CO-LDA: A SEMI-SUPERVISED APPROACH TO AUDIO-VISUAL PERSON RECOGNITION

Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay

Department of Multimedia Communications, EURECOM
2229 Route des Cretes, BP 193, F-06560 Sophia-Antipolis Cedex, France
{xuran.zhao, nick.evans, jld}@eurecom.fr

ABSTRACT

Client models used in Automatic Speaker Recognition (ASR) and Automatic Face Recognition (AFR) are usually trained with labelled data acquired in a small number of enrolment sessions. The amount of training data is rarely sufficient to reliably represent the variation which occurs later during testing. Larger quantities of client-specific training data can always be obtained, but manual collection and labelling is often cost-prohibitive. Co-training, a paradigm of semi-supervised machine learning, which can exploit unlabelled data to enhance weakly learned client models. In this paper, we propose a co-LDA algorithm which uses both labelled and unlabelled data to capture greater intersession variation and to learn discriminative subspaces in which test examples can be more accurately classified. The proposed algorithm is naturally suited to audio-visual person recognition because vocal and visual biometric features intrinsically satisfy the assumptions of feature sufficiency and independency which guarantee the effectiveness of co-training. When tested on the MOBIO database, the proposed co-training system raises a baseline identification rate from 71% to 99% while in a verification task the Equal Error Rate (EER) is reduced from 18% to about 1%. To our knowledge, this is the first successful application of co-training in audio-visual biometric systems.

***Index Terms*—** Biometrics, speaker recognition, face recognition, co-training, audio-visual person recognition, semi-supervised learning

1. INTRODUCTION

Biometric systems exploit physiological and/or behavioural traits to recognize individuals. Popular traits or modalities include fingerprints, hand-geometry, face, voice, iris, retina, gait, signature, palmprint, ear, etc. Among them, face and voice features have the advantages of non-intrusiveness, easy acquisition and also the possibility of non-cooperative acquisition. Automatic Speaker Recognition (ASR) and Automatic Face Recognition (AFR) have thus attracted a high degree of research interest in the last decade.

ASR and AFR systems generally share the same operational paradigm. During enrolment, training data is collected and client models are learnt or adapted, while under normal use or testing new samples are compared to a single model (verification) or to a group of models (identification). Under well controlled conditions performance is typically acceptable. In real operational scenarios, however, test data can exhibit substantial differences to that collected during enrolment. In the case of face recognition, so-called intersession variability may come from differences in illumination or pose, the presence of facial accessories (glasses or piercings), and ageing over an extended time period. Voice features may vary as a consequence of environmental noise or changes to the vocal tract as a consequence of illness or ageing. Unless such variations are captured and represented in the client models, or unless suitably robust features or normalization approaches are applied, recognition performance can deteriorate drastically.

The use of more robust features can ameliorate this problem to some extent. In AFR, for example, Local Binary Pattern (LBP) features [1] are among the most robust to illumination changes, while SIFT-like features [2] are robust to geometrical transformations. To date, however, there are no "perfect features" universally robust to every foreseeable variation. Another approach involves the decomposition of observed features into session-dependent and session-independent components and the only the later are used for recognition. Decomposition and transformation typically require large quantities of data to learn and some important information is often lost. One such example is Joint Factor Analysis (JFA) [3], which is popular in ASR.

Semi-supervised learning (SSL) is another popular approach to the data insufficiency problem and has experienced a surge in research interest in the machine learning community during the last decade [4]. Compared to supervised learning (learning from labelled data) and unsupervised learning (clustering unlabelled data), SSL uses a small amount of labelled data and a larger pool of unlabelled data to learn models, thereby avoiding costly manual labelling. SSL can be used to solve the problem of scarce labelled data in AFR and ASR: models weakly trained during enrolment can be enhanced by learning from abundant unlabelled data obtained during normal use or testing, which is inherently rich in variation. Several semi-supervised AFR and ASR systems have been proposed and show the capacity for increasing the performance of supervised systems [5][6].

