ENHANCING CLASS DISCRIMINATION IN KERNEL DISCRIMINANT ANALYSIS

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
54124, Thessaloniki, Greece
Email: {tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper, we propose an optimization scheme aiming at optimal nonlinear data projection, in terms of Fisher ratio maximization. To this end, we formulate an iterative optimization scheme consisting of two processing steps: optimal data projection calculation and optimal class representation determination. Compared to the standard approach employing the class mean vectors for class representation, the proposed optimization scheme increases class discrimination in the reduced-dimensionality feature space. We evaluate the proposed method in standard classification problems, as well as on the classification of human actions and face, and show that it is able to achieve better generalization performance, when compared to the standard approach.

Index Terms— Kernel Discriminant Analysis, Optimized Class Representation, Nonlinear data projection

1. INTRODUCTION

Kernel Discriminant Analysis (KDA) is a well-known algorithm for supervised feature extraction and dimensionality reduction. It aims at the determination of an optimal subspace for nonlinear data projection, in which the classes are better discriminated [1, 2, 3, 4, 5, 6, 7]. It exploits data representations in an arbitrary-dimensional feature space determined by applying a non-linear data mapping process (and exploiting the so-called kernel trick [8, 9, 10]). 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. The adopted class discrimination criterion is the ratio of the between-class scatter to the within-class scatter in the reduced-dimensionality feature space, which is usually referred as the Fisher ratio.

KDA optimality is based on the assumptions of: a) normal class distributions with the same covariance structure in the kernel space and b) class representation by the corresponding class mean vector (determined in the kernel space). Under these assumptions, the maximization of the Fisher ratio leads to maximal class discrimination in the reduced-dimensionality feature space. Under the assumption of normal class distributions in the kernel space, the assumption that each class should be represented by the class mean vector seems reasonable. However, the normality assumption is restrictive and difficult to be met. Recently, by observing that the between-class and within-class scatter matrices employed for linear data projection in Linear Discriminant Analysis (LDA) can be considered to be functions of the adopted class representation, it has been 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 [11]. In addition, it has been proven 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 for the case of LDA, an iterative optimization scheme has been proposed [11]. The outcomes of [11] have also been verified in [12], where Particle Swarm Optimization-based Fisher ratio maximization has been employed for the maximization of the LDA criterion.

In this paper, 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 feature space of increased discrimination power is subsequently calculated. We prove that the determination of the optimal class representation in the arbitrary-dimensional feature space has a closed form solution and formulate an iterative optimization scheme for the determination of both the optimal class representations and the optimal nonlinear data projection, in terms of Fisher ratio maximization. The proposed method is evaluated on standard classification problems, as well as on two computer vision problems problems, i.e., the recognition of human actions and face. Experimental results show that the proposed method is able to enhance class discrimination and achieve better performance.

The rest of the paper is structured as follows. The proposed method is described in detail in Section 2. Experimental results on two human action recognition and three face recognition datasets are provided in Section 3. Finally, con-

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

clusions are drawn in Section 4.

2. PROPOSED METHOD

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 $\mathcal{F}$ (having the properties of Hilbert spaces) [8, 9, 10] 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 $\mathcal{F}$.

Let us denote by $\mathbf{\Phi}_i \in \mathbb{R}^{|\mathcal{F}| \times N_i}$ a matrix containing the samples belonging to class i (represented in $\mathcal{F}$). 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 the proposed method, each class i is represented by a vector $\phi(\boldsymbol{\mu}_i)$. We do not set the assumption that the class representation must be the class mean vector in $\mathcal{F}$. $\phi(\boldsymbol{\mu}_i)$ 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 $\phi(\boldsymbol{\mu}_i)$, $i = 1, \ldots, C$, we propose to maximize the following criterion with respect to both $\mathbf{P}$ and $\boldsymbol{\mu}_i$:

$$\mathcal{J}(\mathbf{P}, \boldsymbol{\mu}_i) = \frac{trace(\mathbf{P}^T \tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i) \mathbf{P})}{trace(\mathbf{P}^T \tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i) \mathbf{P})}, \tag{1}$$

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

$$\tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i) = \sum_{i=1}^{C} \sum_{j=1}^{N_i} \Big(\phi(\mathbf{x}_{ij}) - \phi(\boldsymbol{\mu}_i)\Big)\Big(\phi(\mathbf{x}_{ij}) - \phi(\boldsymbol{\mu}_i)\Big)^T, \tag{2}$$

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

$\phi(\mathbf{m})$ is the mean vector of the entire dataset in $\mathcal{F}$. In the following, we assume that the data set is centered in $\mathcal{F}$. 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.

