16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

GENDER DETERMINATION USING A SUPPORT VECTOR MACHINE VARIANT

Stefanos Zafeiriou, Anastasios Tefas, and Ioannis Pitas

Artificial Intelligence and Information Analysis Lab/Department of Informatics, Aristotle University of Thessaloniki
University Campus, Agiou Dimitriou , 54124, Thessaloniki, Greece
phone: +302310996304, fax: +302310996304,
email: dralbert@.aiia.csd.auth.gr, tefas@aiia.csd.auth.gr, pitas@aiia.csd.auth.gr
web: www.aiia.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. Moreover, we present a novel class of nonlinear decision surfaces by solving the proposed optimization problem in arbitrary Hilbert spaces defined by Mercer's kernels. 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 gender determination.

1. INTRODUCTION

Human Computer Interaction (HCI) is an area of research that has witnessed a tremendous development over the past decade. Examples include human face detection, face recognition and verification, face and facial feature tracking, human action and gesture recognition, age estimation from face and gender determination from facial images or video. The gender information can be used for enhancing the performance of current face recognition and verification applications and especially facial expression recognition systems. Moreover, the gender information can be used in various other applications like restricting the access to certain areas based on gender or more frequently can be used for collecting valuable statistical information like how much women or men have been using a particular system.

Since, gender determination has been widely accepted as an important problem in computer vision and pattern recognition various methods have been proposed and applied in order to treat it. The interested reader may refer to [1, 2] and in the references therein for more information concerning the various approaches for gender determination. As pointed out in [1, 2] a Support Vector Machine (SVM) system applied directly to the image pixels has been the best classifier among many others [3]. In this paper we propose a new classifier that combines the properties of Fisher's Linear Discriminant Analysis (FLDA) [4] and SVMs [3] and apply it for gender determination.

In detail, motivated by the fact that the Fisher's discriminant optimization problem for two classes is a constraint least-squares optimization problem [5], 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 [5]. It has been shown that it outperforms the typical maximum margin SVMs [5].

In [5], 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 we generally define and solve the problem in cases that the within class scatter matrix is singular (such cases include the definition of the problem in dot product Hilbert spaces). It will be proven that the non-linear optimization problem of the Support Vectors Machine Variant (SVMV) is equivalent to a linear one, subject to an initial Kernel Principal Component Analysis (KPCA) embedding of the training data. The proposed method is applied in gender determination from facial images.

2. 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 [5]. 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 [3].

2.1 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:

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

$\mathbf{m}_{\mathcal{C}_+}$ 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:

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

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 $\mathcal{U}$. 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.

2.2 Support Vector Machines (SVMs)

In the SVMs case, the optimal separating hyperplane is the one which separates the training data with the maximum margin [3]. 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:

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

2.3 A Support Vector Machine Variant (SVMV)

In [5], inspired by the maximization of the Fisher's discriminant ratio (2) and the SVMs separability constraints, the SVMVs 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 $\mathbf{w}^T\mathbf{S}_w\mathbf{w}>0$.

In the case where the training vectors are not linearly separable the optimum decision hyperplane is found by using the *soft margin* formulation [5, 3] and solving the following optimization problem:

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

subject to the separability constraints:

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

where $\mathbf{v}=[v_1,\ldots,v_N]$ 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:

$$\begin{aligned}&L(\mathbf{w},b,\mathbf{a},\mathbf{b},\mathbf{v})=\mathbf{w}^T\mathbf{S}_w\mathbf{w}+C\textstyle\sum_{i=1}^{N}v_i \\ &-\textstyle\sum_{i=1}^{N}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i+b)-1+v_i]-\sum_{i=1}^{N}\beta_i v_i\end{aligned} \tag{10}$$

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

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

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

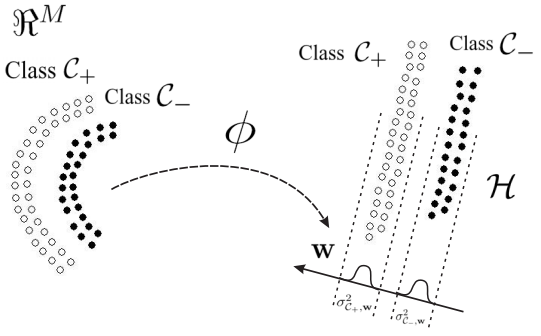


Figure 1: Illustration of the non-linear SVMVs. Search for a direction $\mathbf{w}$ in the feature space $\mathcal{H}$, such that samples projected onto this dimension are separable and the variances ($\sigma^2_{C_k,\mathbf{w}}$ and $\sigma^2_{C_t,\mathbf{w}}$) of the projected samples are minimized.