Co-training is one of the most successful examples of SSL and was proposed by Blum and Mitchell [7] in 1998. The basic assumption is that each data sample can be represented by two independent features, each of which is generally sufficient for correct classification. First, two classifiers are weakly trained using a small number of labelled examples on two different feature sets respectively. Each classifier is then used to classify a larger pool of unlabelled auxiliary data. The most positive examples are then used to train the other classifier. The process is iterative and is repeated several times. Consequently, both classifiers become more robust with the accumulation of new training data. Blum and Mitchell demonstrated that if

the two following assumptions are verified, co-training guarantees improved performance over supervised learning [7]: (i) sufficiency, which requires each classifier feeds to the other more correctly labelled samples than incorrectly labelled samples, (ii) independency, which requires that samples confidently classified by one classifier are fully informative to train the other.

One of the first applications of co-training to AFR is proposed in [8], but based on two different facial features. The two features are extracted from the same image and thus the assumption of independency is not satisfied; unlabelled samples confidently classified by one system may not help to improve the other, and thus improvements in performance are modest. A template co-update biometric system based on two independent biometric features, face and fingerprints, is proposed in [9]. This combination of modalities requires special equipment and thus application is limited.

In this paper, we propose a co-training type algorithm which exploit the natural co-occurrence of audio-visual data, namely co-Linear Discriminant Analysis (co-LDA), which uses both labelled and unlabelled data to learn discriminative subspaces in which test examples can be better classified. In this paper we report its application to audio-visual person recognition in videos. The scenario involves a very limited number of labelled videos and a larger auxiliary pool of unlabelled videos. Each video contains images and audio from a single person, and is parametrized by face and voice feature vectors of high dimension. For each feature, a LDA-based classifier is learnt with the small number of labelled samples and is used to classify the unlabelled samples. The most confident classification results (samples) identified by one classifier are added to the labelled data set, and the corresponding features are then used to train the other LDA subspace and classifier, and vice versa. After several iterations and the accumulation of automatically labelled data, we obtain more reliable subspaces for both face and voice classification.

The remainder of this paper is organized as follows. In Section 2, the principles of co-training are described and the co-LDA framework are presented and analysed. The application of the proposed algorithm in audio-visual person recognition is described in Section 3. Experiments and results are detailed in Section 4 before our conclusions are presented in Section 5.

2. CO-TRAINING AND CO-LDA

In this section, we first briefly introduce the principles of co-training and LDA in Section 2.1 and 2.2 respectively, and then present the semi-supervised discriminant subspace learning problem, propose and analyse the co-LDA algorithm in Section 2.3.

2.1. Principle of Co-Training

Co-training belongs to a class of algorithms which combine semi-supervised learning and multi-view learning into one unified framework. The basic assumption of co-training is that the data samples can be presented with 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 $\mathbf{L}$: $\{x_{i1}; x_{i2}, l_i | i = 1, 2, ..., N\}$ where l is the class label, and a large amount of unlabelled data $\mathbf{U}$: $\{x'_{i1}; x'_{i2} | i = 1, 2, ..., M\}$, where N and M denote the size of labelled and unlabelled dataset respectively. At each iteration, the algorithm incorporates samples from the unlabelled set $\mathbf{U}$ into the pool of labelled data $\mathbf{L}$. 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 unlabelled auxiliary data is incorporated into labelled dataset. Finally, the outputs of the two classifiers C_1 and C_2 can be weighted and give a single-view classifier C. The intuition of co-training is that each classifier can provide the other with additional, automatically labelled data which might be as informative as some random noisy labelled examples. Based on the analysis of Nigam et al [7], co-training requires the two views to be conditionally independent in order that each classifier provides informative data to the other.

2.2. Principle of LDA

Linear Discriminant Analysis is a well-known simple and efficient approach to dimensionality reduction, and is widely used in various classification problems. It aims to find an optimised projection $\boldsymbol{W_{opt}}$ which projects t dimensional data vectors $\boldsymbol{x}$ into a f dimensional space by $\boldsymbol{y} = \boldsymbol{W_{opt}x}$, in which intra-class scatter (S_W) is minimized while the intra-class scatter (S_B) is maximized. S_W and S_B are determined according to:

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{l_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T, \tag{1}$$

and

$$S_B = \sum_{j=1}^{c} l_j (\mu_j - \mu)(\mu_j - \mu)^T, \tag{2}$$

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

$$\boldsymbol{W_{opt}} = \arg\max_W \frac{W^T S_B W}{W^T S_W W} = [w_1, \ldots, w_g] \tag{3}$$

where $\{w_i | i = 1, \ldots, g\}$ are the eigenvectors of S_B and S_W which correspond to the g largest generalized eigenvalues according to:

$$S_B w_i = \lambda_i S_W w_i, i = 1, \ldots, g \tag{4}$$

Note that there are at most $c-1$ non-zero generalized eigenvalues, so m is upper-bounded by $c - 1$. Since S_W is often singular, it is common to first apply Principal Component Analysis (PCA) to reduce the dimension of the original vector. LDA has been applied to AFR and ASR and is often referred to as *Fisherface* [10] and *Fishervoice* [11].