The maximization of (1) 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 vector $\tilde{\boldsymbol{\mu}}_i = \mathbf{P}^T \phi(\boldsymbol{\mu}_i)$ is minimized, while the dispersion of the class vectors belonging to different classes from the total mean is maximized. In order to determine both the optimal data projection $\mathbf{P}$ and the optimal class vectors $\phi(\boldsymbol{\mu}_i)$ we employ an iterative optimization scheme formed by two processing steps. In the following, we describe them in detail.

2.1. Calculation of $\mathbf{P}$

In order to determine the optimal data projection matrix $\mathbf{P}$ we work as follows [4]. 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 $\mathcal{F}$) [8, 9, 10], i.e., $\mathbf{p} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} a_{ij} \phi(\mathbf{x}_{ij}) = \mathbf{\Phi}\mathbf{a}$, 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., $\phi(\boldsymbol{\mu}_i) = \sum_{j=1}^{N_i} b_{ij} \phi(\mathbf{x}_{ij}) = \mathbf{\Phi}_i \mathbf{b}_i$, where $\mathbf{b}_i \in \mathbb{R}^{N_i}$. By setting $\mathbf{K}\mathbf{a} = \mathbf{u}$, the aforementioned eigenproblem can be transformed to the following equivalent eigenproblem:

$$\mathbf{B}(\mathbf{b}_i)\mathbf{u} = \lambda \mathbf{W}(\mathbf{b}_i)\mathbf{u}, \tag{4}$$

where $\mathbf{B}(\mathbf{b}_i) = blockdiag(N_1 \mathbf{b}_1 \mathbf{b}_1^T, \ldots, N_C \mathbf{b}_C \mathbf{b}_C^T)$ and $\mathbf{W}(\mathbf{b}_i) = blockdiag(\mathbf{W}_1, \ldots, \mathbf{W}_C)$, with $\mathbf{W}_i = \mathbf{I}_{N_i} - \mathbf{1}_{N_i} \mathbf{b}_i^T - \mathbf{b}_i \mathbf{1}_{N_i}^T + N_i \mathbf{b}_i \mathbf{b}_i^T$. Both $\mathbf{B}(\mathbf{b}_i), \mathbf{W}(\mathbf{b}_i) \in \mathbb{R}^{N \times N}$.

Thus the maximization of (1) can be approximated by applying a two step process:

- Solution of the eigenproblem $\mathbf{B}(\mathbf{b}_i)\mathbf{u} = \lambda \mathbf{W}(\mathbf{b}_i)\mathbf{u}$. By keeping the eigenvectors corresponding to the d maximal eigenvalues, a matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ is obtained.
- Calculation of the projection 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.

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

2.2. Calculation of $\phi(\boldsymbol{\mu}_i)$, $i = 1, \ldots, C$

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

$$\tilde{\mathcal{J}}(\mathbf{A}, \mathbf{b}_i) = \frac{trace(\mathbf{A}\mathbf{B}\mathbf{A}^T)}{trace(\mathbf{A}\mathbf{W}\mathbf{A}^T)}, \tag{5}$$

where $\mathbf{B} = \sum_{i=1}^{C} N_i \mathbf{\Phi}^T \mathbf{\Phi}_i \mathbf{b}_i \mathbf{b}_i^T \mathbf{\Phi}_i^T \mathbf{\Phi}$ and $\mathbf{W} = \sum_{i=1}^{C} \Big(\mathbf{K}_i \mathbf{K}_i^T - \mathbf{K}_i \mathbf{1}_{N_i} \mathbf{b}_i^T \mathbf{K}_i^T - \mathbf{K}_i \mathbf{b}_i \mathbf{1}_{N_i}^T \mathbf{K}_i^T + N_i \mathbf{K}_i \mathbf{b}_i \mathbf{b}_i^T \mathbf{K}_i^T\Big)$.

