CLASS-SPECIFIC NONLINEAR SUBSPACE LEARNING BASED ON OPTIMIZED CLASS REPRESENTATION

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

In this paper, a new nonlinear subspace learning technique for class-specific data representation based on an optimized class representation is described. An iterative optimization scheme is formulated where both the optimal nonlinear data projection and the optimal class representation are determined at each optimization step. This approach is tested on human face and action recognition problems, where its performance is compared with that of the standard class-specific subspace learning approach, as well as other nonlinear discriminant subspace learning techniques. Experimental results denote the effectiveness of this new approach, since it consistently outperforms the standard one and outperforms other nonlinear discriminant subspace learning techniques in most cases.

***Index Terms*—** Class-specific discriminant learning, Nonlinear subspace learning, Action recognition, Face recognition

1. INTRODUCTION

Standard Discriminant Learning techniques, like Linear Discriminant Analysis (LDA) [1, 2], Kernel Discriminant Analysis (KDA) [3], (kernel) Spectral Regression (KSR) [4] and Class-specific (kernel) Discriminant Analysis (CSKDA) [5], represent classes by adopting the corresponding class mean vectors. Thus, they inherently set the assumption that the classes forming the classification problem follow unimodal normal distributions having the same covariance structure [2]. However, these are two strong assumptions that are difficult to be met in real classification problems. It has been recently shown that, when these assumptions are not met, the adoption of optimized class representations, other than the class mean vectors, leads to the determination of a discriminant subspace of increased class discrimination power [6, 7]. In this paper, we follow this line of work and describe an optimization scheme for the determination of such an optimized class representation for class-specific nonlinear data projection that leads to the determination of a discriminant subspace having increased class discrimination power.

In detail, in this paper we describe a new class-specific discrimination criterion which is used to optimize both the data projections and the class representation for the determination of a low-dimensional feature space of increased discrimination power. This class-specific criterion is formulated so that to exploit data representations in arbitrary-dimensional Hilbert spaces for nonlinear data projection and classification [8–11]. An iterative optimization schemes is applied to this end, which optimizes the class-specific criterion with respect to both the data projection matrix and the class representation. For the calculation of the optimal data projection matrix, an optimization process based on the Spectral Regression framework [4] is adopted in order to obtain a fast optimization method, when compared to the standard approach [3, 5]. We compare the performance of the Class-specific Reference Discriminant Analysis (CSRDA) algorithm with that of other Discriminant Analysis-based classification schemes, i.e., KDA, KSR and CSKDA, as well as with the performance of the Kernel Support Vector Machine (KSVM) classifier, which is one of standard choices in nonlinear classification problems. Experiments are conducted on six publicly available datasets, namely the ORL [12], AR [13] and Extended YALE-B [14] for face recognition and Hollywood2 [15], Olympic Sports [16] and ASLAN [17] datasets for human action recognition.

The rest of the paper is organized as follows. In Section 2, an overview of related work is provided. The CSRDA method is described in Section 3. Experimental results evaluating its performance are provided in Section 4. Finally, conclusions are drawn in Section 5.

2. RELATED WORK

Let us denote by $\mathbf{x}_i \in \mathbb{R}^D,\ i = 1, \dots, N$ a set of N vectors, each belonging to a class appearing in a class set $\mathcal{C} = \{1, \dots, C\}$. Let us also denote by $\mathbf{c}_j \in \mathbb{R}^N,\ j = 1, \dots, C$, C binary vectors having elements equal to $c_{ji} = 1$ in the case where $\mathbf{x}_i$ belongs to class j and to $c_{ji} = 0$, otherwise. We use N_{j0} and N_{j1} in order to denote the number of zeros and ones in $\mathbf{c}_j$, respectively. By using $\mathbf{x}_i,\ i = 1, \dots, N$ and $\mathbf{c}_j,\ j = 1, \dots, C$, a feature space of reduced dimensionality $d < D$ can be determined by learning a nonlinear data projection of the vectors $\mathbf{x}_i$ to vectors $\mathbf{z}_i \in \mathbb{R}^d$.