While LDA can extract discriminant information from high dimensional feature vectors when labelled training data is abundant, but when training data is scarce, the projections can be significantly biased, which generally leads to reduced performance.

2.3. Co-LDA

In many practical AFR and ASR applications, but unlabelled test data is often abundant, ie. obtained during testing. It typically contains a high degree of intersession variations, from which much more reliable LDA projections can be learnt. We propose a novel

co-training framework which is applied to in the discriminant dimensionality reduction problem in two distinct feature spaces, where each classifier iteratively and automatically labels and provide new training data to another.

As illustrated in Fig.1, the input of the co-LDA algorithm is a small amount of labelled data and a large pool of unlabelled data, while each sample can be represented with two features, $\boldsymbol{x_1}$ (left in Fig.1)and $\boldsymbol{x_2}$ (right), which are assumed to be independent and sufficient for classification. An LDA projection is learnt on each view respectively. As shown in Fig.1 (a), the labelled dataset is small and is not representative of the general class distribution, so S_B and S_W in Equation (1) and (2) are not well estimated. The LDA projection ($\boldsymbol{W_{opt}}$) learned from this data is illustrated by a solid line. It is biased and leads to an ineffective classification boundary (dashed line). The LDA space of view 1, a classifier is then applied to classify all the unlabelled data, one (or a few) sample that is farthest from the classification boundary is added to the labelled set, and the LDA projection for view 2 is relearned, as shown in Fig.1 (b). Note that, since the two views are assumed to be independent from each other, one point confidently classified in view 1 is highly informative in view 2 (otherwise if the two views are correlated, that point will be also far from the classification boundary in view 2), and is able to correct to improve the corresponding LDA. In the same way, unlabelled data in view 2 is also classified, and the most confident samples are added to the labelled dataset before the LDA projection for view 1 is also relearned. The process is iterative and as more labelled data is accumulated, the LDA projections are improved and give better results. Of course, one view may feed misclassified samples to the other but according to the sufficiency assumption, classifiers will feed more correctly labelled data than mislabelled data to the other classifier, and thus performance ultimately improves.