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):

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

By replacing (12) into (10) and using the KKT conditions (11), the constraint optimization problem (8) is reformulated to the Wolf dual problem:

$$\begin{aligned}&\max_{\mathbf{a}}f(\mathbf{a})=\mathbf{1}_N^T\mathbf{a}-\tfrac{1}{2}\mathbf{a}^T\mathbf{Q}\mathbf{a} \\ &\text{subject to } 0\leq\alpha_i\leq C,\ i=1,\ldots,N,\ \ \mathbf{a}^T\mathbf{y}=0\end{aligned}$$

where $\mathbf{1}_N$ is a N-dimensional vector of ones and $[\mathbf{Q}]_{i,j}=\frac{1}{2}y_iy_j\mathbf{x}_i^T\mathbf{S}_w^{-1}\mathbf{x}_j$. It is worth noting here that, for the typical maximum margin SVMs problem [3], 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:

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

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:

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

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

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

3. SVMV NONLINEAR DECISION SURFACES

In this Section, the optimization problem of the nonlinear SVMV 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.

Let us define the non-linear mapping $\phi : \Re^M \to \mathcal{H}$ 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 [3] will be considered. In the space $\mathcal{H}$ the within-class scatter is defined as:

$$\begin{aligned} \mathbf{S}_w^{\Phi} &= \textstyle\sum_{\mathbf{x}\in\mathcal{C}_-}(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_-}^{\Phi})(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_-}^{\Phi})^T \\ &+ \textstyle\sum_{\mathbf{x}\in\mathcal{C}_+}(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_+}^{\Phi})(\phi(\mathbf{x}) - \mathbf{m}_{\mathcal{C}_+}^{\Phi})^T \end{aligned} \quad (16)$$

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:

$$\min_{\mathbf{w},b,\mathbf{v}} \mathbf{w}^T\mathbf{S}_w^{\Phi}\mathbf{w} + C\sum_{i=1}^{N} v_i,\ \mathbf{w}^T\mathbf{S}_w^{\Phi}\mathbf{w} > 0 \quad (17)$$

subject to the constraints:

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - v_i\ , v_i \geq 0,\ \ i = 1, \ldots, N. \quad (18)$$

Figure 1 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 (17) subject to the constraints (18) 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:

$$\mathbf{S}_w^{\Phi}\mathbf{w}_o = \frac{1}{2}\sum_{i=1}^{N}\alpha_{i,o} y_i \phi(\mathbf{x}_i). \quad (19)$$

It can be proven that there is a solution to the optimization problem (17) subject to the constraints (18), 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:

$$\mathbf{S}_t^{\Phi} = \sum_{i=1}^{N}(\phi(\mathbf{x}_i) - \mathbf{m}^{\Phi})(\phi(\mathbf{x}_i) - \mathbf{m}^{\Phi})^T = \mathbf{S}_w^{\Phi} + \mathbf{S}_b^{\Phi}. \quad (20)$$

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 [4], 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 $\mathbf{w} = \mathbf{f} + \mathbf{z}$ with $\mathbf{f} \in \mathcal{B}^{\Phi}$ and $\mathbf{z} \in \mathcal{B}_{\perp}^{\Phi}$.

It can be proven that the optimal decision surface for the optimization problem (17) subject to the constraints (18) 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 [4] the space $\mathcal{B}^{\Phi}$ is isomorphic to the $(N-1)$-dimensional Euclidean space $\Re^{N-1}$. The isomorphic mapping is:

$$\mathbf{f} = \mathbf{P}\mathbf{h},\ \mathbf{h} \in \Re^{N-1}, \quad (21)$$

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:

$$\min_{\mathbf{h},b,\mathbf{v}} \mathbf{h}^T\tilde{\mathbf{S}}_w\mathbf{h} + C\sum_{i=1}^{N} v_i,\ \mathbf{h}^T\tilde{\mathbf{S}}_w\mathbf{h} > 0,\ \mathbf{h} \in \Re^{N-1} \quad (22)$$

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:

$$\begin{aligned} &y_i(\mathbf{h}^T\tilde{\mathbf{x}}_i + b) \geq 1 - v_i \\ &v_i \geq 0,\ \ i = 1, \ldots, N,\ \mathbf{h} \in \Re^{N-1} \end{aligned} \quad (23)$$

