Iosifidis, A., Tefas, A., & Pitas, I. (2014). Kernel Reference Discriminant Analysis. Pattern Recognition Letters, 49, 85-91. DOI: 10.1016/j.patrec.2014.06.013

Kernel Reference Discriminant Analysis

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, 54124 Thessaloniki, Greece

{*aiosif,tefas,pitas*}*@aiia.csd.auth.gr*

Abstract

Linear Discriminant Analysis (LDA) and its nonlinear version Kernel Discriminant Analysis (KDA) are well-known and widely used techniques for supervised feature extraction and dimensionality reduction. They determine an optimal discriminant space for (non)linear data projection based on certain assumptions, e.g. on using normal distributions (either on the input or in the kernel space) for each class and employing class representation by the corresponding class mean vectors. However, there might be other vectors that can be used for classes representation, in order to increase class discrimination in the resulted feature space. In this paper, we propose an optimization scheme aiming at the optimal class representation, in terms of Fisher ratio maximization, for nonlinear data projection. Compared to the standard approach, the proposed optimization scheme increases class discrimination in the reduced-dimensionality feature space and achieves higher classification rates in publicly available data sets.

Keywords: Kernel Discriminant Analysis, Kernel Spectral Regression, Class Representation

1. Introduction

Linear Discriminant Analysis (LDA) is a well-known algorithm for supervised feature extraction and dimensionality reduction. It aims at the determination of an optimal subspace for linear data projection, in which the classes are better discriminated. Non-linear extensions ([**? ? ? ? ? ? ?**]) exploit data representations in arbitrary-dimensional feature spaces (determined by applying a non-linear data mapping process). After the determination of the data representation in the

arbitrary-dimensional feature space, a linear projection is calculated, which corresponds to a non-linear projection of the original data. In both cases, the adopted criterion is the ratio of the between-class scatter to the within-class scatter in the reduced-dimensionality feature space, which is usually referred to as Fisher ratio.

LDA optimality is based on the assumptions of: a) normal class distributions with the same covariance structure and b) class representation by the corresponding class mean vector. Under these assumptions, the maximization of the Fisher ratio leads to maximal class discrimination in the reduced-dimensionality feature space. Although relying on rather strong assumptions, both LDA and its kernel extensions have proven very powerful and they have been widely used in many applications, including face recognition/verification ([**? ? ? ? ?**]), human action recognition ([**? ?**]), person identification ([**? ? ?**]) and speech recognition ([**?**]).

By observing that the between-class and within-class scatter matrices employed for the determination of the optimal data projection matrix in LDA can be considered to be functions of the class representation, it has been recently shown that, when the two aforementioned assumptions are not met, the adoption of class representations different from the class mean vectors leads to increased class discrimination in the reduced-dimensionality feature space ([**?**]). In addition, it has been shown that, given a data projection matrix determined by maximizing the criterion adopted in LDA, the optimal class representations can be analytically calculated. In order to determine both the optimal data projection matrix and the optimal class representations, an iterative optimization scheme has been proposed ([**?**]).

In this paper, we extend the method in ([**?**]) in order to operate in arbitrary-dimensional feature spaces for non-linear supervised feature extraction and data projection. We formulate an optimization problem that exploits a non-linear data mapping process to an arbitrary-dimensional feature space, in which optimized class representations are determined. By employing such optimized class representations, a linear data projection from the arbitrary-dimensional feature space to a reduced-dimensionality discriminant feature space is subsequently calculated. We show that, the determination of the optimal class representation in the arbitrary-dimensional feature space has a closed form solution, similar to the linear case. For the determination of the optimal data projection exploiting the optimal class representations, we introduce the proposed criterion to the Spectral Regression framework ([**?**]) and we describe an efficient algorithm to this end. Finally, we combine the two aforementioned optimization processes and propose an iterative optimization scheme for the determination of both the optimal class representation and the optimal (non-linear) data projection. The proposed crite-

rion is evaluated on standard classification problems, as well as on human action and face recognition problems. It is shown that, by exploiting optimized class representations, increased class discrimination can be achieved in the decision space leading to enhanced classification performance.

The rest of the paper is structured as follows. We briefly describe the nonlinear version of LDA, i.e., the Kernel Discriminant Analysis (KDA), in Section 2. The proposed Kernel Reference Discriminant Analysis (KRDA) algorithm is described in detail in Section 3. Experimental results comparing its performance with the standard approach are provided in Section 4. Finally, conclusions are drawn in Section 5.

2. Kernel Discriminant Analysis

Let us denote by $\mathbf{x}_{ij} \in \mathbb{R}^D$, $i = 1, \ldots, C$, $j = 1, \ldots, N_i$ a set of D-dimensional data, each belonging to one of C classes. The number of samples belonging to class i is equal to N_i. In order to determine a nonlinear data projection, the input space $\mathbb{R}^D$ is mapped to an arbitrary-dimensional feature space $\mathcal{F}$ (usually having the properties of Hilbert spaces) ([**? ? ?**]) by employing a function $\phi(\cdot) : \mathbf{x}_{ij} \in \mathbb{R}^D \rightarrow \phi(\mathbf{x}_{ij}) \in \mathcal{F}$ determining a nonlinear mapping from the input space $\mathbb{R}^D$ to the arbitrary-dimensional feature space $\mathcal{F}$. $\phi(\cdot)$ can either be chosen based on the properties of the problem at hand, e.g. for histogram-based data representations the RBF-χ^2 kernel has been proven to be the state-of-the-art choice ([**?**]), or can be determined by applying kernel selection methods. In the second case, a linear combination of a priori chosen kernel functions is usually learned based on optimization, e.g., as in ([**?**]). In $\mathcal{F}$, we would like to determine a data projection matrix $\mathbf{P}$ that can be used to map a given sample $\phi(\mathbf{x}_{ij})$ to a low-dimensional feature space $\mathbb{R}^d$ of increased class discrimination power:

$$\mathbf{y}_{ij} = \mathbf{P}^T \phi(\mathbf{x}_{ij}), \quad \mathbf{y}_{ij} \in \mathbb{R}^d. \tag{1}$$

