Minimum Class Variance Support Vector Machines

Stefanos Zafeiriou[†], Anastasios Tefas[†,⋆] and Ioannis Pitas[†]

[†]Aristotle University of Thessaloniki

Department of Informatics

Box 451

54124 Thessaloniki, Greece

[⋆]Technological Educational Institute of Kavala

Department of Information Management

65404 Kavala, Greece

Address for correspondence :

Professor Ioannis Pitas

Aristotle University of Thessaloniki

54124 Thessaloniki

GREECE

Tel. ++ 30 231 099 63 04

Fax ++ 30 231 099 63 04

email: pitas@zeus.csd.auth.gr

Abstract

In this paper a modified class of Support Vector Machines (SVMs) inspired from the optimization of Fisher's discriminant ratio is presented, the so-called Minimum Class Variance SVMs (MCVSVMs). The MCVSVMs optimization problem is solved in cases in which the training set contains less samples that the dimensionality of the training vectors using dimensionality reduction through Principal Component Analysis (PCA). Afterwards, the MCVSVMs are extended in order to find nonlinear decision surfaces by solving the optimization problem in arbitrary Hilbert spaces defined by Mercer's kernels. In that case, it is shown that, under Kernel Principal Component Analysis (KPCA), the nonlinear optimization problem is transformed into an equivalent linear MCVSVMs problem. The effectiveness of the proposed approach is demonstrated by comparing it with the standard SVMs and other classifiers, like Kernel Fisher Discriminant Analysis (KFDA) in facial image characterization problems like gender determination, eyeglass and neutral facial expression detection.

Index Terms

Support Vector Machines, Fisher's discriminant analysis, Principal Component Analysis, kernel methods, facial images.

I. Introduction

Pattern recognition systems employing Support Vector Machines (SVMs) [1] have drawn much attention due to their good performance in practical applications and their solid theoretical foundations. The applications of SVMs span several disciplines such as object recognition [2], speech and speaker recognition and verification [3], face verification, face detection and gender determination from facial images [4]-[6] and spam mail identification [7].

In binary classification problems, SVMs try to find a separating decision hyperplane with the maximum margin. The margin is defined as the minimum distance of the class sample distances to the decision hyperplane. The property that distinguishes SVMs from other nonparametric techniques,

like nearest-neighbor classification or neural networks, is that it is based on structural risk minimization [1], [8], [9]. Typical pattern recognition methods attempt to minimize the misclassification errors on the training set (empirical risk minimization). Instead, SVMs minimize the structural risk, that is the probability of misclassifying a previously unseen sample drawn randomly from a fixed but unknown probability distribution. If the Vapnik-Chervonenkis (VC)-dimension [10] of the family of decision surfaces is known, the theory of SVMs provides an upper bound for the probability of misclassification of the test set for any possible probability distributions of the data points [1]. The main reason that has made SVMs so popular is that they consist of quadratic optimization problems which can be solved very efficiently and it is guaranteed that they will find a global minimum.

Another aspect of SVMs is that they can be used in order to construct non-linear decision surfaces. In order to find such surfaces a non-linear function ϕ is firstly used in order to project the samples to a very high dimensional feature space (this space has often the structure of a *Hilbert space*), where the vectors are linearly or near-linearly separable and a maximum margin hyperplane is found. The decision surface can be found without having to compute explicitly the mapping ϕ, but by only computing dot products in the Hilbert space by means of the *kernel trick* [8], as long as the mapping ϕ satisfies the Mercer's conditions [11], [12]. The interested reader may refer to [13] for details on the geometry of Hilbert spaces (also referred as feature spaces).

The *kernel trick* procedure has been used to create the nonlinear generalizations of linear techniques, like Principal Component Analysis (PCA) [14] into Kernel-PCA (KPCA) [15] for non-linear component analysis, Fisher's Linear Discriminant Analysis (FLDA) [16], [17] into Kernel-Fisher's Discriminant Analysis (KFDA) [18], [19] and recently into the so-called Complete Kernel Fisher's Discriminant Analysis (CKFDA) algorithm [20] for discriminant learning and recognition, and Independent Component Analysis (ICA)[21] into Kernel-ICA [22] for signal separation. The interested reader may refer to [8], [20], [23] and to references therein for details about kernel based algorithms.

In [18] a unified framework in terms of a nonlinearized variant of the Rayleigh coefficients has been proposed and has been applied in order to formulate nonlinear generalizations of Fishers discriminant analysis and oriented PCA with kernel functions. In order to overcome the fact that both calculation and eigenanalysis of covariance matrices in arbitrary dimensional Hilbert spaces are generally ill-posed problems, regularization parameters have been incorporated in the optimization problem.

An effort to merge Fisher's discriminant and SVMs has been done in [6], where a modified class of SVMs has been constructed, inspired by the optimization of the Fisher's discriminant ratio [24]. In detail, motivated by the fact that the Fisher's discriminant optimization problem for two classes is a constraint least-squares optimization problem [6], [23], [18], the problem of minimizing the within-class variance has been reformulated, so that it can be solved by constructing the optimal separating hyperplane for both separable and nonseparable cases. In the face verification problem, the modified class of SVMs has been applied successfully in order to weight the local similarity value of the elastic graphs nodes according to their corresponding discriminant power for frontal face verification [6]. It has been shown that it outperforms the typical maximum margin SVMs [6].

In [6], only the case where the number of training vectors was larger than the feature dimensionality was considered (i.e., when the within-class scatter matrix of the samples is not singular). In this paper the method is extended in problems where the feature vector dimensionality is larger than the number of available samples, forming in that way the proposed Minimum Class Variance Support Vector Machines (MCVSVMs). It will be proven that the solution of MCVSVM problems in such cases can be found through PCA dimensionality reduction.

Afterwards, in order to define non-linear decision surfaces obtained through the MCVSVMs optimization, the problem will be generalized in dot product Hilbert spaces. It will be proven that the non-linear MCVSVMs problem is equivalent to a linear one, subject to an initial KPCA embedding of the training data. The proposed methods have been inspired from the recent advances in solving the Fisher's discriminant optimization problem in cases where the training set contains less samples than

the feature dimensionality [20], [25], [26], where it has been proven that, under KPCA, the KFDA is reformulated into an equivalent linear FLDA. Moreover, we will show that MCVSVMs have both the advantages of FLDA and SVMs. That is, MCVSVMs consider class distribution characteristics in their optimization problem but at the same time ensures separability. In contrast to FLDA that does not ensures separability and to maximum margin SVMs that take into consideration only the samples that are in the class boundaries.

The proposed methods have been applied to facial image characterization problems like gender determination, eyeglass and neutral state detection. The experiments indicate the power of the proposed approach against other techniques like maximum margin SVMs [1] and CKFDA [20]. As will be shown in the paper in small sample size problems (e.g., image classification problems) the MCVSVMs should be defined and solved in spaces defined from PCA or KPCA embeddings. The motivations to apply the proposed method in image processing applications and especially to facial image characterization problems is that PCA and KPCA spaces have been proven very rich in information for the specific applications and that classifiers and feature extraction methods based on the minimization of within-class-variance (e.g., FLDA and KFDA) have been very successfully applied. This was firstly shown in the pioneer work of Turk and Pentland [27] and Kirby and Sirovich [28] where PCA has been applied for facial feature extraction, face recognition and face detection. Since then, PCA plus LDA classifiers has been used for facial image retrieval [16] and face recognition [17]. Moreover, PCA plus two-class LDA classifiers have been used for eyeglass detection, in [17]. This is similar to the proposed approach where a PCA plus MCVSVMs classifiers have been tested for eyeglass detection.

In order to capture nonlinearities in facial image modelling KPCA has been widely used. In [29] KPCA plus SVM classifiers have been used for recognition. This is very similar to our approach where KPCA plus MCVSVMs have been used in various facial image characterization applications. Moreover, in [20] it has been proven that the KFDA is equivalent to firstly applying KPCA and

afterwards performing LDA. Moreover, it has been shown that this scheme is very successful for facial feature extraction and face recognition. In [30] Gabor-based KPCA spaces have given very good results in face recognition. Finally, one of the best gender determination algorithm is the one presented in [4], where SVMs have been applied directly to facial images.

Summarizing the contributions of this paper are:

- The presentation of the MCVSVMs in their general form, for the cases where the training set contains more samples than the dimensionality of the samples and for the cases where the training set contains less samples than the samples dimensionality.
- The generalization of MCVSVMs in arbitrary Hilbert spaces, using Mercer's kernels in order to define non-linear decision surfaces.
- The theoretical and experimental investigation of the relationship of MCVSVMs with SVMs and CKFDA.

The rest of this paper is organized as follows. The problem will be outlined in Section II. In Section III, the linear case of MCVSVMs is treated for the case where the number of the training vectors is smaller than the samples dimension. In Section IV the problem will be defined and solved in reproducing *Hilbert spaces* in order to find the non-linear decision surfaces. In Section V, a discussion is carried out about the relationship of the proposed decision surfaces with maximum margin SVMs, CKFDA, and the surfaces proposed in [6]. The experimental results are discussed in Section VI. Finally, conclusions are drawn in Section VII.

II. Problem Statement

Let a training set with finite number of elements $\mathcal{U} = \{\mathbf{x}_i, i = 1, \ldots, N\}$, be separated into two different classes $\mathcal{C}_+$ and $\mathcal{C}_-$, with training samples $\mathbf{x}_i \in \Re^M$ and labels $y_i \in \{1, -1\}$. The simplest way to separate these two classes is by finding a separating hyperplane:

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{1}$$

where $\mathbf{w} \in \Re^M$ is the normal vector to the hyperplane and $b \in \Re$ is the corresponding scalar term of the hyperplane, also known as bias term [6]. The decision whether a test sample $\mathbf{x}$ belongs to one of the different classes $\mathcal{C}_+$ and $\mathcal{C}_-$ is taken by using the linear decision function $g_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$, also known as canonical decision hyperplane [1].

A. Fisher's Linear Discriminant Analysis

The best studied linear pattern classification algorithm for separating these classes is the one that finds a decision hyperplane that maximizes the Fisher's discriminant ratio, also known as Fisher's Linear Discriminant Analysis (FLDA):

$$\max_{\mathbf{w},b} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}, \tag{2}$$

where the matrix $\mathbf{S}_w$ is the within-class scatter matrix defined as:

$$\mathbf{S}_w = \sum_{\mathbf{x}\in\mathcal{C}_-} (\mathbf{x} - \mathbf{m}_{\mathcal{C}_-})(\mathbf{x} - \mathbf{m}_{\mathcal{C}_-})^T + \sum_{\mathbf{x}\in\mathcal{C}_+} (\mathbf{x} - \mathbf{m}_{\mathcal{C}_+})(\mathbf{x} - \mathbf{m}_{\mathcal{C}_+})^T, \tag{3}$$

$\mathbf{m}_{\mathcal{C}_+}$ and $\mathbf{m}_{\mathcal{C}_-}$ are the mean sample vectors for the classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively. The matrix $\mathbf{S}_b$ is the between class scatter matrix defined in the two class case as:

$$\mathbf{S}_b = N_{\mathcal{C}_+}(\mathbf{m} - \mathbf{m}_{\mathcal{C}_+})(\mathbf{m} - \mathbf{m}_{\mathcal{C}_+})^T + N_{\mathcal{C}_-}(\mathbf{m} - \mathbf{m}_{\mathcal{C}_-})(\mathbf{m} - \mathbf{m}_{\mathcal{C}_-})^T \tag{4}$$

where $N_{\mathcal{C}_+}$ and $N_{\mathcal{C}_-}$ are the cardinalities of the classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively and $\mathbf{m}$ is the total mean vector of the set $\mathcal{U}$. The solution of the optimization problem (2) can be found in [24]. It can be proven that the corresponding separating hyperplane is the optimal Bayesian solution when the samples of each class follow Gaussian distributions with same covariance matrices [24]. The decision hyperplane that is derived from the FLDA optimization problem (2) does not separate the data, using the FLDA hyperplane, even though the training samples are linearly separable [24]. This fact is illustrated in Figure 1, where it is shown that FLDA leads to a decision hyperplane that does not separates the data even though the data are indeed linear separable. The SVM and MCVSVM solution, that will be presented in the following, find a decision hyperplane, which in this case the two solutions coincide, that separates linear the data.

