Unsupervised Multi-view Dimensionality Reduction with Application to Audio-Visual Speaker Retrieval

Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay

Multimedia Communication Department, EURECOM
Campus SophiaTech, 450 Route des Chappes, 06410 Biot Sophia Antipolis, France
{xuran.zhao, evans, jld}@eurecom.fr

***Abstract*—This paper presents a novel approach to unsupervised multi-view dimensionality reduction and reports its application to multi-modal biometrics retrieval, specifically audio-visual speaker retrieval. We propose a new concept referred to as *multi-view subspace agreement*, which aims to learn a subspace for each view which respects the similarity relationships between data points in the other view. The proposed algorithm is unsupervised but exhibits discriminative characteristics and is thus well suited to applications such as retrieval and clustering where class labels are generally unavailable. The effectiveness of the proposed algorithm is evaluated under an audio-visual speaker retrieval experiment with the MOBIO database. The retrieval performance of the proposed approach out-performs other single-view or multi-view dimensionality reduction methods with a significant margin.**

I. Introduction

Biometric systems exploit physiological and/or behavioural traits to recognize individuals. Popular traits or modalities include fingerprints, face, voice, iris, retina, gait, signature, palmprint, ear, etc. Most biometric systems share two common operation modules, namely feature extraction and matching. In the feature extraction module, each biometric sample is represented by a numerical feature, which is often a high-dimensional vector. In the matching module, the feature vector of a query sample is compared to a template sample to generate a distance or similarity score. If the similarity score is greater than a certain threshold, the two samples are considered belonging to the same subject. A major challenge in biometrics research is the compensation of intra-class variation and often with only a small inter-class variation. For example, in face recognition problems, two facial images of the same person with different poses and/or expressions are visually different and their distance in the feature space can be significant. A typical solution to this problem involves discriminative feature extraction (or dimensionality reduction) such as Linear Discriminant Analysis (LDA) [1], which finds an optimal projection such that, in the lower-dimensional projected feature space, the intra-class scatter is minimized while the inter-class scatter is maximized. These methods are typically supervised and require a large amount of labelled training data, which is typically expensive to obtain. In some unsupervised tasks such as clustering and retrieval, where labelled data is entirely absent, supervised feature extraction methods cannot be applied. Unsupervised feature extraction methods such as Principle Component Analysis (PCA) [2] are of potential use, but commonly lack discriminative power.

Biometric data can have multiple representations. This is the case with multibiometric systems [3] where different features could be extracted from multiple biometric traits, e.g. the human face and voice for example. The fusion of modalities remains a challenging problem and is generally treated in isolation to that of dimensionality reduction [4].

The problem of extracting discriminative features from multi-view, unlabelled data is referred to as unsupervised multi-view dimensionality reduction (UMVDR). Most existing UMVDR algorithms are based on Canonical Component Analysis (CCA) and its variants. Suppose data samples can be represented by two different features $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. CCA is applied jointly to the two features to learn projections $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$, so that the correlation between projections of two features $\mathbf{P}^{(1)}\mathbf{X}^{(1)}$ and $\mathbf{P}^{(2)}\mathbf{X}^{(2)}$ is maximized. The work in [5] and [6] has shown that, given conditional independent feature pairs, CCA can extract class-discriminative features in an unsupervised manner. Kernel CCA (KCCA) [7] is able to cope with non-linear conditions and has been applied in content-based image retrieval (CBIR) with image and associated text. In [8], the authors approached the UMVDR problem by extracting shared features between multiple views through a matrix factorization approach.

