Merging SVMs with Linear Discriminant Analysis: A Combined Model

Symeon Nikitidis[1] , Stefanos Zafeiriou[1] and Maja Pantic[1,2]
[1]Department of Computing, Imperial College London, United Kingdom
[2]EEMCS, University of Twente, Netherlands
{s.nikitidis,s.zafeiriou,m.pantic}@imperial.ac.uk

Abstract

A key problem often encountered by many learning algorithms in computer vision dealing with high dimensional data is the so called "curse of dimensionality" which arises when the available training samples are less than the input feature space dimensionality. To remedy this problem, we propose a joint dimensionality reduction and classification framework by formulating an optimization problem within the maximum margin class separation task. The proposed optimization problem is solved using alternative optimization where we jointly compute the low dimensional maximum margin projections and the separating hyperplanes in the projection subspace. Moreover, in order to reduce the computational cost of the developed optimization algorithm we incorporate orthogonality constraints on the derived projection bases and show that the resulting combined model is an alternation between identifying the optimal separating hyperplanes and performing a linear discriminant analysis on the support vectors. Experiments on face, facial expression and object recognition validate the effectiveness of the proposed method against state-of-the-art dimensionality reduction algorithms.

1. Introduction

Two of the most crucial problems that every learning algorithm in Computer Vision (CV) often encounters are the high dimensionality of the input data, which yields several problems in subsequently performed statistical learning algorithms due to the so-called "curse of dimensionality" and the "small sample size problem" which arises when the number of data samples is less than the data sample dimensionality. To overcome these problems various techniques have been proposed for efficient data embedding (or dimensionality reduction) aiming to obtain a more manageable problem and alleviate computational complexity. More precisely, research in the field has primarily revolved around providing efficient and effective solutions to the following problem: given a set of training samples of a high-dimensional space, estimate a low-dimensional space where either the intrinsic structure of the input data is preserved or discrimination between different classes is enhanced. To accomplish these goals various approaches have been proposed in the literature where arguably the most popular ones are the so-called Principal Component Analysis (PCA) [22], Linear Discriminant Analysis (LDA) [1] and the quite related class of Graph Embedding techniques [25]. Moreover, in applications involving a recognition phase, classification is typically performed by projecting the test samples onto the identified low-dimensional space and applying off-the-shelf classifiers such as SVMs. Hence, the task of designing dimensionality reduction or feature extraction methodologies and the design of classifiers are most commonly treated independently, as different modules, in the pipeline of the general framework of recognition applications.

Joint dimensionality reduction and classification has only recently received some attention mainly within the Non-negative Matrix Factorization (NMF) framework [7, 17]. In particular, in [7, 17] joint generative-discriminant frameworks were proposed where a set of projection bases that best reconstruct the data are derived using NMF or Semi-NMF, while the weights that are assigned to these bases are evaluated such as the projected low dimensional samples form classes that are well separated with maximum margin. Support vector machines (SVMs) is probably the most widely used classifier in CV applications [23, 20]. For instance, among the current state-of-the-art approaches for pedestrian detection are SVM with χ^2 square kernels and Histogram of Oriented Gradients (HOG) descriptors [23], SVMs with Gaussian RBF (GRBF) kernels using additive distances are some of the best classifiers in vision [23] and also structured SVM approaches are among the state-of-the-art in object detection [9, 28]. Another reason that SVMs are very popular in vision applications is that recently packages for solving the Quadratic Programming (QP) optimization problem for SVM training in linear time, with respect to the number of training samples, have been proposed and publicly released [12, 13].

In this paper, we follow a different line of research and

propose a pure discriminative framework. That is, we propose a combined framework of dual discriminative dimensionality reduction and classification within the maximum margin framework of SVMs. We build our method by defining a joint optimization problem for finding both a set of low-dimensional projections and the separating hyperplanes. However, since, the dual optimization problem with respect to the projection bases is computationally expensive, we also propose an algorithmically efficient approach resulting by introducing orthogonality constraints on the identified projection bases. For the latter case, we demonstrate that the alternative optimization procedure is equivalent to finding the maximum margin separating hyperplane in the low-dimensional space defined by performing LDA explicitly on the support vectors. Summarizing the novel contributions of the paper are the following:

- We propose to the best of our knowledge, the first[1] joint dimensionality reduction and classification method developed within a maximum margin framework. Our methodology is radically different than the maximum margin projections in [15], since in that work dimensionality reduction was treated as a purely classification problem. That is, the set of projections were produced by solving a number of SVM optimization problems (equal to the number of retained dimensions) and removing at each step the learned hyperplane from the data (a procedure called deflation). The last hyperplane learned from the deflation approach is the final hyperplane that can be used for classification. A similar to our methodology line of research is presented in [11] where dimensionality reduction is attempted in the context of multi-label classification and the projection directions are derived by considering only binary classification problems one for each label. However, in this paper we drawn radically different conclusions from [11]. More precisely, in [11] is stated that the joint framework is equivalent to the separate application of LDA for dimensionality reduction and SVM for classification and thus performance is not improved by the joint framework. As we show in our theory the joint framework, first is not feasible for binary classification problems, since in this case the resulting low dimensional projection matrix is a degenerate rank one matrix and second, in the general case of multiclass classification problems the joint framework is not equivalent to the separate approach, since the covariance matrix is explicitly evaluated on the support vectors. Finally, we also experimentally verify the superiority of the joint approach on different recognition problems on various datasets.
- In the proposed approach we do not need to resort to sub-optimal approaches such as deflation, since we can jointly compute the low dimensional projections and the separating hyperplanes using alternative optimization. Furthermore, we can reduce the computational cost by incorporating orthogonality constraints on our projection bases and show that in this case the projection bases are given by the largest eigenvectors of a between-class scatter matrix defined on the support vectors.
- Finally, our methodology is radically different to methods that use off-the self dimensionality reduction and feature extraction algorithms such as PCA and Kernel PCA (KPCA) [8, 5, 24] or use a first ad-hoc step of data dependent transformation by projecting on the non-null space of data covariance matrices (e.g., the within-class scatter matrix [26, 16]), where there is no connection between the data transformation and classification steps. In this paper we formulate a joint optimization problem where there is a natural interplay between dimensionality reduction/data transformation and identifying the optimal classification hyperplanes.

[1]The methodology proposed in [4] although it is referred as a margin discrimination approach it follows a totally different approach than ours. That is a non-parametric LDA was proposed using weights from minimum bounding hyperdisks.

2. Maximum Margin Projections

Given a set $\mathcal{X} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ of N training data pairs, where $\mathbf{x}_i \in \mathcal{R}^F, i = 1, ..., N$ are the F-dimensional input feature vectors each assigned a class label $y_i \in \{1, \ldots, K\}$ with K denoting the total number of classes, a multiclass SVM classifier [6] attempts to determine a set of K separating hyperplanes $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K\}$ where $\mathbf{u}_p \in \mathcal{R}^F, p = 1, ..., K$ is the normal vector of the p-th hyperplane that separates the training vectors of the p-th class from all the others with maximum margin. Thus, the decision whether a test sample $\mathbf{x}$ belongs to one of the K different classes is derived by projecting the test sample on the normal vectors of each decision hyperplane and using the decision function

$$y = \arg\max_p \mathbf{u}_p^T \mathbf{x} + b^p, \quad p = 1, \ldots, K. \tag{1}$$

where $b^p \in \mathcal{R}$ is the bias term associated with the p-th class separating hyperplane.

In this work we assume that the intrinsic data dimensionality is much less than the input feature space and that the problem at hand can be efficiently described using a smaller number of degrees of freedom. Thus, we express the separating hyperplanes normal vectors as an appropriately linear combination of the columns of a projection matrix $\mathbf{R} \in \mathcal{R}^{F \times M}$ ($M \ll F$) as $\mathbf{u}_p = \mathbf{R}\mathbf{w}_p$. Consequently, the decision function in the projection subspace is:

$$y = \arg\max_p \mathbf{w}_p^T \mathbf{R}^T \mathbf{x} + b^p = 0, \quad p = 1, \ldots, K \tag{2}$$

which can be also interpreted as exploiting the normal vectors $\mathbf{w}_p \in \mathcal{R}^M$ of the appropriate separating hyperplanes in the low dimensional space of the projection matrix $\mathbf{R}$, determined using the low dimensional data representations derived by performing the linear transformation $\acute{\mathbf{x}}_i = \mathbf{R}^T\mathbf{x}_i$.

Inspired by the multiclass SVM optimization problem proposed in [6] we aim to jointly learn the optimal projection matrix $\mathbf{R}$ such that the training samples of different classes are projected in a subspace, where they are separated with maximum margin (i.e. are better discriminated) and also to determine the optimal decision hyperplanes in the respective projection subspace. To do so we form the following cost function aiming to simultaneously maximize the separating margin both in the initial high dimensional and the reduced dimensional space and minimize the classification error defined according to which side of the decision hyperplane training samples of each class fall in. Moreover, in the cost function we also incorporate minimization of term $\text{Tr}[\mathbf{R}^T\mathbf{R}]$ in order to avoid data scaling in the projection space, regularize between different terms in the cost function to improve optimization stability and also facilitate our subsequent mathematical derivations. Thus, our optimization problem is defined as:

$$\min_{\mathbf{w}_p,\xi_i,\mathbf{R}} \quad \frac{1}{2}\sum_{p=1}^{K}\mathbf{w}_p^T\mathbf{w}_p + \frac{1}{2}\sum_{p=1}^{K}\mathbf{w}_p^T\mathbf{R}^T\mathbf{R}\mathbf{w}_p + \frac{1}{2}\text{Tr}[\mathbf{R}^T\mathbf{R}] + C\sum_{i=1}^{N}\xi_i, \tag{3}$$

subject to the constraints:

$$\mathbf{w}_{y_i}^T\mathbf{R}^T\mathbf{x}_i - \mathbf{w}_p^T\mathbf{R}^T\mathbf{x}_i \geq b_i^p - \xi_i, \qquad i = 1,\ldots,N \quad p = 1,\ldots,K. \tag{4}$$

where Tr[.] is the matrix trace operator, $\mathbf{w}_p \in \mathcal{R}^M$ is the M-dimensional normal vector of the p-th hyperplane, $\boldsymbol{\xi} = [\xi_1,\ldots,\xi_N]^T$ are the slack variables, each one associated with a training sample, C is the term that penalizes the training error and $\mathbf{b}$ is the bias vector defined as $b_i^p = 1-\delta_{y_i}^p$ where $\delta_{y_i}^p$ is the Kronecker delta function.