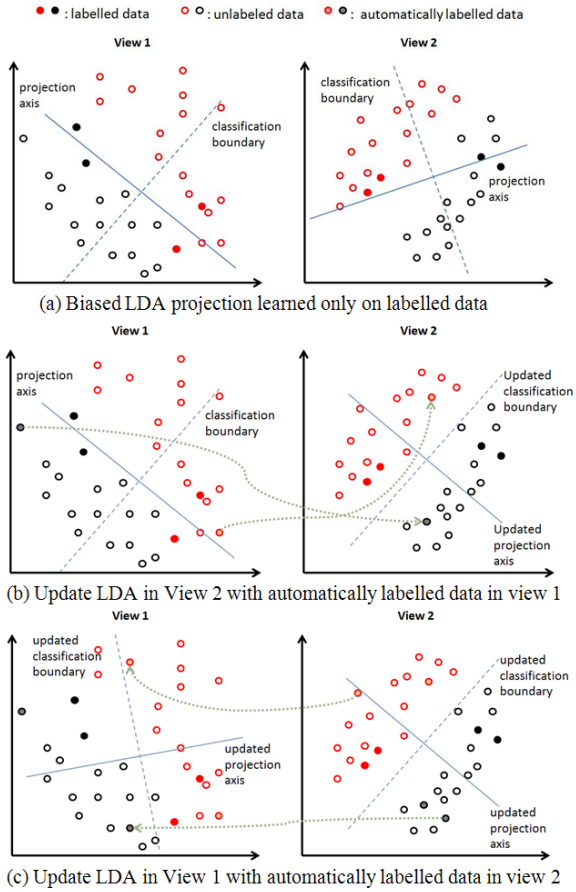


Fig. 1. Illustration of Co-LDA algorithm

3. APPLICATION TO AUDIO-VISUAL PERSON RECOGNITION

It is well known that better recognition performance can be achieved through the combination of multiple biometric modalities, through so-called multi-modal systems [12]. With both traits available with standard commercial video capturing devices and on account of their non-intrusive nature, audio-visual person recognition is of natural appeal to both commercial clients and end-users and thus attracted considerable research interest in recent years. Such systems generally involves the score level fusion of AFR and ASR systems. Both are vulnerable to inter-session variations discussed in Section 1, and the proposed co-LDA approach has natural application in audio-visual person recognition scenario: (1) Labelled data is limited while abundant unlabelled data is available during the normal system operation; (2) Peoples's face and voice are naturally available in videos and are independent from each other; (3) Many state-of-the-art ASR & AFR implementations use high-dimensional feature vectors so dimensionality reduction is needed.

The proposed co-LDA audio-visual person recognition system is composed of three steps. First, a facial feature vector and a vocal feature vector are extracted from each video; second, two discriminant subspaces are learned with both labelled and unlabelled face and voice data respectively; third, verification is achieved with accepting or rejecting the claim, while in the identification task, there is no identity claim, and the system is required to establish their identity.

3.1. Feature vector extraction

The process of feature extraction is illustrated in Fig.2. For the face modality, face detection is first applied and detected faces are aligned according to detected facial landmark positions. For each video, Local Binary Pattern (LBP) feature vector [1] is extracted from the most confident detected face. LBP feature extraction divides faces into sub-regions and LBP histograms, which reflect the local texture are extracted from each region and concatenated to form a high dimensional vector. For the voice modality, voice detection is first applied to eliminate non-speech frames. MFCC coefficients are then extracted from each audio frame and used to determine a Gaussian Mixture Model (GMM) through the Maximum A Posteriori (MAP) adaptation of a speaker independent world model. The means of the GMM model are concatenated into a high-dimensional supervector [13]. Accordingly each video is represented by a facial feature vector f_{face} and a voice feature vector f_{voice}.

3.2. Subspace learning

The co-LDA system is supplied with a small set of labelled training data acquired during the enrolment session, and a large set of

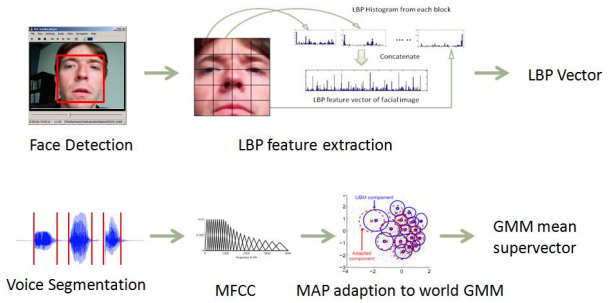


Fig. 2. Feature vector extraction for face and voice

unlabelled data acquired during a period of normal system operation. The dimensionality of the original face and voice feature vector f_{face} and f_{voice} is too great to perform LDA, so a PCA step is first applied to reduce the dimension to n, (x_{face}, x_{voice}) represents the two features in the PCA space. As illustrated in Fig.3, the labelled training samples are first used to learn LDA projections with face and voice feature vectors respectively, and then to learn two classifiers C_{face} and C_{voice}. Here we simply use a nearest-template classifier, where a template for each class is calculated as the within-class mean, and the test samples are assigned the label of the closest template according to the label of a test data is determined according to the normalized correlation metrics, which has been demonstrated to be an appropriate similarity measure for LDA space [14]. The similarity between a test point $\boldsymbol{x}$ and a template $\boldsymbol{\mu}$ is defined as:

$$S_N = \frac{\|\boldsymbol{x}^T \boldsymbol{\mu}\|}{\sqrt{\boldsymbol{x}^T \boldsymbol{x} \boldsymbol{\mu}^T \boldsymbol{\mu}}} \tag{5}$$

All unlabelled face and voice samples are projected into their LDA spaces respectively, and classified by C_{face} and C_{voice}. For each classifier and each class, the unlabelled samples closest to the the template are moved from the unlabelled dataset to the labelled dataset with the automatically determined label. We refer to this auxiliary training datas pseudo-labelled data. With the increased pool of labelled data, the two LDA subspaces are relearned, and the templates are recalculated. This process is iterative and is repeated until the unlabelled dataset is empty.

3.3. Identification and verification

Both identification and verification tasks can be accomplished using the LDA projections and client templates learned according to the above procedure.