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

$$\mathbf{b}_i = \frac{\gamma}{N_i}\mathbf{1}_{N_i}. \tag{6}$$

In the above, $\mathbf{1}_{N_i} \in \mathbb{R}^{N_i}$ is a vector of ones. γ is given by:

$$\gamma = \frac{trace\left(\sum_{i=1}^{C}\mathbf{A}\mathbf{K}_i\mathbf{K}_i^T\mathbf{A}^T\right)}{trace\left(\sum_{i=1}^{C}\frac{1}{N_i}\mathbf{A}\mathbf{K}_i\mathbf{1}_{N_i}\mathbf{1}_{N_i}^T\mathbf{K}_i^T\mathbf{A}^T\right)}. \tag{7}$$

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

2.3. Iterative Optimization Scheme

Taking into account that $\mathbf{A}$ is a function of $\mathbf{b}_i,\ i = 1,\dots,C$ and that $\mathbf{b}_i$ is a function of $\mathbf{A}$, a direct maximization of $\mathcal{J}$ with respect to both $\mathbf{A}$ and $\mathbf{b}_i$ is difficult. In order to maximize $\mathcal{J}$ with respect to both $\mathbf{A}$ and $\mathbf{b}_i$, we employ the following iterative optimization scheme. Let us denote by $\mathbf{b}_{i,t},\ i = 1,\dots,C$ the class vectors 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 2.1. After the calculation of $\mathbf{A}_t$, $\mathbf{b}_{i,t+1}$ can be calculated by using (6). The above described process is initialized by using the class mean vectors, i.e., $\mathbf{b}_i = \frac{1}{N_i},\ i = 1,\dots,C$ and is terminated when $(\mathcal{J}(t+1)-\mathcal{J}(t))/\mathcal{J}(t) < \epsilon$, where ϵ is a small positive value, equal to $\epsilon = 10^{-6}$ in our experiments.

3. EXPERIMENTS

In this Section we describe experiments conducted in order to compare the performance of the proposed method with that of KDA [4] employing the class mean vectors for class representation. We have applied the two algorithms on standard classification problems, as well as on human action and face recognition problems. Experiments conducted on standard classification problems are described in Subsection 3.1. Experiments conducted on publicly available action and face recognition databases will be described in Subsections 3.2 and 3.3, respectively. In all the experiments we have employed the proposed method and KDA-based data projection 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 KDA-based classification scheme. For the proposed classification scheme, classification is performed by using the class reference vectors.

3.1. Experiments on Standard Classification Problems

We have conducted experiments on eight publicly available classification datasets coming from the machine learning repository of University of California Irvine (UCI) ([13]). On each dataset, the 5-fold cross-validation procedure has been performed by using the same data partitioning for the two classification schemes. 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 $\left[\mathbf{K}\right]_{l,m} = exp\left(-g\|\mathbf{x}_l - \mathbf{x}_m\|_2^2\right)$. The value of parameter g has been automatically chosen in each fold from a set $g = 10^r,\ r = -6,\dots,6$, by applying 5-fold cross-validation on the corresponding training set.

Table 1. Mean classification rate and standard deviation (%) on UCI datasets.

Dataset	KDA	**Proposed**
Abalone	52.85 (±0.69)	**54.19 (±0.4)**
German	70.65 (±1.18)	**72.16 (±1.25)**
Glass	67.66 (±3.6)	**68.36 (±3.39)**
Indians	72.24 (±1.04)	**74.61 (±1.66)**
Iris	80.53 (±2.79)	**85.07 (±2.71)**
Spect	79.59 (±1.9)	**81.09 (±1.17)**
SpectF	77.87 (±1.73)	**79.14 (±1.46)**
TeachAss	56.03 (±6.35)	**58.28 (±2.84)**

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 method outperforms KDA in all datasets, enhancing its performance by $1-5\%$.

3.2. 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. 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 [14]. Following [14], we set the number of codebook vectors for each descriptor type equal to $D_k = 4000$ and employ the χ^2 kernel function $\left[\mathbf{K}\right]_{l,m}^{k} = exp\left(\frac{1}{\sigma_k}\sum_{n=1}^{D_k}\frac{\left(x_{ln}^k - x_{mn}^k\right)^2}{x_{ln}^k + x_{mn}^k}\right)$. 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 $\sigma = 2^r,\ r = 0,\dots,3$. Different descriptors are finally combined by exploiting a multi-channel approach [15], i.e., $\left[\mathbf{K}\right]_{l,m} = \prod_{k=1}^{K}\left[\mathbf{K}\right]_{l,m}^{k}$.

The Olympic Sports dataset consists of 783 videos depict-

Table 2. Performance (%) on Human Action Recognition.

	KDA	**Proposed**
Olympic Sports	81.54	**83.35**
Hollywood2	58.63	**61.22**

ing athletes practicing 16 sports [16]. 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. 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 [16].