To solve the optimization problem in (3) we consider an alternative optimization framework where we first compute the optimal decision hyperplanes for an initialized projection matrix $\mathbf{R}$ and subsequently, solve (3) for $\mathbf{R}$ so that the identified projection matrix improves the objective function i.e., it projects the training samples in a subspace where the margin that separates the training samples of each class from all the others, is maximized. Next, we first demonstrate the optimization process with respect to the normal vectors of the separating hyperplanes in the projection subspace of $\mathbf{R}$ and subsequently, we discuss the projection matrix $\mathbf{R}$ evaluation, while keeping the optimal normal vectors $\mathbf{w}_{p,o}$ fixed.

2.1. Finding the optimal $\mathbf{w}_{p,o}$ in the projection subspace determined by $\mathbf{R}$

To solve the constrained optimization problem in (3) for $\mathbf{w}_p$ we introduce positive Lagrange multipliers α_i^p, each associated with one inequality constraint in (4) and formulate the Lagrangian function $\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha})$:

$$\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{p=1}^{K}\mathbf{w}_p^T\left(\mathbf{I}_M + \mathbf{R}^T\mathbf{R}\right)\mathbf{w}_p + \frac{1}{2}\text{Tr}[\mathbf{R}^T\mathbf{R}] + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\sum_{p=1}^{K}\alpha_i^p\left[\left(\mathbf{w}_{y_i}^T - \mathbf{w}_p^T\right)\mathbf{R}^T\mathbf{x}_i + \xi_i - b_i^p\right], \tag{5}$$

where $\mathbf{I}_M$ is an $M \times M$ dimensional identity matrix. To find the minimum over the primal variables $\mathbf{w}_p$ and $\boldsymbol{\xi}$ we require that the partial derivatives of $\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\xi}$ and $\mathbf{w}_p$ vanish, which yields the following equalities:

$$\frac{\partial\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha})}{\partial\xi_i} = 0 \Rightarrow \sum_{p=1}^{k}\alpha_i^p = C, \tag{6}$$

$$\frac{\partial\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha})}{\partial\mathbf{w}_p} = 0 \Rightarrow$$

$$\mathbf{w}_{p,o} = \left(\mathbf{I}_M + \mathbf{R}^T\mathbf{R}\right)^{-1}\sum_{i=1}^{N}\left(\alpha_i^p - \sum_{p=1}^{K}\alpha_i^p\delta_{y_i}^p\right)\mathbf{R}^T\mathbf{x}_i. \tag{7}$$

Substituting terms from (6) and (7) into (5) and expressing the corresponding to the i-th training sample bias terms and Lagrange multipliers in a vector form as $\mathbf{b}_i = [b_i^1,\ldots,b_i^K]^T$ and $\boldsymbol{\alpha}_i = [\alpha_i^1,\ldots,\alpha_i^K]^T$, respectively and performing the substitution $\mathbf{n}_i = C\mathbf{1}_{y_i} - \boldsymbol{\alpha}_i$, (where $\mathbf{1}_{y_i}$ is a K-dimensional vector with all its components equal to zero except of the y_i-th, which is equal to one) the saddle point of the Lagrangian can be found by the minimization of the following Wolfe dual problem:

$$\min_{\mathbf{n}} \quad \frac{1}{2}\sum_{i,j}^{N}\left[\mathbf{x}_i^T\mathbf{R}\left(\mathbf{I}+\mathbf{R}^T\mathbf{R}\right)^{-\frac{1}{2}}\left(\mathbf{I}+\mathbf{R}^T\mathbf{R}\right)^{-\frac{1}{2}}\mathbf{R}^T\mathbf{x}_j\right] \times \mathbf{n}_i^T\mathbf{n}_j + \frac{1}{2}\text{Tr}[\mathbf{R}^T\mathbf{R}] + \sum_{i=1}^{N}\mathbf{n}_i^T\mathbf{b}_i, \tag{8}$$

subject to the constraints:

$$\sum_{p=1}^{K} n_i^p = 0, \quad n_i^p \leq \begin{cases} 0 & \text{, if } y_i \neq p \\ C & \text{, if } y_i = p \end{cases} \quad \forall \; i = 1,\ldots,N \;, \; p = 1,\ldots,K. \tag{9}$$

The optimal variables $\mathbf{n}$ can be found by solving the above QP problem with the linear kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{R} \left(\mathbf{I} + \mathbf{R}^T \mathbf{R}\right)^{-\frac{1}{2}} \left(\mathbf{I} + \mathbf{R}^T \mathbf{R}\right)^{-\frac{1}{2}} \mathbf{R}^T \mathbf{x}_j$ thus practically by feeding to a linear SVM classifier the transformed training samples $\acute{\mathbf{x}}_i$ derived as: $\acute{\mathbf{x}}_i = \left(\mathbf{I} + \mathbf{R}^T \mathbf{R}\right)^{-\frac{1}{2}} \mathbf{R}^T \mathbf{x}_i$. Subsequently the normal vectors of the optimal separating hyperplanes can be derived from (7).

