OPEN-SET SEMI-SUPERVISED AUDIO-VISUAL SPEAKER RECOGNITION USING CO-TRAINING LDA AND SPARSE REPRESENTATION CLASSIFIERS

Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay

Department of Multimedia Communication, EURECOM
Campus SophiaTech, 450 Route des Chappes, 06410 Biot FRANCE
E-mail: {zhaox, evans, dugelay}@eurecom.fr

ABSTRACT

Semi-supervised learning is attracting growing interest within the biometrics community. Almost all prior work focuses on closed-set scenarios, in which samples labelled automatically are assumed to belong to an enrolled class. This is often not the case in realistic applications and thus open-set alternatives are needed. This paper proposes a new approach to open-set, semi-supervised learning based on co-training, Linear Discriminant Analysis (LDA) subspaces and Sparse Representation Classifiers (SRCs). Experiments on the standard MOBIO dataset show how the new approach can utilize automatically labelled data to augment a smaller, manually labelled dataset and thus improve the performance of an open-set audio-visual person recognition system.

***Index Terms*—** Semi-supervised learning, open-set identification, multimodal biometrics, co-training

1. INTRODUCTION

Most biometric systems follow a supervised learning paradigm where client models are learned with labelled training samples acquired during enrollment and latter compared to test samples to establish their identity. In many, real operational scenarios, however, test data can exhibit substantial differences to that collected during enrollment. This so-called inter-session variability can cause significant degradation in biometric recognition performance.

As a result, recent decades have seen a tremendous amount of research in discriminant feature extraction and more robust approaches to modelling and classification to improve recognition performance. The general approach to discriminant feature extraction involves the decomposition of observations into session-dependent and session-independent components and the use only of the later for recognition. Joint factor analysis (JFA) [1] is one such example which has dominated the field of speaker recognition over recent years. Another class of approaches involves Semi-supervised dimensionality reduction, semi-supervised discriminant analysis (SDA) [2] for example as an example, aim to learn discriminant low dimensional discriminant subspaces using both labelled and unlabelled data. Those feature extraction approaches generally served as a pre-processing step before classification, and the client model itself is not enhanced.

While there are many different strategies to improve the robustness of modelling and classification, improvements are limited by the quantity of available data. As a result, semi-supervised learning approaches [3] have attracted significant attention in the recent past. The widely acclaimed self-training and co-training algorithm [4] is perhaps the most popular. The aim here is to augment a manually labelled training set with abundant, but automatically labelled data alternatively acquired and thus to improve recognition performance through the better modelling of intersession variability using larger datasets. Self-training and co-training algorithms have been applied extensively in face identification and verification [5, 6, 7], speaker identification and verification [8, 9] and multi-modal biometrics [10]. While all of this **prior work** uses larger quantities of data for more reliable modelling, the feature space remains unchanged and therefore it still contains intersession variation. Moreover, the prior work generally considers only closed-set scenarios by assuming that unlabelled samples belong to one of the pre-enrolled clients. Out-of-class samples must be expected, however, and they will degrade recognition performance if not properly handled. Virtual label regression [11] is one of the very few semi-supervised learning methods which independently models out-of-class samples and excluded them from the unlabelled samples to train classifiers, but it cannot deal with multi-view learning problems where the input is constitute of multiple modalities.

Our own **prior work** in multimodal person recognition [12] introduced a new approach to combine the learning of discriminant features with more robust modelling and classification in a unified co-training framework. However, it also assumes a closed-set scenario. This paper presents our latest work to extend our co-training algorithm to open-set scenarios. The new algorithm combines linear discriminant analysis (LDA) with a sparse representation classifier (SRC) [13]. While SRC has shown to give state-of-the-art performance in face recognition [13] and speaker recognition [14], it depends upon the availability of large quantities of data, hence its combination with co-training. A sparsity concentration index (SCI) is also effective in rejecting out-of-class data, hence its suitability to open-set problems.

The remainder of this paper is organized as follows. In Section 2, we briefly summarize the three central components of the proposed algorithm: LDA, SRC and co-training. In Section 3, we introduce the new open-set co-training LDA and SRC algorithm (co-LDA-SRC). Experiments in open-set, semi-supervised audio-visual speaker identification are reported in Section 4 before our conclusions are presented in Section 5.