In this paper, we treat the UMVDR problem from a different angle based on a *multi-view subspace structure agreement* assumption. Let data samples be again represented in two views, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, which exhibit a certain level of conditional independence, as is often the case with multibiometrics. Suppose also that a similarity graph $S^{(v)}(v = 1, 2)$ can be constructed from the v-th view of the data where the nodes represent data samples and the edges $s_{ij}^{(v)}$ represent a similarity measure between the i^{th} and j^{th} sample in the v-th view. Since the two features are corrupted by different intra-class variation, two samples considered "similar" in one feature space may not be considered "similar" in the other. Now assume that, there exist optimal projections $\mathbf{P}_{opt}^{(1)}$ and

WIFS'2013, November 18-21, 2013, Guangzhou, China.
ISBN 978-1-4673-5593-3

$\mathbf{P}_{opt}^{(2)}$ such that in the two projected subspaces, the intra-class variation is successfully suppressed. Then a pair of closely-located samples in one projected space should also be closely-located in the other, since they belong to the same class. If $S^{(1)}$ and $S^{(2)}$ are constructed with the projected samples $\mathbf{P}_{opt}^{(1)T}\ \mathbf{X}^{(1)}$ and $\mathbf{P}_{opt}^{(2)T}\mathbf{X}^{(2)}$, they are expected to be similar. Based on this logic, we propose to approximate $\mathbf{P}_{opt}^{(1)}$ and $\mathbf{P}_{opt}^{(2)}$ by finding the $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ which minimize the difference between $S^{(1)}$ and $S^{(2)}$. In this paper, we show that this objective could be achieved through the graph-based co-training of locality preserving projection (LPP) [9]. We refer to this approach as the Co-LPP algorithm.

We report the application of the proposed Co-LPP algorithm in an audio-visual speaker retrieval problem. In this task, we have an unlabelled video database where each video sample contains the facial and vocal traits of a single subject. Given a query video sample, the retrieval system is expected to return t samples from the database containing the same subject as the query. The Co-LPP algorithm is applied to the facial and vocal features of the unlabelled video database to obtain projections $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$. The facial and vocal feature vectors of the query sample are then projected onto their corresponding subspaces and a cosine distance is calculated between the query and each of the samples in the database. The t closest samples are returned as retrieved results. The proposed method is evaluated using MOBIO audio-visual database [10]. We observe significant improvements in retrieval accuracy compared to other single-view or multi-view dimensionality reduction methods such as PCA, LPP, CCA and KCCA. With the help of data-visualization tools, we observe that the proposed Co-LPP algorithm has significantly higher discriminative power than competing approaches.

The remainder of this paper is organized as follows. Section 2 presents two essential components of the proposed algorithm: co-training and LPP. Section 3 presents the new algorithm itself and its application to multibiometric retrieval. Section 4 presents experimental evaluations and Section 5 presents our conclusions.

II. Locality Preserving Projection and Co-training

In this section we describe two essential components of the proposed algorithm: LPP and Co-training.

A. Locality Preserving Projection

LPP belongs to the family of manifold (or local) dimensionality reduction techniques, and seeks to preserve intrinsic geometric structure by learning a locality preserving sub-manifold [9]. In simpler words, LPP seeks to find an optimal projection $\boldsymbol{P}$ such that the neighbouring samples in the original space remain closely located in the projected space. The objective function of LPP is formulated as:

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

where S is a local similarity matrix which reflects the similarity 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:

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

or heat kernel weight:

$$S_{ij} = \begin{cases} exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}), \text{ if } \boldsymbol{x}_i \in KNN(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in KNN(\boldsymbol{x}_i) \\ 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{3}$$

where $KNN(\boldsymbol{x}_i)$ denotes the set of K nearest neighbours of sample $\boldsymbol{x}_i$. From Equation 1 we see that, if two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are considered similar in the original space ($S_{ij} > 0$), projecting them far apart will incur a high penalty.

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 [9] for details), the objective function in Equation 1 can be re-written as:

$$\arg\min_{P} (\boldsymbol{P}^T \boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^T \boldsymbol{P}), \tag{4}$$

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

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

in which $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. The projection is obtained by $\boldsymbol{P} = [\boldsymbol{p_1}, \ldots, \boldsymbol{p_k}]$ where $\boldsymbol{p_1}, \ldots, \boldsymbol{p_k}$ are the eigenvectors corresponding to the k smallest eigenvalues of the generalized eigenvalue problem:

$$\boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^T \boldsymbol{p} = \lambda \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^T \boldsymbol{p}. \tag{6}$$

Although LPP has been successfully applied in automatic face recognition problems [11], it has relatively low discriminative power. Biometric data often contain significant intra-class variations, causing data sample from the same subject located far-apart in the original feature space. This is likely to be the same in the projected space, due to the data structure preserving nature of LPP.