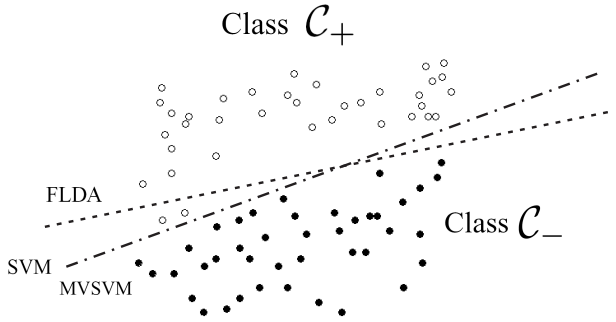


Fig. 1. An FLDA decision hyperplane that cannot separate linearly the data even though the data are linear separable. The MCVSVMs and SVMs solutions lead to a hyperplane that fully separates the data

B. Support Vector Machines (SVMs)

In the SVMs case, the optimal separating hyperplane is the one which separates the training data with the maximum margin [1]. The SVMs optimization problem is defined as:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{5}$$

subject to the separability constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \;\; i = 1, \ldots, N. \tag{6}$$

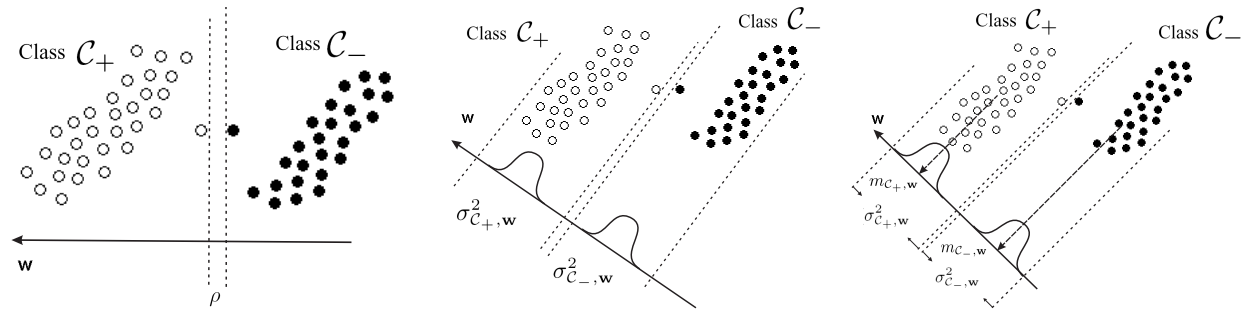
C. Minimum Class Variance Support Vector Machines (MCVSVMs)

In [6], inspired by the maximization of the Fisher's discriminant ratio (2) and the SVMs separability constraints, the MCVSVMs have been introduced. Their optimization problem is defined as:

$$\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{S}_w\mathbf{w}, \;\; \mathbf{w}^T\mathbf{S}_w\mathbf{w} > 0 \tag{7}$$

subject to the separability constraints (6). It is required that the normal vector $\mathbf{w}$ satisfies the constraint $\mathbf{w}^T\mathbf{S}_w\mathbf{w} > 0$. A detailed discussion about this constraint will be given in Section V. It is interesting to note here that, since the matrix $\mathbf{S}_w$ is positive semi-definite (i.e., $\forall\, \mathbf{w} \in \Re^M, \;\; \mathbf{w}^T\mathbf{S}_w\mathbf{w} \geq 0$) and, in particular, if the within-class scatter matrix $\mathbf{S}_w$ is not singular, then $\nexists \mathbf{w} \in \Re^M \;\; : \;\; \mathbf{w}^T\mathbf{S}_w\mathbf{w} = 0$. Thus,

when $\mathbf{S}_w$ is invertible, no solutions with $\mathbf{w}^T\mathbf{S}_w\mathbf{w} = 0$ can be found. Figure 2 describes pictorially the solution of the optimization problems of SVMs, MCVSVMs and FLDA where $m^2_{\mathcal{C}_+,\mathbf{w}}, m^2_{\mathcal{C}_-,\mathbf{w}}$ and $\sigma^2_{\mathcal{C}_+,\mathbf{w}}, \sigma^2_{\mathcal{C}_-,\mathbf{w}}$ are the means and the variances of the classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively along the projection $\mathbf{w}$. As can be seen from the case illustrated in Figure 2 the SVMs solution does not take into consideration the class distribution and results to a non-robust solution. On the other hand the solution of the MCVSVMs takes into consideration both the samples in the boundaries and the distribution of the classes and gives a robust solution. FLDA gives a robust solution in this problem, as well. Now, by examining Figures 1 and 2 we have a first experimental indication that MCVSVMs is a compromise between SVMs and FLDA.



(a) (b) (c)

Fig. 2. Illustration of the SVM, MCVSVM and FLDA optimization problems (a) search for a direction $\mathbf{w}$, such that the projected samples are separable with the maximum possible margin ρ; (b) search for a direction $\mathbf{w}$, such that samples projected onto this dimension are separable and the variances ($\sigma^2_{\mathcal{C}_+,\mathbf{w}}$ and $\sigma^2_{\mathcal{C}_-,\mathbf{w}}$) of the projected samples is minimized; (c) search for a direction $\mathbf{w}$, such that the distance of the centers of the classes projected onto this dimension ($m_{\mathcal{C}_+,\mathbf{w}}$ and $m_{\mathcal{C}_-,\mathbf{w}}$) is maximized while the variances ($\sigma^2_{\mathcal{C}_+,\mathbf{w}}$ and $\sigma^2_{\mathcal{C}_-,\mathbf{w}}$) of the projected samples is minimized;

In the case where the training vectors are not linearly separable the optimum decision hyperplane is found by using the *soft margin* formulation [6], [1] and solving the following optimization problem:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \mathbf{w}^T\mathbf{S}_w\mathbf{w} + C\sum_{i=1}^{N}\xi_i, \;\; \mathbf{w}^T\mathbf{S}_w\mathbf{w} > 0 \tag{8}$$

subject to the separability constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \;, \xi_i \geq 0, \;\; i = 1, \ldots, N \tag{9}$$

where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]$ is the vector of the non-negative slack variables and C is a given constant that defines the cost of the errors after the classification. Larger values of C correspond to higher penalty assigned to errors. The linearly separable case can be achieved when choosing $C = \infty$.

The solution of the minimization of (8), subject to the constraints (9), is given by the saddle point of the Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) = \mathbf{w}^T \mathbf{S}_w \mathbf{w} + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{N} \beta_i \xi_i \qquad (10)$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]^T$ are the vectors of the Lagrangian multipliers for the constraints (9). The Karush-Kuhn-Tucker (KKT) conditions [1] [34] imply that for the saddle point of $\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, b, \boldsymbol{\xi}$ the following hold:

$$\begin{aligned}
\nabla_{\mathbf{w}} L|_{\mathbf{w}=\mathbf{w}_o} &= \mathbf{0} \Leftrightarrow \mathbf{S}_w \mathbf{w}_o = \tfrac{1}{2} \textstyle\sum_{i=1}^{N} \alpha_{i,o} y_i \mathbf{x}_i \\
\tfrac{\partial L}{\partial b}|_{b=b_o} &= 0 \Leftrightarrow \boldsymbol{\alpha}_o^T \mathbf{y} = 0 \\
\tfrac{\partial L}{\partial \xi_i}|_{\xi_i=\xi_{i,o}} &= 0 \Leftrightarrow \beta_{i,o} = C - \alpha_{i,o} \\
\beta_{i,o} \geq 0, 0 \leq \alpha_{i,o} &\leq C, \xi_{i,o} \geq 0, \beta_{i,o} \xi_{i,o} = 0 \\
y_i(\mathbf{w}_o^T \mathbf{x}_i + b_o) - 1 + \xi_{i,o} &\geq 0, \alpha_{i,o}\{y_i(\mathbf{w}_o^T \mathbf{x}_i + b_o) - 1 + \xi_{i,o}\} = 0
\end{aligned} \qquad (11)$$

the subscript o denotes the optimal case and $\mathbf{y} = \{y_1, \ldots, y_N\}$ is the vector denoting the class labels.

If the matrix $\mathbf{S}_w$ is invertible, i.e. feature dimensionality is less or equal to the number of samples minus two ($M \leq N - 2$), the optimal normal vector $\mathbf{w}$ of the hyperplane is given by (11):

$$\mathbf{S}_w \mathbf{w}_o = \frac{1}{2} \sum_{i=1}^{N} \alpha_{i,o} y_i \mathbf{x}_i \Leftrightarrow \mathbf{w}_o = \frac{1}{2} \mathbf{S}_w^{-1} \sum_{i=1}^{N} \alpha_{i,o} y_i \mathbf{x}_i. \qquad (12)$$

By replacing (12) into (10) and using the KKT conditions (11), the constraint optimization problem

[1]KKT conditions are necessary for a solution in nonlinear programming to be optimal. The necessary conditions for inequality constrained problem were first published in the Masters thesis of W. Karush [31], although they became renowned after a seminal conference paper by Harold W. Kuhn and Albert W. Tucker [32]. For SVM based optimization problems the interested reader may refer the tutorial paper [33].