This can be achieved by maximizing the following criterion:

$$\mathcal{J}_{KDA}(\mathbf{P}) = \frac{trace(\mathbf{P}^T \mathbf{S}_b \mathbf{P})}{trace(\mathbf{P}^T \mathbf{S}_w \mathbf{P})}, \tag{2}$$

where the matrices $\mathbf{S}_b$, $\mathbf{S}_w$ are given by:

$$\mathbf{S}_b = \sum_{i=1}^{C} N_i \Big(\phi(\mathbf{m}_i) - \phi(\mathbf{m})\Big)\Big(\phi(\mathbf{m}_i) - \phi(\mathbf{m})\Big)^T, \tag{3}$$

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{j=1}^{N_i} \Big(\phi(\mathbf{x}_{ij}) - \phi(\mathbf{m}_i)\Big)\Big(\phi(\mathbf{x}_{ij}) - \phi(\mathbf{m}_i)\Big)^T. \quad (4)$$

In (3), (4), $\phi(\mathbf{m}_i)$ is the mean vector of class i in $\mathcal{F}$, i.e., $\phi(\mathbf{m}_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\phi(\mathbf{x}_{ij})$. $\phi(\mathbf{m})$ is the mean vector of the entire set in $\mathcal{F}$, i.e., $\phi(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}\phi(\mathbf{x}_{ij})$, where $N = \sum_{i=1}^{C} N_{ij}$. The direct maximization of (2) is intractable, since $\mathbf{S}_b$, $\mathbf{S}_w$ are matrices with arbitrary (possibly infinite) dimensions. In practice we overcome this problem by exploiting the so-called kernel trick ([**? ? ?**]). That is, the maximization of (2), as well as the multiplication in (1), are inherently computed by using dot-products in $\mathcal{F}$.

The maximization of (2) with respect to $\mathbf{P}$ leads to the determination of a data projection that can be used to map the original data to a reduced-dimensionality feature space where the data dispersion from the corresponding class mean vectors is minimized and the dispersion of class mean vectors from the total mean is maximized. In the cases where the classes (when represented in $\mathcal{F}$) follow normal distributions with the same covariance structure, by maximizing (2) maximal class discrimination can be achieved. However, this is a strong assumption which may not be met in many real problems. As has been shown in ([**?**]), the determination of optimized class representations enhances class discrimination in the projection space in the cases where the assumptions of LDA are not met. In the following, we describe an iterative optimization scheme that can be exploited in order to determine both the optimal class representations in $\mathcal{F}$ and the optimal projection for nonlinear data mapping exploiting such optimized representations.

3. Kernel Reference Discriminant Analysis

In this Section we describe in detail the proposed Kernel Reference Discriminant Analysis algorithm. Let us denote by $\mathbf{\Phi}_i$ a matrix containing the samples belonging to class i (represented in $\mathcal{F}$), i.e., $\mathbf{\Phi}_i = [\phi(\mathbf{x}_{i1}), \ldots, \phi(\mathbf{x}_{iN_i})]$. By using $\mathbf{\Phi}_i$, $i = 1, \ldots, C$ we can construct the matrix $\mathbf{\Phi} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_C]$ containing the representations of the entire data set in $\mathcal{F}$. The so-called kernel matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$ is given by $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$. Let us denote by $\mathbf{K}_i \in \mathbb{R}^{N\times N_i}$ a matrix containing the columns of $\mathbf{K}$ corresponding to the samples belonging to class i. That is, $\mathbf{K} = [\mathbf{K}_1, \ldots, \mathbf{K}_C]$, where $\mathbf{K}_i = \mathbf{\Phi}^T\mathbf{\Phi}_i$.

In KRDA, each class i is represented by a vector $\phi(\boldsymbol{\mu}_i)$. $\phi(\boldsymbol{\mu}_i)$ is not restricted to be the class mean in $\mathcal{F}$, but can be any vector enhancing class discrimination in the projection space $\mathbb{R}^d$. In order to determine both the optimal data projection

matrix $\mathbf{P}$ and the optimal class representations $\phi(\boldsymbol{\mu}_i)$, we propose to maximize the following criterion with respect to both $\mathbf{P}$ and $\boldsymbol{\mu}_i$:

$$\mathcal{J}_{KRDA}(\mathbf{P}, \boldsymbol{\mu}_i) = \frac{trace(\mathbf{P}^T \tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)\mathbf{P})}{trace(\mathbf{P}^T \tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i)\mathbf{P})}, \tag{5}$$

where the matrices $\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)$, $\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)$ are given by:

$$\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i) = \sum_{i=1}^{C} N_i \Big(\phi(\boldsymbol{\mu}_i) - \phi(\mathbf{m})\Big)\Big(\phi(\boldsymbol{\mu}_i) - \phi(\mathbf{m})\Big)^T, \tag{6}$$

$$\tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i) = \sum_{i=1}^{C} \sum_{j=1}^{N_i} \Big(\phi(\mathbf{x}_{ij}) - \phi(\boldsymbol{\mu}_i)\Big)\Big(\phi(\mathbf{x}_{ij}) - \phi(\boldsymbol{\mu}_i)\Big)^T. \tag{7}$$