B. Co-training

Co-training [12] is one of the most acclaimed approaches to semi-supervised learning. In co-training, data samples are assumed to be represented by two conditionally independent features $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. Two predictors $f^{(1)}$ and $f^{(2)}$ assign to each $\boldsymbol{X}$ a class label $\boldsymbol{Y}$ ($f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$) and are trained according to each view using a small pool of labelled data. The two predictors are used to assign labels to a larger pool of unlabelled data. A subset of samples with which the predictors have the most confidence in label assignments is added to the pool of labelled data. The predictors are then iteratively re-learned and applied to the remaining unlabelled data. The objective of co-training is essentially to learn two different predictors $f^{(1)}$ and $f^{(2)}$ such that:

1) $f^{(1)}(\boldsymbol{X}^{(1)}) = f^{(2)}(\boldsymbol{X}^{(2)})$ for all labelled data;
2) the number of samples for which $f^{(1)}(\boldsymbol{X}^{(1)}) \neq f^{(2)}(\boldsymbol{X}^{(2)})$ is minimized for unlabelled data;

This objective is based on a *predictor agreement assumption*. Suppose two optimal predictors $f_{opt}^{(1)}$ and $f_{opt}^{(2)}$ can reliably predict the label $\boldsymbol{Y}$ from two features $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ respectively, their predictions should be the same for all unlabelled data. As a result, these optimal predictors can be approximated by requiring weak predictors $f^{(1)}$ and $f^{(2)}$ to agree on unlabelled samples.

III. Proposed approach

In this section, we apply the idea of co-training to multi-view unsupervised subspace learning. Analogous to the *predictor agreement assumption* in co-training, we propose a *subspace data structure agreement assumption* in the UMVDR problem. Given paired features $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ and assuming that there exist optimal projections $\mathbf{P}_{opt}^{(1)}$ and $\mathbf{P}_{opt}^{(2)}$ which can remove the intra-class variation and kept inter-class variation, two closely-located data samples in one projected space should be also closely-located in the other projected space. Since the similarity relationships between data samples can be represented by similarity graphs, we propose to approximate $\mathbf{P}_{opt}^{(1)}$ and $\mathbf{P}_{opt}^{(2)}$ by $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ which minimize the difference between the two similarity graphs constructed in the subspace of two views.

A. Objective function

Consider a set of n samples represented in two views: $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$, where $\boldsymbol{X}^{(v)} = \{\boldsymbol{x}_1^{(v)}, \ldots, \boldsymbol{x}_n^{(v)}\}$, $v = 1, 2$. $\boldsymbol{X}^{(v)}$ is first centred so that $\bar{\boldsymbol{X}}^{(v)} = \sum_i \boldsymbol{x}_i^{(v)}/n = 0$. Given two projections $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$, we define local similarity matrices $S^{(1)}$ and $S^{(2)}$ in subspaces for each view such that:

$$S_{ij}^{(v)} = \begin{cases} 1, & \text{if } \boldsymbol{P}^{(v)T}\boldsymbol{x}_i^{(v)} \in KNN(\boldsymbol{P}^{(v)T}\boldsymbol{x}_j^{(v)}) \\ & \text{or } \boldsymbol{P}^{(v)T}\boldsymbol{x}_j^{(v)} \in KNN(\boldsymbol{P}^{(v)T}\boldsymbol{x}_i^{(v)}) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

which is similar to Equation 2 but the KNN function is performed in the subspace. For simplicity, here we employ binary weight rather than a heat kernel weight to prevent the optimization of parameter σ in Equation 3. $S^{(v)}$ encodes local similarity relationships between data samples in the subspace of the v-th view. The i-th and j-th sample are considered similar in the v-th view if $S_{ij}^{(v)} = 1$ and dissimilar otherwise.