2. BACKGROUND

In this section, we briefly review the three central components of the proposed algorithm: LDA for feature extraction, SRC for classification and the rejection of out-of-class samples, and co-training for semi-supervised learning.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

2.1. LDA

LDA seeks an optimised transform $\boldsymbol{P_{opt}}$ which projects t dimensional data vectors $\boldsymbol{x}$ into a $g < t$ dimensional subspace according to $\boldsymbol{y} = \boldsymbol{P_{opt}x}$. The projection aims to minimize within-class scatter (S_W) while maximising between-class scatter (S_B) where:

$$\begin{aligned} S_W &= \sum_{j=1}^{c}\sum_{i=1}^{l_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^T \\ S_B &= \sum_{j=1}^{c} l_j(\mu_j - \mu)(\mu_j - \mu)^T \end{aligned} \quad (1)$$

and where x_i^j is the i^{th} training sample of of class j, μ_j is the class mean, l_j is the number of samples and c is the number of classes. $\boldsymbol{P_{opt}}$ is obtained according to the objective function:

$$\boldsymbol{P_{opt}} = \arg\max_P \frac{P^T S_B P}{P^T S_W P} = [p_1, \ldots, p_g] \quad (2)$$

where each p_i is an eigenvector of S_B and S_W which corresponds to the g largest generalized eigenvalues of:

$$S_B p_i = \lambda_i S_W p_i \ (i = 1, \ldots, g), \quad (3)$$

LDA is still used widely for discriminant feature extraction. Its application in face and speaker recognition is referred to as ”Fisherface” [15] and ”Fishervoice” [16] respectively.

2.2. SRC

Suppose we have c classes, and let $\mathbf{A} = [\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_c}]$ be a set of training samples, where $\mathbf{A_i} = \{\mathbf{v_{i,1}}, \ldots, \mathbf{v_{i,n_i}}\}$ indicates the subset of training samples for class i. A single testing sample $\mathbf{y}$ could be well approximated by the linear combination of training samples from $\mathbf{A_i}$, which could be written as

$$\mathbf{y} = \sum_{j=1}^{n_i} \alpha_{i,j}\mathbf{v_{i,j}}. \quad (4)$$

Since $\mathbf{A}$ is the dictionary which includes all the training samples, Equation 4 can be rewritten in the form $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha}_0$ where $\boldsymbol{\alpha}_0 = \{0, \ldots, 0, \alpha_{i,1}, \ldots, \alpha_{i,n_i}, 0, \ldots, 0\}^T$ is the coefficient vector in which most coefficients are zero except the ones associated with class i. Due to the fact that a valid test sample $\mathbf{y}$ can be sufficiently represented only using the training samples from the same class, and this representation is the sparsest among all others, to find the identity of $\mathbf{y}$ then equals to find the sparsest solution of $\boldsymbol{\alpha}$. So The four main steps involved in the application of SRC are outlined in the following.

1. Normalize each entry in $\mathbf{A}$ to have unit $l2$-norm;
2. Sparsely code $\mathbf{y}$ on $\mathbf{A}$ via $l1$-norm minimization:

$$\hat{\boldsymbol{\alpha}} = \arg\min \parallel \boldsymbol{\alpha} \parallel_1 \text{ subject to } \parallel \mathbf{y} - \mathbf{A}\boldsymbol{\alpha} \parallel_\mathbf{2} < \epsilon \quad (5)$$

3. Compute the residuals of each class by:

$$r_i(\mathbf{y}) = \parallel \mathbf{y} - \mathbf{A}\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}} \parallel \ \ (\mathbf{i = 1, \ldots, c}) \quad (6)$$

where $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with class i, and $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_\mathbf{1}, \ldots, \hat{\boldsymbol{\alpha}}_\mathbf{c}]$,

4. Classification according to:

$$\text{Identify}(\mathbf{y}) = \arg\min_\mathbf{i}(\mathbf{r_i(y)}) \quad (7)$$

SRC was originally developed for face identification [13] and has since been applied in speaker identification [14]. Comparative experiments show that SRC outperforms Nearest Neighbor (NN) and Support Vector Machine (SVM) classifiers.

