j) = \tanh\left(\mathbf{c}_t^T\mathbf{S}_W^{-1}\mathbf{c}_j - \theta\right)$

TABLE II

EQUAL ERROR RATES ACHIEVED BY SEVERAL WEIGHTING METHODS.

Authentication algorithm	EER (%)
Standard MEGM [13] (Baseline algorithm)	9.2
Weighted MEGM by using constrained least squares (Section III)	8.2
Weighted MEGM by the standard Support Vector Machines	6.4
Weighted MEGM by the proposed Support Vector Machines (Section IV-B)	5.6
Weighted MEGM by the standard nonlinear Support Vector Machines	4.5
Weighted MEGM by the proposed nonlinear SVMs (Section IV-C)	**2.4**

TABLE III

NUMBER OF SUPPORT VECTORS FOUND USING THE STANDARD SVM AND THE PROPOSED SVM METHODS.

SVM method	Number of support vectors	
	Minimum	Maximum
standard SVMs	35	43
proposed SVMs	34	40
standard nonlinear SVMs	40	50
proposed nonlinear SVMs	41	51

TABLE IV

COMPARISON OF EQUAL ERROR RATES FOR SEVERAL AUTHENTICATION TECHNIQUES IN THE M2VTS DATABASE.

Authentication Technique	EER (%)
MEGM with linear/nonlinear SVMs	**2.4-5.6**
MEGM	9.2
Gray level frontal face matching [35]	8.5
Discriminant GDLA [24]	6.0-9.2
GDLA [24]	10.8-14.4

Fig. 1. The response of multiscale erosion dilation for scales $-9, \ldots, 9$ used in MEGM.

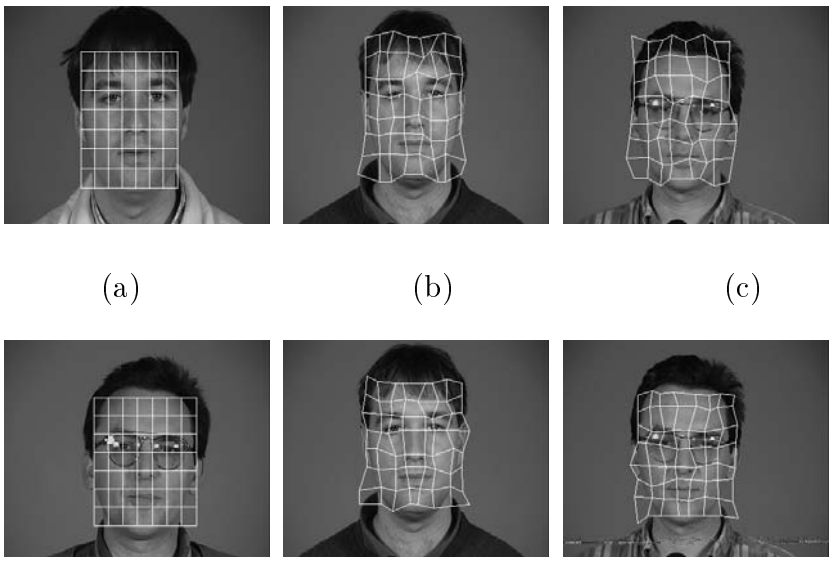

Fig. 2. Grid matching procedure in MEGM: (a) Model grid for person BS. (b) Best grid for test person BS after elastic graph matching with the model grid. (c) Best grid for test person LV after elastic graph matching with the model grid for person BS. (d) Model grid for person LV. (e) Best grid for test person BS after elastic graph matching with the model grid for LV. (f) Best grid for test person LV after elastic graph matching with the model grid.

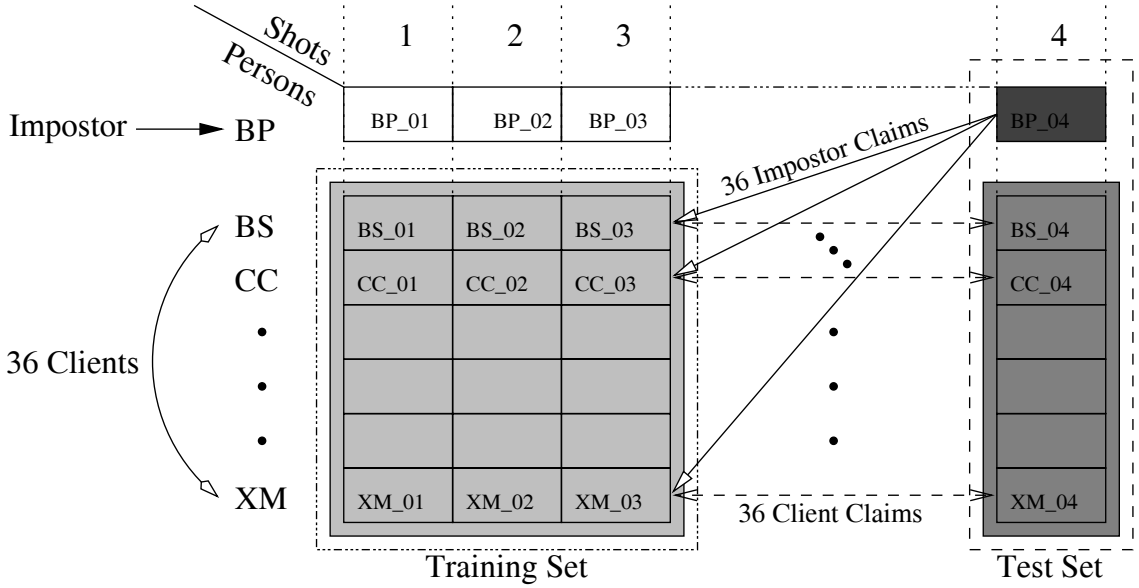

Fig. 3. Experimental Protocol.