We further define a multi-view local structure agreement index as:

$$Agr(S^{(1)}, S^{(2)}) = 1 - \frac{2 \times \sum_{ij} |S_{ij}^{(1)} - S_{ij}^{(2)}|}{\sum_{ij} S_{ij}^{(1)} + \sum_{ij} S_{ij}^{(2)}} \tag{8}$$

which is upper-bounded by 1 if $S^{(1)} = S^{(2)}$ and lower-bounded by 0 if $S_{ij}^{(1)} \neq S_{ij}^{(2)}$ for all pairs of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We seek projections $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ such that the agreement in the local data structure is maximized. The objective function thus is given by:

$$\underset{\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}}{\arg\max}\ Agr(S^{(1)}, S^{(2)}) \tag{9}$$

Algorithm 1 Co-LPP

Input: A set of n multi-view samples $\boldsymbol{X} = \{\boldsymbol{X}^{(v)} | v = 1, 2\}$, number of neighbourhood K.
Output: Projection matrices $\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}$
Initialize:

- Center the feature vectors in each view and apply PCA if the dimensionality of the feature space is too high;
- Constructed Similarity graphs $S^{(1)}$ and $S^{(2)}$,

repeat
  for $v = 1$ **to** 2 **do**
- Use $\boldsymbol{X}^{(v)}$ and $S^{(3-v)}$ to train LPP projections $\boldsymbol{P}^{(v)}$ and project $\boldsymbol{X}^{(v)}$ into this subspace;
- Update $S^{(v)}$ with projected samples $\boldsymbol{P}^{(v)T}\boldsymbol{X}^{(v)}$

  end for
until $Agr(S^{(1)}, S^{(2)})$ does not reach a new maximum within a fixed number of iterations;

Note that $S^{(v)}$ is solely determined by $\boldsymbol{P}^{(v)}$ if the number of neighbours K in the KNN function is fixed.

B. Subspace learning by co-training LPP

In the following we propose an algorithm that optimizes $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ by a cross-view training of LPP. The main steps of the iterative algorithm are as follows:

1) Fix $\boldsymbol{P}^{(1)}$ and construct $S^{(1)}$ according to Equation 7. Solve for $\boldsymbol{P}^{(2)}$ according to Equation 5 and 6 by setting $\boldsymbol{X} = \boldsymbol{X}^{(2)}$ and $S = S^{(1)}$;
2) Fix $\boldsymbol{P}^{(2)}$ and construct $S^{(2)}$ according to Equation 7. Solve for $\boldsymbol{P}^{(1)}$ according to Equation 5 and 6 by setting $\boldsymbol{X} = \boldsymbol{X}^{(1)}$ and $S = S^{(2)}$;
3) Go back to step 1 and iterate. At the end of each iteration, calculate the agreement score using Equation 8. Stop when the agreement score converges or does not reach a new maximum within a fixed number of iterations;

In other words, we iteratively use similarity matrix generated in the subspace of one view as constraint to train LPP projections in the other view. Sometimes the dimensionality of original features is very high (more than several thousand), PCA can be applied to each view as a preprocessing step as suggested in Laplacianface method [11]. Since the cross-view training process of LPP projections is similar to the co-training process of predictors, we refer the new approach as Co-LPP algorithm, which is formally summarized in Algorithm 1.