(8) is reformulated to the Wolf dual problem:

$$\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = \mathbf{1}_N^T \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha}$$

$$\text{subject to } 0 \leq \alpha_i \leq C,\ i = 1, \ldots, N,\ \ \boldsymbol{\alpha}^T \mathbf{y} = 0$$

where $\mathbf{1}_N$ is a N-dimensional vector of ones and $[\mathbf{Q}]_{i,j} = \frac{1}{2} y_i y_j \mathbf{x}_i^T \mathbf{S}_w^{-1} \mathbf{x}_j$. It is worth noting here that, for the typical maximum margin SVMs problem [1], the matrix $\mathbf{Q}$ is $[\mathbf{Q}]_{i,j} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$. The corresponding decision surface is:

$$g(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x} + b) = \text{sign}\left(\frac{1}{2} \sum_{i=1}^{N} \alpha_{i,o} y_i \mathbf{x}_i^T \mathbf{S}_w^{-1} \mathbf{x} + b_o\right). \tag{13}$$

The optimal threshold b_o can be found by exploiting the fact that for all support vectors $\mathbf{x}_i$ with $0 < \alpha_{i,o} < C$, their corresponding slack variables are zero, according to the KKT condition (11). Thus, for any support vector $\mathbf{x}_i$ with $i \in \mathcal{S} = \{i : 0 < \alpha_i < C\}$ the following holds:

$$y_i(\frac{1}{2} \sum_{j=1}^{N} y_j \alpha_{j,o} \mathbf{x}_j^T \mathbf{S}_w^{-1} \mathbf{x}_i + b_o) = 1. \tag{14}$$

Averaging over these patterns yields a numerically stable solution for the bias term:

$$b_o = \frac{1}{N} \sum_{i \in \mathcal{S}} \left(y_i - \frac{1}{2} \sum_{j=1}^{N} y_j \alpha_{j,o} \mathbf{x}_j^T \mathbf{S}_w^{-1} \mathbf{x}_i\right). \tag{15}$$

As can be seen, an analytical solution for the optimal vector $\mathbf{w}_o$ is given only when the matrix $\mathbf{S}_w$ is invertible. In the following two Sections it will be shown that:

- solutions for the MCVSVMs can be found when the matrix $\mathbf{S}_w$ is singular, which is the typical case in small sample size problems (e.g., facial image classification problems) where the dimensionality is much larger than the number of available samples ($N \ll M$),
- the MCVSVMs can be defined and solved in reproducing Hilbert spaces in order to find the corresponding non-linear decision surfaces.

III. MCVSVM Hyperplanes in Small Sample Size Problems

When $\mathbf{S}_w$ is singular, the optimal normal vector $\mathbf{w}$ cannot be found directly from (12). In this case, it will be proven that, through dimensionality reduction using PCA [27], the optimization problem (8)

under the separability constraints (9) is reformulated into an equivalent one in a lower dimensional space, where the MCVSVMs optimization problem can be solved.

Let the total scatter matrix be defined as:

$$\mathbf{S}_t = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T = \mathbf{S}_w + \mathbf{S}_b. \tag{16}$$

It can be proven that $\mathbf{S}_t$ is bounded, compact, self-adjoint and positive operator in $\Re^M$ [20]. Thus, according to the Hilbert-Schmidt Theorem [35], its eigenvectors system is an orthonormal basis of $\Re^M$. Let $\mathcal{B}$ and $\mathcal{B}_\perp$ be the complementary M-dimensional spaces spanned by the orthonormal eigenvectors of $\mathbf{S}_t$ that correspond to non-zero eigenvalues and to zero eigenvalues, respectively. Thus, each vector $\mathbf{w} \in \Re^M$ can be represented as $\mathbf{w} = \boldsymbol{\varphi} + \boldsymbol{\zeta}$ with $\boldsymbol{\varphi} \in \mathcal{B}$ and $\boldsymbol{\zeta} \in \mathcal{B}_\perp$ [25], [20]. Let the linear mapping $L : \Re^M \rightarrow \mathcal{B}$ be defined as:

$$\mathbf{w} = \boldsymbol{\varphi} + \boldsymbol{\zeta} \rightarrow \boldsymbol{\varphi}. \tag{17}$$

It will be shown below that the optimization problem (8) subject to the constraints (9) can be solved in $\mathcal{B}$ instead of $\Re^M$.

Theorem. Under the mapping L the optimization problem (8) subject to the constraints (9) is equivalent to:

$$\min_{\boldsymbol{\varphi},b,\boldsymbol{\xi}} \boldsymbol{\varphi}^T \mathbf{S}_w \boldsymbol{\varphi} + C \sum_{i=1}^{N} \xi_i, \;\; \boldsymbol{\varphi}^T \mathbf{S}_w \boldsymbol{\varphi} > 0, \;\; \boldsymbol{\varphi} \in \mathcal{B} \tag{18}$$

subject to the constraints:

$$y_i(\boldsymbol{\varphi}^T \mathbf{x}_i + b) \geq 1 - \xi_i \;, \xi_i \geq 0, \;\; i = 1, \ldots, N, \;\; \boldsymbol{\varphi} \in \mathcal{B}. \; \square \tag{19}$$

A proof of the above Theorem can be found in Appendix I.

Thus, the above problem can be solved in a subspace isomorphic to $\mathcal{B}$. In order to find this subspace, the matrix $\boldsymbol{\Pi}$, with columns the orthonormal eigenvectors of $\mathbf{S}_t$ that correspond to non-null eigenvalues will be used. The number of these eigenvectors is $K \leq N - 1$. In case that the training samples are linearly independent, $K = N - 1$. In many problems (e.g., facial image characterization problems) it can be safely assumed that the initial training vectors are linearly independent [20], [27].

Since the columns of $\mathbf{\Pi}$ form an orthonormal basis of $\Re^{N-1}$, the space $\mathcal{B}$ is isomorphic to the space $\Re^{N-1}$, under the PCA transform $\mathbf{\Pi}$:

$$\boldsymbol{\varphi} = \mathbf{\Pi}\boldsymbol{\eta},\ \boldsymbol{\eta} \in \Re^{N-1}, \tag{20}$$

which is an one-to-one mapping from $\mathcal{B}$ to $\Re^{N-1}$. Under this mapping the optimization problem (18) is equivalent to:

$$\min_{\boldsymbol{\eta},b,\boldsymbol{\xi}} \boldsymbol{\eta}^T \acute{\mathbf{S}}_w \boldsymbol{\eta} + C\sum_{i=1}^{N} \xi_i,\ \boldsymbol{\eta}^T \acute{\mathbf{S}}_w \boldsymbol{\eta} > 0,\ \boldsymbol{\eta} \in \Re^{N-1} \tag{21}$$

where $\acute{\mathbf{S}}_w$ is the within-class scatter matrix of the projected samples in $\Re^{N-1}$ and is given by $\acute{\mathbf{S}}_w = \mathbf{\Pi}^T \mathbf{S}_w \mathbf{\Pi}$. The separability constraints are reformulated as:

$$y_i(\boldsymbol{\eta}^T \acute{\mathbf{x}}_i + b) \geq 1 - \xi_i\ , \xi_i \geq 0,\ \ i = 1, \ldots, N,\ \boldsymbol{\eta} \in \Re^{N-1} \tag{22}$$

where $\acute{\mathbf{x}}_i = \mathbf{\Pi}^T \mathbf{x}_i$ are the projected training vectors in $\Re^{N-1}$. Thus, without losing any information it is feasible to solve the constraint optimization problem in $\Re^{N-1}$ and then move to $\Re^M$ using (20). Although, the new total scatter matrix $\acute{\mathbf{S}}_t = \mathbf{\Pi}^T \mathbf{S}_t \mathbf{\Pi}$, is not singular, the new within-class scatter matrix $\acute{\mathbf{S}}_w$ may be still singular, containing one null eigenvector. This happens due to the fact that in small sample size problems the rank of $\acute{\mathbf{S}}_t$ is $N-1$ while the rank of $\acute{\mathbf{S}}_w$ is $N-2$. Thus, in the $N-1$ space the $\acute{\mathbf{S}}_w$ is not invertible and contains one eigenvector that corresponds to null eigenvalue. The matrix $\acute{\mathbf{S}}_w$ should become invertible in order to find the MCVSVMs hyperplane. There are two alternatives to achieve this. In the first case, in order to satisfy the invertibility of the matrix $\acute{\mathbf{S}}_w$, the matrix $\mathbf{\Pi}$ is formed using the $N-2$ eigenvectors of $\mathbf{S}_t$. That is, along with the eigenvectors that correspond to null eigenvalues only the eigenvector that corresponds to the lowest non-zero eigenvalue is discarded. The alternative is to perform eigenanalysis to the singular $\acute{\mathbf{S}}_w$ and to remove the eigenvector that corresponds to null eigenvalue.

The optimization problem (21) subject to the separability constraints (22) can be solved using the KKT conditions and the Wolf dual problem (13) having now as matrix $[\mathbf{Q}]_{i,j} = \frac{1}{2} y_i y_j \acute{\mathbf{x}}_i \acute{\mathbf{S}}_w^{-1} \acute{\mathbf{x}}_j$, since the matrix $\acute{\mathbf{S}}_w$ is not singular. The optimal normal vector in $\Re^{N-2}$ is $\boldsymbol{\eta}_o = \frac{1}{2}\sum_{i=1}^{N} \alpha_{i,o} y_i \acute{\mathbf{S}}_w^{-1} \acute{\mathbf{x}}_i$. The

final decision hyperplane in $\Re^M$ is given by:

$$\begin{aligned} g(\mathbf{x}) &= \text{sign}(\mathbf{w}_o^T\mathbf{x} + b_o) = \text{sign}(\boldsymbol{\varphi}_o^T\mathbf{x} + b_o) = \text{sign}(\boldsymbol{\eta}_o^T\mathbf{\Pi}^T\mathbf{x} + b_o) = \\ &= \text{sign}\left(\tfrac{1}{2}\sum_{\alpha_{i,o}=0}^{N}\alpha_{i,o}y_i\mathbf{x}_i^T\mathbf{\Pi}\acute{\mathbf{S}}_w^{-1}\mathbf{\Pi}^T\mathbf{x} + b_o\right). \end{aligned} \tag{23}$$

For the choice of b_o, a strategy similar to the one used Section II can be followed.

Summarizing the procedure, the training phase includes an initial projection of the training samples to $\Re^{N-2}$ using $\mathbf{\Pi}$; the MCVSVMs optimization problem is solved in this reduced space; for the test phase when a test vector arrives for classification, it should be first projected to $\Re^{N-2}$ (using $\mathbf{\Pi}$) and finally classified using (23).

IV. MCVSVM Nonlinear Decision Surfaces

In this Section, the optimization problem of the nonlinear MCVSVM decision surfaces will be defined and solved. These decision surfaces are derived from the minimization of the within-class variance in a dot product Hilbert space $\mathcal{H}$ subject to separability constraints. The space $\mathcal{H}$ will be called feature space while the original $\Re^M$ space will be called input space [13].

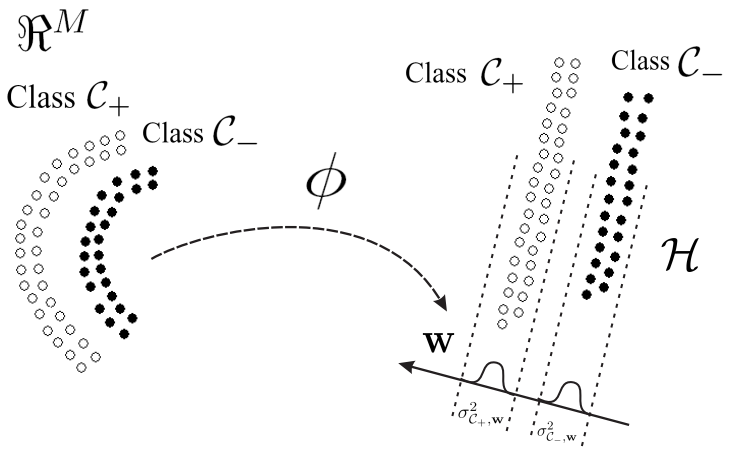


Fig. 3. Illustration of the non-linear MCVSVMs. Search for a direction $\mathbf{w}$ in the feature space $\mathcal{H}$, such that samples projected onto this dimension are separable and the variances ($\sigma^2_{C_k,\mathbf{w}}$ and $\sigma^2_{C_t,\mathbf{w}}$) of the projected samples are minimized.