2.2. Finding the maximum margin projection matrix $\mathbf{R}$ considering fixed separating hyperplanes $\mathbf{w}_{p,o}$

To learn the optimal projection matrix $\mathbf{R}$ we consider the normal vectors $\mathbf{w}_{p,o}$ fixed and similarly require the partial derivatives of the cost function in (3) with respect to $\mathbf{R}$ to vanish:

$$\mathbf{R} = \sum_{i=1}^{N} \sum_{p=1}^{k} \alpha_i^p \mathbf{x}_i \left(\mathbf{w}_{y_i,o}^T - \mathbf{w}_{p,o}^T\right) \left(\sum_{p=1}^{k} \mathbf{w}_{p,o} \mathbf{w}_{p,o}^T + \mathbf{I}_M\right)^{-1} \tag{10}$$

Substituting terms from (10) into (3) we derive the following QP optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \sum_{i,j}^{N} \sum_{p,l}^{K} \alpha_i^p \alpha_j^l \left(\text{vec}\left(\mathbf{x}_i \mathbf{w}_{y_i,o}^T\right) - \text{vec}\left(\mathbf{x}_i \mathbf{w}_{p,o}^T\right)\right) \\ & \left(\mathbf{I}_{MF} + \sum_{p=1}^{K} \mathbf{w}_{p,o} \mathbf{w}_{p,o}^T \otimes \mathbf{I}_F\right)^{-1} \times \\ & \left(\text{vec}\left(\mathbf{x}_j \mathbf{w}_{y_j,o}^T\right) - \text{vec}\left(\mathbf{x}_j \mathbf{w}_{l,o}^T\right)\right)^T + \sum_{i=1}^{N} \sum_{p=1}^{K} \alpha_i^p b_i^p. \end{aligned} \tag{11}$$

subject to the constraints:

$$\sum_{p=1}^{K} \alpha_i^p = C \quad \text{and} \quad \alpha_i^p \geq 0, \quad \forall \quad i = 1, \ldots, N, \quad p = 1, \ldots, K. \tag{12}$$

where vec(.) denotes an operator that converts a matrix into a vector by stacking its columns and $\otimes$ the Kronecker product operation. Solving (11) for the Lagrange multipliers $\boldsymbol{\alpha}$ we can subsequently derive the optimal projection matrix $\mathbf{R}$ from (10).

Unfortunately the size of the generated QP optimization problem in (11) may become extremely large, since the number of the optimized variables is proportional to the product of the number of training samples multiplied by the number of classes in the classification task at hand. This can be impractical for training tasks involving a large number of classes, as for instance, in face recognition where the number of different persons involved can reach to several hundreds. Moreover, the QP problem in (11) requires huge amounts of memory in order to store and handle the dense kernel matrix of dimensionality $MF \times MF$ which may become infeasible when dealing with high dimensional image data where the number of extracted features can range from several hundreds to thousands.

3. Orthogonal Maximum Margin Projections and its Relation to LDA

To overcome the above mentioned algorithmic limitations we modify the considered QP optimization problem in (3) by incorporating additional constraints. More precisely, we require the projection matrix $\mathbf{R}$ to be semiorthogonal imposing orthogonality constraints on its columns. However, according to the optimization problem in (3) and the involved constraints, the projection matrix $\mathbf{R}$ cannot be uniquely determined. To overcome this problem we adopt a robust optimization strategy formulating a minimax optimization problem [2] where we attempt to minimize the multiclass SVM cost function for the worst case projection matrix $\mathbf{R}$:

$$\min_{\mathbf{w}_p, \xi_i} \max_{\mathbf{R}} \quad \frac{1}{2} \sum_{p=1}^{K} \mathbf{w}_p^T \mathbf{w}_p + C \sum_{i=1}^{N} \xi_i, \tag{13}$$

subject to the constraints in (4) and:

$$\mathbf{R}^T \mathbf{R} = \mathbf{I}_M. \tag{14}$$

To solve the new constrained optimization problem we similarly introduce positive Lagrange multipliers α_i^p, and $\boldsymbol{\Lambda} \in \mathcal{R}^{M \times M}$ each associated with one of the constraints in (4) and (14) and formulate the Lagrangian function $\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha}, \boldsymbol{\Lambda})$:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}) = & \frac{1}{2} \sum_{p=1}^{K} \mathbf{w}_p^T \mathbf{w}_p + C \sum_{i=1}^{N} \xi_i \\ & - \text{Tr}[\boldsymbol{\Lambda}(\mathbf{R}^T \mathbf{R} - \mathbf{I}_M)] \\ & - \sum_{i=1}^{N} \sum_{p=1}^{K} \alpha_i^p \left[\left(\mathbf{w}_{y_i}^T - \mathbf{w}_p^T\right) \mathbf{R}^T \mathbf{x}_i + \xi_i - b_i^p\right]. \end{aligned} \tag{15}$$