Here we justify the proposed co-training approach to optimize the objective function in Equation 9. Step 1 of the co-training process optimizes the following objective function:

$$\arg\min_{\boldsymbol{P}^{(2)}} \sum_{i,j} (\boldsymbol{P}^{(2)T}\boldsymbol{x}_i^{(2)} - \boldsymbol{P}^{(2)T}\boldsymbol{x}_j^{(2)})^2 S_{ij}^{(1)} \tag{10}$$

Accordingly, if two samples are considered similar in view 1 ($S_{ij}^{(1)} = 1$), they are required to be projected close to each other in view 2, otherwise a penalty is incurred. As a result, the new similarity matrix $S^{(2)}$ determined from

$\boldsymbol{P}^{(2)T}\boldsymbol{X}^{(2)}$ is forced to have a similar structure to $S^{(1)}$. The same logic applies in step 2 where $\boldsymbol{P}^{(1)}$ is obtained by training LPP with $\boldsymbol{X}^{(1)}$ and $S^{(2)}$. We acknowledge that the proposed optimization approach is heuristic, as is in the case of the original co-training method [12]. While we do not present a strict proof of convergence, we did not observed divergence in any case of our experiments.

IV. Application to Multibiometric Retrieval

Because of the unsupervised nature of the proposed Co-LPP algorithm, it is particularly suitable for biometric applications where no class labels are available, retrieval and clustering for instance. Here we discuss its applications to a multibiometric data retrieval problem.

Given a pool of unlabelled biometric database consisting of n samples represented in two views $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ where $\boldsymbol{x}_i = \{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}\}$ and one query sample $\boldsymbol{q} = \{\boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}\}$, a retrieval algorithm is expected to return t (retrieval window size) samples from $\boldsymbol{X}$ which are considered containing the same subject as in the query $\boldsymbol{q}$. In our approach, PCA is first performed separately on each view of $\mathbf{X}$ as a preprocessing step, the obtained PCA projection matrix for two views are noted as $\boldsymbol{P}_{pca}^{(1)}$ and $\boldsymbol{P}_{pca}^{(2)}$ respectively. The Co-LPP projections matrix $\boldsymbol{P}_{colpp}^{(1)}$ and $\boldsymbol{P}_{colpp}^{(2)}$ are then jointly learned with PCA embeddings of both views. The final embedding of the i-th sample in the v-th view is:

$$\mathbf{y}_i^{(v)} = \boldsymbol{P}_{colpp}^{(v)T}\boldsymbol{P}_{pca}^{(v)T}(\boldsymbol{x}_i^{(v)} - \boldsymbol{\mu}^{(v)}),\ (v = 1, 2) \tag{11}$$

where $\boldsymbol{\mu}^{(v)}$ is the mean of $\boldsymbol{x}^{(v)}$. Similarly, the two view of the query sample are projected into the obtained subspaces by:

$$\mathbf{z}^{(v)} = \boldsymbol{P}_{colpp}^{(v)T}\boldsymbol{P}_{pca}^{(v)T}(\boldsymbol{q}^{(v)} - \boldsymbol{\mu}^{(v)}),\ (v = 1, 2) \tag{12}$$

Then the cosine similarity score between $\boldsymbol{z}$ and each target $\boldsymbol{y}_i$ in the v-th view is calculated as:

$$s_i^{(v)} = \frac{\boldsymbol{z}^{(v)} \cdot \boldsymbol{y}_i^{(v)}}{||\boldsymbol{z}^{(v)}||\ ||\boldsymbol{y}_i^{(v)}||},\ (v = 1, 2) \tag{13}$$

This similarity score is bounded between -1 and 1. In each view, t samples in the database with the largest similarity score are returned. Since our approach learns similar local data structure across different views, the retrieval results in each view also tend to be similar. However, they are not necessarily the same. A fusion could be performed to further improve the performance. Since this paper is focused on discriminant feature extraction rather than fusion, we apply a simple score level fusion by weighted sum:

$$s_i = \alpha s_i^{(1)} + (1 - \alpha)s_i^{(2)} \tag{14}$$

where $0 \leq \alpha \leq 1$ is a weighting parameter.

V. Experimental Results

In this section, the effectiveness of the proposed algorithm is evaluated with an audio-visual speaker retrieval experiment, where each data sample is represented by a person's facial and vocal feature.