The original work [13] proposed a sparsity concentration index (SCI) which aims to reject invalid test samples. We propose its use to reject out-of-class data. Since the aim in SRC is to represent each test sample according to a sparse, weighted set of training samples, the representation of within-class samples should be concentrated on a single class. The representation of out-of-class samples, however, is more dispersed. The SCI score of a coefficient vector $\hat{\boldsymbol{\alpha}}$ is defined as:

$$SCI(\hat{\boldsymbol{\alpha}}) = \frac{c * \max_i \parallel \hat{\boldsymbol{\alpha}}_{\boldsymbol{i}} \parallel_1 / \parallel \hat{\boldsymbol{\alpha}} \parallel_1 - 1}{c - 1} \quad (8)$$

and is bounded between 0 and 1. Out-of-class samples can thus be rejected according to a threshold $\tau \in (0, 1)$ where $SCI(\hat{\boldsymbol{\alpha}}) < \tau$.

2.3. Co-training

Co-training belongs to a class of algorithms which combine semi-supervised learning and multi-view learning into one unified framework. In co-training, data samples are assumed to be represented by two disjoint views $\mathbf{x_1}$ and $\mathbf{x_2}$. Two classifiers $C_1(\mathbf{x_1})$ and $C_2(\mathbf{x_2})$ are initially learnt with a small set of labelled data **L**: $\{x_{i1}; x_{i2}, l_i | i = 1, 2, ..., m\}$ where l is the class label, and a large amount of unlabelled data **U**: $\{x'_{i1}; x'_{i2} | i = 1, 2, ..., n\}$, where n and m denote the size of labelled and unlabelled datasets respectively. At each iteration, the algorithm incorporates samples from the unlabelled set **U** into the pool of labelled data **L**. Typically the selected data are those with the highest prediction confidence for each view. Each classifier is then updated using the augmented labelled data set. The process can be repeated iteratively until all auxiliary data is incorporated. In [4], co-training was shown to require two conditionally independent views in order that each classifier provides informative data to the other.

3. CO-LDA-SRC

This section describes our core contribution, namely a new approach to open-set, semi-supervised learning.

3.1. Overview

As shown in [17], LDA projections can be unrepresentative of intersession variations when learned on smaller datasets and thus give unsatisfactory performance. SRC also requires abundant labelled training data so that test samples can be reliably reconstructed from a linear combination of same-class training samples [13]. In most biometric applications, however, labelled data acquired during enrollment is generally limited in quantity and the acquisition of more, manually labelled data is usually costly or impractical. In the following we show how both LDA and SRC can be integrated within a unified co-training framework thereby exploiting abundant, unlabelled data to improve performance.

Consider a multi-modal biometric system where different biometric modalities can be considered as independent views of the same data. Also assume that abundant auxiliary data can be acquired over an extended period so that it is representative of intersession variations. According to a general co-training scheme, a classifier in one view can be used to provide automatically labelled, new training data to another, and vis-versa.

Algorithm 1 Co-LDA-SRC

```
Input:
    • Labelled dataset $\mathbf{L}$ from $c$ classes and unlabelled dataset $\mathbf{U}$;
    • SCI Threshold $\tau$ and number of samples $N$ to be incorporated into the set of labelled samples.
Output:
    • Projection matrix $\mathbf{P_1}$ and $\mathbf{P_2}$;
    • Increased labelled training set $\mathbf{L}$.
Initialization: Center $\mathbf{L}$ and $\mathbf{U}$ in both view, apply PCA if the dimensionality is too high;
repeat
  for $v = 1, 2$ do
    •  Train LDA projections $\mathbf{P_v}$ with samples in the $v_{th}$ view of $\mathbf{L}$ and project samples according to $\mathbf{P_v}$ to form $\mathbf{A_v}$;
    •  Project the $v$-th view of $\mathbf{U}$ into $\mathbf{P_v}$, noted as $\mathbf{Y_v}$;
    •  Run SRC on each entry of $\mathbf{Y_v}$ with training set $\mathbf{A_v}$, discard entries with SCI lower than $\tau$.
    •  $\mathbf{L_v} \leftarrow \emptyset$
    for $i = 1$ to $c$ do
      for each class $i$, add to $\mathbf{L_v}$ the single sample in $\mathbf{U}$ most confidently labelled (lowest $r_i(\mathbf{y})$).
    end for
  end for
  $\mathbf{L} \leftarrow \mathbf{L} \cup \mathbf{L_1} \cup \mathbf{L_2}$; $\mathbf{U} \leftarrow \mathbf{U} - \mathbf{L_1} - \mathbf{L_2}$
until $N$ pseudo-labelled samples are incorporated into the training set
```