In the identification scenario, facial and vocal feature vectors are extracted from each test video in the manner as described in Section 3.1, and each of them is first projected into their PCA subspaces, and then into their LDA subspaces respectively. In each space, the projected point is compared to each of the c templates according to the normalized correlation similarity measure as described above, thus resulting in two sets of c similarity scores $(S^1_{face}, S^2_{face}, \ldots, S^c_{face})$ and $(S^1_{voice}, S^2_{voice}, \ldots, S^c_{voice})$ Corresponding face and voice similarity scores are then averaged to obtain a fused score:

$$S^i_{fused} = \frac{S^i_{face} + S^i_{voice}}{2}, \tag{6}$$

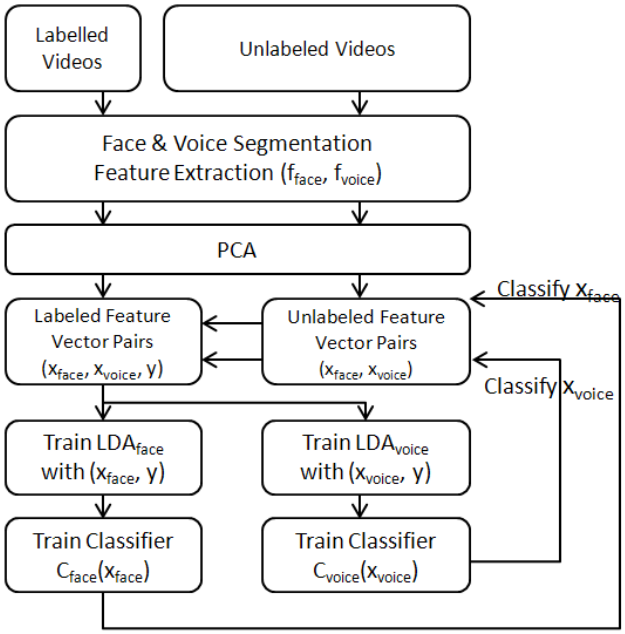


Fig. 3. Illustration of co-LDA subspace learning

and the test sample is assigned the label of the template whose similarity score is highest. The recognition performance is evaluated with the top 1 identification rate.

In the verification scenario, the face and voice feature vectors of a test data sample are extracted and projected into the same LDA space as before, but are compared only to the template corresponding to the claimed identity. Face and voice similarity scores are fused in the same way. The verification performance is evaluated with the Detection Error Trade-offs (DET) plot acquired with client and imposter scores.

4. EXPERIMENT AND RESULTS

4.1. Database

The experiments reported here aim to evaluate the capability of the co-LDA audio-visual person recognition algorithm to use inter-session variations contained in unlabelled data to enhance models which are weakly learned with limited labelled data. All experiments were conducted with the MOBIO database [15]. It contains videos of 150 subjects captured in real-world challenging conditions in 12 sessions. Recordings come from a mobile phone camera over a one-and-a-half-year period, and each session contains 11-21 videos. Fig.4 shows example images which demonstrate typical pose and illumination variability. Similar variability is also presented in the audio streams which contain different environmental noise. We selected 30 subjects with which to train a GMM world model for speaker recognition, another 30 subjects to conduct co-training experiments, and 15 subjects are selected as imposters in the verification experiment. For subspace learning, one session is randomly selected and used as labelled training data for enrolment, another session is randomly chosen as test data, and the other 10 sessions are

Fig. 4. Image examples of MOBIO database

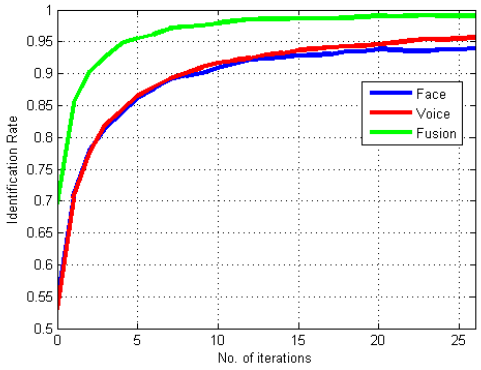
used as unlabelled data.