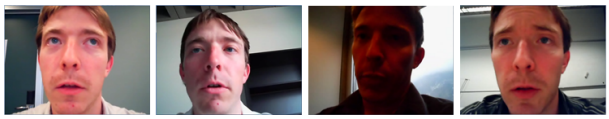


Fig. 1. Sample faical images of a subject in different sessions in MOBIO database

A. Database and feature extraction

The experiment is carried out on with the standard MOBIO database [10] which contains videos of 150 subjects captured in real-world, challenging conditions. Recordings come from a mobile phone camera and are captured in 12 different sessions over a 18-month period where each session contains 11-21 videos. Figure 1 illustrates the level of inter-session variation in a set of example frames for a given subject. For computational efficiency, we test our algorithm using a subset of data from 40 male subjects and for each of them, 5 videos are selected from each of the 12 sessions. This results in a pool of 2400 video samples.

We use cropped face images provided with the MOBIO database, one image per video sample. All images are resized to 50×43 pixels and then histogram equalized. Rows of pixel intensities are concatenated to form feature vectors of 2150 dimensions. The speech signal is split into frames of 20ms duration before the extraction of features composed of 26 Mel-scaled frequency cepstral coefficients (MFCCs), their 26 derivatives and the delta energy. Energy-based voice activity detection is then applied to disguard non-speech frames. A 64-component Gaussian mixture model (GMM) is then fitted to remaining speech data through the maximum a posteriori (MAP) adaptation of a speaker-independent world model. The means of the GMM model are then concatenated to form a 3392-dimensional GMM supervector [13]. As a result, each video is represented by a face feature vector and a vocal feature vector.

B. Evaluation Protocol

In our experiment, we adopted a *leave-one-out* strategy. Each time, one video sample is randomly selected as a query while the left 2399 samples are used as the target database. Note the query sample is not included in the subspace training process. Commonly used evaluation metrics for an information retrieval system involves *Precision* (P) and *Recall* (R). Let a be the total number samples of the same class as the query in the database, b be the number of correctly retrieved samples and t be the retrieval window size, then $P = b/t$ and $R = b/a$. The larger is the retrieval window, the lower is the precision score and the higher is the recall score. In our experiment, we choose a window size $t = a = 59$. In this case $P = R$, which is similar to the concept of equal error rate (EER) in biometric verification setting. We use the corresponding precision/recall score as an evaluation metric. The experiment is repeated 50 times with the random selection of the query sample and the mean precision is reported.

TABLE I
AVERAGE RETRIEVAL PRECISION COMPARISON

Precision	Face	Speech	Fusion
PCA	0,478	0,452	0,615
LPP	0,710	0,675	0,847
CCA	0,858	0,784	0,898
KCCA	0,879	0,796	0,910
Co-LPP	**0.984**	**0,952**	**0,994**

The performance of the proposed Co-LPP algorithm is compared to four alternative dimensionality reduction approach: single-view approach PCA [2], LPP [9], as well as multi-view approach CCA [6] and KCCA [7]. Note that different dimensionality reduction algorithms are used to determine subspace projections, while the retrieval processes follow the same protocol as presented in IV. Here we declare the parameter selections in our experiments:

Dimensionality of subspaces: According to the analysis in [6], in the case of CCA, most discriminative information resides in the first *number of classes - 1* eigenvectors corresponding to the largest eigenvalues. For simplicity, in all compared methods, the dimensionality of subspaces is set to be *number of classes - 1*.

Number of neighbours in KNN graphs: Both LPP and Co-LPP need to specify the number of neighbours in KNN graphs. Here we adopt a rule of thumb $K = log(n)$ where n is the total number of samples, as suggested in [14]. In our experiments, this choice leads to reasonable performance for both LPP and Co-LPP.

PCA pre-preprocessing: As discussed in Section III-B, in all our experiment, the original features are pre-processed by PCA while keeping 90% information in the sense of reconstruction error.

Fusion: In the score level fusion process, for each method, the weighting parameter α in Equation 14 is set to 100 values equally distributed in $[0, 1]$ region and the best accuracy is reported.

