MULTI-VIEW SEMI-SUPERVISED DISCRIMINANT ANALYSIS: A NEW APPROACH TO AUDIO-VISUAL PERSON RECOGNITION

Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay

Multimedia Communication Department, EURECOM,
2229 Route des Cretes, BP 193, F-06560 Sophia-Antipolis Cedex, France
E-mail: {zhaox, evans, dugelay}@eurecom.fr

ABSTRACT

Many state-of-the-art biometric systems use feature vectors of high dimension and call for dimensionality reduction techniques to avoid the co-called 'curse of dimensionality.' Supervised approaches such as Linear Discriminant Analysis can extract discriminative features and is used widely, but suffers from over-fitting when used with small datasets. Through the imposition of local adjacency constraints, semi-supervised dimensionality reduction techniques can make use of abundant, unlabelled data to improve classification performance. This paper reports a new multi-view, semi-supervised discriminant analysis (MSDA) algorithm and its application in audio-visual person recognition. In contrast to existing approaches which typically utilize a single view, MSDA determines a more reliable neighbourhood constraint built jointly from multiple views of the same data. Experimental results on the standard MOBIO database show that our algorithm not only outperforms baseline supervised and unsupervised methods, but that it also outperforms single-view semi-supervised dimension reduction techniques in single view.

Index Terms— Audio-visual person recognition, semi-supervised learning, discriminant analysis

1. INTRODUCTION

Many state-of-the-art biometric systems make use feature vectors of high dimension. Classification performance, typically degrades, however, when faced with high-dimensional feature vectors which often contain redundant components not necessary for classification. A straight forward solution involves the use of dimensionality reduction techniques which project the original data vector into a lower-dimensional discriminant subspace which is more favourable for classification.

One of the most popular dimensionality reduction techniques is Linear Discriminant Analysis (LDA) [1]. LDA is a supervised method which searches for projection axes which maximise the distance between samples belonging to different classes while minimising the distance between samples of the same class. LDA can achieve significantly better performance than Principle Component Analysis (PCA) [2] which is an unsupervised alternative. However, in practical scenarios and especially biometric applications, manually labelled samples are often difficult or expensive to collect leaving the application of supervised approaches to dimensionality reduction potentially problematic when labelled training data is scarce. In particular, class covariance matrices may not be reliably estimated when the number of training samples is insufficient relative to the number of dimensions and. Classification performance can then be worse than that achieved with unsupervised methods [3].

While manually labelled training samples acquired during enrolment are often limited in number, unlabelled data is often available in abundance or can be collected easily. In such situations it is often possible to augment the pool of manually labelled training samples with unlabelled samples and thus to improve classification performance. Semi-supervised dimensionality reduction techniques, e.g. [4, 5, 6, 7] have attracted significant attention over recent years. Despite some major differences between such techniques, the common idea involves the avoidance of over-fitting through the imposition of local adjacency constraints on the pool of unlabelled data. For example, Semi-supervised Discriminant Analysis (SDA) [4] imposes on the conventional LDA objective function a local adjacency constraint of Locality Preserving Projection (LPP) [8]. It ensures that unlabelled samples which are close to each other in the original feature space remain close to each other in the lower-dimensional, discriminant subspace.

Practical data is often noisy and samples from the same class may lie far from each other in feature space and still distributed far apart after projection. Neighbourhood links learnt from single views are often sparse and the resulting improvement in classification performance is often modest. In some applications such as multi-modal biometrics for example, a data sample is represented by more than a single view. Most approaches to semi-supervised learning, including all those mentioned above, focus only on a single view and are not capable of exploiting multiple views. The standard approach in this case involves the independent learning of projections for each view. Assuming that each view is uncorrelated, then two same-class samples belonging to the same class which are far from each other in one view could be very close to each other in another view. Local adjacency information in one view can therefore help in learning more reliable projections in another view.

