Introducing Crossmodal Biometrics: Person Identification from Distinct Audio & Visual Streams

Anindya Roy and Sébastien Marcel

Abstract— **Person identification using audio or visual biometrics is a well-studied problem in pattern recognition. In this scenario, both training and testing are done on the same modalities. However, there can be situations where this condition is not valid, i.e. training and testing has to be done on different modalities. This could arise, for example, in covert surveillance. Is there any person specific information common to both the audio and visual (video-only) modalities which could be exploited to identify a person in such a constrained situation? In this work, we investigate this question in a principled way and propose a framework which can perform this task consistently better than chance, suggesting that such crossmodal biometric information exists.**

I. Introduction

Conventional biometric systems use person-specific models tested on the same modalities on which they are trained. The modality can be audio [1], visual [2][3] or a fusion of audio and visual (bimodal) [4][5]. Such systems do not exploit person-specific information which might be embedded "crossmodally", i.e. in *both* the modalities.

Let us first define any such person-specific information which exists jointly in two modalities as a "crossmodal biometric" and any system able to exploit such information for the purpose of person identification as a "crossmodal biometric system". Essentially, it means that the training and test data are from distinct modalities.

A necessary criterion for a crossmodal biometric is that, like conventional biometrics, it should not vary with time. This means that its value should remain unchanged even when the audio and visual data of a person are recorded separately at completely non-overlapping times. We term this the Audio-Visual Mismatch criterion. This means, any correlation or mutual information based on audio-visual synchrony which could arise if the audio and visual data were extracted at the same time [6][7][8] cannot be treated as a crossmodal biometric. Additionally, it is preferable that such crossmodal information be robust to variations in the lexical content of speech.

The primary significance of such crossmodal biometric systems is in the context of surveillance [9]. Let us imagine a surveillance system which has collected speech data uttered by a set of persons. In this phase (the 'training phase') no visual information was available perhaps because the speech was recorded from telephone conversations. Presently, the system is in the 'test phase', i.e. it is observing a person talking whom it should identify as one out of the set in the training phase (closed-set identification). Identity information of this person might be useful in planning how to interact with this person or whether to interact at all. However, due to either the distance of the system from the person (a common occurence in covert surveillance), or due to a noisy acoustic environment, only the visual data (dynamic facial appearance) is available. In such a scenario, a crossmodal biometric system could provide important information until a conventional biometric system could be employed. A similar scenario could be imagined by interchanging the audio and visual modalities, where prior visual data of a person has been collected and presently the person should be identified using audio data alone.

Before approaching the problem from a purely pattern recognition perspective, it is worthwhile to note that we, as humans, often perform this task. We often create a mental image of a person whose voice is familiar (from telephone conversations, for example) but whom we have never seen. We also often create a mental "voice model" from visual information (either static or dynamic) of persons we have never heard.

Recent studies have investigated these phenomena from the viewpoint of human perception and psychophysics [10][11][12][13][14]. In these studies, human observers were asked to match an audio recording of an unknown voice X to two video (visual-only) recordings of two unknown speakers, A and B, one of which is X, and vice versa, under a variety of experimental conditions, a task termed as the XAB task [10]. Lachs et al. [11], Rosenblum et al.[14] and Kamachi et al. [10] reported human observers correctly matching X to A or B around 65% of the times compared to the chance value of 50%. This was shown to be statistically significant given the number of independent test cases considered. Krauss et al. have shown similar matching performance using static instead of dynamic visual information [13]. These studies suggest that crossmodal biometric information exists.

In a previous work, we presented a preliminary system able to perform such XAB matching tasks better than chance [15]. In this work, we extended this system to a proper crossmodal biometric system, going from a 2:1 matching problem to an N:1 identification problem. In addition, we also provide here a more thorough and detailed description

This work was supported by the Swiss National Science Foundation, projects MultiModal Interaction and MultiMedia Data Mining (MULTI, 200020-122062) and Interactive Multimodal Information Management (IM2, 51NF40-111401) and the FP7 European MOBIO project (IST-214324).