Let us define the non-linear mapping $\phi : \Re^M \rightarrow \mathcal{H}$ that maps the training samples to the arbitrary dimensional feature space. In this paper, only the case in which the mapping ϕ satisfies the Mercer's

condition [1] will be considered. In the space $\mathcal{H}$ the within-class scatter is defined as:

$$\mathbf{S}_w^\Phi = \sum_{\mathbf{x}\in\mathcal{C}_-}(\phi(\mathbf{x})-\mathbf{m}_{\mathcal{C}_-}^\Phi)(\phi(\mathbf{x})-\mathbf{m}_{\mathcal{C}_-}^\Phi)^T + \sum_{\mathbf{x}\in\mathcal{C}_+}(\phi(\mathbf{x})-\mathbf{m}_{\mathcal{C}_+}^\Phi)(\phi(\mathbf{x})-\mathbf{m}_{\mathcal{C}_+}^\Phi)^T \tag{24}$$

the mean vector $\mathbf{m}_{\mathcal{C}_-}^\Phi$ is $\mathbf{m}_{\mathcal{C}_-}^\Phi = \frac{1}{N_{\mathcal{C}_-}}\sum_{\mathbf{x}\in\mathcal{C}_-}\phi(\mathbf{x})$ and the mean vector $\mathbf{m}_{\mathcal{C}_+}^\Phi$ is $\mathbf{m}_{\mathcal{C}_+}^\Phi = \frac{1}{N_{\mathcal{C}_+}}\sum_{\mathbf{x}\in\mathcal{C}_+}\phi(\mathbf{x})$.

The problem (8), in the feature space is to find a vector $\mathbf{w}\in\mathcal{H}$ such that:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \mathbf{w}^T\mathbf{S}_w^\Phi\mathbf{w} + C\sum_{i=1}^N \xi_i,\ \mathbf{w}^T\mathbf{S}_w^\Phi\mathbf{w} > 0 \tag{25}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i\ , \xi_i \geq 0,\ \ i=1,\ldots,N. \tag{26}$$

Figure 3 demonstrates the optimization problem in the feature space. The optimal decision surface is given by the minimization of a Lagrangian similar to the one in the linear case (10). The KKT conditions for the optimization problem (25) subject to the constraints (26) are similar to (11) (use $\mathbf{S}_w^\Phi$ instead of $\mathbf{S}_w$ and $\phi(\mathbf{x}_i)$ instead of $\mathbf{x}_i$). Since the feature space is of arbitrary dimension, the matrix $\mathbf{S}_w^\Phi$ is almost always singular. Thus, the optimal normal vector $\mathbf{w}_o$ cannot be directly found from:

$$\mathbf{S}_w^\Phi\mathbf{w}_o = \frac{1}{2}\sum_{i=1}^N \alpha_{i,o}y_i\phi(\mathbf{x}_i). \tag{27}$$

It will be proven, as in the linear case, that there is a solution to the optimization problem (25) subject to the constraints (26), by demonstrating that there is a mapping that makes this solution feasible. This mapping is the Kernel PCA (KPCA) transform.

Let us define the total scatter matrix $\mathbf{S}_t^\Phi$ in the feature space $\mathcal{H}$ as:

$$\mathbf{S}_t^\Phi = \sum_{i=1}^N(\phi(\mathbf{x}_i)-\mathbf{m}^\Phi)(\phi(\mathbf{x}_i)-\mathbf{m}^\Phi)^T = \mathbf{S}_w^\Phi + \mathbf{S}_b^\Phi. \tag{28}$$

where $\mathbf{m}^\Phi = \frac{1}{N}\sum_{\mathbf{x}}\phi(\mathbf{x})$. The matrix $\mathbf{S}_t^\Phi$ is bounded, compact, positive and self-adjoint operator in the Hilbert space $\mathcal{H}$. Thus, according to the Hilbert-Schmidt Theorem [35], its eigenvectors system is an orthonormal basis of $\mathcal{H}$. Let $\mathcal{B}^\Phi$ and $\mathcal{B}_\perp^\Phi$ be the complementary spaces spanned by the orthonormal

eigenvectors of $\mathbf{S}_t^{\Phi}$ that correspond to non-zero eigenvalues and to zero eigenvalues, respectively. Thus, any arbitrary vector $\mathbf{w} \in \mathcal{H}$, can be uniquely represented as $\mathbf{w} = \boldsymbol{\varphi} + \boldsymbol{\zeta}$ with $\boldsymbol{\varphi} \in \mathcal{B}^{\Phi}$ and $\boldsymbol{\zeta} \in \mathcal{B}_{\perp}^{\Phi}$.

It can be proven, using the same reasoning as in the linear case, that the optimal decision surface for the optimization problem (25) subject to the constraints (26) can be found in the reduced space $\mathcal{B}^{\Phi}$ spanned by the non-zero eigenvectors of $\mathbf{S}_t^{\Phi}$. The number of the non-zero eigenvectors of $\mathbf{S}_t^{\Phi}$ is $K \leq N-1$ thus, the dimensionality of $\mathcal{B}^{\Phi}$ is $K \leq N-1$ and according to the functional analysis theory [36] the space $\mathcal{B}^{\Phi}$ is isomorphic to the $(N-1)$-dimensional Euclidean space $\Re^{N-1}$. The isomorphic mapping is:

$$\boldsymbol{\varphi} = \mathbf{P}\boldsymbol{\eta},\ \boldsymbol{\eta} \in \Re^{N-1}, \tag{29}$$

where $\mathbf{P}$ is the matrix with columns the eigenvectors of $\mathbf{S}_t^{\Phi}$ that correspond to non-null eigenvalues and is an one-to-one mapping from $\Re^{N-1}$ onto $\mathcal{B}$.

Under this mapping the optimization problem is reformulated as:

$$\min_{\boldsymbol{\eta},b,\boldsymbol{\xi}} \boldsymbol{\eta}^T \tilde{\mathbf{S}}_w \boldsymbol{\eta} + C \sum_{i=1}^{N} \xi_i,\ \boldsymbol{\eta}^T \acute{\mathbf{S}}_w \boldsymbol{\eta} > 0,\ \boldsymbol{\eta} \in \Re^{N-1} \tag{30}$$

where $\tilde{\mathbf{S}}_w$ is the within-class scatter matrix of the projected vectors in $\Re^{N-1}$ given by $\tilde{\mathbf{S}}_w = \mathbf{P}^T \mathbf{S}_w^{\Phi} \mathbf{P}$ (KPCA transform). The equivalent separability constraints are:

$$y_i(\boldsymbol{\eta}^T \tilde{\mathbf{x}}_i + b) \geq 1 - \xi_i\ , \xi_i \geq 0,\ \ i = 1, \ldots, N,\ \boldsymbol{\eta} \in \Re^{N-1} \tag{31}$$

where $\tilde{\mathbf{x}}_i = \mathbf{P}^T \phi(\mathbf{x}_i)$ are the projected vectors in $\Re^{N-1}$ using the KPCA transform. For details on the calculation of the projections using the KPCA transform someone can refer to [15]. Under the projection to KPCA mapping, the optimal decision surface for the optimization problem (25) subject to (26) in $\mathcal{H}$ can be found by solving the optimization problem (30) subject to (31) in $\Re^{N-1}$. It is very interesting to notice here that now the problem falls in the linear MCVSVMs case (i.e., a linear MCVSVMs optimization should be solved) with dimensionality K equal to $N-1$. The problem here is that the matrix $\tilde{\mathbf{S}}_w$ may still be singular since the rank of $\tilde{\mathbf{S}}_t$ is at most $N-1$ and the rank of $\tilde{\mathbf{S}}_w$ is at most $N-2$. But, if the matrix $\tilde{\mathbf{S}}_w$ is singular it contains only one null dimension. Thus, in order

to satisfy the invertibility of $\tilde{\mathbf{S}}_w$ along with the null eigenvectors of $\mathbf{P}$, only one more eigenvector is discarded, which corresponds to lowest non-zero eigenvalue (as in the linear case).

Now that $\tilde{\mathbf{S}}_w$ is not singular the solution is derived in the same manner as in Section II. That is, the optimization problem (30) subject to the constraints (31) can be found by solving the Wolf dual problem (13) having as $[\mathbf{Q}]_{i,j} = \frac{1}{2} y_i y_j \tilde{\mathbf{x}}_i^T \tilde{\mathbf{S}}_w^{-1} \tilde{\mathbf{x}}_j$. The optimal normal vector of this problem is $\boldsymbol{\eta}_o = \frac{1}{2} \sum_{i=1}^{N} \alpha_{i,o} y_i \tilde{\mathbf{S}}_w^{-1} \tilde{\mathbf{x}}_i$. The decision surface in $\mathcal{H}$ is given by:

$$\begin{aligned} g(\mathbf{x}) &= \operatorname{sign}(\mathbf{w}_o^T \phi(\mathbf{x}_i) + b_o) = \operatorname{sign}(\boldsymbol{\varphi}_o^T \phi(\mathbf{x}) + b_o) = \operatorname{sign}(\boldsymbol{\eta}_o^T \mathbf{P}^T \phi(\mathbf{x}) + b_o) = \\ &= \operatorname{sign}\left(\tfrac{1}{2} \sum_{i=1}^{N} \alpha_{i,o} y_i \phi(\mathbf{x}_i)^T \mathbf{P} \tilde{\mathbf{S}}_w^{-1} \mathbf{P}^T \phi(\mathbf{x}) + b_o \right) \end{aligned} \tag{32}$$

for the optimal choice of b_o a similar strategy to Section II can be followed.

Summarizing, in order to find the optimal decision surface derived from the optimization problem (25) subject to the constraints (26), the training samples should be projected to $\Re^{N-2}$ using the KPCA transform (matrix $\mathbf{P}$) and solve a linear MCVSVMs problem there; for the test phase when a sample $\mathbf{x}$ arrives for classification it should be first projected to $\Re^{N-2}$ using the KPCA transform (matrix $\mathbf{P}$) and afterwards classified using (32).

V. Relationship with Other Decision Surfaces

In this Section a discussion about the relationship of the proposed approach with other classifiers like SVMs [1], CKFDA [20] and the decision surfaces proposed in [6] will be given. This discussion will also lead to some explanations about the constraint $\mathbf{w}^T \mathbf{S}_w \mathbf{w} > 0$ that has been employed in the optimization problem (7).