4.2. Experimental work

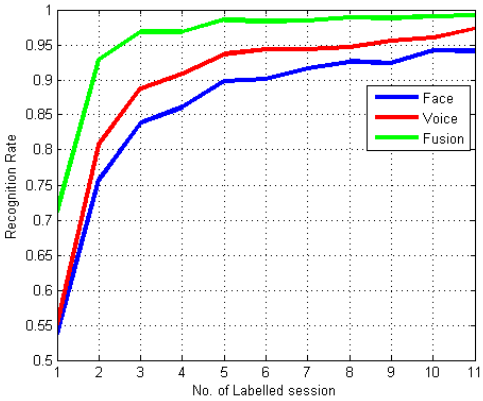
In each video, face images are detected automatically with an OpenCV based face detector. It incorporates eye and nose detection which help to crop detected faces according to facial landmark coordinates. Cropped face images are then resized to 144×128 pixels. For each video, the single most confidently detected face is selected. This face image is divided into 9×8 blocks and $LBP^{u2}_{(8,2)}$ features are extracted from each block and concatenated into a 4248-dimensional vector. MFCC acoustic features are extracted over 20ms Hamming windowed frames at a 10ms frame rate. Features are composed of 26 MFCC coefficients augmented with their 26 delta coefficients and the delta energy, resulting in acoustic vectors of 53 coefficients. Informative speech frames are extracted with an acoustic energy based speech detector described in [16] and non-speech frames are discarded. A 64-component speaker model is then adapted from the world model trained with an EM algorithm of the wold model subset. MAP adaptation is performed with a relevance factor of 14 and only means were adapted. The GMM means are concatenated to form a 3392-dimensional supervector. Each video is thus represented by an LBP feature vector and a GMM voice supervector.

Following co-training as described in Section 3, initial LDA projections and classifiers are learned on the labelled dataset, and iteratively updated with automatically labelled data. After the learning process, data is projected into the learnt LDA spaces and both identification and verification experiments were conducted. The identification rate reported is the average of 50-fold cross-validation. In verification experiment, following the protocol for LDA face verification described in [14], we used an imposter set containing 15 subjects which is independent from the training set used to learn the projections and models. Thus client scores are calculated by comparing the test data of the 30 clients to their true identity models, and imposter scores are calculated by comparing 15 imposters to 30 client models in an exhaustive way. we and the verification performance is reported in terms of Detection Error Trade-off (DET) curves which correspond to these client and imposter scores.

We first report results for the identification task. The identification rate attained by independent face and voice classifiers and their fusion is shown in Fig.5 (a). In all cases, performance is shown as function of number of iterations of co-training. 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: face identification rate increases from 53% to 96% while the voice identification rate increases from 55% to 94%; and the identification rate for the fused system increases from 70% to 99%. Among the automatically labelled data samples, 98.5% of them are correctly labelled.



(a) Identification rate as a function of co-training iterations



(b) Identification rate of as a function of labelled training sessions for baseline system

Fig. 5. Results for identification task

We may wonder with purely supervised learning method, how many sessions of labelled data we need in order to achieve the same performance. So we randomly select 1-11 sessions as labelled training data to train the LDA spaces and models, and another session as test data, each experiment is repeated 50 times and the average identification rate with respect to the different number of labelled training sessions is shown in Fig.5 (b). The result shows that, with supervised method, at least 10 labelled training sessions are needed to reach the performance of the proposed co-training method, which uses only 1 labelled session accompanied with 10 unlabelled sessions.

In a verification scheme, test data vectors are projected into the LDA subspaces learnt through co-training and are compared to all the client models. The DET curves for Face/Voice/Fusion verification systems before and after co-training are shown in Fig. 6. The performance for these systems without co-training is generally low due to the large inter-session variations which are not represented in the low quantity of training data (AFR and ASR verification rates are around 20%). Similar results were reported in [17]. However, after co-training, both single systems achieve below 5% EER while the fusion system achieves an EER of 1.4%. These results demonstrate the effectiveness of the proposed method.