The standard co-training algorithm assumes a closed-set scenario, where all unlabelled data belong to one of the registered classes. In practical scenarios, however, and particularly for biometric systems, data acquired automatically during regular use may often contain out-of-class samples (persons not pre-enrolled). Out-of-class samples should not be incorporated into the labelled training set.It is thus necessary to adapt the standard co-training algorithm to reject out-of-class samples. This facility is provided readily through a threshold SCI as discussed in Section 2.2.

3.2. Algorithm

We assume each data sample is represented by two feature vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ extracted from two independent biometric traits. A small labelled training set of n samples $\mathbf{L}$: $\{\mathbf{x_{i1}}, \mathbf{x_{i2}}; l_i | i = 1, 2, ..., n\}$ is acquired during an enrollment session, while a larger unlabelled dataset of m samples $\mathbf{U}$: $\{\mathbf{x'_{i1}}, \mathbf{x'_{i2}} | i = 1, ..., m\}$ is obtained over an extended period of normal use. The entire training set is noted by $\mathbf{X} = \mathbf{L} \cup \mathbf{U}$.

$\mathbf{X}$ is first centred so that $\overline{\mathbf{x}}^{(\mathbf{v})} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_{iv}} = 0, (v = 1, 2)$, and optionally treated conventionally with principal component analysis (PCA) to reduce its dimensionality if to high to be treated directly with LDA. Then, LDA projections $\mathbf{P_1}$ and $\mathbf{P_2}$ are determined for each view using only the set of labelled samples $\mathbf{L}$. The same set is then projected into the new subspaces according to $\mathbf{A_v} = \mathbf{P_v}^T \mathbf{x_v}$. The result forms training examples for SRC in the v-th view.

Both views $\mathbf{x'_1}$ and $\mathbf{x'_2}$ of the set of unlabelled samples U are then projected onto their respective subspaces according to $\mathbf{Y_v} = \mathbf{P_v}^T \mathbf{x'_v}$. Each entry $\mathbf{y}$ of $\mathbf{Y_v}$ is sparsely coded on $\mathbf{A_v}$ according to Equation 5, and the reconstruction residues $r_i(\mathbf{y})$ and SCI score are determined according to Equations 6 and 8 respectively. Those entries whose SCI score is less than a threshold τ are labelled as out-of-class samples, whereas the remaining in-class samples are assigned to one of the known classes according to Equation 7. For each view and each class, the single in-class sample most confidently labelled (with the lowest $r_i(\mathbf{y})$) is removed from $\mathbf{U}$ and incorporated into $\mathbf{L}$. Projections $\mathbf{P_1}$ and $\mathbf{P_2}$ are then re-trained with the now-larger labelled dataset. This process is repeated until a pre-specified number of labelled samples are gathered. The algorithm is summarized in Algorithm 1. In the test phase, the v-th view of a test sample is projected onto $\mathbf{P_v}$ and classified by SRC with the increased training set $\mathbf{A_v}$.