$\tilde{\mathbf{S}}_w$ describes the class dispersion with respect to $\phi(\boldsymbol{\mu}_i)$ in $\mathcal{F}$. That is, the maximization of (5) leads to the determination of a data projection that can be used to map the original data to a reduced-dimensionality feature space $\mathbb{R}^d$, where the data dispersion from the corresponding class reference vector $\tilde{\boldsymbol{\mu}}_i = \mathbf{P}^T\phi(\boldsymbol{\mu}_i)$ is minimized, while the dispersion of the class reference vectors from the total mean is maximized. In the following, we assume that the data set is centered in $\mathcal{F}$[1].

3.1. Calculation of $\mathbf{P}$

In order to determine the optimal data projection matrix $\mathbf{P}$ we work as follows. Let us denote by $\mathbf{p}$ an eigenvector of the problem $\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)\mathbf{p} = \lambda\tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i)\mathbf{p}$ with eigenvalue λ. $\mathbf{p}$ can be expressed as a linear combination of the data (representated in $\mathcal{F}$) ([**?** **?** **?**]), i.e., $\mathbf{p} = \sum_{i=1}^{C}\sum_{j=1}^{N_i} a_{ij}\phi(\mathbf{x}_{ij}) = \boldsymbol{\Phi}\mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^N$. In addition, we can express $\phi(\boldsymbol{\mu}_i)$ as a linear combination of the samples belonging to class i, i.e., $\phi(\boldsymbol{\mu}_i) = \sum_{j=1}^{N_i} b_{ij}\phi(\mathbf{x}_{ij}) = \boldsymbol{\Phi}_i\mathbf{b}_i$, where $\mathbf{b}_i \in \mathbb{R}^{N_i}$. As it will be described in Appendix A, by setting $\mathbf{K}\mathbf{a} = \mathbf{u}$, the aforementioned eigenproblem can be transformed to the following equivalent eigenproblem:

$$\mathbf{B}(\mathbf{b}_i)\mathbf{u} = \lambda\mathbf{W}(\mathbf{b}_i)\mathbf{u}. \tag{8}$$

In (8), $\mathbf{B}(\mathbf{b}_i) = blockdiag(N_1\mathbf{b}_1\mathbf{b}_1^T, \ldots, N_C\mathbf{b}_C\mathbf{b}_C^T)$ and $\mathbf{W}(\mathbf{b}_i) = blockdiag(\mathbf{W}_1, \ldots, \mathbf{W}_C)$, where $\mathbf{W}_i = \mathbf{I}_{N_i} - \mathbf{1}_{N_i}\mathbf{b}_i^T - \mathbf{b}_i\mathbf{1}_{N_i}^T + N_i\mathbf{b}_i\mathbf{b}_i^T$

Thus the maximization of (5) can be approximated by applying a two step process:

[1]This can always be done by using $\tilde{\phi}(\mathbf{x}_{ij}) = \phi(\mathbf{x}_{ij}) - \phi(\mathbf{m})$, leading to a centered version of the kernel matrix given by $\tilde{\mathbf{K}} = \frac{1}{N}\mathbf{K}\mathbf{1} - \frac{1}{N}\mathbf{1}\mathbf{K} + \frac{1}{N^2}\mathbf{1}\mathbf{K}\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{N\times N}$ is a matrix of ones.

- Solution of the eigenproblem $\mathbf{B}(\mathbf{b}_i)\mathbf{u} = \lambda\mathbf{W}(\mathbf{b}_i)\mathbf{u}$, which is tractable since $\mathbf{B}(\mathbf{b}_i), \mathbf{W}(\mathbf{b}_i) \in \mathbb{R}^{N\times N}$. The solution of this problem leads to the determination of a matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$, where $\mathbf{u}_j$ is the eigenvector corresponding to the j-th non-zero eigenvalue.

- Determination of a matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_d]$, where $\mathbf{K}\mathbf{a}_j = \mathbf{u}_j$. In the case where $\mathbf{K}$ is non-singular, the vectors $\mathbf{a}_j,\ j = 1, \ldots, d$ are given by $\mathbf{a}_j = \mathbf{K}^{-1}\mathbf{u}_j$. When $\mathbf{K}$ is singular, the vectors $\mathbf{a}_j,\ j = 1, \ldots, d$ can be approximated by $\mathbf{a}_j = (\mathbf{K} + c\mathbf{I})^{-1}\,\mathbf{u}_j$, where c is a small positive value and $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix.

That is, $\mathbf{P}$ can be inherently determinated through the calculation of the reconstruction weights $\mathbf{A}$. After the calculation of $\mathbf{A}$, a vector $\mathbf{x}_t \in \mathbb{R}^D$ can be projected to the discriminant space $\mathbb{R}^d$ by applying:

$$\mathbf{y}_t = \mathbf{A}^T\mathbf{k}_t, \tag{9}$$

where $\mathbf{k}_t \in \mathbb{R}^N$ is a vector given by $\mathbf{k}_t = \mathbf{\Phi}^T\phi(\mathbf{x}_t)$.