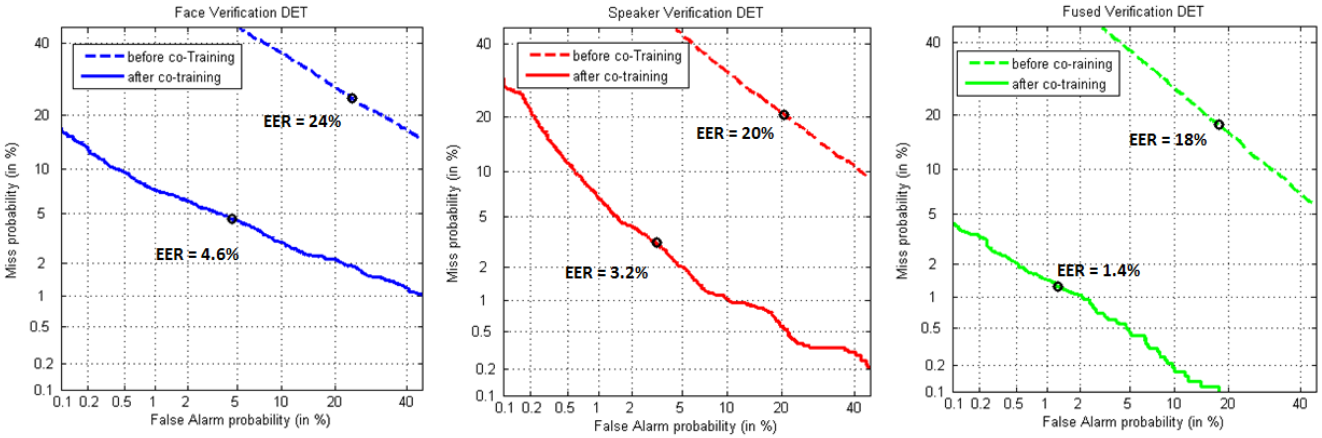


Fig. 6. DET curves for face (left) voice (middle) and fused (right) verification system

5. CONCLUSION

This paper proposes a new semi-supervised, linear dimensionality reduction algorithm, referred to as co-LDA, which allows two independent biometric systems to train each other using a large pool of automatically labelled auxiliary training data while equally applicable to any combination of biometric modalities. In this paper we demonstrate its utility in the scenario of audio-visual person recognition in videos. Automatic speaker and face recognition systems are shown to make efficient use of both labelled and unlabelled data, where unlabelled data are added iteratively to the labelled dataset and are used to improve the discriminative power of LDA. Experimental results on both identification and verification tasks show significant improvements in performance and demonstrate the effectiveness of our algorithm.

6. REFERENCES

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

[2] M. Bicogo, A. Lagorio, E. Grosso, Tistarelli, M."On the Use of SIFT Features for Face Authentication," in CVPR Workshop, 2006.

[3] P. Kenny, Boulianne, G. Boulianne, P. Ouellet, P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(4), pp. 1435- 447, 2007.

[4] X. Zhu, "Semi-supervised Learning Literature Survey, " *Technical report*, Univ. Wisconsin, Madison, USA, Jan. 2006.

[5] F. Wang, C. Zhang, H.C. Shen, and J. Wang, "Semi-Supervised Classification Using Linear Neighborhood Propagation, " in *CVPR 2006*.

[6] M. Yamada, M. Sugiyama, T. Matsui, "Semi-supervised speaker identification under covariate shift," *Signal Processing*, vol. 90(8), pp. 2353- 2361, 2010.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann, pp. 92-100, 1998

[8] X. Zhao, N. Evans, J.L., Dugelay, "A Co-training Approach to Automatic Face Recognition," in *EUSIPCO 2011*, Spain, Barcelona, 2011.

[9] F. Rol, L. Didaci, G.L. Marcialis, "Template Co-update in Multimodal Biometric Systems," in *Advances in Biometrics*, LNCS, Volume 4642/2007, pp.1194-1202, 2007.

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

[11] Z. Li, W. Jiang, H. Meng, "Fishervioce: A discriminant subspace framework for speaker recognition " IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010

[12] J. Kittler, M. Hatef, R.P.W.Duin, and J. Matas, "On combining classifiers," in *IEEE Transaction on PAMI* vol. 20, no. 3, pp. 226-239, 1998

[13] D. Reynolds, T. Quatieri, R. Dunn, "Speaker verification using adapted gaussian mixture models" *Digital Signal Processing*, vol. 10(1), (January 2000).

[14] J. Kittler, Y.P. Li, J. Matas, "On matching scores for LDA-based face verification " *British Machine Vision Conference*, 2000

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

[16] L. Besacier, J.-F. Bonastre, C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling" *Speech Communication*, 31(2-3), pp.89106, 2000.

[17] S. Marcel, etc., "Mobio Biometry (MOBIO) Face and Speaker Verification Evaluation" *IDIAP research report*, Idiap-RR-09-2010.