In this paper, we propose a new algorithm referred to as Multi-view Semi-supervised Discriminant Analysis (MSDA) as a multi-view extension to SDA. It imposes a multi-view local adjacency constraint on the conventional LDA objective function and requires that two neighbouring samples in the original feature space of one view lie near to each other in the projected, lower-dimensional and discriminative space of the other view. The approach better exploits the information in unlabelled data to augment small, manually labelled training sets and thus to enhance the performance of supervised LDA. The algorithm is assessed in an audio-visual person identification scenario where both speaker and face recognition models make use of high-dimensional features. Experimental results show that the proposed MSDA system not only enhanced supervised and unsupervised baseline methods (LDA, PCA, LPP) by a large margin but that it also outperforms single-view SDA.

The remainder of this paper is organized as follows. Section 2 presents a review of several single view dimensionality reduction methods which are related to the proposed multi-view approach. The new MSDA algorithm is presented in Section 3. Section 4 presents experimental work and results which demonstrate the effectiveness of our approach. Finally, we provide some conclusions in Section 5.

2. DIMENSIONALITY REDUCTION

Since they are closely related to the new approach proposed in this paper, we include here a brief overview of the Locality Preserving Projection (LPP) [8] and Semi-supervised Discriminant Analysis (SDA) [4].

2.1. Locality preserving projection

LPP belongs to the family of manifold (or local) dimensionality reduction techniques which have together received considerable research interest over the last decade. Compared to global dimensionality reduction techniques, such as PCA and LDA, LPP seeks to preserve intrinsic geometric structure by learning a locality preserving sub-manifold. The objective of LPP is to find an optimal projection $\boldsymbol{a}_{opt}$ such that the neighbouring samples in the original space remain closely located in the projected space, where

$$\boldsymbol{a_{opt}} = \arg\min_{\boldsymbol{a}} \sum_{i,j} (\boldsymbol{a}^T \boldsymbol{x_i} - \boldsymbol{a}^T \boldsymbol{x_j})^2 S_{ij}, \tag{1}$$

and where S is local adjacency matrix which reflects of any pair of samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. This commonly involves a simple weight function; two common examples are a binary weight:

$$S_{ij} = \begin{cases} 1, \text{ if } \boldsymbol{x}_i \in N_p(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_p(\boldsymbol{x}_i) \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

or heat kernel weight:

$$S_{ij} = \begin{cases} exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|), & \text{if } \boldsymbol{x}_i \in N_p(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_p(\boldsymbol{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $N_p(\boldsymbol{x}_i)$ denotes the set of p nearest neighbours in the vicinity of sample $\boldsymbol{x}_i$. A projection is then sought which minimises the sum of distances between linked samples.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ be a matrix of n samples. Through some straightforward algebraic manipulation (interested readers are referred to [8] for details), the objective function in Eq. (1) can be re-written as:

$$\boldsymbol{a_{opt}} = \arg\min_{a} (\boldsymbol{a}^T \boldsymbol{XLXa}), \tag{4}$$

where $\boldsymbol{L}$ is the graph Laplacian matrix and where:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}, \tag{5}$$

in which D is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. The projection $\boldsymbol{a}$ can be obtained by solving the generalized eigenvalue problem:

$$\boldsymbol{XLX}^T \boldsymbol{a} = \lambda \boldsymbol{XDX}^T \boldsymbol{a}. \tag{6}$$

LPP has been successfully applied in automatic face recognition problems and is commonly referred to as Laplacianface [9].

2.2. Semi-supervised discriminant analysis

LPP is an unsupervised method and thus lacks discriminant power favourable for classification. While supervised approaches such as LDA have discriminant power they often require impractical quantities of training data. SDA [4] harnesses the benefits of discriminant LDA and unsupervised LPP to learn reliable projections from both labelled and unlabelled data and prevent over-fitting. As presented in [4], the LDA objective function is first reformulated in a form compatible with LPP:

$$\boldsymbol{a_{opt}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^T X W_{l\times l} X^T \boldsymbol{a}}{\boldsymbol{a}^T X X^T \boldsymbol{a}}, \tag{7}$$