3.2. Optimized class representation in $\mathcal{F}$

In order to maximize (5) with respect to the class representations $\boldsymbol{\mu}_i,\ i = 1, \ldots, C$, we also exploit the fact that $\mathbf{p} = \mathbf{\Phi}\mathbf{a}$ and $\phi(\boldsymbol{\mu}_i) = \mathbf{\Phi}_i\mathbf{b}_i$. As will be described in Appendix B, the optimization problem in (5) can be transformed to the following equivalent optimization problem:

$$\tilde{\mathcal{J}}_{KRDA}(\mathbf{A}, \mathbf{b}_i) = \frac{trace(\mathbf{A}\mathbf{B}(\mathbf{b}_i)\mathbf{A}^T)}{trace(\mathbf{A}\mathbf{W}(\mathbf{b}_i)\mathbf{A}^T)}. \tag{10}$$

By solving for $\nabla_{\mathbf{b}_i}\left(\tilde{\mathcal{J}}_{KRDA}\right) = 0$ we obtain:

$$\mathbf{b}_i = \frac{\gamma}{N_i}\mathbf{1}_{N_i}, \tag{11}$$

where $\mathbf{1}_{N_i} \in \mathbb{R}^{N_i}$ is a vector of ones. As will be described in Appendix C, γ is given by:

$$\gamma = \frac{trace\left(\sum_{i=1}^{C}\mathbf{A}\mathbf{K}_i\mathbf{K}_i^T\mathbf{A}^T\right)}{trace\left(\sum_{i=1}^{C}\frac{1}{N_i}\mathbf{A}\mathbf{K}_i\mathbf{1}_{N_i}\mathbf{1}_{N_i}^T\mathbf{K}_i^T\mathbf{A}^T\right)}. \tag{12}$$

After the calculation of $\mathbf{b}_i,\ i = 1, \ldots, C$, class i is represented in $\mathcal{F}$ by using $\phi(\boldsymbol{\mu}_i) = \sum_{j=1}^{N_i} b_{ij}\phi(\mathbf{x}_{ij})$.

3.3. Optimization with respect to both $\mathbf{A}$ *and* $\mathbf{b}_i$

Since $\mathcal{J}_{KRDA}$ is a function of both $\mathbf{A}$ and $\mathbf{b}_i$, we would like to determine a combination $\{\mathbf{A}, \mathbf{b}_i\}$ maximizing $\mathcal{J}_{KRDA}$. Taking into account that $\mathbf{A}$ is a function of $\mathbf{b}_i$ and that $\mathbf{b}_i$ is a function of $\mathbf{A}$ (as has been explained in Subsections 3.1 and 3.2, respectively), a direct maximization of $\mathcal{J}_{KRDA}$ with respect to both $\mathbf{A}$ and $\mathbf{b}_i$ is difficult. In order to maximize $\mathcal{J}_{KRDA}$ with respect to both $\mathbf{A}$ and $\mathbf{b}_i$, we employ an iterative optimization scheme. In the following, we introduce a index t denoting the iteration of the adopted iterative optimization scheme.

Let us denote by $\mathbf{b}_{i,t},\ i = 1, \ldots, C$ the class vectors that have been calculated at the t-th iteration of the optimization scheme. By using $\mathbf{b}_{i,t}$, the data projection matrix $\mathbf{A}_t$ can be calculated by following the process described in subsection 3.1. After the calculation of $\mathbf{A}_t$, $\mathbf{b}_{i,t+1}$ can be calculated by using (11). The above described process is initialized by using the class mean vectors, i.e., $\mathbf{b}_i = \frac{1}{N_i},\ i = 1, \ldots, C$ and is terminated when $(\mathcal{J}_{KRDA}(t+1) - \mathcal{J}_{KRDA}(t))/\mathcal{J}_{KRDA}(t) < \epsilon$, where ϵ is a small positive value, which is set equal to $\epsilon = 10^{-6}$ in our experiments.

4. Experiments

In this Section we describe experiments conducted in order to compare the performance of the proposed KRDA algorithm with that of SRKDA ([**?**]) employing the class mean vectors for class representation. We also compare the proposed approach with that of Kernel Local Fisher Discriminant Analysis (KLFDA) algorithm ([**?**]). We have applied the three algorithms on standard classification problems coming from the UCI repository ([**?**]). These experiments will be described in Subsection 4.1. We have also applied the algorithms on human action and face recognition problems. These experiments will be described in Subsections 4.2 and 4.3, respectively. In all the experiments we have employed SRKDA, KRDA and KLFDA in order to map the data to the corresponding discriminant subspace $\mathbb{R}^d$. Subsequently, classification is performed by using the class mean vectors for the SRKDA and KLFDA-based classification schemes. For the proposed KRDA-based classification scheme, classification is performed by using the class reference vectors.

4.1. Experiments on Standard Classification Problems

We have conducted experiments on twenty publicly available classification datasets coming from the machine learning repository of University of California Irvine (UCI) ([**?**]). On each dataset, the 5-fold cross-validation procedure has

been performed for the SRKDA, KLFDA and the proposed KRDA algorithms by using the same data partitioning. The mean classification rate over all folds has been used to measure the performance of each algorithm in one experiment. Ten experiments have been performed for each data set. The mean classification rate and the observed standard deviation over all experiments have been used to measure the performance of each algorithm. In all these experiments we have employed the RBF kernel function, i.e.:

$$\left[\mathbf{K}\right]_{l,m} = exp\left(-g\|\mathbf{x}_l - \mathbf{x}_m\|_2^2\right). \tag{13}$$

The value of parameter g has been automatically chosen in each fold from a set $g = 10^r$, $r = -6, \ldots, 6$, by applying 5-fold cross-validation on the corresponding training set. It should be noted here that since the value of g is empirically chosen based on performance criteria, its optimal value may be different for KRDA, KLFDA and SRKDA.