In order to exploit kernel techniques for 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 [8–11, 18]) by employing a function $\phi(\cdot) : \mathbf{x}_i \in \mathbb{R}^D \rightarrow \phi(\mathbf{x}_i) \in \mathcal{F}$ determining a nonlinear mapping from the input space $\mathbb{R}^D$ to the arbitrary-dimensional space $\mathcal{F}$. In this space, we would like to determine a data projection matrix $\mathbf{W}$ that will be used to map a given sample $\mathbf{x}_i$ to a low-dimensional feature space $\mathbb{R}^{d_j}$ of increased discrimination power:

$$\mathbf{z}_i = \mathbf{W}^T \phi(\mathbf{x}_i), \quad \mathbf{z}_i \in \mathbb{R}^{d_j}. \tag{1}$$

In practice, since the multiplication in (1) can not be directly computed, the so-called *kernel trick* [8, 9] is adopted. That is, the multiplication in (1) is inherently computed by using dot-products in $\mathcal{F}$.

Standard nonlinear Discriminant Learning techniques, like KDA [3] and KSR [4], solve an optimization problem involving relations between the within-class and between-class scatters of the training data in $\mathcal{F}$. That is, they employ the class mean vectors:

$$\phi(\mathbf{m}_j) = \frac{1}{N_{j1}} \sum_{i, c_{ji}=1} \phi(\mathbf{x}_i) \tag{2}$$

in order to calculate the within-class and between-class scatter matrices:

$$\mathbf{S}_w = \sum_{j=1}^{C} \sum_{i=1}^{N} c_{ji} \left(\phi(\mathbf{x}_i) - \phi(\mathbf{m}_j) \right)^T \left(\phi(\mathbf{x}_i) - \phi(\mathbf{m}_j) \right), \tag{3}$$

$$\mathbf{S}_b = \sum_{j=1}^{C} N_{j1} \left(\phi(\mathbf{m}_j) - \phi(\mathbf{m}) \right)^T \left(\phi(\mathbf{m}_j) - \phi(\mathbf{m}) \right), \tag{4}$$

where $\phi(\mathbf{m}) = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i)$ is the mean of the entire training set in $\mathcal{F}$, and calculate the data projection matrix $\mathbf{W}$ by solving an optimization problem that is function of $\mathbf{S}_w$, $\mathbf{S}_b$, e.g., the trace ratio optimization problem [1, 19]:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \frac{trace\{\mathbf{W}^T \mathbf{S}_w \mathbf{W}\}}{trace\{\mathbf{W}^T \mathbf{S}_b \mathbf{W}\}}. \tag{5}$$

While the multi-class discriminant learning approach described above is able to determine a reduced-dimensionality feature space of increased class discrimination, it has been shown that class-specific discriminant learning methods are able to outperform multi-class ones in several tasks, like facial image classification [5]. In this case, the objective is the determination of a reduced-dimensionality feature space $\mathbb{R}^{d_j}$, $d_j < D$, where class j is better discriminated from all others. This is achieved by optimizing the trace ratio criterion using the following scatter matrices:

$$\mathbf{S}_w = \sum_{i=1}^{N} c_{ji} \left(\phi(\mathbf{x}_i) - \phi(\mathbf{m}_j) \right)^T \left(\phi(\mathbf{x}_i) - \phi(\mathbf{m}_j) \right), \tag{6}$$

$$\mathbf{S}_b = \sum_{k \neq j} \sum_{i=1}^{N} c_{ki} \left(\phi(\mathbf{x}_i) - \phi(\mathbf{m}_j) \right)^T \left(\phi(\mathbf{x}_i) - \phi(\mathbf{m}_j) \right), \tag{7}$$

where the class mean vector $\phi(\mathbf{m}_j)$ is employed for the representation of class j in $\mathcal{F}$. It has been recently shown that, for the multi-class subspace learning problem, the adoption of optimized class representations increases class discrimination in the reduced-dimensionality feature space, leading to enhanced performance [6, 7]. In the following Section, we describe a class-specific optimization scheme that can be employed for the determination of both optimized class representation and data projection.