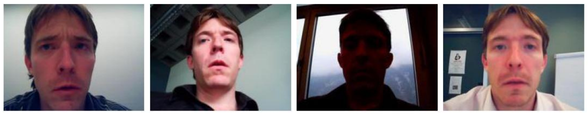

Fig. 1: Sample faical images of a subject in different sessions in MOBIO database

4. EXPERIMENTAL WORK

In this section, we report an evaluation of the proposed Co-LDA-SRC algorithm through experiments in audio-visual persons identification where the task is to identify the speaker in a video sequence according to acoustic and facial observations. A small sum of labelled training data collected during a single enrollment session is used as labelled data for initial modelling. Comparisons against a baseline system using supervised LDA feature extraction and SRC classification show how learning from a larger pool of unlabelled data acquired during normal system use is effective in capturing intersession variation. We stress, however, that the framework is general and can be applied to any multi-view problems.

4.1. Database

Experiments were conducted with the standard MOBIO database [18] 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. A typical example of the inter-session variation, which necessitates the modelling of inter-session variation, is illustrated in Figure 1.

4.2. Feature extraction

Experiments are conducted with largely standard speaker and face recognition systems which represent each video sequence by a GMM speaker supervector [19] and a simple holistic face feature vector based on intensity. Both are of high dimensionality.

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

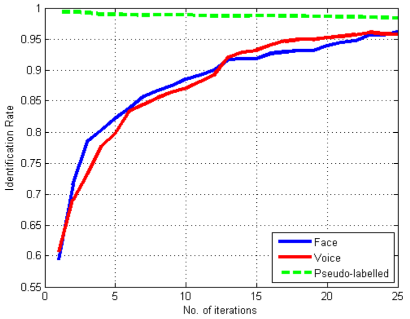

Fig. 2: Identification rate and pseudo-labelled data accuracy as a function of the iteration number.

(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. Both face and speech feature vectors are first reduced to 100 dimensions through the application of PCA.

4.3. Protocols

To create a pool of in-class samples, we selected only 20 subjects as registered clients. For each subject, 5 videos are selected from each of the 12 sessions, which results in 60 videos per subject. For each registered subject, 5 videos in one randomly selected session are used as labelled training data for enrolment, 5 videos from another randomly selected session are used as test data, and the 50 videos in the other 10 sessions are used as unlabelled data. A random selection of 300 videos of the remaining, different subjects were added to the unlabelled dataset as out-of-class samples. Thus, from an unlabelled pool of 1300 samples, just under 25% are out-of-class. The SCI threshold τ is empirically set to 0.4, and the number of pseudo-labelled samples N to be incorporated into the labelled set is set to 90% of the expected number of in-class samples in the unlabelled dataset.

The evaluation is two-fold: first, we report a top-1, closed-set identification experiment performed on the independent test dataset; second, we report the labelling accuracy of automatically labelled data. All results are averaged through 20-fold cross-validation.

4.4. Results

Figure 2 shows the identification rate of an SRC classifier applied to face and voice observations independently, in addition to the accuracy of the increasing number of pseudo-labelled samples added to the labelled dataset. Between each iteration the size of the labelled dataset increases by about of $20 \times 2 = 40$ samples. While the labelling accuracy of pseudo-labelled samples is shown to decrease to $98,5\%$, the effect of labelling errors does not outweigh the benefit of modelling intersession variations through the use of additional, automatically labelled data. Profiles show that the identification rate for both face and voice classifiers increases when a greater number of unlabelled samples is incorporated into the training set through co-training.

Table 1 shows the mean value and standard deviation of identification rate over 20 runs of different algorithms. The baseline approaches are the SRC classifiers applied to features in PCA and LDA-derived subspaces, where the training samples only include the original, manually labelled dataset. The performance of LDA is even worse than that of unsupervised PCA, most probably due to the effect of over-fitting. We also report results for Semi-supervised Discriminant Analysis (SDA) [2] and Virtual Label Regression (VLR) [11], two semi-supervised feature extraction methods trained on both labelled and unlabelled data. Due to the use of single views in each case, however, both approaches yield only modest improvements over the PCA and LDA systems. VRL outperforms SDA since it is one of the very few semi-supervised learning approaches where out-of-class samples are modelled independently and excluded from the in-class data to train the projection. Our own previous approach, Co-LDA [12], out-performs all single view methods on account of the the co-training framework. Finally, the proposed multi-view, co-training algorithm out-performs co-LDA by a large margin. The significant improvement in performance is attributed to the use of an SRC classifier and its capacity to reject out-of-class samples. Compared to the co-LDA algorithm, the error rate is reduced by over 60% relative. The experiment demonstrate the effectiveness of the proposed algorithm to use unlabelled data to enhance the recognition performance of traditional supervised multi-modal biometric systems.