The mean classification rates and the observed standard deviations over all experiments for each data set are illustrated in Table 1. By observing this Table, it can be seen that the proposed KRDA algorithm provides the highest performance in twelve, out of twenty, classification problems. In addition, it can be seen that KRDA outperforms SRKDA in fourteen datasets.

4.2. Experiments on Human Action Recognition

We have conducted experiments on two publicly available action recognition datasets, namely the Hollywood2 and the Olympic Sports datasets. A brief description of the datasets and the experimental protocols used in our experiments is given in the following Subsections. We have employed the Bag-of- Words (BoW)-based video representation by using HOG, HOF, MBHx, MBHy and Trajectory descriptors evaluated on the trajectories of densely sampled interest points ([**?**]). Following ([**?**]), we set the number of codebook vectors for each descriptor type equal to $D_k = 4000$ and employ the χ^2 kernel function:

$$\left[\mathbf{K}\right]_{l,m}^{k} = exp\left(\frac{1}{\sigma_k}\sum_{n=1}^{D_k}\frac{\left(x_{ln}^k - x_{mn}^k\right)^2}{x_{ln}^k + x_{mn}^k}\right). \tag{14}$$

The value of parameter σ_k has been determined by applying 5-fold cross validation on the training vectors of descriptor k using the values $\sigma = 2^r$, $r = 0, \ldots, 3$. Different descriptors are finally combined by exploiting a multi-channel approach ([**?**]), i.e., $\left[\mathbf{K}\right]_{l,m} = \prod_{k=1}^{K}\left[\mathbf{K}\right]_{l,m}^{k}$.

Table 1: UCI datasets details and mean classification rate and standard deviation (%) for 5-fold cross-validation.

Dataset	KLFDA	SRKDA	**KRDA**
Abalone	**54.29 (±0.12)**	52.85 (±0.69)	54.19 (±0.4)
Australian	67.83 (±1.9)	77.65 (±1.78)	**77.87 (±1.56)**
Columns2C	80.77 (±1.39)	82.61 (±1.7)	**82.71 (±2.33)**
Columns3C	76.74 (±2.15)	83.42 (±1.16)	**83.58 (±1.42)**
German	70.46 (±0.6)	70.65 (±1.18)	**72.16 (±1.25)**
Glass	67.66 (±1.57)	67.66 (±3.6)	**68.36 (±3.39)**
Heart	76.04 (±1.19)	76.04 (±2.84)	**76.15 (±2.07)**
Indians	68.41 (±1.8)	72.24 (±1.04)	**74.61 (±1.66)**
Ionosphere	89.63 (±1.86)	**90.23 (±1.03)**	90.03 (±0.93)
Iris	**86.2 (±2.13)**	80.53 (±2.79)	85.07 (±2.71)
Libras	82.53 (±1.77)	**85.03 (±1.48)**	84.69 (±1.54)
Madelon	**67.18 (±0.8)**	**67.18 (±0.83)**	**67.18 (±0.83)**
OptDigits	**99.06 (±0.01)**	98.64 (±0.24)	98.66 (±0.23)
Relax	**73.39 (±1.41)**	72.71 (±0.1)	71.43 (±0.43)
Segmentation	91.46 (±0.63)	95.66 (±0.48)	**95.73 (±0.69)**
Spect	80.19 (±1.86)	79.59 (±1.9)	**81.09 (±1.17)**
SpectF	77.98 (±1.63)	77.87 (±1.73)	**79.14 (±1.46)**
SynthCon	98.3 (±0.66)	99.45 (±0.11)	**99.57 (±0.25)**
TeachAss	56.89 (±4.69)	56.03 (±6.35)	**58.28 (±2.84)**
Tic-tac-toe	**99.13 (±0.45)**	**99.13 (±1.41)**	**99.13 (±0.41)**

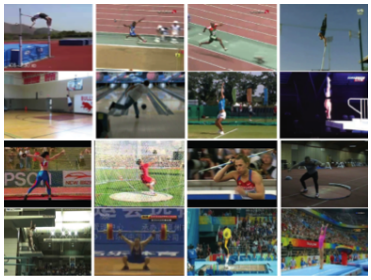

Figure 1: *Video frames from the Olympic Sports dataset depicting instances of all the sixteen actions.*

4.2.1. The Olympic Sports dataset

This dataset consists of 783 videos depicting athletes practicing 16 sports ([**?**]). The actions appearing in the dataset are: high-jump, long-jump, triple-jump, pole-vault, basketball lay-up, bowling, tennis-serve, platform, discus, hammer, javelin, shot-put, springboard, snatch, clean-jerk and vault. Example video frames from the dataset are illustrated in Figure 1. We used the standard training-test split provided by the dataset (649 videos are used for training and performance is measured in the remaining 134 videos). The performance is evaluated by computing the average precision (AP) for each action class and reporting the mean AP over all classes (mAP), as suggested in ([**?**]).

4.2.2. The Hollywood2 dataset

This dataset consists of 1707 videos depicting 12 actions ([**?**]). The videos have been collected from 69 different Hollywood movies. The actions appearing in the dataset are: answering the phone, driving car, eating, fighting, getting out of car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. Example video frames from this dataset are illustrated in Figure 2. We used the standard training-test split provided by the dataset (823 videos are used for training and performance is measured in the remaining 884 videos). Training and test videos come from different movies. The performance is evaluated by computing the mean Average Precision (mAP) over all classes, as suggested in ([**?**]).