A. Roy is with Idiap Research Institute, Martigny, and École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. Anindya.Roy@idiap.ch

S. Marcel is with Idiap Research Institute, Martigny, Switzerland Sebastien.Marcel@idiap.ch

of the underlying concept and algorithm. We report identification experiments on a standard multimodal database involving a large number of tests under several experimental conditions.

The rest of the paper is organized as follows. In Sec.II, we give a general overview of the proposed approach, which we describe in greater detail in Sec.III. We describe our crossmodal person identification experiments in Sec.IV. In Sec.V, we discuss the results of our experiments and highlight certain aspects of our method. Finally, Sec.VI outlines the main conclusions of our work.

II. General overview

The main challenge of crossmodal biometrics is that the datasets used to train and test the person-specific models are from different modalities. Our approach is to use a suitable mapping framework to transform the person-specific information present in the testing modality to the training modality and then match this transformed information against the models.

The parameters of this mapping framework is to be learnt from a synchronized audio-visual dataset which we denote as the *learning dataset*, $\mathbf{D}_\mathrm{L}$. The learning dataset comprises of an audio part, $\mathbf{D}_\mathrm{L}^a$ and a visual part, $\mathbf{D}_\mathrm{L}^v$. These two parts are ordered such that the i-th element $\mathbf{x}_i^a \in \mathbf{D}_\mathrm{L}^a$ is synchronous to the i-th element, $\mathbf{x}_i^v \in \mathbf{D}_\mathrm{L}^v$. We term the data to be used in the training phase as the *train dataset* and the data to be used in the test phase as the *test dataset*. It is to be noted that persons in the learning dataset are all distinct from those in the train and test datasets (to preserve the Audio-Visual Mismatch criterion, ref. Sec.I).

The crossmodal mapping can be carried out at two distinct levels: 1) feature level and 2) model level.

In feature level mapping, the feature vectors from one modality are directly transformed to feature vectors in the other modality using a mapping function, exploiting the correlation which exists between them [16]. However, unlike other applications [6] [7] [8][17], feature-level mapping has not performed well in our task mainly due to the Audio-Visual Mismatch criterion: the learnt mapping parameters are highly person-specific and cannot generalize from the learning dataset to the train and test datasets. Even with nonlinear mapping techniques like Support Vector Regression [18], no improvement was obtained.

In the second approach, instead of trying to directly map features from one modality to another, a statistical model of the features in one modality is mapped to a statistical model of the features in the other modality using a model-mapping framework whose parameters are learned from the learning dataset. More precisely, a feature point in one modality is mapped to a "probability density cloud" in the other modality instead of a precise point. Such clouds are then summed up to form the equivalent model in the other modality.

Thus, in this approach, both train and test data are first used to generate models in their respective modalities, termed the *train model* and *test model* respectively. To identify a person, the test model in the testing modality is first transformed to its equivalent model in the training modality using the model-mapping framework. This transformed test model is matched with all the available train models, using a suitable model similarity measure and the person is identified as the one whose train model shows maximum similarity with the test model. Unlike feature-level mapping, this technique has proved to be much more robust to generalization and has achieved significantly better results in the task of crossmodal identification. We discuss this approach in subsequent sections.

III. Crossmodal person identification system

In this work, we used non-parametric density estimation [18] involving smoothed probability mass functions (PMF) for creating the speaker-specific models and Hebbian projection matrices [19] for the model-mapping framework. Various model similarity measures were explored to match the models. We discuss each of these concepts in more detail as follows.