3. CLASS-SPECIFIC PROJECTIONS BASED ON OPTIMIZED REPRESENTATION

Let us denote by $\phi(\boldsymbol{\mu}_j) \in \mathcal{F}$ a so-called reference vector that will be used in order to represent class j. $\phi(\boldsymbol{\mu}_j)$ is not restricted to be the class mean vector in $\mathcal{F}$. $\phi(\boldsymbol{\mu}_j)$ can be any vector that enhances the discrimination of class j from the remaining ones in the discriminant space $\mathbb{R}^{d_j}$. As has been previously described, we would like to learn a data projection matrix $\mathbf{W}$ which maps $\mathcal{F}$ to a low-dimensional discriminant space $\mathbb{R}^{d_j}$ where the samples belonging to class j are as close as possible to the image of $\phi(\boldsymbol{\mu}_j)$ in $\mathbb{R}^{d_j}$, i.e., $\mathbf{z}_j = \mathbf{W}^T \phi(\boldsymbol{\mu}_j)$, while the samples belonging to the remaining classes are as far as possible from it. That is, we would like to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{F}| \times d_j}$ minimizing:

$$D_j = \sum_{i, c_{ji}=1} \|\mathbf{W}^T \phi(\mathbf{x}_i) - \mathbf{W}^T \phi(\boldsymbol{\mu}_j)\|_2^2 \tag{8}$$

and maximizing:

$$D_0 = \sum_{i, c_{ji}=0} \|\mathbf{W}^T \phi(\mathbf{x}_i) - \mathbf{W}^T \phi(\boldsymbol{\mu}_j)\|_2^2. \tag{9}$$

$\mathbf{W}$ can be determined by solving for:

$$\begin{aligned} \mathcal{J}(\mathbf{W}) &= \frac{\sum_{i, c_{ji}=0} \|\mathbf{W}^T \phi(\mathbf{x}_i) - \mathbf{W}^T \phi(\boldsymbol{\mu}_j)\|_2^2}{\sum_{i, c_{ji}=1} \|\mathbf{W}^T \phi(\mathbf{x}_i) - \mathbf{W}^T \phi(\boldsymbol{\mu}_j)\|_2^2} \\ &= \frac{trace\left(\mathbf{W}^T \mathbf{S}_0 \mathbf{W}\right)}{trace\left(\mathbf{W}^T \mathbf{S}_j \mathbf{W}\right)}, \end{aligned} \tag{10}$$

where $\mathbf{S}_j$, $\mathbf{S}_0$ are defined by:

$$\mathbf{S}_j = \sum_{i, c_{ji}=1} \Big(\phi(\mathbf{x}_i) - \phi(\boldsymbol{\mu}_j) \Big) \Big(\phi(\mathbf{x}_i) - \phi(\boldsymbol{\mu}_j) \Big)^T \tag{11}$$

$$\mathbf{S}_0 = \sum_{i, c_{ji}=0} \Big(\phi(\mathbf{x}_i) - \phi(\boldsymbol{\mu}_j) \Big) \Big(\phi(\mathbf{x}_i) - \phi(\boldsymbol{\mu}_j) \Big)^T \tag{12}$$

The direct maximization of (10) is intractable since $\mathbf{S}_j$, $\mathbf{S}_0$ express the intra-class and out-of-class variances of the training samples with respect to $\phi(\boldsymbol{\mu}_j)$, respectively ($\mathbf{S}_j$, $\mathbf{S}_0$ are matrices of arbitrary dimensions). In the following subsection, we describe an optimization process that can be used in order to maximize (10) for the determination of the optimal data projection $\mathbf{W}$, which is based on kernel Spectral Regression [4]. Subsequently, we describe an optimization process that can be used in order to determine the optimal class representation $\phi(\boldsymbol{\mu}_j)$ (given $\mathbf{W}$) and the iterative optimization process that can be used in order to optimize $\mathcal{J}$ with respect to both $\mathbf{W}$ and $\phi(\boldsymbol{\mu}_j)$. Finally, we describe a classification process that can be employed in combination with this method.