C. Results and analysis

The retrieval precision in single face and speech modality and score-level fusion scheme is reported in TABLE I. We observed that the performance of all multi-view approaches (CCA, KCCA, and Co-LPP) out-performs single-view approaches (PCA and LPP) in each single modality and the fusion scheme. If for each approach, we compare the score-level fusion accuracy (the third column) with the best single-view accuracy (the first column), it is noticed that the fusion process provides less relative gain in multi-view approaches than in single-view approaches. This could be expected since the information fusion process has already been integrated in the multi-view dimensionality reduction process, and the extracted features in two views become correlated. Nevertheless, the best single view accuracy (face modality) of CCA, KCCA, and Co-LPP out-performs the score-level fusion accuracy in PCA and LPP, which demonstrates the effectiveness of multi-view dimensionality reduction in exploring multiple information sources. The performance of KCCA is slightly better than CCA, but it needs careful tuning of kernel parameters. Finally, the proposed Co-LPP algorithm out-performs the closest-performing KCCA approach with a significant margin. Even using the weaker single view (speech), a better retrieval accuracy is obtained than the fusion scheme in KCCA. The score-level fusion scheme in Co-LPP subspaces obtained near perfect retrieval performance ($Precision = Recall > 99\%$). Figure 2 shows the variation in retrieval accuracy in and agreement scores as a function of the number of iterations. The agreement score is seen to stabilized after approximately 10 iterations.

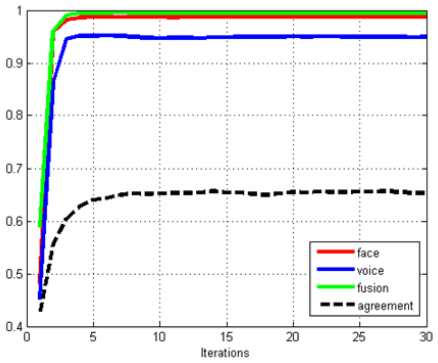
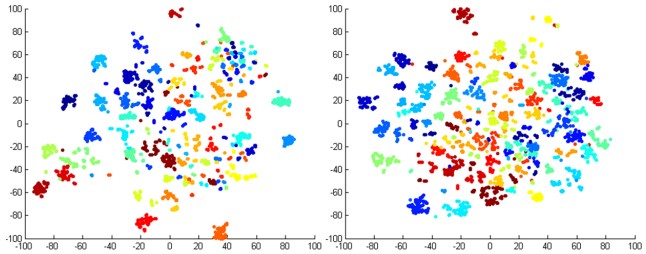


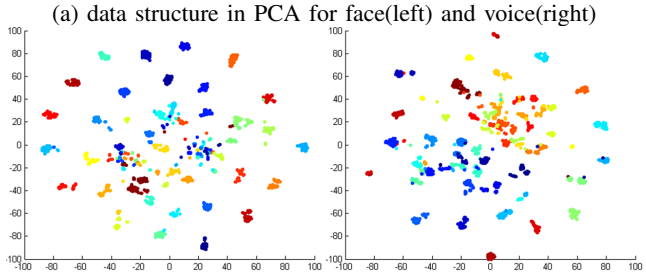
Fig. 2. Retrieval precision for face (red), voice (blue), fusion (green) scheme and agreement score (black) as a function of number of iterations of co-training