Requiring that the partial derivatives of $\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\xi}$ and $\mathbf{w}_p$ vanish, we derive the equality in (6) and:

$$\mathbf{w}_p = \sum_{i=1}^{N} \left(\alpha_i^p - \sum_{p=1}^{k} \alpha_i^p \delta_{y_i}^p\right) \mathbf{R}^T \mathbf{x}_i. \tag{16}$$

By substituting terms from (6) and (16) into (15), and expressing the bias terms and Lagrange multipliers in a vector form as in 2.1 the saddle point of the Lagrangian function

$\mathcal{L}(\mathbf{w}_p, \boldsymbol{\xi}, \mathbf{R}, \boldsymbol{\alpha}, \boldsymbol{\Lambda})$ can be found by solving the equivalent minimax optimization problem:

$$\min_{\mathbf{n}} \max_{\mathbf{R}} \quad \frac{1}{2}\sum_{i,j}^{N} \mathbf{x}_i^T \mathbf{R}\mathbf{R}^T \mathbf{x}_j \mathbf{n}_i^T \mathbf{n}_j - \mathrm{Tr}[\boldsymbol{\Lambda}(\mathbf{R}^T\mathbf{R} - \mathbf{I}_M)] + \sum_{i=1}^{N} \mathbf{n}_i^T \mathbf{b}_i, \tag{17}$$

subject to the constraints in (9).

To solve the QP problem in (17) we similarly use alternative optimization thus solving for one variable, while keeping the other fixed. More precisely, we first optimize (17) with respect to $\mathbf{n}$, for a randomly initialized orthogonal projection matrix $\mathbf{R}$, which is essentially the conventional multiclass SVM training problem performed in the projection subspace determined by $\mathbf{R}$ using the linear kernel function of the form $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{R}\mathbf{R}^T \mathbf{x}_j$. This can be easily performed by feeding to the SVM classifier the projected training samples $\acute{\mathbf{x}}_i = \mathbf{R}^T \mathbf{x}_i$. Subsequently, the normal vectors of the optimal separating hyperplanes can be derived from (16).

To optimize for $\mathbf{R}$ we remove term $\sum_{i=1}^{N} \mathbf{n}_i^T \mathbf{b}_i$ from (17), since it is independent of the optimized variable and solve the equivalent trace optimization problem:

$$\max_{\mathbf{R}} \mathrm{Tr}[\mathbf{R}^T \sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T \mathbf{R}] - \mathrm{Tr}[\boldsymbol{\Lambda}(\mathbf{R}^T\mathbf{R} - \mathbf{I}_M)]. \tag{18}$$

Computing the derivative of the maximized cost function in (18) with respect to $\mathbf{R}$ and setting it equal to zero the optimization problem leads to the following generalized eigenvalue problem:

$$(\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T)\mathbf{R} = \mathbf{R}\boldsymbol{\Lambda}. \tag{19}$$

Thus, the projection bases of $\mathbf{R}$ correspond to the $K-1$ eigenvectors of matrix $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T$ associated with the $K-1$ largest eigenvalues. Matrix $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T$ has a similar form to the LDA between class covariance matrix, since it can be written as a covariance matrix $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T = \mathbf{X}\mathbf{M}\mathbf{M}^T\mathbf{X}^T = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A} \in \mathcal{R}^{F\times K}$, $\mathbf{X} \in \mathcal{R}^{F\times N}$ is a data matrix created by stacking the training samples $\mathbf{x}_i$ column-wise, while $\mathbf{M} = [\mathbf{n}_1, \ldots, \mathbf{n}_K]^T \in \mathbf{R}^{N\times K}$ is created by stacking the vectors of the optimal Lagrange multipliers for each training sample row-wise. Thus, $\sum_{i,j}^{N} \mathbf{n}_i^T \mathbf{n}_j \mathbf{x}_i \mathbf{x}_j^T$ encodes the between class scatter evaluating the weighted by the Lagrange multipliers mean for each class.