3.1. Spectral Regression-based optimization of (10)

In order to directly optimize $\mathcal{J}$ in (10), we express $\mathbf{W}$ as a linear combination of the training data (represented in $\mathcal{F}$) [8, 9, 18], i.e.,:

$$\mathbf{W} = \sum_{i=1}^{N} \phi(\mathbf{x}_i)\boldsymbol{\alpha}_i^T = \mathbf{\Phi}\mathbf{A}. \tag{13}$$

$\mathbf{A} \in \mathbb{R}^{N \times d_j}$ is a matrix containing the reconstruction weights of $\mathbf{W}$, with respect to the training data in $\mathcal{F}$. $\mathbf{\Phi}$ is a matrix containing the data representations in $\mathcal{F}$. Without loss of generality, we assume that the data are ordered so that $\mathbf{\Phi} = [\mathbf{\Phi}_j \mathbf{\Phi}_0]$, where $\mathbf{\Phi}_j$ is a matrix containing the training data belonging to class j and $\mathbf{\Phi}_0$ is a matrix containing the remaining samples.

Let us denote by $\mathbf{v}$ an eigenvector of the problem $\mathbf{S}_0\mathbf{v} = \lambda\mathbf{S}_j\mathbf{v}$ with eigenvalue λ. $\mathbf{v}$ can be expressed as a linear combination of the training data in $\mathcal{F}$, i.e., $\mathbf{v} = \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i)$. By setting $\mathbf{K}\mathbf{a} = \mathbf{q}$, this eigenanalysis problem can be transformed to the following equivalent problem:

$$\mathbf{P}_0\mathbf{q} = \lambda\mathbf{P}_j\mathbf{q}. \tag{14}$$

Thus, the reconstruction weights matrix $\mathbf{A}$ can be performed by applying a two step procedure:

- Solution of the eigenproblem $\mathbf{P}_0\mathbf{q} = \lambda\mathbf{P}_j\mathbf{q}$, which is tractable since $\mathbf{P}_0, \mathbf{P}_j \in \mathbb{R}^{N \times N}$. The solution of this problem leads to the determination of a matrix $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{d_j}]$, where $\mathbf{q}_i$ is the eigenvector corresponding to the i-th largest eigenvalue.
- Determination of the matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{d_j}]$, where $\mathbf{K}\mathbf{a}_i = \mathbf{q}_i$. In the case where $\mathbf{K}$ is non-singular, the vectors $\mathbf{a}_i$ are given by $\mathbf{a}_i = \mathbf{K}^{-1}\mathbf{q}_i$. When this is not true, the vectors $\mathbf{a}_i$ can be obtained by solving the following set of linear equations:

$$(\mathbf{K} + \delta\mathbf{I})\,\mathbf{a}_i = \mathbf{q}_i. \tag{15}$$

where $\delta > 0$ is a regularization parameter. Thus, $\mathbf{a}_i$ is given by $\mathbf{a}_i = (\mathbf{K} + \delta\mathbf{I})^{-1}\,\mathbf{q}_i$.

As can be seen, the above-described optimization process requires the solution of one eigenanalysis problem (14) and the inversion of a $N \times N$ matrix, leading to a time complexity equal to $O(N^3)$.

Table 1. Performance for different training percentage on the face recognition datasets.

AR	KSVM	KSR	KDA	CSKDA	**CSRDA**
10%	20.22%	35.74%	**37.26**%	27.13%	29.39%
20%	27.86%	42.48%	44.24%	44.1%	**44.71**%
30%	46.56%	66.67%	67.22%	68.56%	**69.17**%
40%	44.13%	65.25%	66.31%	66.81%	**68**%
50%	61.62%	85.69%	86.77%	87.77%	**88.54**%
ORL	KSVM	KSR	KDA	CSKDA	**CSRDA**
10%	44.44%	57.22%	56.67%	51.67%	**54.17**%
20%	49.06%	86.56%	**87.81**%	70.94%	75.31%
30%	76.07%	89.64%	90.36%	90.71%	**91.43**%
40%	84.17%	92.22%	93.75%	93.75%	**94.58**%
50%	88%	93.5%	93%	93%	**96**%
Yalle	KSVM	KSR	KDA	CSKDA	**CSRDA**
10%	53.9%	69.69%	70.01%	70.51%	**71.69**%
20%	75.13%	85.35%	81.84%	86.02%	**87.56**%
30%	81.11%	87.89%	79.06%	88.77%	**89.94**%
40%	89.4%	93.56%	88.99%	93.63%	**94.18**%
50%	93.34%	96.38%	94.65%	96.79%	**97.62**%