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 [4] and in the references therein. Under the projection to KPCA mapping, the optimal decision surface for the optimization problem (17) subject to (18)

in $\mathcal{H}$ can be found by solving the optimization problem (22) subject to (23) in $\Re^{N-1}$. It is very interesting to notice here that now the problem falls in the linear SVMVs case (i.e., a linear SVMVs 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 2. That is, the optimization problem (22) subject to the constraints (23) 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 $\mathbf{h}_o = \frac{1}{2} \sum_{i=1}^{N} \alpha_{i,o} y_i \tilde{\mathbf{S}}_w^{-1} \tilde{\mathbf{x}}_i$. The decision surface in $\mathcal{H}$ is given by:

$$\begin{aligned} &g(\mathbf{x}) = \text{sign}(\mathbf{w}_o^T \phi(\mathbf{x}_i) + b_o) = \text{sign}(\mathbf{f}_o^T \phi(\mathbf{x}) + b_o) \\ &= \text{sign}(\mathbf{h}_o^T \mathbf{P}^T \phi(\mathbf{x}) + b_o) = \\ &= \text{sign}\left(\frac{1}{2} \sum_{i=1}^{N} \alpha_{i,o} y_i \phi(\mathbf{x}_i)^T \mathbf{P} \tilde{\mathbf{S}}_w^{-1} \mathbf{P}^T \phi(\mathbf{x}) + b_o \right) \end{aligned} \tag{24}$$

for the optimal choice of b_o a similar strategy to Section 2 can be followed.

Summarizing, in order to find the optimal decision surface derived from the optimization problem (17) subject to the constraints (18), the training samples should be projected to $\Re^{N-2}$ using the KPCA transform (matrix $\mathbf{P}$) and solve a linear SVMVs 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 (24).

4. EXPERIMENTS ON GENDER DETERMINATION USING THE XM2VTS DATABASE

Experiments were conducted using data from the XM2VTS database [6] 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. The resolution of the resulting "face-prints" was 85×156. As in the gender determination experiments in [1], 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 [3] and the Complete Kernel Fisher Discriminant (CKFDA) framework proposed in [4].

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 [1], where it has been shown that maximum margin SVMs outperform several other classifiers in this problem. The interested reader may refer to [1] 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 [1] 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:

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \phi(\mathbf{x})^T \phi(\mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^d \\ k(\mathbf{x}, \mathbf{y}) &= \phi(\mathbf{x})^T \phi(\mathbf{y}) = e^{-\gamma (\mathbf{x}-\mathbf{y})^T (\mathbf{x}-\mathbf{y})} \end{aligned} \tag{25}$$

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 [1]. 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 Section 3 and afterwards the quadratic optimization problem of SVMVs is solved. For CKFDA the regular and the irregular discriminant projections are found using the method proposed in [4]. 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.

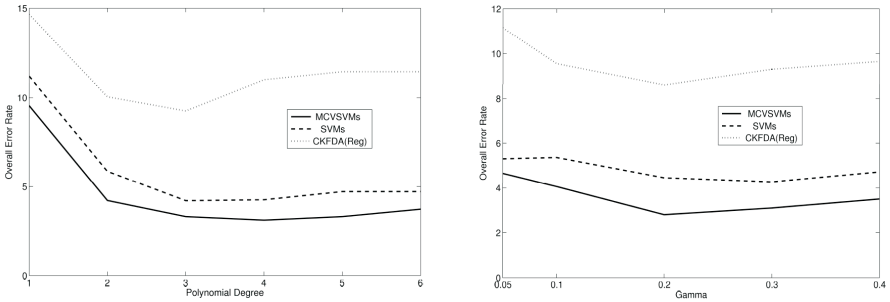
The experimental results with various kernels and parameters are shown in Figure 3. As can been seen in this Figure the error rates for the SVMVs are constantly lower than those achieved for the other tested classifiers for all the tested kernels and parameters. Some of the support faces used for constructing the non-linear SVMV surfaces are shown in Figure 2. The lowest error rates for the tested classifiers are summarized in Table 1. The best error rate for the SVMVs have been 2.86% while for SVMs have been 4.4%.

5. CONCLUSIONS

In this paper an SVM variant has been proposed based on the minimization of within class variance subject to separability constraints. We have provided solution when defining nonlinear surfaces by means of Mercer's kernels. The proposed classifiers have been applied to

(a) (b)

Figure 2: Some of the Support faces used by the polynomial SVMVs of degree 3 a) Support men; b) Support women.



(a) (b)

Figure 3: Average error rates for gender determination using various kernels; a) polynomial kernel b) RBF kernel.

Table 1: The best error rates of the tested classifiers at Gender Determination.

Algorithm	Overall %	Male %	Female %
SVMVs-RBF	**2.86**	**2.19**	**3.5**
SVMVs-cubic	3.28	2.98	3.62
SVMs-RBF	4.4	3.48	5.43
SVMs-cubic	4.4	3.48	5.43
Regular CKFDA-RBF	8.58	8.17	9
Regular CKFDA-cubic	9.27	8.17	10.4

the gender determination method and we have shown that this classifier outperform maximum margin SVMs in this problem.

REFERENCES

[1] 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.

[2] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541–547, 2008.

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

[4] 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.

[5] 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.

[6] 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.