A. Relation with SVMs

Let the within-class scatter matrix $\mathbf{S}_w$ for a certain training set be invertible, then by letting $\boldsymbol{\tau} = \sqrt{2} \mathbf{S}_w^{\frac{1}{2}} \mathbf{w}$ the optimization problem (8) is equivalent to:

$$\min_{\boldsymbol{\tau}, b, \boldsymbol{\xi}} \frac{1}{2} \boldsymbol{\tau}^T \boldsymbol{\tau} + C \sum_{i=1}^{N} \xi_i, \ (||\boldsymbol{\tau}||^2 > 0) \tag{33}$$

as can be seen the constraint $||\boldsymbol{\tau}||^2 > 0$ is equivalent to $\mathbf{w}^T\mathbf{S}_w\mathbf{w} > 0$ in (7). The separability constraints are:

$$y_i(\boldsymbol{\tau}^T\mathbf{z}_i + b) \geq 1 - \xi_i \ , \xi_i \geq 0, \ \ i = 1, \ldots, N \tag{34}$$

where $\mathbf{z}_i = \frac{1}{\sqrt{2}}\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_i$ and $\mathbf{S}_w^{\frac{1}{2}}, \mathbf{S}_w^{-\frac{1}{2}} \in \Re^{M\times M}$ since the matrices $\mathbf{S}_w, \mathbf{S}_w^{-1}$ are real and positive definite matrices. Then, the solution of the optimization (8) subject to the constraints (9) is found by using the Wolf dual problem (13) having as:

$$[\mathbf{Q}]_{i,j} = y_i y_j \frac{1}{2}\mathbf{x}_i^T\mathbf{S}_w^{-1}\mathbf{x}_j = y_i y_j (\frac{1}{\sqrt{2}}\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_i)^T(\frac{1}{\sqrt{2}}\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_j) = y_i y_j \mathbf{z}_i^T\mathbf{z}_j \tag{35}$$

which is a Wolf dual problem of the maximum margin SVMs [1].

It can be easily verified that the within-class scatter matrix of the $\mathbf{z}_i$ is equal to $\frac{1}{2}\mathbf{I}$ where $\mathbf{I}$ is the $M \times M$ identity matrix. From the above analysis it can be verified that the problem (33) subject to the constraints (34) is equivalent to a maximum margin SVMs problem [1] in a transformed space with within-class scatter matrix equal to $\frac{1}{2}\mathbf{I}$. Thus, MCVSVMs converge to maximum margin SVMs when the within-class scatter matrix of the data tends to $\frac{1}{2}\mathbf{I}$. Hence all the useful theoretical properties (i.e., minimization of the structural risk, unique solution) of the typical linear SVMs hold as well for the MCVSVMs.

It should be noted here that, if the condition $\mathbf{w}^T\mathbf{S}_w\mathbf{w} = 0$ holds for the normal vector $\mathbf{w}$, then the previous analysis does not hold for the decision hyperplanes/surfaces that are defined by these normal vectors (i.e., they cannot be fitted in the SVMs framework).

B. Relationship with Complete Kernel Fisher Discriminant Analysis

In this section, the relationship of the proposed decision hyperplanes/surfaces with the ones derived through CKFDA [20] is analyzed. Moreover, we will indicate some important aspects of CKFDA that has not been treated in [20]. Only the linear case will be considered, in our discussion, since the non-linear case is a direct generalization of the linear one using Mercer's kernels.

As it has been proven in the Theorem in Section III, in order to solve the linear or the generalized non-linear constraint optimization problems of MCVSVMs, the solution space can be mapped in $\Re^{N-1}$ using PCA or KPCA in the linear or the non-linear case, respectively. Afterwards, a linear optimization problem is solved.

In the linear case, presented in Section III, in order to move from $\Re^{N-1}$ to $\Re^{N-2}$ we have removed one column from the matrix $\mathbf{\Pi}$ which is the eigenvector that corresponds to the lowest non-zero eigenvalue of $\mathbf{S}_t$. If this column is not removed from $\mathbf{\Pi}$, then $\acute{\mathbf{S}}_w = \mathbf{\Pi}^T \mathbf{S}_w \mathbf{\Pi}$ contains one eigenvector $\boldsymbol{\rho}$ that corresponds to a null eigenvalue. Let $\boldsymbol{\upsilon} \in \Re^M$ be $\boldsymbol{\upsilon} = \mathbf{\Pi}\boldsymbol{\rho}$, then, under the projection to $\boldsymbol{\upsilon}$, all the training samples are separated without an error, while $\boldsymbol{\upsilon}^T \mathbf{S}_w \boldsymbol{\upsilon} = 0$. In other words, the canonical decision hyperplane $g(\mathbf{x}) = \text{sign}(\boldsymbol{\upsilon}^T \mathbf{x} - c)$ (where $c = (\boldsymbol{\upsilon}^T \mathbf{x}_i + \boldsymbol{\upsilon}^T \mathbf{x}_j)/2$ with $\mathbf{x}_i \in \mathcal{C}_+$ and $\mathbf{x}_j \in \mathcal{C}_-$) satisfies the separability criterion (6) while for the normal vector $\boldsymbol{\upsilon}$, $\boldsymbol{\upsilon}^T \mathbf{S}_w \boldsymbol{\upsilon} = 0$ and $\boldsymbol{\upsilon}^T \mathbf{S}_t \boldsymbol{\upsilon} > 0$. That is, $\boldsymbol{\upsilon}$ is a solution of the optimization problem (7) subject to separability constraints (6) if the constraint $\boldsymbol{\upsilon}^T \mathbf{S}_w \boldsymbol{\upsilon} > 0$ has been removed. This fact is proven in Appendix II. Figure 4 describes pictorially the effects of the vectors $\mathbf{w}$ for the case, $\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 0$ and $\mathbf{w}^T \mathbf{S}_t \mathbf{w} > 0$.

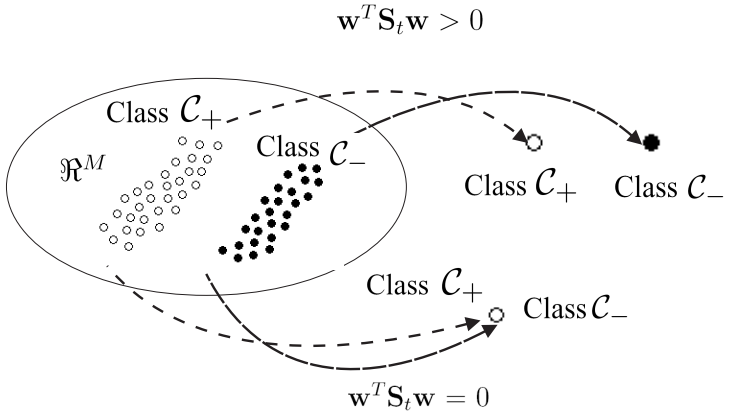


Fig. 4. Illustration of the effect of the projection to a vector $\mathbf{w}$ with $\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 0$. If $\mathbf{w}^T \mathbf{S}_t \mathbf{w} > 0$ is valid for the vector $\mathbf{w}$ then all the training vectors of the different classes are projected to one vector different for each class, while if $\mathbf{w}^T \mathbf{S}_t \mathbf{w} = 0$ all the training vectors are projected to the same point.

It is interesting to notice that the vector $\boldsymbol{\upsilon}$ is the one given by the irregular discriminant projection

defined in [20], [25] in case of a two class problem. That is, the vector $\boldsymbol{\upsilon}$ is the solution of the optimization problem:

$$\begin{aligned} &\max_{\mathbf{w}\in\Re^M} \mathbf{w}^T\mathbf{S}_b\mathbf{w}\,(||\mathbf{w}||=1) \\ &\text{subject to } \mathbf{w}^T\mathbf{S}_w\mathbf{w}=0, \end{aligned} \tag{36}$$

which is also a maximization point of the Fisher's discriminant ratio:

$$J(\mathbf{w})=\frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{37}$$

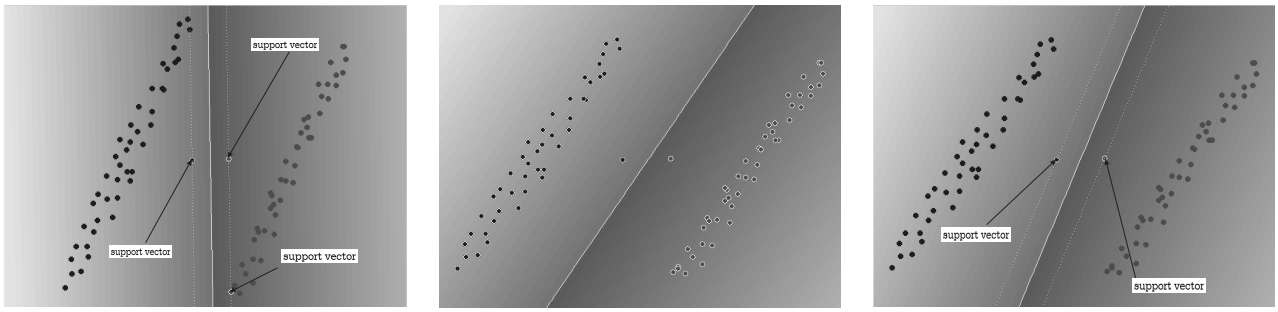
that makes $J(\boldsymbol{\upsilon})\rightarrow+\infty$. This interesting attribute of the irregular discriminant projections (i.e., the ones that satisfy $\mathbf{w}^T\mathbf{S}_w\mathbf{w}=0$ while $\mathbf{w}^T\mathbf{S}_t\mathbf{w}\neq 0$) that provide perfect separability in the training set has not been discussed in [20]. Summarizing the constraint $\mathbf{w}^T\mathbf{S}_w\mathbf{w}>0$ is included in the MCVSVMs optimization problem (7) and (8) due to the fact that:

1) The vectors $\mathbf{w}$, with $\mathbf{w}^T\mathbf{S}_w\mathbf{w}=0$ cannot be fitted in the SVMs framework (Section V-A).
2) The interesting vector $\mathbf{w}$ with $\mathbf{w}^T\mathbf{S}_w\mathbf{w}=0$ that satisfies the separability criteria (6) can be found by eigenanalysis only (Section V-B) and not by solving a quadratic optimization problem.

We can now conclude that MCVSVMs method is a compromise between FLDA and maximum margin SVMs.