3.2. Reference Class Vector calculation

By observing that $\mathbf{S}_j$, $\mathbf{S}_0$ are functions of $\phi(\boldsymbol{\mu}_j)$, as detailed in (10), and by using $\phi(\boldsymbol{\mu}_j) = \mathbf{\Phi}_j\mathbf{b}_j$ [8, 9, 18], $\phi(\boldsymbol{\mu}_j)$ can be inherently determined by maximizing $\mathcal{J}$ with respect to $\mathbf{b}_j$, i.e.,:

$$\mathbf{b}_j^* = \arg\max_{\mathbf{b}_j} \mathcal{J}(\mathbf{W}, \mathbf{b}_j). \tag{16}$$

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

$$\mathbf{b}_j^* = \frac{h + \left(h^2 + 4q\,(b+f)\,N_{j1}e\right)^{1/2}}{2qeN_{j1}} \mathbf{1}_{N_{j1}}. \tag{17}$$

where $h = fN_{j0} - bN_{j1}$, $q = N_{j1}^2 + N_{j1}N_{j0}$, $b = tr\left(\mathbf{A}^T\mathbf{K}_0\mathbf{K}_0^T\mathbf{A}\right)$, $e = tr\left(\mathbf{A}^T\mathbf{\Phi}^T\phi(\boldsymbol{\mu}_j)\phi(\boldsymbol{\mu}_j)^T\mathbf{\Phi}\mathbf{A}\right)$, $f = tr\left(\mathbf{A}^T\mathbf{K}_j\mathbf{K}_j^T\mathbf{A}\right)$ and $\mathbf{K}_j = \mathbf{\Phi}^T\mathbf{\Phi}_j$, $\mathbf{K}_0 = \mathbf{\Phi}^T\mathbf{\Phi}_0$.

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

Taking into account that $\mathbf{A}$ is a function of $\mathbf{b}_j$ and that $\mathbf{b}_j$ is a function of $\mathbf{A}$, a direct maximization of $\mathcal{J}$ with respect to both $\mathbf{A}$ and $\mathbf{b}_j$ is difficult. In order to maximize $\mathcal{J}$ with respect to both $\mathbf{A}$ and $\mathbf{b}_j$, we employ an iterative optimization scheme, where $\mathbf{A}$ and $\mathbf{b}_j$ are iteratively updated until $(\mathcal{J}(t+1) - \mathcal{J}(t))/\mathcal{J}(t) < \epsilon$, where ϵ is a small positive value (equal to $\epsilon = 10^{-6}$ in our experiments).

3.4. Classification (test phase)

In order to perform classification, we work as follows. After the determination of the discriminant space $\mathbb{R}^{d_j}$, both the training data $\mathbf{x}_i,\ i = 1, \ldots, N$ and the reference class vector $\phi(\boldsymbol{\mu}_j)$ are mapped to that space and $\mathbf{z}_i,\ i = 1, \ldots, N$, $\mathbf{z}_j$ are obtained. Subsequently, we calculate distance vectors $\mathbf{d}_i \in \mathbb{R}^{d_j}$ having elements equal to:

$$d_{ik} = |\mathbf{z}_{ik} - \mathbf{z}_{jk}|,\ \ k = 1, \ldots, d_j, \tag{18}$$

where $\mathbf{z}_{ik}, \mathbf{z}_{jk}$ are the k-th elements of $\mathbf{z}_i$ and $\mathbf{z}_j$, respectively. $|\cdot|$ denotes the absolute value operator. By using $\mathbf{d}_i$, classification can be performed based on a linear classifier, e.g., linear SVM. In case of multi-class classification, we train C linear SVM classifiers in an one-versus-rest manner using the above described process. A test sample is introduced to all the C classifiers and is assigned to the class providing the maximal probability, similar to [20, 21].

4. EXPERIMENTS

In this section, we present experiments conducted in order to compare the performance of the two class-specific discriminant learning approaches. We have employed six publicly available datasets to this end. These are: the ORL [12], AR [13] and Extended YALE-B [14] (face recognition) and the Hollywood2 [15], Olympic Sports [16] and ASLAN [17] (human action recognition) datasets. In all our experiments we compare the performance of the Class-Specific Reference Discriminant Analysis (CSRDA) with that of the Class-Specific Kernel Discriminant Analysis (CSKDA) [5], as well as with Kernel Spectral Regression (KSR) [4], Kernel Discriminant Analysis (KDA) [3] and kernel Support Vector Machine (SVM)-based classification [22].