4.2.3. Experimental Results

The AP values for different actions on the Olympic Sports and Hollywood2 datasets are illustrated in Tables 2 and 3, respectively. As it can be seen in these Tables, KRDA provides the highest AP value in nine and seven class-specific clas-

Figure 2: *Video frames from the Hollywood2 dataset depicting instances of all the twelve actions.*

sification problems on the Olympic Sports and the Hollywood2 datasets, respectively. Overall, KRDA outperforms both SRKDA and KLFDA in both datasets providing mAP values equal to 83.35% and 61.22% on the Olympic Sports and Hollywood2 datasets, respectively.

4.3. Experiments on Face Recognition

We have conducted experiments on three publicly available face recognition datasets, namely the ORL, AR and Extended YALE-B datasets. A brief description of the datasets is given in the following Subsections. We have used the facial images provided by the databases and resized them to fixed size images of 40×30 pixels. The resized facial images have been vectorized to produce 1200-dimensional facial vectors. The dimensionality of the facial vectors has been further reduced by applying PCA so that 90% of the dataset energy is preserved. The 5-fold cross-validation procedure has been performed for the SRKDA, KRDA and KLFDA algorithms. The mean classification rate over all folds has been used to measure the performance of each algorithm in one experiment. Ten experiments have been performed in total for each dataset. In all the experiments we have employed the RBF kernel function (13). The value of parameter g has been automatically chosen for each fold from a set $g = 10^r,\ r = -6, \ldots, 6$, by applying 5-fold cross-validation on the corresponding training set.

4.3.1. The ORL dataset

This dataset contains 10 images of 40 persons, leading to a total number of 400 images ([**?**]). The images were captured at different times and with different conditions, in terms of lighting, facial expressions (smiling/not smiling) and facial details (open/closed eyes, with/without glasses). Facial images were taken in frontal position with a tolerance for face rotation and tilting up to 20 degrees.

4.3.2. The AR dataset

This dataset contains over 4000 images depicting 70 male and 56 female faces ([**?**]). In our experiments we have used the preprocessed (cropped) facial images

Table 2: Average precisions on the Olympic Sports dataset.

	KLFDA	SRKDA	**KRDA**
Basketball lay-up	91.31%	93.94%	**96.25%**
Bowling	79.85%	81.48%	**82.86%**
Clean and Jerk	**78.72%**	78.41%	78.41%
Discus Throw	**84.18%**	82.53%	83.79%
Diving 3m	**100%**	**100%**	**100 %**
Diving 10m	95.46%	96.69%	**97.27%**
Hammer Throw	88.79%	88.78%	**90.19%**
High Jump	63.44%	64.92%	**65.81%**
Javelin Throw	**100%**	**100%**	**100 %**
Long Jump	**88.31%**	**88.31%**	**88.31%**
Pole Vault	81.01%	81.01%	**84.41%**
Shot Put	**74.35%**	68.89%	69.76%
Snatch	65.35%	63.18%	**68.11%**
Triple Jump	33.33%	39.50%	**48.14%**
Tennis Serve	95.96%	92.09%	**97.72%**
Vault	**83.97%**	80.66%	82.55%
Mean	81.54%	81.27%	**83.35%**

Table 3: Average precisions on the Hollywood2 dataset.

	KLFDA	SRKDA	**KRDA**
Answer Phone	34.1%	27.46%	**36.09%**
Drive Car	87.8%	89.23%	**90.56%**
Eat	62.74%	67.41%	**69.23%**
Fight	**82.04%**	80.09%	79.55%
Get Out of Car	51.71%	**60.52%**	59.56%
Hand Shake	32.23%	36.71%	**38.73%**
Hug Person	51.52%	49.97%	**53.93%**
Kiss	**66.7%**	59.74%	62.05%
Run	**83.1%**	82.37%	81.77%
Sit Down	64.06%	**68.45%**	67.57%
Sit up	19.88%	22.2%	**22.43%**
Stand up	67.63%	64.56%	**73.22%**
Mean	58.63%	59.06%	**61.22%**

provided by the database, depicting 100 persons (50 males and 50 females) having a frontal facial pose, performing several expressions (anger, smiling and screaming), in different illumination conditions (left and/or right light) and with some occlusions (sun glasses and scarf). Each person was recorded in two sessions, separated by two weeks.

4.3.3. The Extended YALE-B dataset

This dataset contains images of 38 persons in 9 poses, under 64 illumination conditions ([**?**]). In our experiments we have used the frontal cropped images provided by the database.

4.3.4. Experimental Results

The mean classification rates and the observed standard deviations over all experiments for each data set are illustrated in Table 4. By observing this Table, it can be seen that the proposed KRDA algorithm outperforms both SRKDA and KLFDA in the ORL dataset, while KLFDA outperforms SRKDA and KRDA in the remaining two datasets. This can be explained by taking into account the multimdality of facial classes in these two datasets. In addition, it can be seen that the proposed KRDA outperforms SRKDA in all the three datasets.

Table 4: Performance (%) on Face Recognition.