C. Relationship with the Decision Surfaces in [6]

Finally for completeness, a note about the decision surfaces proposed in [6] will be made. These decision surfaces have been inspired by the solution of the linear case where the term $\mathbf{x}_i^T\mathbf{S}_w^{-1}\mathbf{x}_j$ is employed in the dual optimization problem (13). This term has been expressed as an inner product of the form $(\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_i)^T(\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_j)$, since $\mathbf{S}_w$ is a positive definite matrix (assuming that the original within-class scatter matrix of the data is not singular). Then, in [6] instead of projecting $\mathbf{x}_i$ using ϕ, the transformed vector $\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_i$ has been projected in the Hilbert space using ϕ and the matrix $[\mathbf{Q}]_{i,j}=\frac{1}{2}y_iy_jk(\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_i,\mathbf{S}_w^{-\frac{1}{2}}\mathbf{x}_j)$ is used for solving the dual optimization problem, where $k(\mathbf{x},\mathbf{y})=\phi(\mathbf{x})^T\phi(\mathbf{y})$ is the kernel function. Of course, the decision surface provided in [6] is not the solution



(a) (b) (c)

Fig. 5. a) The maximum margin SVM hyperplane; b) The hyperplane of FLDA; c) the MCVSVM hyperplane

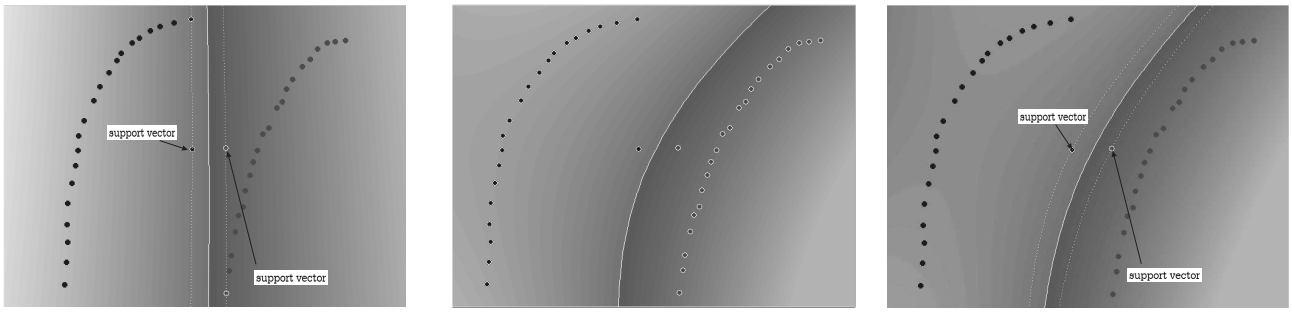
of the optimization problem of MCVSVMs in Hilbert spaces (optimization problem (25) subject to (26)).

VI. Experimental Results

A. Experiments with Artificial Data

Artificial data have been used in order to show that the proposed MCVSVM hyperplanes and surfaces are not so sensitive to outliers as the ones defined by the maximum margin SVMs. A comparison of the linear maximum margin SVMs against the linear MCVSVMs in the separable case is shown in Figure 5. The advantage of the MCVSVMs method is that it takes into account both the class distribution statistics and the vectors that are in the boundaries, in contrast to the maximum margin SVMs that considers only the vectors that lie in the boundaries.

In the case of a non-linear decision surface the suitability of the proposed approach against the maximum margin SVMs can be seen in Figure 6. The SVMs approach totally failed to capture the nonlinearity of the data (Figure 6a). The KFDA based surface (Figure 6b) that considers the class distribution captured the nonlinearity of the data. The proposed MCVSVMs captured the underlying non-linearity of the data (Figure 6c).



(a) (b) (c)

Fig. 6. The optimal decision surface using second order polynomial kernel and (a) maximum margin SVM, (b) regular CKFDA in [20] and (c) the proposed MCVSVM.

B. Experiments on Gender Determination using the XM2VTS database

Experiments were conducted using real data from the XM2VTS database [37] for testing the proposed algorithm to the gender determination problem. The luminance information at a resolution of 720×576 has been considered in our experiments. The images were aligned using fully automatic alignment according to the eyes position coordinates that have been derived by the method reported in [38]. The facial region has been detected using the face localization and normalization method proposed in [39]. The resolution of the resulting "face-prints" was 85×156. As in the gender determination experiments in [4], little or no hair information has been present in the training and the test facial images. The power of the proposed approach is demonstrated against the maximum margin SVMs [1] and the CKFDA framework proposed in [20].

A total of 2360 "face-prints" (1256 males and 1104 females images) have been used for our experiments. For each classifier, the average error rate was estimated with five-fold cross validation. That is, a five-way data set split with $\frac{4}{5}$-th used for training and $\frac{1}{5}$-th used for testing, with four subsequent non-overlapping data permutations. The average size of the training set has been 1888 facial images (1005 male images and 883 female images) and the average size of the test set has been 472 images (251 male images and 221 female images). The persons that have been included in the training set has been excluded from the test set. The overall error rate has been measured as

$E = \frac{N_e}{N_t}$ where N_e is the total number of classification errors for the test sets in all data permutations and N_t is the total number of the test images (here $N_t = 4 \times 472$).

A similar experimental setup has been used in gender determination experiments in [4], where it has been shown that maximum margin SVMs outperform several other classifiers in this problem. The interested reader may refer to [4] and to the references therein for more details on the gender determination problem. For the experiments using the maximum margin SVMs, the methodology presented in [4] has been used. That is, several kernels have been used in the experiments and the parameter C has been set to infinity so that no training errors were allowed. The typical kernels that have been used in our experiments have been polynomial and Radial Basis Functions (RBF) kernels:

$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^d \tag{38}$$

$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}) = e^{-\gamma (\mathbf{x}-\mathbf{y})^T (\mathbf{x}-\mathbf{y})}$$

where d is the degree of the polynomial and γ is the spread of the Gaussian kernel.

The quadratic optimization problem of SVMs has been solved using a decomposition similar to [5]. For the proposed method the original $85 \times 156 = 13260$ dimensional facial image space has been projected to a lower dimensional image space using the strategy described in Sections III and IV and afterwards the quadratic optimization problem of MCVSVMs is solved. For CKFDA the regular and the irregular discriminant projections are found using the method proposed in [20]. That is, two classifiers were obtained, one that corresponds to regular discriminant information and another one that corresponds to the irregular discriminant information. In the conducted experiments the irregular discriminant information, even though it has no errors in the training set it has lead to over 15% overall error rate in the test sets. Thus, irregular discriminant information has not been used in the CKFDA method.

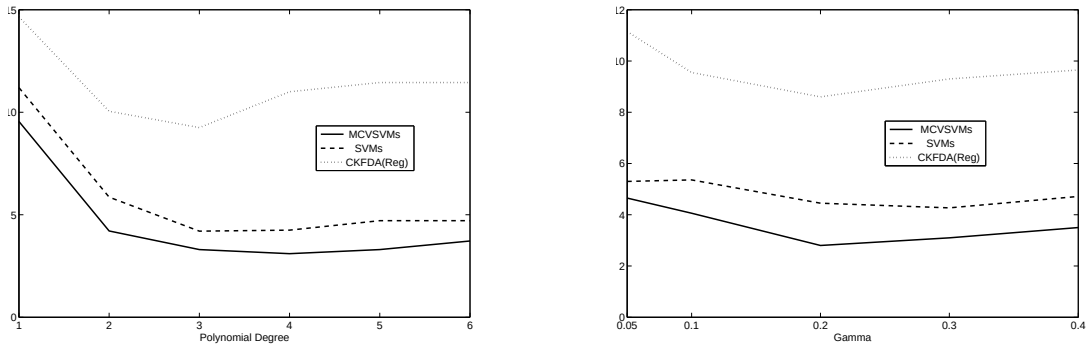
The experimental results with various kernels and parameters are shown in Figure 7. As can been seen in this Figure the error rates for the MCVSVMs are constantly lower than those achieved for the other tested classifiers for all the tested kernels and parameters. Some of the support faces used

TABLE I

THE BEST ERROR RATES OF THE TESTED CLASSIFIERS AT GENDER DETERMINATION.

Algorithm	Overall %	Male %	Female %
MCVSVMs with Gaussian RBF kernel	**2.86**	**2.19**	**3.5**
MCVSVMs with cubic polynomial kernel	3.28	2.98	3.62
SVMs with Gaussian RBF kernel	4.4	3.48	5.43
SVMs with cubic polynomial kernel	4.4	3.48	5.43
Regular CKFDA with Gaussian RBF kernel	8.58	8.17	9
Regular CKFDA with cubic polynomial kernel	9.27	8.17	10.4

for constructing the non-linear MCVSVM surfaces are shown in Figure 8. The lowest error rates for the tested classifiers are summarized in Table I. The best error rate for the MCVSVMs have been 2.86% while for SVMs have been 4.4%. Confusion matrices for the best case of MCVSVMs and SVMs can be found in Tables IV and V, respectively. Finally, statistical analysis of the results can be found in Section VI-E.



(a) (b)

Fig. 7. Average error rates for gender determination using various kernels; a) polynomial kernel b) RBF kernel.

(a) (b)

Fig. 8. Some of the Support faces used by the polynomial MCVSVMs of degree 3 a) Support men; b) Support women.

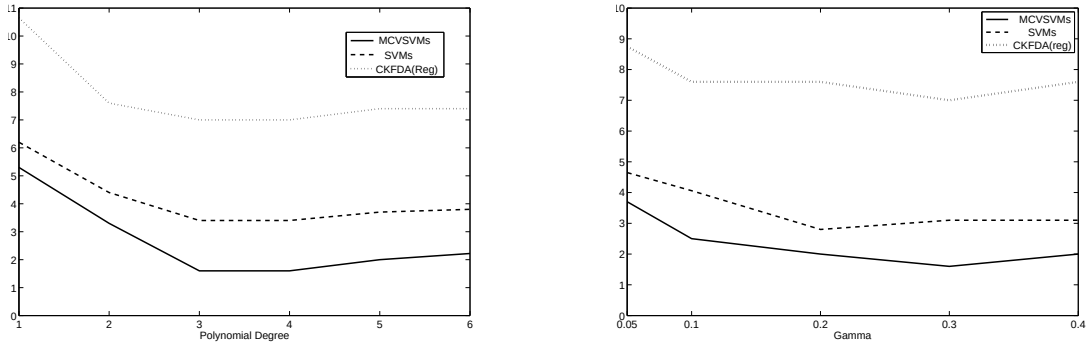
C. Eyeglass Detection using the XM2VTS database

The proposed algorithm has been also tested in eye-glass detection from facial images. The output of the eye-glass detection can be used in order to assist eye-glass removal algorithms [40] and/or in order to assist face verification systems in reducing their false rejections, by asking the client to remove his eyeglasses during the verification procedure. The procedure described for the gender determination experiments has been also followed in eyeglass detection. From the total of 2360 "face-prints" of the XM2VTS database, 1518 are facial images with eye-glasses and the 842 without eye-glasses. The average size of the training set has been 1888 facial images (1215 images with eye-glasses and 673 images without eye-glasses) and the average size of the test set has been 472 images (303 facial images with eye-glasses and 169 without eye-glasses).

Figure 9 shows the experimental results with various kernels and parameters. The best experimental results for the tested classifiers are summarized in Table II. As can be seen, the proposed non-linear MCVSVMs technique outperforms all the other tested classifiers in eyeglass detection as well. Confusion matrices for the best case of MCVSVMs and SVMs can be found in Tables IV and V, respectively. Finally, statistical analysis of the results can be found in Section VI-E.

D. Neutral Facial Expression Detection using Cohn-Kanade database

The final experiment illustrates the application of the MCVSVMs to the neutral facial expression detection problem. Gabor-based feature have been used for this specific problem [30]. The recognition



(a) (b)

Fig. 9. Experimental results for eyeglass detection using various kernels; a) polynomial kernel b) RBF kernel.

TABLE II

THE BEST ERROR RATES OF THE TESTED CLASSIFIERS AT EYEGLASS DETECTION.

Algorithm	Overall Error Rate%
MCVSVMs with Gaussian RBF kernel	**1.6**
MCVSVMs with 4-th degree polynomial kernel	**1.6**
SVMs with Gaussian RBF kernel	2.8
SVMs with 4-th degree polynomial kernel	3.4
Regular CKFDA with Gaussian RBF kernel	7
Regular CKFDA with 4-th degree polynomial kernel	7

of the neutral facial expression can be used in order to assist face verification algorithms [41], that, in general, are sensitive to the change of facial expressions and ask the client to have a neutral facial expression when using the verification system.