3.1. Orthogonal maximum margin projections for binary problems

To better demonstrate the relation between the proposed orthogonal maximum margin projection method and performing LDA explicitly on the support vectors let us consider a binary separation problem of two classes $\mathcal{C}_+$ and $\mathcal{C}_-$. The corresponding minimax optimization problem is formulated as:

$$\min_{\mathbf{w},\xi_i} \max_{\mathbf{R}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i \tag{20}$$

subject to the constraints:

$$\begin{aligned} y_i(\mathbf{w}^T\mathbf{R}^T\mathbf{x}_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \\ \mathbf{R}^T\mathbf{R} &= \mathbf{I}_M, \end{aligned} \tag{21}$$

where $y_i \in \{-1,\ 1\}$ is the class label associated with each sample $\mathbf{x}_i$. Consequently the optimization problem with respect to the projection matrix $\mathbf{R}$ can be summarized as follows:

$$\max_{\mathbf{R}} \quad \frac{1}{2}\mathrm{Tr}[\mathbf{R}^T \sum_{i,j}^{N} \alpha_i\alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j^T \mathbf{R}] - \mathrm{Tr}[\boldsymbol{\Lambda}(\mathbf{R}^T\mathbf{R} - \mathbf{I}_M)], \tag{22}$$

which can be similarly solved by performing eigenanalysis on matrix $\sum_{i,j}^{N} \alpha_i\alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j^T$ which can be expressed as:

$$\begin{aligned} &\Big(\sum_{\mathbf{x}_i\in\mathcal{C}_+} \alpha_i\mathbf{x}_i - \sum_{\mathbf{x}_j\in\mathcal{C}_-} \alpha_j\mathbf{x}_j\Big)\Big(\sum_{\mathbf{x}_i\in\mathcal{C}_+} \alpha_i\mathbf{x}_i - \sum_{\mathbf{x}_j\in\mathcal{C}_-} \alpha_j\mathbf{x}_j\Big)^T \\ &= (\mathbf{m}_{\mathcal{C}_+} - \mathbf{m}_{\mathcal{C}_-})(\mathbf{m}_{\mathcal{C}_+} - \mathbf{m}_{\mathcal{C}_-})^T, \end{aligned} \tag{23}$$

where $\mathbf{m}_{\mathcal{C}_+}$ and $\mathbf{m}_{\mathcal{C}_-}$ denote the weighted mean vectors of the two classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively evaluated explicitly on the support vectors. However, in this case $\mathbf{R}$ becomes a degenerate matrix of rank 1, hence it is not possible to find both variables $\mathbf{w}$ and $\mathbf{R}$.

4. Experimental Results

We compare the performance of the proposed method with that of several state-of-the-art dimensionality reduction techniques, such as PCA, LDA, Subclass Discriminant Analysis (SDA) [27], Locality Preserving Projections (LPP) [10] and Orthogonal Locality Preserving Projections (OLPP) [3]. Moreover, in our comparison, we also directly feed the initial high dimensional samples to a linear multiclass SVM classifier, to serve as our baseline testing method. Experiments have been performed for facial expression recognition on the Cohn-Kanade database [14], for face recognition on the Extended M2VTS (XM2VTS) database [21] and for object recognition on the ETH-80 image dataset [18].

On the experiments for facial expression recognition as our classification features, we either considered only the facial image intensity information or its augmented Gabor wavelet representation, which provides robustness to illumination variations [19]. To create the augmented Gabor feature vectors we convolved each facial image with Gabor kernels considering 5 different scales and 8 directions. Hence, for each facial image, and for each Gabor kernel a complex vector containing a real and an imaginary part was generated. Based on these parts we computed the Gabor magnitude information creating in total 40 feature vectors for each facial image. Each such feature vector was subsequently downsampled, in order to reduce its dimension and normalized to zero mean and unit variance. Thus, for each facial image we derived its augmented Gabor wavelet representation by concatenating the 40 feature vectors into a single vector. Moreover, for the face recognition experiments on XM2VTS database we only used the facial image intensity information as our underlying features and did not exploit more complex representations such as the Gabor features, since the derived recognition rates where already sufficiently high. Finally, on the experiments for object recognition we used the cropped and scaled to a fixed size of 128×128 pixels binary images of ETH-80 containing the contour of each object,

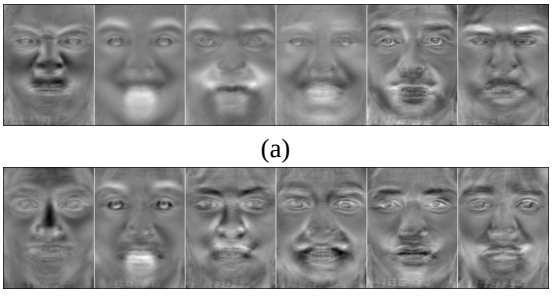
4.1. Facial Expression Recognition in the Cohn-Kanade Database

The Cohn-Kanade AU-Coded facial expression database is among the most popular databases for benchmarking methods that perform facial expression recognition. To form our data collection we discarded the video frames depicting subjects performing each facial expression in increasing intensity level and considered only the last video frame depicting each formed facial expression at its highest intensity. Thus, in our experiments, we used in total 407 images depicting 100 subjects, posing 7 different expressions (anger, disgust, fear, happiness, sadness, surprise and the neutral emotional state). The extracted facial images were manually aligned with respect to the eyes position, anisotropically scaled to a fixed size of 150×200 pixels and converted to grayscale.