	KLFDA	SRKDA	**KRDA**
ORL	95.05 ($\pm$0.81)	96.43 ($\pm$0.42)	**96.5 ($\pm$0.37)**
AR	**83.65 ($\pm$0.81)**	81.96 ($\pm$0.6)	82.22 ($\pm$0.4)
YALE	**94.02 ($\pm$0.25)**	91.95 ($\pm$0.5)	92.14 ($\pm$0.79)

5. Conclusions

In this paper, we described an optimization scheme aiming at determining the optimal class representation in arbitrary-dimensional Hilbert spaces for KDA-based data projection. By optimizing the KDA criterion with respect to both the data projection matrix and the class representation in the projection space, the optimal discriminant projection space, in terms of Fisher ratio maximization, is obtained. Experimental results on standard classification problems, as well as on human action and face recognition problems denote that the adopted approach increases class discrimination, when compared to the standard KDA approach.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART).

Appendix A. Derivation of $\mathbf{B}(\mathbf{b}_i)\mathbf{u} = \boldsymbol{\lambda}\mathbf{W}(\mathbf{b}_i)\mathbf{u}$

We expand $\tilde{\mathbf{S}}_b\mathbf{p}$:

$$\tilde{\mathbf{S}}_b\mathbf{p} = \Big(\sum_{i=1}^{C} N_i\boldsymbol{\Phi}_i\mathbf{b}_i\mathbf{b}_i^T\boldsymbol{\Phi}_i^T\Big)\Big(\boldsymbol{\Phi}\mathbf{a}\Big) = \sum_{i=1}^{C} N_i\boldsymbol{\Phi}_i\mathbf{b}_i\mathbf{b}_i^T\boldsymbol{\Phi}_i^T\boldsymbol{\Phi}\mathbf{a}. \tag{A.1}$$

By using (A.1), we obtain:

$$\begin{aligned}\boldsymbol{\Phi}^T\tilde{\mathbf{S}}_b\mathbf{p} &= \sum_{i=1}^{C} N_i\boldsymbol{\Phi}^T\boldsymbol{\Phi}_i\mathbf{b}_i\mathbf{b}_i^T\boldsymbol{\Phi}_i^T\boldsymbol{\Phi}\mathbf{a} = \sum_{i=1}^{C} N_i\mathbf{K}_i\mathbf{b}_i\mathbf{b}_i^T\mathbf{K}_i^T\mathbf{a} \\ &= \sum_{i=1}^{C}\mathbf{K}_i\mathbf{B}_i\mathbf{K}_i^T\mathbf{a} = \mathbf{KBKa},\end{aligned} \tag{A.2}$$

where $\mathbf{B}_i = N_i \mathbf{b}_i \mathbf{b}_i^T$ and $\mathbf{B} = blockdiag(N_1 \mathbf{b}_1 \mathbf{b}_1^T, \ldots, N_C \mathbf{b}_C \mathbf{b}_C^T)$.

We expand $\tilde{\mathbf{S}}_w \mathbf{p}$:

$$\begin{aligned} \tilde{\mathbf{S}}_w \mathbf{p} &= \sum_{i=1}^{C} \Big(\mathbf{\Phi}_i \mathbf{\Phi}_i^T \mathbf{\Phi a} - \mathbf{\Phi}_i \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{\Phi}_i^T \mathbf{\Phi a} \\ &- \mathbf{\Phi}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{\Phi}_i^T \mathbf{\Phi a} + \mathbf{\Phi}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{\Phi}_i^T \mathbf{\Phi a} \Big). \end{aligned} \tag{A.3}$$

By using (A.3), we obtain:

$$\begin{aligned} \mathbf{\Phi}^T \tilde{\mathbf{S}}_w \mathbf{p} &= \sum_{i=1}^{C} \Big(\mathbf{\Phi}^T \mathbf{\Phi}_i \mathbf{\Phi}_i^T \mathbf{\Phi a} - \mathbf{\Phi}^T \mathbf{\Phi}_i \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{\Phi}_i^T \mathbf{\Phi a} \\ &- \mathbf{\Phi}^T \mathbf{\Phi}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{\Phi}_i^T \mathbf{\Phi a} + \mathbf{\Phi}^T \mathbf{\Phi}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{\Phi}_i^T \mathbf{\Phi a} \Big) \\ &= \sum_{i=1}^{C} \Big(\mathbf{K}_i \mathbf{K}_i^T \mathbf{a} - \mathbf{K}_i \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{K}_i^T \mathbf{a} \\ &- \mathbf{K}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{K}_i^T \mathbf{a} + \mathbf{K}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{K}_i^T \mathbf{a} \Big) \\ &= \sum_{i=1}^{C} \mathbf{K}_i \mathbf{W}_i \mathbf{K}_i^T \mathbf{a} = \mathbf{KWKa} \end{aligned} \tag{A.4}$$

where $\mathbf{W}_i = \mathbf{I}_{N_i} - \mathbf{1}_{N_i} \mathbf{b}_i^T - \mathbf{b}_i \mathbf{1}_{N_i}^T + N_i \mathbf{b}_i \mathbf{b}_i^T$ and $\mathbf{W} = blockdiag(\mathbf{W}_1, \ldots, \mathbf{W}_C)$.

By using $\mathbf{Ka} = \mathbf{u}$, and (A.2), (A.4) the result follows.