The Cohn-Kanade database [42] was used for the facial expression recognition in 6 basic facial expressions (anger, disgust, fear, happiness, sadness and surprise) classes. This database, is anottated with Facial Action Units (FAUs). These combinations of FAUs were translated into facial expressions, in order to define the corresponding ground truth for the facial expressions. In order to form the dataset to be used for the experiments, every image sequence available was taken under consideration, for

every subject (96 subjects in total). One image for the neutral state and one image for the fully intensed facial expression were chosen from each image sequence (first and last frame of the image sequence respectively). Not all six facial expressions were present for every subject. For example a subject may have three video sequences posing happiness and none posing sadness, thus creating 3 samples for the happiness facial expression and 3 samples for the neutral facial expression, but none for the sadness facial expression. The chosen images were used to build the database, consisting of 704 images (equal number of samples for the neutral and fully expressive images). In Figure 10, a sample of image sequences of one poser from this database, is shown.

The same procedure, as in the previous experiments, has been used for measuring the performance of the tested classifiers. That is, from the total of 704 ”face-prints” of the Cohn-Kanade database the 352 are neutral facial images while the remaining 352 are expressive images. The average size of the training set has been 564 facial images (282 expressive and 282 neutral images) and the average size of the test set has been 141 images (70.5 neutral and 70.5 expressive images).

Fig. 10. Neutral Vs Expressive Images of a poser of Kanade database

Figure 11 shows the results of the regular CKFDA, SVMs, and MCVSVMs approach for the polynomial kernel and for various degrees. As can be seen MCVSVMs approach is constantly better than SVMs and CKFDA for all the tested polynomial kernels. The lowest error rates are summarized in Table III. The Confusion matrices for MCVSVMs and SVMs in neutral state detection can be found in Tables IV and V, respectively. Finally, statistical analysis of the results can be found in Section VI-E.

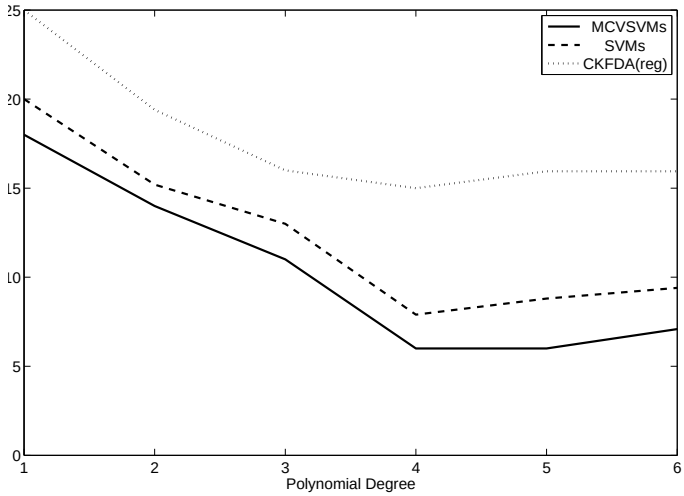


Fig. 11. Experimental results for neutral detection determination using polynomial kernel with various degrees.

TABLE III

THE BEST ERROR RATES OF THE TESTED CLASSIFIERS FOR NEUTRAL STATE DETECTION.

Algorithm	Overall Error Rate%
MCVSVMs with 4-th degree polynomial kernel	**6**
SVMs with 4-th degree polynomial kernel	7.9
Regular CKFDA 4-th degree polynomial kernel	14

E. Statistical Significance of Results

In order to calculate if the difference in performance is not just numerical but also statistically significant, the McNemar's test [43], [44] has been used. McNemar's test is a null hypothesis statistical test based on a Bernoulli model. If the resulting p-value is below a desired significance level (for example, 0.02), the null hypethesis is rejected and the performance difference between two algorithms is considered to be statistically significant. The McNemar's test has been widely used in order to estimate the statistical significance between recognition algorithms [20], [45]. We have used the best cases of SVMs and MCVSVMs in all experiments in order to measure the significance and it has

TABLE IV

CONFUSION MATRICES FOR THE BEST RESULTS OF MCVSVMS FOR A) GENDER DETERMINATION B)EYEGLASS DETECTION C) NEUTRAL STATE DETERMINATION

$lab_{cl} \backslash lab_{ac}$	male	female
male	982	22
female	31	853

$lab_{cl} \backslash lab_{ac}$	eyeglass	no-eyeglass
eyeglass	1193	22
no-eyeglass	8	665

$lab_{cl} \backslash lab_{ac}$	neutral	expressive
neutral	268	14
expressive	20	262

TABLE V

CONFUSION MATRICES FOR THE BEST RESULTS OF SVMS FOR A) GENDER DETERMINATION B) EYEGLASS DETECTION C) NEUTRAL STATE DETERMINATION

$lab_{cl} \backslash lab_{ac}$	male	female
male	969	35
female	48	836

$lab_{cl} \backslash lab_{ac}$	eyeglass	no-eyeglass
eyeglass	1172	43
no-eyeglass	10	663

$lab_{cl} \backslash lab_{ac}$	neutral	expressive
neutral	257	25
expressive	20	262

been calculated that $p \ll 0.02$. Thus, the difference in performance, for the best cases, is statistically significant.

Apart from measuring the significance of the best results we have measured the significance in terms of mean classification rate. To do so, we have used the method in [46]. We have measured that there is statistical significant difference between the mean classification rate of SVMs and MCVSVMs in the gender determination experiments for the tested parameters in the nonlinear case (all polynomial kernels with degrees from 2 to 6 and RBF kernel parameters). This also holds for eyeglass detection for all the tested parameters (all polynomials and RBF kernel parameters). According to the presented experiments we could not conclude that the difference in performance, according to mean recognition rate, between MCVSVMs and SVMs is statistical significant for the neutral state recognition experiments.

Finally, we have measured the sparseness of the MCVSVMs solution. A machine learning algorithm yields a sparse result when, among all the coefficients that describe the model, only a small number are non-zero [1], [47]. In statistical learning theory, sparsity is related to statistical robustness and fast optimization. In order to have insights concerning the sparsity of the approaches we have measured the minimum and maximum number of Support Vectors (SVs) in every experimental setup for SVMs and MCVSVMs. For MCVSVMs the number of SVs is measured from the solution of their optimization problem, i.e. after the application of PCA or KPCA. From the conducted experiments it has been verified that MCVSVMs are as sparse as SVMs in the specific applications.

VII. Conclusions

A novel class of decision hyperplanes and surfaces, the so-called MCVSVMs, inspired from the Fisher's discriminant ratio and SVMs has been proposed. Solutions for the MCVSVMs in cases when the training set contains less and more samples that the feature dimensionality have been described. Moreover, kernels have been employed in order to define MCVSVM nonlinear decision surfaces. The relationship of MCVSVMs with SVMs and FDA has been discussed and it has been indicated both theoretically and by using artificial data that MCVSVMs are a compromise between maximum margin SVMs and FDA classifiers. It is believed that the proposed classifiers have the advantages of both SVMs and FDA Finally, the described experiments have shown that the proposed class of decision surfaces outperforms SVMs and CKFDA in gender determination, eyeglass and neutral state detection from facial images. Topics for further research on this subject include the incorporation of robust statistics [48], [49], [50] for the calculation of the within-class scatter matrix in order to cope with the presence of possible outliers in the class distributions. Another potential topic for further research is to meticulously study the generalization ability of the proposed classifiers by carefully combining the results in [51], where the generalization ability of KPCA is discussed, with the results in [52], where the generalization of soft-SVM classifiers is measured.

VIII. Acknowledgment

This work was supported by the project 03ED849 co-funded by the European Union and the Greek Secretariat of Research and Technology (Hellenic Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support Framework.

Appendix I

Proof of Theorem in Section III

Since $\mathbf{S}_b$ and $\mathbf{S}_w$ are both positive and $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$, it is easy to verify that: $\boldsymbol{\zeta}^T\mathbf{S}_t\boldsymbol{\zeta} = 0$ for $\boldsymbol{\zeta} \in \Re^M$ if and only if $\boldsymbol{\zeta}^T\mathbf{S}_w\boldsymbol{\zeta} = 0$ and $\boldsymbol{\zeta}^T\mathbf{S}_b\boldsymbol{\zeta} = 0$ (or equivalently $\mathbf{S}_t\boldsymbol{\zeta} = \mathbf{0}$ if and only if $\mathbf{S}_w\boldsymbol{\zeta} = \mathbf{0}$ and $\mathbf{S}_b\boldsymbol{\zeta} = \mathbf{0}$). Let $\mathcal{B}$ and $\mathcal{B}_\perp$ be the complementary spaces spanned by the orthonormal eigenvectors of $\mathbf{S}_t$ that correspond to non-zero to zero eigenvalues, respectively. Since $\mathcal{B}_\perp$ is the null space of $\mathbf{S}_t$ for every $\boldsymbol{\zeta} \in \mathcal{B}_\perp$ it is valid that $\boldsymbol{\zeta}^T\mathbf{S}_t\boldsymbol{\zeta} = 0$ (every $\boldsymbol{\zeta}$ can be written, in a unique manner, as a linear combination of the orthonormal eigenvectors of $\mathbf{S}_t$ that correspond to zero eigenvalues).

Since, $\mathbf{S}_t$ is a compact self-adjoint and positive operator in $\Re^M$ any $\mathbf{w} \in \Re^M$ can be written as $\mathbf{w} = \boldsymbol{\varphi} + \boldsymbol{\zeta}$. Hence,

$$\mathbf{w}^T\mathbf{S}_w\mathbf{w} = \boldsymbol{\varphi}^T\mathbf{S}_w\boldsymbol{\varphi} + 2\boldsymbol{\zeta}^T\mathbf{S}_w\boldsymbol{\varphi} + \boldsymbol{\zeta}^T\mathbf{S}_w\boldsymbol{\zeta} = \boldsymbol{\varphi}^T\mathbf{S}_w\boldsymbol{\varphi} \tag{39}$$

Using the previous facts the Lagrangian (10) can be written as:

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) &= \mathbf{w}^T\mathbf{S}_w\mathbf{w} + C\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^N \beta_i\xi_i \\ &= \boldsymbol{\varphi}^T\mathbf{S}_w\boldsymbol{\varphi} + C\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i[y_i(\boldsymbol{\varphi}^T\mathbf{x}_i + \boldsymbol{\zeta}^T\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^N \beta_i\xi_i. \end{aligned} \tag{40}$$