A. Modelling the data

Let $\mathbf{R}^a$ denote the feature space corresponding to the audio modality. Given a finite set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbf{R}^a$ of feature points extracted from the audio data of a certain person, the aim is to estimate the probability density function (PDF) which generated these points [18]. In this work, techniques like Gaussian Mixture Models (GMM) [1][2] conventionally used to model audio and visual data are not suitable because it is difficult to map such models between modalities. Instead, we chose to represent the PDF non-parametrically as a piecewise linear approximation, i.e., a probability mass function (PMF) [18].

Let $\mathbf{M} = \{\mu_k\}_{k=1}^K$ be a set of representative points in $\mathbf{R}^a$. In practice, these points $\{\mu_k\}_{k=1}^K$ are chosen by K-Means clustering of the learning dataset $\mathbf{D}_\mathrm{L}^a \in \mathbf{R}^a$ for the audio modality. This ensures that they are probabilistically evenly distributed in the space, already following a "background" distribution of points for that modality (ref. background models, [1]). These points $\mathbf{M}$ decompose the space $\mathbf{R}^a$ into a set of K disjoint regions, $\{\mathbf{R}_k^a\}_{k=1}^K$ which is termed a Voronoi tessellation of $\mathbf{R}^a$ [18]. All points in a particular region $\mathbf{R}_k^a$ are nearer to μ_k according to a certain metric d, among all points in $\mathbf{M}$. This metric d can be the Euclidean distance, the Mahalanobis distance [18] or the Cosine distance with little effect on the final results although Euclidean distance performs slightly better. Given this decomposition of $\mathbf{R}^a$, the PMF of $\mathbf{X}$, $\mathbf{p}_\mathbf{X}^a$ can be estimated as $\mathbf{p}_\mathbf{X}^a = [\mathbf{p}(1)\ \mathbf{p}(2)\ \cdots\ \mathbf{p}(K)]^T$, where

$$\mathbf{p}(k) = \Pr(\mathbf{x} \in \mathbf{R}_k^a | \mathbf{x} \in \mathbf{X}) \quad (1)$$

$$= \frac{1}{|\mathbf{X}|} \sum_{\forall \mathbf{x} \in \mathbf{X}} \mathbf{1}_{\{\mathbf{x} \in \mathbf{R}_k^a\}} \quad (2)$$

where $1 \leq k \leq K$ and $|\cdot|$ denotes size of a countable set. This PMF is the model of $\mathbf{X}$ and hence the model of the given person. In practice, better results are obtained with higher values of K [18], and in fact, best results are obtained when K approaches $|\mathbf{D}_\mathrm{L}^a|$, the size of the learning set. However, typically $|\mathbf{D}_\mathrm{L}^a| \gg N$ where $N = |\mathbf{X}|$. This implies that $K \gg N$,

i.e. the number of regions far outnumbers N, the number of data points in $\mathbf{X}$. Hence, the PMF estimation is extremely sparse which is not desirable. This problem is solved by smoothing the PMF. The simplest way is to replace the probability estimate $\mathbf{p}(k)$ for a particular region $\mathbf{R}_k^a$ by the average of the probabilities of its κ nearest regions, where κ is the smoothing parameter.[1] The κ nearest regions are those whose centroids $\{\mu_{k'}\}$ are the κ-nearest neighbours of μ_k in $\mathbf{M}$. A similar procedure can be followed to generate models in the visual modality. Let us denote such a PMF in the visual feature space as $\mathbf{p}^v$.