where $W_{l\times l}$ is a $l \times l$ matrix:

$$W_{l\times l} = \begin{pmatrix} W^{(1)} & 0 & \cdots & 0 \\ 0 & W^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W^{(c)} \end{pmatrix}, \tag{8}$$

and where $W^{(k)}$ is a $l_k \times l_k$ matrix with all the elements equal to $1/l_k$, and where l and l_k are the total number of labelled samples and number of samples in the k-th class respectively. The SDA objective function is defined as:

$$\boldsymbol{a}_{opt} = \arg\max_{a} \frac{\boldsymbol{a}^T X W_{l\times l} X^T \boldsymbol{a}}{\boldsymbol{a}^T X X^T \boldsymbol{a} + \alpha \boldsymbol{a}^T X L X^T \boldsymbol{a}^T} \tag{9}$$

$$= \arg\max_{a} \frac{\boldsymbol{a}^T X W X^T \boldsymbol{a}}{\boldsymbol{a}^T (\tilde{I} + \alpha L) \boldsymbol{a}^T}, \tag{10}$$

where:

$$\tilde{I} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \text{ and} \tag{11}$$

$$W = \begin{pmatrix} W_{l\times l} & 0 \\ 0 & 0 \end{pmatrix} \tag{12}$$

where I is an identity matrix of size $l \times l$, and α is a parameter which weights the contribution of labelled and unlabelled information. The projection $\boldsymbol{a}$ is obtained by solving the generalized eigenvalue problem:

$$XWX^T\boldsymbol{a} = \lambda X(\tilde{I} + \alpha L)X^T\boldsymbol{a}, \tag{13}$$

SDA exploits labelled data to cluster same-class data together and to better separate different-class data, while retaining the local structure of unlabelled data to prevent overfitting.

3. MULTI-VIEW SEMI-SUPERVISED DISCRIMINANT ANALYSIS

According to a recent, comparative study [10], dimensionality reduction techniques based on local manifold learning work very well for artificial datasets where samples are densely linked within local manifold structures. With real datasets, however, which often contain substantial noise, same-class data in the original feature space can be dispersed and sample links are sparse. In such situations the performance of local manifold learning techniques can degrade significantly. In this section we show how this problem can be alleviated by extending the approach to exploit multiple, uncorrelated views of the same data.

3.1. Multi-view local adjacency constraint

The inspiration for our work stems from a multi-view semi-supervised learning method referred to as co-training [11]. With co-training two classifiers are first weakly trained with a small number of labelled data in two different views. Each classifier is used to label a large pool of auxiliary, unlabelled data and the most confidently labelled data is then used to augment the training set and thus to enhance the other classifier. The co-training algorithm is based on view sufficiency and view agreement assumptions which imply that, given well-trained classifiers, information in each view is sufficient to predict the labels of unlabelled data and that the two classifiers should generally agree on a common labelling decision.

We extend this idea to multi-view dimensionality reduction. Given two well-trained projections in uncorrelated views, two data samples belonging to a sample manifold in one projected space should belong to the corresponding manifold in the other projected space. Accordingly, we construct the multi-view local adjacency matrix as follows: if a link between two data samples is established in the original space of either view, a corresponding link is also established in the other view. This idea is illustrated in Fig. 1 in which blue and white circles represent samples corresponding to two different classes. Left and right plots illustrate two different

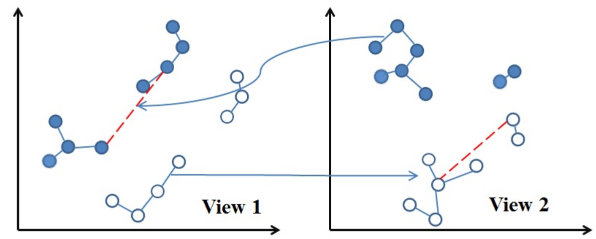

Fig. 1. An illustration of multi-view adjacency constraints

views of the same data. The solid lines represent the links between samples which are deemed automatically to belong to the same class. Samples in each class lie in disjoint manifolds and with conventional manifold learning methods, it is unlikely that they will lie near to each other in the projected, lower dimensional space. Assuming that the two views are conditionally independent from each other, two different but same-class manifolds in view 1 may be in the same manifold in view 2, and vice versa. By updating the neighbourhood constraints in one view according to the corresponding adjacency constraints in the other, a more reliable adjacency graph (dashed red lines in Fig. 1) can be constructed and used to improve learnt projections.