Id. Rate(std.)	Face	Speech
PCA + SRC	0,670(0,035)	0,652(0,030)
LDA + SRC	0,590(0,046)	0,611(0,048)
SDA[2] + SRC	0,725(0,029)	0,759(0,032)
VLR[11]	0.772(0,036)	0,863(0,033)
Co-LDA[12]	0.902(0.032)	0.891(0.034)
Co-LDA-SRC	**0,961(0,051)**	**0,962(0,054)**

Table 1: Comparison of identification rate and standard deviation of different algorithms on MOBIO database

5. CONCLUSIONS

This paper reports a new open-set, semi-semi-supervised learning framework capable of the simultaneous extraction of discriminant features and the learning of robust classifiers using both labelled and unlabelled data. In contrast to prior work and traditional co-training algorithms, the proposed Co-LDA-SRC algorithm is able to filter out-of-class samples typical in many realistic applications. Experiments in open-set, audio-visual person identification using the MOBIO database show relative improvements of over 60% compared to several competitive baseline algorithms.

6. REFERENCES

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435 –1447, may 2007.

[2] D. Cai, X. He, and J. Han, “Semi-supervised discriminant analysis,” in *IEEE 11th International Conference on Computer Vision (ICCV)*, oct. 2007.

[3] X. Zhu, “Semi-supervised learning literature survey,” Tech. Rep., Univ. Wisconsin, Madison.

[4] Avrim Blum and Tom Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA, 1998, COLT’ 98, pp. 92–100, ACM.

[5] X Zhao, N Evans, and J.L. Dugelay, “Semi-supervised face recognition with lda self-training,” in *IEEE International Conference on Image Processing (ICIP)*, sept. 2011, pp. 3041 – 3044.

[6] X Zhao, N Evans, and J.L. Dugelay, “A co-training approach to automatic face recognition,” in *19 the European Signal Processing Conference (EUSIPCO)*, Aug. 2011.

[7] H.S. Bhatt and et al., “On co-training online biometric classifiers,” in *2011 International Joint Conference on Biometrics (IJCB),*, Oct. 2011.

[8] Makoto Yamada, Masashi Sugiyama, and Tomoko Matsui, “Semi-supervised speaker identification under covariate shift,” *Signal Processing*, vol. 90, no. 8, pp. 2353 – 2361, 2010.

[9] C. Fredouille and et al., “Behavior of a bayesian adaptation method for incremental enrollment in speaker verification,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),*, 2000.

[10] F. Roli, L. Didaci, and G.L. Marcialis, “Template co-update in multimodal biometric systems,” *Advances in Biometrics, LNCS*, vol. 4642/2007, pp. 11941202, 2007.

[11] F. Nie, D. Xu, X. Li, and S. Xiang, “Semisupervised dimensionality reduction and classification through virtual label regression,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,*, vol. 41, no. 3, pp. 675 –685, june 2011.

[12] X. Zhao, N Evans, and J.L. Dugelay, “Co-lda: a semi-supervised approach to audio-visual person recognition,” in *IEEE International Conference on Multimedia Exposition (ICME),*, July. 2012.

[13] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 31, no. 2, pp. 210–227, feb. 2009.

[14] I. Naseem, R. Togneri, and M. Bennamoun, “Sparse representation for speaker identification,” in *International Conference on Pattern Recognition (ICPR)*, aug. 2010, pp. 4460 –4463.

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

[16] S.M. Chu, T. Hao, and T.S. Huang, “Fishervoice and semi-supervised speaker clustering,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, april 2009, pp. 4089 –4092.

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

[18] Chris McCool and et al., “Bi-modal person recognition on a mobile phone: using mobile phone data,” in *IEEE ICME Workshop on Hot Topics in Mobile Mutlimedia*, July 2012.

[19] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker verification using adapted gaussian mixture models,” in *Digital Signal Processing*, 2000, p. 2000.