To measure the facial expression recognition accuracy, we randomly partitioned the available samples into 5 approximately equal sized subsets (folds) and a 5-fold cross-validation has been performed by feeding the projected discriminant facial expression representations to the linear SVM classifier. This resulted into such a test set formation, where some expressive samples of an individual were left for testing, while his rest expressive images (depicting other facial expressions) were included in the training set. This fact significantly increased the difficulty of the expression recognition problem, since person identity related issues arose.

Table 1 summarizes the best average facial expression recognition rates achieved by each examined embedding method, both for the considered facial image intensity and the augmented Gabor features. The mean facial expression recognition rates attained by directly feeding the initial high dimensional data to the linear SVM classifier are also provided in Table 1. Considering the facial image intensity as the chosen classification features, the proposed method outperforms all other competing embedding algorithms. The best average expression recognition rate attained by the joint framework is 80.4% extracting 6-dimensional discriminant representations of the initial 30,000-dimensional input samples. Exploiting the augmented Gabor features significantly improved the recognition performance of all examined methods, verifying the appropriateness of these descriptors in the task compared against the image intensity features. The proposed algorithm attained the highest average expression recognition rate outperforming the second best method (LDA) by 2.7%.

Figure 1 compares the basis images generated from training on Cohn-Kanade database the proposed joint dimensionality reduction and classification method and LDA. As can be seen, the basis images extracted by the proposed method better highlight facial parts around mouth, eyes and eyebrows characteristic for each facial expression, such as the mouth shape at disgust and surprise expression (bases 1 and 2) the raised or lowered lip corners characteristic of the happiness or sadness facial expression (bases 4 and 5) or the mouth stretch and eyebrows movement (bases 3 and 6) evident in fear and anger facial expressions, respectively.



(a)

(b)

Figure 1. Basis images derived from training on Cohn-Kanade database: a) the proposed joint dimensionality reduction and classification framework and b) LDA.

4.2. Face Recognition in XM2VTS Database

XM2VTS database contains 8 shots of each of the 295 subjects captured at four recording sessions over a period of four months. For our face recognition experiments we acquired a single facial image from each shot depicting each subjects face at a frontal position in a neutral emotional

Table 1. Best average expression recognition accuracy rates (%) in Cohn-Kanade database. In parentheses it is shown the dimension that results in the best performance for each method.

	SVM	PCA	LDA	SDA	LPP	OLPP	Proposed
Intensity	$73.4(30,000)$	$74.5(260)$	$74.2(6)$	$76.4(55)$	$76.6(6)$	$75.2(6)$	$\mathbf{80.4(6)}$
Gabor	$77.8(48,000)$	$84.6(150)$	$86.5(6)$	$86.1(69)$	$85.5(6)$	$83.3(6)$	$\mathbf{89.2(6)}$

state. Thus, in total our dataset is comprised of $2,360$ images which have been grayscaled, aligned and scaled to a fix size of 40×30 pixels using their facial landmarks annotations. To form our training set we used the six facial images of each subject captured during the first three recording sessions, while for testing we used the remaining 2 images for each subject captured during the last session. Table 2 summarizes the highest face recognition rate attained by each method in the comparison and the respective projection subspace dimensionality. The proposed joint framework outperformed all other linear dimensionality reduction algorithms achieving a highest recognition rate equal to 97.5%.

In order to investigate our algorithms performance with respect to the projection subspace dimensionality we performed experiments on XM2VTS extracting a varying number of discriminant features. Figure 2 plots the number of extracted features with respect to the face recognition accuracy rate attained by the proposed joint framework and the common separate application of LDA for dimensionality reduction and SVM for classification. As can be observed the proposed method not only achieved a highest recognition rate for the optimal 294-dimensional projection subspace but also constantly outperformed LDA for low dimensional projection spaces where less features with higher discriminant information were extracted.

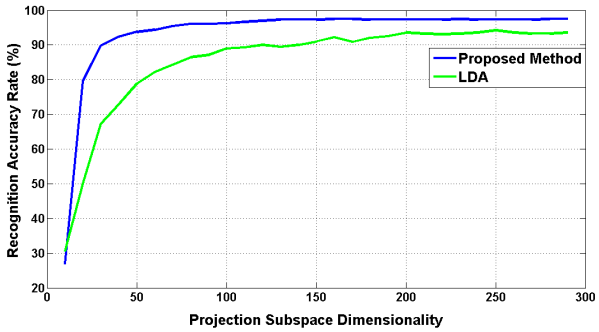


Figure 2. Face recognition accuracy rate versus the dimensionality of the projection subspace on XM2VTS database.

4.3. Object Recognition in the ETH-80 Image Dataset

ETH-80 image dataset [18] depicts 80 objects divided into 8 different classes, where for each object 41 images have been captured from different view points, spaced equally over the upper viewing hemisphere. Thus, the database contains 3,280 images in total. For this experiment we used the cropped and scaled to a fixed size of 128×128 pixels binary images containing the contour of each object. In order to form our training set we randomly picked 25 binary images of each object, while the rest were used for testing. Table 3 shows the highest attained object recognition accuracy rate by each method and the respective subspace dimensionality. The proposed algorithm attained the highest object recognition rate equal to 84.6% outperforming all other methods in the comparison.