3.2. Projection algorithm

We propose a multi-view extension of SDA, which we refer to a Multi-view Semi-supervised Discriminant Analysis (MSDA). While this paper focuses the use only on two views, the algorithm can be applied to an unlimited number of different views and consequently the algorithm is presented below in its most general form (not limited to two views).

We assume a small set of l labelled n-view data samples, $\{x_{i1}, x_{i2} \ldots, x_{in}; y_i | i = 1 \ldots l\}$ and a set of u unlabelled data samples $\{x_{i1}, x_{i2} \ldots, x_{in}; | i = 1 \ldots u\}$, where x_{ij} is the j-th view of the i-th sample and y_i is the label for the i-th sample. The algorithm is applied as follows:

1. **Construct the adjacency graph for each view:** For each view x_n, construct the p-nearest neighbour graph S_n according to Eq. (2) or (3).

2. **Construct the multi-view adjacency graph:** the multi-view graph S is constructed according to:

$$S = \max_n(S_n), \tag{14}$$

and the graph Laplacians are determined according to $L_n = D_n - S_n$.

3. **Construct the labelled graph:** weight matrices W and $\tilde{I}$ are determined according to Eq. (11) and (12) respectively.

4. **Solve the eigenvalue problem:** for each view x_n the eigenvectors are determined for all non-zero eigenvalues according to the generalized eigenvalue problem:

$$X_n W X_n^T \boldsymbol{a}_n = \lambda_n X_n(\tilde{I} + \alpha L_n) X_n^T \boldsymbol{a}_n, \quad (15)$$

where $X_n = [\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nl}, \boldsymbol{x}_{n(l+1)}, \ldots, \boldsymbol{x}_{n(l+u)}]$ is the whole data matrix which pools together both labelled and unlabelled samples. If W is of rank c then there will be c eignenvectors with non-zero eigenvalues. We denote them as $\boldsymbol{a}_{n1}, \ldots, \boldsymbol{a}_{nc}$.

5. **MSDA embedding:** let $A_n = [\boldsymbol{a}_{n1}, \boldsymbol{a}_{n2}, \ldots, \boldsymbol{a}_{nc}]$ such that the samples in the n-th view can be embedded into a c-dimensional subspace according to:

$$\boldsymbol{z}_n = A_n^T \boldsymbol{x}_n. \quad (16)$$

In the following we show how the MSDA approach can be applied in audio-visual person recognition, i.e. using $n = 2$ views.

4. EXPERIMENTS IN AUDIO-VISUAL PERSON RECOGNITION

We now describe the application and assessment of the proposed MSDA algorithm in audio-visual person recognition. MSDA has natural appeal in such a scenario since: (1) labelled data is limited while abundant unlabelled data can often be collected with ease; (2) face and voice are natural biometrics in the case of videos and are independent from each other, and (3) most state-of-the-art, automatic speaker and face recognition systems utilise high-dimensional feature vectors which require dimensionality reduction.

4.1. Database

All experiments were conducted with the MOBIO database[1]. It contains videos of 150 subjects from whom video data is captured in 12 sessions over a one-and-a-half year period. Each session contains 11-21 videos of spoken data captured in variable, challenging conditions recorded with a mobile phone video camera. Data from a random subset of 30 subjects was used to train a world Gaussian Mixture Model (GMM) for speaker recognition and data from another random and non-overlapping subset of 30 subjects was used for recognition experiments in an identification framework. a number of $l = 1, 2, 3$ sessions are randomly selected as labelled training data for enrolment. Another single, random session is set aside as test data. The remaining $11 - l$ sessions are used as unlabelled data. The total amount of training data thus remains constant and the parameter l controls the fraction of it which is manually labelled.