Fig. 4. Weighting coefficients for the grid nodes in morphological elastic graph matching. The brighter a node is, the bigger discriminatory power possesses. The intensity of the nodes is normalized for visualization purposes.

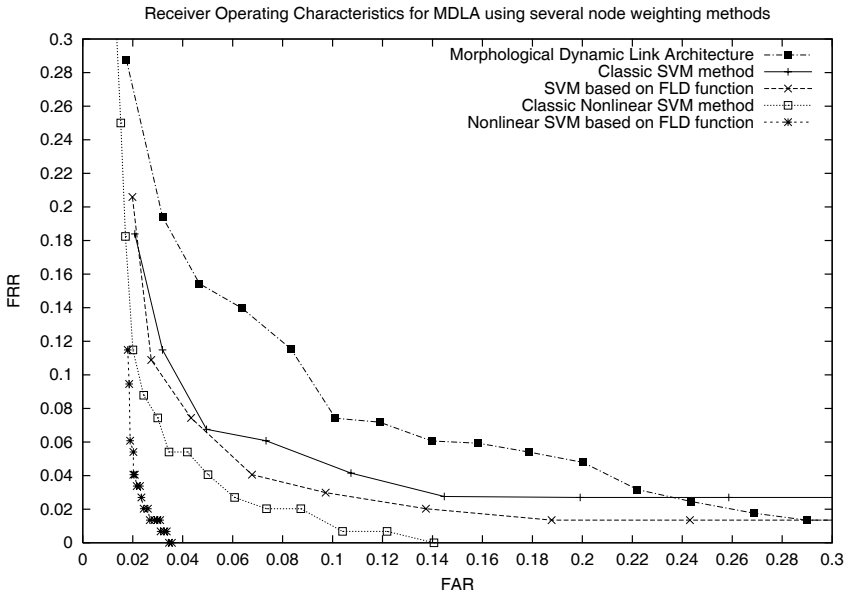

Fig. 5. Receiver Operating Characteristics for MEGM with discriminatory analysis for several node weighting methods.

Biographies

Anastasios Tefas received the Diploma degree in Informatics in 1997 from Aristotle University of Thessaloniki, Greece. He is currently a researcher and teaching assistant and he is studying towards a PhD at the Department of Informatics at the University of Thessaloniki. His current research interests lie in the areas of signal and image processing, pattern recognition and computer vision. He is a student member of the IEEE.

Constantine Kotropoulos received the Diploma degree with honors in electrical engineering in 1988 and the Ph.D. degree in electrical & computer engineering in 1993, both from the Aristotle University of Thessaloniki. From 1989 to 1993 he was a research and teaching assistant in the Department of Electrical & Computer Engineering at the same university. In 1995, after his military service in the Greek Army, he joined the Department of Informatics at the Aristotle University of Thessaloniki as a senior researcher. He currently is a lecturer in the same department. He has also conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland during the summer of 1993.

His current research interests include nonlinear digital signal processing, detection and estimation theory, pattern recognition, neural networks, and computer vision.

Dr. Kotropoulos has been a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a member of the IEEE, SPIE and the Technical Chamber of Greece.

Ioannis Pitas received the Diploma of Electrical Engineering in 1980 and the PhD degree in Electrical Engineering in 1985 both from the University of Thessaloniki, Greece. Since 1994, he has been a Professor at the Department of Informatics, University of Thessaloniki. From 1980 to 1993 he served as Scientific Assistant, Lecturer, Assistant Professor, and Associate Professor in the Department of Electrical and Computer Engineering at the same University. He served as a Visiting Research Associate at the University of Toronto, Canada, University of Erlangen-Nuernberg, Germany, Tampere University of Technology, Finland and as Visiting Assistant Professor at the University of Toronto. He was lecturer in short courses for continuing education.

His current interests lie in the areas of digital image processing, multidimensional signal processing and computer vision. He has published over 350 papers and contributed in 19 books in his area of

interest.

Dr Pitas has been member of the European Community ESPRIT Parallel Action Committee. He has also been an invited speaker and/or member of the program committee of several scientific conferences and workshops. He was Associate Editor of the IEEE Transactions on Circuits and Systems and co-editor of Multidimensional Systems and Signal Processing and he is currently an Associate Editor of the IEEE Transactions on Neural Networks. He was chair of the 1995 IEEE Workshop on Nonlinear Signal and Image Processing (NSIP95). He was technical chair of the 1998 European Signal Processing Conference (EUSIPCO98). He is general chair of the 2001 IEEE International Conference on Image Processing (ICIP2001).