The Hollywood2 dataset consists of 1707 videos depicting 12 actions [17]. 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. 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 [17].

The performance obtained by applying the two classification schemes on each data set is illustrated in Table 2. By observing this Table, it can be seen that the proposed method outperforms KDA in both databases, enhancing its performance by $2-3\%$.

3.3. 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. 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. Since there is not a widely adopted experimental protocol for these datasets, the 5-fold cross-validation procedure has been adopted. We have employed the RBF kernel function, similar to the UCI datasets. 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.

Table 3. Performance (%) on Face Recognition.

	KDA	**Proposed**
ORL	96.43 (±0.42)	**96.5 (±0.37)**
AR	81.96 (±0.6)	**82.22 (±0.4)**
YALE	91.95 (±0.5)	**92.14** (±0.79)

The ORL dataset contains 10 images of 40 persons, leading to a total number of 400 images ([18]). 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.

The AR dataset contains over 4000 images depicting 70 male and 56 female faces ([19]). In our experiments we have used the preprocessed (cropped) facial images 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.

The Extended YALE-B dataset contains images of 38 persons in 9 poses, under 64 illumination conditions ([20]). In our experiments we have used the frontal cropped images provided by the database.

The mean classification rates and the observed standard deviations over all experiments for each data set are illustrated in Table 3. It can be seen that the proposed method outperforms KDA in all the three datasets. It should be noted here that

4. CONCLUSIONS

In this paper, we described an optimization scheme aiming at optimal nonlinear data projection, in terms of Fisher ratio maximization. By optimizing the Fisher ratio with respect to both the data projection matrix and the class representation in the projection space, the optimal discriminant projection space is obtained. Experimental results on standard classification problems, as well as on human action and face recognition problems show 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).

5. REFERENCES

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

[2] H. Wang, Z. Hu, and Y. Zhao, “An efficient algorithm for generalized discriminant analysis using incomplete cholesky decomposition,” *Pattern Recognition Letters*, vol. 28, no. 2, pp. 254–259, 2007.

[3] M. Sugiyama, “Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis.,” *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[4] D. Cai, X. He, and J. Han, “Speed up kernel discriminant analysis,” *International Journal on Very Large Data Bases*, vol. 20, no. 1, pp. 21–33, 2011.

[5] I. Rodriguez-Lujan, C. Santa Crouz, and R. Huerta, “On the equivalence of kernel fisher discriminant analysis and kernel quadratic programming feature selection,” *Pattern Recognition Letters*, vol. 32, no. 11, pp. 1567–1571, 2011.

[6] J. Liu, F. Zhao, and Y. Liu, “Learning kernel parameters for kernel fisher discriminant analysis,” *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1026–1031, 2013.

[7] A. Diaf, B. Boufama, and R. Benlamri, “Non-parametric fisher’s discriminant analysis with kernels for data classification,” *Pattern Recognition Letters*, vol. 34, no. 5, pp. 552–558, 2013.

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

[9] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[10] A. Argyriou, C. Micchelli, and M. Pontil, “When is there a representer theorem? vector versus matrix regularizers,” *Journal of Machine Learning Research*, vol. 10, pp. 2507–2529, 2009.

[11] A. Iosifidis, A. Tefas, and I. Pitas, “On the optimal class representation in linear discriminant analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1491–1497, 2013.

[12] E. Gopi and P. Palanisamy, “Formulating particle swarm optimization based membership linear discriminant analysis,” *Swarm and Evolutionalry Computation*, vol. 12, pp. 65–73, 2013.

[13] A. Frank and A. Asuncion, “Uci machine learning repository,” 2010.

[14] H. Wang and C. Schmid, “Action recognition with improved trajectories,” *IEEE International Conference on Computer Vision*, pp. 3551–3558, 2013.

[15] J. Zhang, M. Marszalek, M. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[16] J. Niebles, C. Chend, and L. Fei-Fei, “Modeling temporal structure of decomposable motion segemnts for activity classification,” *European Conference on Computer Vision*, pp. 392–405, 2010.

[17] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936, 2009.

[18] F. Samaria and A. Harter, “Parameterisation of a stochastic model for human face identification,” *IEEE Workshop on Applications of Computer Vision*, pp. 138–142, 1994.

[19] A. Martinez and A. Kak, “Pca versus lda,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[20] K.C. Lee, J. Ho, and D. Kriegman, “Acquiriing linear subspaces for face recognition under varialbe lighting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.