D. Data structure visualization

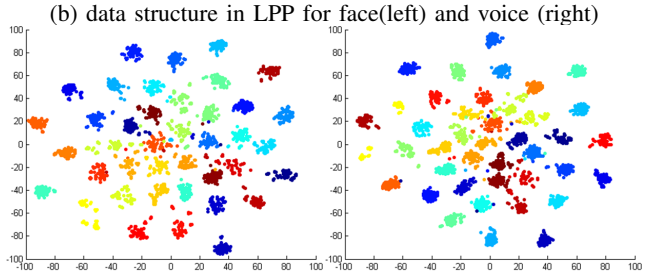
All the approaches compared above embed data samples into lower dimensional spaces. It is of interest to visualize the embedded data structure and thus to observe the relationship between the embedded structure and retrieval performance. However, the embedded subspaces are still high-dimensional and cannot be visualized directly. T-distributed Stochastic Neighbour Embedding (t-SNE) [15] is a powerful tool used to visualize high-dimensional data via embedding into a 2-D or 3-D space while respecting relative distances between data samples. Figure 3 illustrates 2-D scatter plots of PCA, LPP, CCA, and Co-LPP embeddings of all 2400 samples after the application of t-SNE. In all cases, samples belonging to different classes are represented by different colours. In PCA subspace (Figure 3(a)), the sample distribution is especially noisy and explains poor retrieval performance in this case. In LPP subspace (Figure 3(b)), some classes form compact clusters while, and some other classes are mixed together. In CCA subspace (Figure 3(c)), the mixing of different classes is significantly reduced, yet intra-class scattering is still relatively large. This observation confirms the analysis in [6], CCA is able to maximize the scattering of the centroids of each underlying classes (inter-class scattering), but is not able to minimize the intra-class scattering. Finally, in the proposed



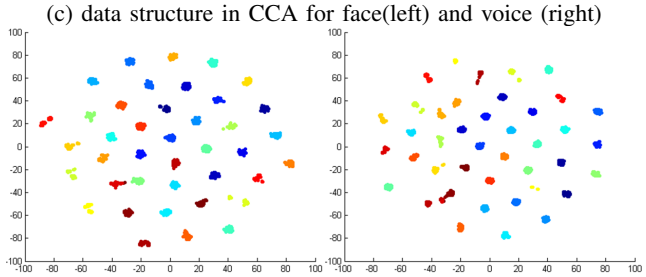
(a) data structure in PCA for face(left) and voice(right)



(b) data structure in LPP for face(left) and voice (right)



(c) data structure in CCA for face(left) and voice (right)



(d) data structure in Co-LPP for face(left) and voice (right).

Fig. 3. 2-D t-SNE visualizations of data structures for PCA, LPP, CCA, and Co-LPP subspaces. Different subjects/classes are represented by different colors.

Co-LPP subspaces (Figure 3(d)), same-class samples are well located in compact clusters with a much higher between-class separation, thereby illustrating its superior performance. In other words, Co-LPP algorithm have high discriminative power despite of its unsupervised nature.

Besides its application in multi-view data retrieval, the proposed Co-LPP algorithm is also well suited to clustering of multi-view data. Still using the same database as in the retrieval experiment, we performed k-means clustering on the 2400 samples in PCA, LPP, CCA and Co-LPP subspaces. The best clustering accuracy is achieved in Co-LPP subspace, which could be expected from the observation of data structure in Figure 3.

VI. Conclusions

This paper proposes a new unsupervised multi-view dimensionality reduction algorithm. Given data with multiple representations, the proposed algorithm aims to learn a subspace projection for each view such that the local data structure in each subspace is in maximal agreement across each view. The method is unsupervised, leading to the potential to avoid expensive and time-consuming manual labelling in scenarios where labelled data is scarce. The new algorithm has high discriminative power compared to other competing approaches.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[2] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.

[3] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of multibiometrics*. Springer Science+ Business Media, 2006, vol. 6.

[4] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," in *Proceedings of 12th European Signal Processing Conference*, 2004, pp. 1221–1224.

[5] D. P. Foster, S. M. Kakade, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," Technical Report TR-2008-4, TTI-Chicago, Tech. Rep., 2008.

[6] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 129–136.

[7] D. R. Hardoon and J. Shawe-Taylor, "Kcca for different level precision in content-based image retrieval," in *Proceedings of Third International Workshop on Content-Based Multimedia Indexing, IRISA, Rennes, France*, 2003.

[8] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang, "Sparse unsupervised dimensionality reduction for multiple view data," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 10, pp. 1485–1496, 2012.

[9] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, p. 153.

[10] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, july 2012, pp. 635 –640.

[11] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, 2005.

[12] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[14] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[15] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.