If for some $\boldsymbol{\zeta} \in \mathcal{R}^M$, $\boldsymbol{\zeta}^T\mathbf{S}_t\boldsymbol{\zeta} = 0$ then under the projection $\boldsymbol{\zeta}$, for all training vectors $\mathbf{x}_i, \mathbf{x}_j$ with $\mathbf{x}_i \neq \mathbf{x}_j$ then $\boldsymbol{\zeta}^T\mathbf{x}_i = \boldsymbol{\zeta}^T\mathbf{x}_j$. In other words, all the training vectors $\mathbf{x}_i$ fall in the same point under the projection $\boldsymbol{\zeta}$. Thus, $r = \boldsymbol{\zeta}^T\mathbf{x}_i$ is a constant $\forall\ \mathbf{x}_i$. Now, using the KKT condition $\boldsymbol{\alpha}_o^T\mathbf{y} = 0$ the following is valid:

$$\sum_{i=1}^N \alpha_i y_i \boldsymbol{\zeta}^T\mathbf{x}_i = \sum_{i=1}^N \alpha_i y_i r = r\sum_{i=1}^N \alpha_i y_i = 0. \tag{41}$$

Hence, the Lagrangian (40) can be written as:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) \quad = \boldsymbol{\varphi}^T \mathbf{S}_w \boldsymbol{\varphi} + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [y_i (\boldsymbol{\varphi}^T \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{N} \beta_i \xi_i. \quad (42)$$

The optimum hyperplane $\mathbf{w}_o \in \Re^M$ can be written, in a unique manner, as $\mathbf{w}_o = \boldsymbol{\varphi}_o + \boldsymbol{\zeta}_o$ ($\boldsymbol{\varphi}_o \in \mathcal{B}$ and $\boldsymbol{\zeta}_o \in \mathcal{B}_\perp$) and then using the chain rule it can be easily shown that:

$$\nabla_{\mathbf{w}} L|_{\mathbf{w}=\mathbf{w}_o} = \nabla_{\boldsymbol{\varphi}} L|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_o} = \mathbf{0} \Leftrightarrow 2\mathbf{S}_w \boldsymbol{\varphi}_o - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \mathbf{0}. \quad (43)$$

Thus, the decision surface depends only on $\boldsymbol{\varphi}_o \in \mathcal{B}$ (an arbitrary vector $\boldsymbol{\zeta}_o$ can be chosen). The separability constraints (9) can be safely replaced by the separability constraints (18) and the Theorem has been proven.

APPENDIX II

PROOF OF PROPOSITION 1

Proposition 1. Let $\mathbf{S}_t$ and $\mathbf{S}_w$ be the total scatter and the within-class scatter matrix of a training set $\mathcal{U} = \{\mathbf{x} : \mathbf{x} \in \Re^M\}$ with finite number of elements. If for some $\boldsymbol{\rho} \in \Re^M$, $\boldsymbol{\rho}^T \mathbf{S}_t \boldsymbol{\rho} > 0$ and $\boldsymbol{\rho}^T \mathbf{S}_w \boldsymbol{\rho} = 0$ then the training samples under the projection $\boldsymbol{\rho}$ are separated without an error.

Proof: Since $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$ is not singular and positive, it follows that $\boldsymbol{\rho}^T \mathbf{S}_t \boldsymbol{\rho} = \boldsymbol{\rho}^T \mathbf{S}_b \boldsymbol{\rho} > 0$. Since, $\boldsymbol{\rho}^T \mathbf{S}_t \boldsymbol{\rho}$ the projection to $\boldsymbol{\rho}$ all the training vectors $\mathbf{x}_i \in \mathcal{C}_+$ fall in the same point, $a = \boldsymbol{\rho}^T \mathbf{x}_i$ and all the training vectors $\mathbf{x}_j \in \mathcal{C}_-$ fall in the point $c = \boldsymbol{\rho}^T \mathbf{x}_j$. Since $\boldsymbol{\rho}^T \mathbf{S}_b \boldsymbol{\rho} > 0$, $a \neq c$. Hence, under the projection $\boldsymbol{\rho}$ all the projected vectors are separated without an error.

REFERENCES

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

[2] M. Pontil and A. Verri, "Support vector machines for 3D object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637–646, 1998.

[3] A. Ganapathiraju, J.E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348 – 2355, 2004.

[4] B. Moghaddam and Y. Ming-Hsuan, "Learning gender with support faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 707–711, 2002.

[5] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Juan, Puerto Rico, 1997, pp. 130–136.

[6] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, 2001.

[7] H. Drucker, W. Donghui, and V.N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048 – 1054, 1999.

[8] B. Scholkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[10] V.N. Vapnik and A.J. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probability Appl.*, vol. 16, pp. 264 – 280, 1971.

[11] S. Saitoh, *Theory of Reproducing Kernels and its Applications*, Harlow, UK: Longman Scientific & Technical, 1988.

[12] R.C. Williamson, A.J. Smola, and B. Scholkopf, "Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators," *IEEE Transactions on Information Theory*, vol. 47, no. 6, pp. 2516–2532, 2001.

[13] B. Scholkopf, S. Mika, C.J.C Burges, P. Knirsch, K.-R. Muller, G Ratsch, and A.J. Smola, "Input space vs. feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[14] K.I. Diamantaras and S.Y. Kung, *Principal Component Neural Networks*, New York: Wiley, 1996.

[15] A. Scholkopf, B. Smola and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

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

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

[18] S. Mika, Ratsch G., J. Weston, B. Scholkopf, A. Smola, and K.-R. Muller, "Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 623 – 628, 2003.

[19] L. Juwei, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 117–126, 2003.

[20] J. Yang, A.F. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

vol. 27, no. 2, pp. 230–244, 2005.

[21] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York:John Wiley & Sons, 2001.

[22] F.R. Bach and M.I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1 – 48, 2002.

[23] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.

[24] K. Fukunaga, *Statistical Pattern Recognition*, CA: Academic, San Diego, 1990.

[25] J. Yang and J.-Y. Yang, "Why can LDA be performed in PCA transformed space?," *Pattern Recognition*, vol. 36, no. 2, pp. 563–566, 2003.

[26] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, 2005.

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

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

[29] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40–42, 2002.

[30] L. Chengjun, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572 – 581, May 2004.

[31] W. Karush, *Minima of Functions of Several Variables with Inequalities as Side Constraints*, M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

[32] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proceedings of 2nd Berkeley Symposium*, Berkeley, 1951, pp. 481–492.

[33] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Minining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[34] R. Fletcher, *Practical Methods of Optimization*, second ed. New York: John Wiley, 1987.

[35] V. Hutson and J.S. Pym, *Applications of Functional Analysis and Operator Theory*, London: Academic Press, 1980.

[36] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, 1978.

[37] K. Messer, J. Matas, J.V. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *AVBPA'99*, Washington, DC, USA, 22-23 March 1999, pp. 72–77.

[38] K. Jonsson, J. Matas, and Kittler, "Learning salient features for real-time face verification," in *Proc. Second International Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA'99)*, Washington D. C. USA,

March 22-23 1999, pp. 60–65.

[39] C. Kotropoulos, A. Tefas, and I. Pitas, "Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions," *Pattern Recognition*, vol. 33, no. 12, pp. 31–43, Oct. 2000.

[40] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xy, and Z. Zhang, "Automatic eyeglasses removal from face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 322 – 336, 2004.

[41] Y. Tian and R. M. Bulle, "Automatic detecting neutral face for face authentication," in *AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management*, California, USA, August 20-23 2003, pp. 24–26.

[42] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive databases for facial expression analysis," in *Proc. IEEE Inter. Conf. on Face and Gesture Recognition*, Grenoble, France, March 2000, pp. 46–53.

[43] J. Devore and R. Peck, *Statistics: The Exploration and Analysis of Data*, third ed. Brooks Cole, 1997.

[44] I. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153 – 157, 1947.

[45] B.A. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge, "Recognizing faces with PCA and ICA," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 115–137, 2003.

[46] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1550–1565, 1996.

[47] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, pp. 1455 – 1480, 1998.

[48] G. Seber, *Multivariate Observations*, New York: Wiley, 1986.

[49] I. Pitas and A.N. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications*, Kluwer Academic Publishers, Norwell, MA, 1990.

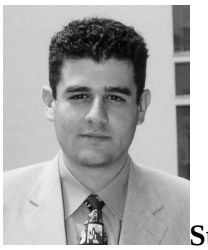
[50] A. G. Bors and I. Pitas, "Median radial basis function neural network," *IEEE Transactions on Neural Networks*, vol. 7, pp. 1351–1364, 1996.

[51] J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, and J. Kandola, "On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2510–2522, 2005.

[52] J. Shawe-Taylor and N. Cristianini, "On the generalization of soft margin algorithms," *IEEE Transactions on Information Theory*, vol. 48, no. 10, pp. 2721–2735, 2002.

LIST OF FIGURES

LIST OF TABLES



Stefanos Zafeiriou was born in Thessaloniki, Greece in 1981. He received the B.Sc, degree in Informatics with highest honors in 2003 and the Ph.D degree in Informatics in 2007, both from the Aristotle University of Thessaloniki, Thessaloniki, Greece. He has received various scholarships and awards during his undergraduate and Ph.D. studies. He has co-authored over than 20 journal and conference publications. He is currently a researcher and teaching assistant at the Department of Informatics at the Aristotle University of Thessaloniki. His current research interests lie in the areas of signal and image processing, computational intelligence, pattern recognition and computer vision.

Anastasios Tefas received the B.Sc. in informatics in 1997 and the Ph.D. degree in informatics in 2002, both from the Aristotle University of Thessaloniki, Greece. Since 2006, he has been an Assistant Professor at the Department of Information Management, Technological Educational Institute of Kavala. From 1997 to 2002, he was a researcher and teaching assistant in the Department of Informatics, University of Thessaloniki. From 2003 to 2004, he was a temporary lecturer in the Department of Informatics, University of Thessaloniki where he is currently, a senior researcher. He has co-authored over 50 journal and conference papers. His current research interests include computational intelligence, pattern recognition, digital signal and image processing, detection and estimation theory, and computer vision.

Ioannis Pitas received the Diploma of Electrical Engineering in 1980 and the PhD degree in Electrical Engineering in 1985 both from the Aristotle University of Thessaloniki, Greece. Since 1994, he has been a Professor at the Department of Informatics, Aristotle 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, as Visiting Assistant Professor at the University of Toronto and as Visiting Professor at the University of British Columbia, Vancouver, Canada. He was lecturer in short courses for continuing education. He has published over 150 journal papers, 440 conference papers and contributed in 22 books in his areas of interest. He is the co-author of the books Nonlinear Digital Filters: Principles and Applications (Kluwer, 1990), 3-D Image Processing Algorithms (J. Wiley, 2000), Nonlinear Model-Based Image/Video Processing and Analysis (J. Wiley, 2001) and author of Digital Image Processing Algorithms and Applications (J. Wiley, 2000). He is the editor of the book Parallel Algorithms and Architectures for Digital Image Processing, Computer Vision and Neural Networks (Wiley, 1993). He has also been an invited speaker and/or member of the program committee of several scientific conferences and workshops. In the past he served as Associate Editor of the IEEE Transactions on Circuits and Systems, IEEE Transactions on Neural Networks, IEEE Transactions on Image Processing, EURASIP Journal on Applied Signal Processing and co-editor of Multidimensional Systems and Signal Processing. He was general chair of the 1995 IEEE Workshop on Nonlinear Signal and Image Processing (NSIP95), technical chair of the 1998 European Signal Processing Conference and general chair of IEEE ICIP 2001. His current interests are in the areas of digital image and video processing and analysis, multidimensional signal processing, watermarking and computer vision.