4.2. Experiments

Experiments are conducted with largely standard speaker and face recognition systems which represent each video by a GMM speaker supervector [12] and an Local Binary Pattern (LBP) [13] face feature vector, both of high dimensionality.

The speech signal is parameterised using Mel-scaled Cepstral Coefficients (MFCCs) which are extracted from 20ms Hamming windowed frames at a 10ms frame rate. Feature vectors are composed of the first 26 MFCC coefficients augmented with their 26 delta coefficients and the delta energy. Non-informative speech frames are identified using energy-based speech detection and discarded. A 64-component GMM world model is learned with an Expectation-Maximization (EM) algorithm and adapted to generate speaker models using Maximum A-Posteriori (MAP) adaptation. Only the GMM means are adapted and are concatenated to form a 3392-dimensional supervector.

Face images are extracted and cropped according to facial landmarks identified using a face detector based on OpenCV[2]. Cropped face images are then resized to 144×128 pixels. The single most confidently detected face is divided into 9×8 blocks and $LBP^{u2}_{(8,2)}$ features are extracted from each block and concatenated into a 4248-dimensional vector.

Experiments were performed with PCA [2], LDA [1], LPP [8], SDA [4] and MSDA algorithms in an otherwise identical setup. Since LDA cannot be applied to data of dimensionality greater than the difference between the number of samples and number of classes, a step of PCA dimensionality reduction is often applied to obtain an intermediate representation upon which LDA can be applied [1]. PCA is therefore used to reduce both speech and face feature vectors to 100 dimensions and all compared algorithms are then applied in the same intermediate feature space.

Projections are learnt with different techniques for both speech and face biometric modalities, or views. Here unsupervised methods PCA and LPP use all 11 training sessions as unlabelled data, supervised method LDA only uses l labelled sessions, while semi-supervised methods SDA and MSDA use both labelled and unlabelled data. LDA, SDA and MSDA methods are applied to reduce the feature dimension to 29 (number of classes - 1) dimensions, and so PCA and LPP are applied to reduce features to the same dimensions. Test data are projected into lower dimensional spaces and are classified with a nearest-neighbour classifier based on Euclidean distances. For a score-level fusion, in each projected space, the distances of a test sample to all n labelled training samples $[d_1, ..., d_n]$ are normalized to reduce the range to the interval $[0, 1]$ according to $d_i^{norm} = (d_i - d_{min})/(d_{max} - d_{min})$, where d_{min} and d_{max} are the minimum and maximum values of these distances. Then the corresponding d_i^{norm}s in two spaces are averaged to get fused distances $[d_1^{fuse}, ..., d_n^{fuse}]$. The test sample is then assigned to the same class as nearest labelled sample.

[1] http://www.idiap.ch/dataset/mobio

[2] http://opencv.willowgarage.com/wiki/

	PCA	LPP	LDA	SDA	MSDA
Face	58.9%	68.1%	43.0%	67.4%	**73.6%**
Voice	68.0%	75.9%	33.7%	76.8%	**87.5%**
Fusion	78.3%	81.6%	50.0%	87.5%	**90.6%**

(a) 1 labelled training session ($l = 1$)

	PCA	LPP	LDA	SDA	MSDA
Face	72.3%	76.4%	79.2%	83.0%	**86.9%**
Voice	79.7%	80.0%	82.5%	88.0%	**91.4%**
Fusion	89.0%	89.2%	92.5%	95.7%	**96.7%**

(b) 2 labelled training sessions ($l = 2$)

	PCA	LPP	LDA	SDA	MSDA
Face	76.7%	77.8%	83.6%	85.7%	**88.4%**
Voice	84.4%	81.8%	92.0%	92.0%	**94.1%**
Fusion	91.9%	91.0%	96.4%	97.0%	**97.4%**

(c) 3 labelled training sessions ($l = 3$)

Table 1. Person recognition performance for different approaches to dimensionality reduction on the MOBIO database

4.3. Results

Average results across 50 iterations of cross-validation are presented in Table 1. Results are presented independently for both speaker and face recognition systems in addition to results after score-level fusion and where the number of labelled training sessions is varied between $l = 1, 2$ and 3.