Appendix B. Derivation of $\tilde{\mathcal{J}}_{KRDA}$

We expand $\mathbf{P}^T \tilde{\mathbf{S}}_b \mathbf{P}$:

$$\begin{aligned} \mathbf{P}^T \tilde{\mathbf{S}}_b \mathbf{P} &= \sum_{i=1}^{C} N_i \mathbf{A} \mathbf{\Phi}^T \mathbf{\Phi}_i \mathbf{b}_i \mathbf{b}_i^T \mathbf{\Phi}_i^T \mathbf{\Phi} \mathbf{A}^T \\ &= \mathbf{A} \Big(\sum_{i=1}^{C} \mathbf{K}_i \mathbf{B}_i \mathbf{K}_i^T \Big) \mathbf{A}^T = \mathbf{ABA}^T. \end{aligned} \tag{B.1}$$

We expand $\mathbf{P}^T\tilde{\mathbf{S}}_w\mathbf{P}$:

$$\begin{aligned}
\mathbf{P}^T\tilde{\mathbf{S}}_w\mathbf{P} &= \sum_{i=1}^{C}\Big(\mathbf{A}\mathbf{\Phi}^T\mathbf{\Phi}_i\mathbf{\Phi}_i^T\mathbf{\Phi}\mathbf{A}^T \\
&- \mathbf{A}\mathbf{\Phi}^T\mathbf{\Phi}_i\mathbf{1}_{N_i}\mathbf{b}_i^T\mathbf{\Phi}_i^T\mathbf{\Phi}\mathbf{A}^T \\
&- \mathbf{A}\mathbf{\Phi}^T\mathbf{\Phi}_i\mathbf{b}_i\mathbf{1}_{N_i}^T\mathbf{\Phi}_i^T\mathbf{\Phi}\mathbf{A}^T \\
&+ \mathbf{A}\mathbf{\Phi}^T\mathbf{\Phi}_i\mathbf{b}_i\mathbf{1}_{N_i}^T\mathbf{1}_{N_i}\mathbf{b}_i^T\mathbf{\Phi}_i^T\mathbf{\Phi}\mathbf{A}^T\Big) \\
&= \sum_{i=1}^{C}\Big(\mathbf{A}\mathbf{K}_i\mathbf{K}_i^T\mathbf{A}^T - \mathbf{A}\mathbf{K}_i\mathbf{1}_{N_i}\mathbf{b}_i^T\mathbf{K}_i^T\mathbf{A}^T \\
&- \mathbf{A}\mathbf{K}_i\mathbf{b}_i\mathbf{1}_{N_i}^T\mathbf{K}_i^T\mathbf{A}^T + \mathbf{A}\mathbf{K}_i\mathbf{b}_i\mathbf{1}_{N_i}^T\mathbf{1}_{N_i}\mathbf{b}_i^T\mathbf{K}_i^T\mathbf{A}^T\Big) \\
&= \mathbf{A}\Big(\sum_{i=1}^{C}\mathbf{K}_i\mathbf{W}_i\mathbf{K}_i^T\Big)\mathbf{A}^T = \mathbf{A}\mathbf{W}\mathbf{A}^T. \qquad \text{(B.2)}
\end{aligned}$$

By using (B.1), (B.2) the result follows.

Appendix C. Derivation of γ

By using (B.1), (B.2) we have:

$$\nabla_{\mathbf{b}_i}\left(trace(\mathbf{A}\mathbf{B}\mathbf{A}^T)\right) = 2N_i\mathbf{K}_i^T\mathbf{A}^T\mathbf{A}\mathbf{K}_i\mathbf{b}_i, \qquad \text{(C.1)}$$

$$\nabla_{\mathbf{b}_i}\left(trace(\mathbf{A}\mathbf{W}\mathbf{A}^T)\right) = 2\mathbf{K}_i^T\mathbf{A}^T\mathbf{A}\mathbf{K}_i\Big(N_i\mathbf{b}_i - \mathbf{1}_{N_i}\Big). \qquad \text{(C.2)}$$

By using (C.1), (C.2) and solving for $\frac{\vartheta\mathcal{J}_{KRDA}}{\vartheta\mathbf{b}_i} = 0$ we obtain:

$$\mathbf{b}_i = \frac{trace(\mathbf{A}\mathbf{B}\mathbf{A}^T)}{N_i(trace(\mathbf{A}\mathbf{B}\mathbf{A}^T) - trace(\mathbf{A}\mathbf{W}\mathbf{A}^T))}\mathbf{1}_{N_i} = \frac{\gamma}{N_i}\mathbf{1}_{N_i}. \qquad \text{(C.3)}$$

By substituting (C.3) to (B.1), (B.2) and solving for $\frac{\vartheta\mathcal{J}_{KRDA}}{\vartheta\gamma} = 0$ we obtain the following two solutions: $\gamma_1 = 0$ and $\gamma_2 = f/c$, where $f = trace\left(\sum_{i=1}^{C}\mathbf{A}\mathbf{K}_i\mathbf{K}_i^T\mathbf{A}^T\right)$ and $c = trace\left(\sum_{i=1}^{C}\frac{1}{N_i}\mathbf{A}\mathbf{K}_i\mathbf{1}_{N_i}\mathbf{1}_{N_i}^T\mathbf{K}_i^T\mathbf{A}^T\right)$. It is straightforward to se that the solution $\gamma_1 = 0$ leads to the minimization of $\mathcal{J}_{KRDA} = 0$ and the solution

$\gamma_2 = f/c$ leads to the maximization of $\mathcal{J}_{KRDA}$. Thus, we employ the following value for the maximization of $\mathcal{J}_{KRDA}$:

$$\gamma = \frac{trace\left(\sum_{i=1}^{C} \mathbf{A}\mathbf{K}_i\mathbf{K}_i^T\mathbf{A}^T\right)}{trace\left(\sum_{i=1}^{C} \frac{1}{N_i}\mathbf{A}\mathbf{K}_i\mathbf{1}_{N_i}\mathbf{1}_{N_i}^T\mathbf{K}_i^T\mathbf{A}^T\right)}. \tag{C.4}$$