In all the experiments involving facial image classification we have employed the RBF kernel function. In human action recognition, we used the state-of-the-art methods proposed in [17, 23] as baseline approaches. On the ASLAN dataset we employ a set of 12 histogram similarity values expressing the similarity of pairs of videos represented by using the BoW model for HOG, HOF and HNF descriptors evaluated on STIP video locations [24] combined with a linear classification scheme. For the remaining datasets, we employ the BoW-based video representation by using HOG, HOF, MBHx, MBHy and (normalized) Trajectory descriptors evaluated on the trajectories of densely sampled interest points [23] and classification is performed by a nonlinear classification scheme using the RBF-χ^2 kernel function.

4.1. Results

We have applied the competing algorithms on the face recognition data sets. Since there is not a widely adopted experimental protocol for these datasets, we randomly partition the datasets in training and test sets as follows: we randomly select a subset of the facial images depicting each of the persons in each dataset in order to form the training set and we keep the remaining facial images for evaluation. Experimental results obtained by applying the competing algorithms are illustrated in Table 1. Class-specific classification schemes outperformed the multi-class ones in all but one cases. By optimizing both the data projection matrix and the class representation, CSRDA enhances class discrimination when compared to CSKDA, leading to enhanced classification performance. Table 2 illustrates the performance obtained by applying the competing classification schemes on the action recognition data sets. It can be seen that CSRDA provides satisfactory performance in all cases.

Table 2. Performance on the action recognition data sets.

	Olympic Sports	Hollywood2	ASLAN
SVM	86.56%	61.51%	60.88 ± 0.77%
KSR	88.35%	61.34%	54.9 ± 0.71%
KDA	88.64%	61.04%	51.20 ± 0.43%
CSKDA	87.65%	60.5%	54.9 ± 0.71%
CSKRDA	**88.89**%	**61.69**%	**61.03**± **0.54**%

5. CONCLUSIONS

In this paper, we described a new nonlinear subspace learning technique for class-specific data representation based on an optimized class representation. An iterative optimization scheme was formulated and evaluated to this end, where both the optimal nonlinear data projection and the optimal class representation are determined at each optimization step. Experimental results on six publicly available data sets denote the effectiveness of this class-specific approach, since it consistently outperforms the standard class-specific one and outperforms other nonlinear discriminant subspace learning techniques in most cases.

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).

REFERENCES

[1] Y. Jia, F. Nie, and C. Zhang, “Trace ratio problem revisited,” *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[2] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification, 2nd ed*, Wiley-Interscience, 2000.

[3] 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.

[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] G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas, “Class-specific kernel-discriminant analysis for face verification,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 570–587, 2007.

[6] 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.

[7] A. Iosifidis, A. Tefas, and I. Pitas, “Kernel reference discriminant analysis,” *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.

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

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

[10] W. Zheng, L. Zhao, and Z. Cairong, “Foley-sammon optimal discriminant vectors using kernel approach,” *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 1–9, 2005.

[11] Q. Liu, X. Tang, H. Lu, and S. Ma, “Face recognition using kernel scatter-difference-based discriminant analysis,” *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 1081–1085, 2006.

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

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

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

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

[16] J.C. Niebles, C.W. Chend, and L. Fei-Fei, “Modeling temporal structure of decomposable mition segements for activity classification,” *European Conference on Computer Vision*, 2010.

[17] O.K. Gross, T. Hassner, and L. Wolf, “The action similarity labeling challenge,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 615–621, 2013.

[18] A. Argyriou, C.A. 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.

[19] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, “Trace ratio vs. ratio trace for dimensionality reduction,” *Computer Vision and Pattern Recognition*, 2007.

[20] R.E. Fan, P.H. Chen, and C. J. Lin, “Working set selection using the second order information for training svm,” *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 1889–1918, 2005.

[21] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” *British Machine Vision Conference*, 2009.

[22] C.C. Chang and C.J. Lin, “Libsvm : a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp. 1–27, 2011.

[23] H. Wang and C. Schmid, “Action recognition with improved trajectories,” *International Conference on Computer Vision*, 2013.

[24] I. Laptev, “On space-time interest points,” *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.