When labelled training data is scarce ($l = 1$), due to over-fitting LDA performance is even worse than for unsupervised methods PCA and LPP. When the training set is augmented with the auxiliary pool of unlabelled data then semi-supervised SDA and the proposed MSDA algorithm lead to significantly improved performance. Improvements are observed for both individual classifiers in addition to the fused, bi-modal system. When sufficient labelled training data is available ($l = 2$), LDA out-performs PCA and LPP, but SDA and MSDA still achieve better performance, though the gain brought through the use of unlabelled data decreases as the labelled training set becomes larger ($l = 3$). Nonetheless, in all cases the proposed MSDA algorithm achieves superior performance to alternative approaches.

These results show that local adjacency constraints in one view can be effectively harnessed to learn more reliable projections in another view of the same data and ultimately give significantly improved classification performance, particularly when initial models are weakly or insufficiently trained.

5. CONCLUSIONS

This paper reports a new multi-view, semi-supervised discriminant analysis algorithm and its application to audio-visual person recognition. A more reliable local adjacency constraint is constructed using unlabelled information in different views and helps to enhance projections learnt from limited labelled data. Results on the standard MOBIO database show that the new algorithm not only outperforms baseline supervised and unsupervised approaches by a significant margin, but that it also outperforms semi-supervised dimensionality reduction techniques applied to a single view. The new algorithm is thus effective in harnessing unlabelled, multi-view biometric data to extract discriminative features favourable for classification.

6. REFERENCES

[1] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection,” in *IEEE Transaction on PAMI*, vol. 19(7), PP. 711-720, 1997.

[2] M. Turk and A. Pentland. “Eigenfaces for recognition,” in *Journal of Cognitive Neuroscience*, 3(1):71-86, 1991.

[3] A. M. Martinez and A.C. Kak. “PCA versus LDA,”in *IEEE Trans. on PAMI*, 23(2), 2001.

[4] D. Cai, X. He and J. Han. “Semi-supervised Discriminant Analysis,” in Procedings of *International Conference on Computer Vision (ICCV)*, 2007 .

[5] D. Zhang, Z.-H. Zhou and S. Chen. “Semi-supervised dimensionality reduction,”in *Proceedings of the 7th SIAM International Conference on Data Mining (SDM'07)*.

[6] M. Sugiyama, S. Nakajima and J. Sese, “Semi-supervised local Fisher discriminant analysis”, in *Machine Learning*, 78(1-2), 2011, pp. 35-61.

[7] X. Zhao, N. Evans and J.L. Dugelay, “Co-LDA: A new approach to audio-visual person recognition in videos,”in *IEEE International Conference on Multimedia and Expo (ICME)*, 2012

[8] X. He and P. Niyogi. “Locality Preserving Projection,”in *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2003.

[9] X. He, S. Yan, P. Niyogi and H. J. Zhang. “Face recognition using Laplacianfaces,”in *IEEE Trans. PAMI*, 27(3):328-340, 2005.

[10] L. J. P. van der Maaten, E. O. Postma and H. J. van den Herik. “Dimensionality Reduction: A Comparative Review,”Tilburg University Technical Report, TiCC-TR 2009-005, 2009.

[11] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the Workshop on Computational Learning Theory*, 1998.

[12] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification”, IEEE Signal Processing Letters 13, 5 (May 2006), pp. 308311.

[13] T. Ahonen, A. Hadid and M. Pietikainen. “Face recognition with local binary patterns,” in *Proc. ECCV 2004*, LNCS, vol. 3021, pp. 469-481. Springer, Heidelberg, 2004.