It is significant to note that all linear dimensionality reduction algorithms in our comparison, based on Fisher discriminant ratio (i.e. LDA, LPP and OLPP) attained a reduced performance compared against the baseline approach which is feeding directly the initial high dimensional feature vectors to the linear SVM for classification. This can be attributed to the fact that since each category in the ETH-80 dataset includes images depicting 10 different objects captured from various view angles, data samples inside classes span large in-class variations. As a result all the aforementioned methods which have the Gaussian data distribution optimality assumption [27] fail to identify appropriate discriminant projection directions. In contrast to the proposed method which depends only on the support vectors and the overall data samples distribution inside classes does not affect its performance.

5. Conclusion

We proposed a combined framework of dual discriminative dimensionality reduction and classification within the maximum margin framework of SVMs. The developed optimization problems are solved using alternative optimization where we jointly compute the low dimensional maximum margin projections and the separating hyperplanes in the respective subspace. In the experimental study we demonstrated that the proposed method outperforms current state-of-the-art linear data embedding methods on challenging computer vision recognition tasks such as face, expression and object recognition on popular datasets.

Acknowledgments

This work has been funded by the EPSRC project EP/J017787/1 (4D-FAB).

Table 2. Face recognition accuracy rates (%) in XM2VTS database. In parentheses it is shown the dimension that results in the best performance for each method.

	SVM	PCA	LDA	SDA	LPP	OLPP	Proposed
Intensity	$90.6(1,200)$	$94.7(200)$	$93.1(294)$	$96.8(300)$	$93.2(294)$	$95.6(250)$	$\mathbf{97.5(160)}$

Table 3. Object recognition accuracy rates (%) in the ETH-80 database. In parentheses it is shown the dimension that results in the best performance for each method.

	SVM	PCA	LDA	SDA	LPP	OLPP	Proposed
Binary Images	$80.3(16,384)$	$81.9(20)$	$74.4(7)$	$79.8(300)$	$74.2(7)$	$74.4(7)$	$\mathbf{84.6(7)}$

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE T-PAMI*, 19(7):711–720, 1997. 4321

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009. 4324

[3] D. Cai, X. He, J. Han, and H. Zhang. Orthogonal laplacianfaces for face recognition. *IEEE T-IP*, 15(11):3608–3614, 2006. 4325

[4] H. Cevikalp, B. Triggs, F. Jurie, and R. Polikar. Margin-based discriminant dimensionality reduction for visual recognition. In *CVPR*, pages 1–8. IEEE, 2008. 4322

[5] Y.-W. Chen and C.-J. Lin. Combining svms with various feature selection strategies. In *Feature Extraction*, pages 315–324. Springer, 2006. 4322

[6] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001. 4322, 4323

[7] M. Das Gupta and J. Xiao. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In *CVPR*, pages 2841–2848. IEEE, 2011. 4321

[8] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations and feature selection for multimedia database search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):911–920, 2003. 4322

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9):1627–1645, 2010. 4321

[10] X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, volume 16, Vancouver, British Columbia, Canada, 2003. 4325

[11] S. Ji and J. Ye. Linear dimensionality reduction for multi-label classification. In *International Joint Conference on Artifical Intelligence (IJCAI)*, pages 1077–1082, 2009. 4322

[12] T. Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, pages 217–226. ACM, 2006. 4321

[13] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009. 4321

[14] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, March 2000. 4325

[15] A. Kocsor, K. Kovács, and C. Szepesvári. Margin maximizing discriminant analysis. In *Machine Learning: ECML 2004*, pages 227–238. Springer, 2004. 4322

[16] I. Kotsia, I. Pitas, and S. Zafeiriou. Novel multiclass classifiers based on the minimization of the within-class variance. *IEEE T-NN*, 20(1):14–34, 2009. 4322

[17] V. B. Kumar, I. Patras, and I. Kotsia. Max-margin semi-nmf. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2011. 4321

[18] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*. IEEE, June 2003. 4325, 4327

[19] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE T-IP*, 11(4):467–476, 2002. 4326

[20] S. Maji and A. C. Berg. Max-margin additive classifiers for detection. In *ICCV*, pages 40–47. IEEE, 2009. 4321

[21] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999. 4325

[22] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 4321

[23] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE T-PAMI*, 34(3):480–492, 2012. 4321

[24] Z.-l. Wu and C.-h. Li. Feature selection using transductive support vector machine. In *Proc. NIPS 2003 Workshop Feature Selection*, 2003. 4322

[25] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE T-PAMI*, 29(1):40–51, 2007. 4321

[26] S. Zafeiriou, A. Tefas, and I. Pitas. Minimum class variance support vector machines. *IEEE T-IP*, 16(10):2551–2564, 2007. 4322

[27] M. Zhu and A. Martinez. Subclass discriminant analysis. *IEEE T-PAMI*, 28(8):1274–1286, August 2006. 4325, 4327

[28] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012. 4321