B. Crossmodal Mapping of Models

Let $\mathbf{R}^a, \mathbf{R}^v$ denote the feature spaces corresponding to the audio and visual modalities respectively. Model mapping between the two modalities is achieved via Hebbian projection matrices [19], which are conditional probability matrices. Let $\mathbf{H}^{av}$ denote the $K \times K$ Hebbian projection matrix from the audio to the visual modality, each of whose elements $\mathbf{H}^{av}(k_a, k_v)$ estimates the conditional probability that a point $\mathbf{x}^v$ belongs to a particular region $\mathbf{R}_{k_v}^v$ in the feature space $\mathbf{R}^v$ of the visual modality, given that its corresponding point $\mathbf{x}^a$ in the audio modality belongs to the region $\mathbf{R}_{k_a}^a$ in feature space $\mathbf{R}^a$, i.e. $\mathbf{H}^{av}(k_a, k_v) = \Pr(\mathbf{x}^v \in \mathbf{R}_{k_v}^v | \mathbf{x}^a \in \mathbf{R}_{k_a}^a)$. Since the points in the learning datasets $\mathbf{D}_{\mathrm{L}}^a, \mathbf{D}_{\mathrm{L}}^v$ have a one-to-one correspondence based on synchrony (ref. Sec.II), the matrix $\mathbf{H}^{av}$ can be estimated using data from the learning dataset as follows,

$$\mathbf{H}^{av}(k_a, k_v) = \frac{A}{B} \tag{3}$$

where

$$A = \sum_{\forall \mathbf{x}_i^a \in \mathbf{D}_{\mathrm{L}}^a} \mathbf{1}_{\{\mathbf{x}_i^v \in \mathbf{R}_{k_v}^v\}} \cdot \mathbf{1}_{\{\mathbf{x}_i^a \in \mathbf{R}_{k_a}^a\}} \tag{4}$$

$$B = \sum_{\forall \mathbf{x}_i^a \in \mathbf{D}_{\mathrm{L}}^a} \mathbf{1}_{\{\mathbf{x}_i^a \in \mathbf{R}_{k_a}^a\}} \tag{5}$$

$1 \leq k_a, k_v \leq K$ and $\mathbf{x}_i^a \in \mathbf{D}_{\mathrm{L}}^a$ is the audio vector synchronous with visual vector $\mathbf{x}_i^v \in \mathbf{D}_{\mathrm{L}}^v$. The inverse Hebbian projection matrix $\mathbf{H}^{va}$ from the visual to the audio modality can be calculated similarly by interchanging the audio and visual modalities in the above equation.

Given a PMF $\mathbf{p}_{\mathbf{X}}^a$ generated from a set of points $\mathbf{X}$ in the audio feature space as in Sec.III-A, we can use the matrix $\mathbf{H}^{av}$ to "project" $\mathbf{p}_{\mathbf{X}}^a$ on the visual feature space,

$$\tilde{\mathbf{p}}_{\mathbf{X}}^v = \mathbf{H}^{av} \mathbf{p}_{\mathbf{X}}^a \tag{6}$$

where $\tilde{\mathbf{p}}_{\mathbf{X}}^v$ is an estimate of the true PMF $\mathbf{p}_{\mathbf{X}}^v$ of the set of visual feature points corresponding to the audio feature points in $\mathbf{X}$. It is to be noted that these visual feature points are actually not available, hence we use $\mathbf{H}^{av}$ to indirectly estimate $\mathbf{p}_{\mathbf{X}}^v$. A PMF $\mathbf{p}_{\mathbf{X}}^v$ in the visual feature space can be similarly "projected" on the audio feature space using $\mathbf{H}^{va}$,

$$\tilde{\mathbf{p}}_{\mathbf{X}}^a = \mathbf{H}^{va} \mathbf{p}_{\mathbf{X}}^v \tag{7}$$

It is to be noted that when K is high, the Hebbian projection matrices become sparse and need to be smoothed out in a similar way as for the PMFs in Sec.III-A. Even then, the estimation becomes gradually poorer since the number of quantities to be estimated (K^2) increases rapidly compared to the size of available data. However, in the extreme case when K approaches $|\mathbf{D}_{\mathrm{L}}^a|$, the size of the learning dataset, each point of the learning dataset becomes a centroid in the set $\mathbf{M} = \{\mu_k\}_{k=1}^K$ (ref. Sec.III-A). Then, due to the one-to-one correspondence between $\mathbf{D}_{\mathrm{L}}^a$ and $\mathbf{D}_{\mathrm{L}}^v$, $\mathbf{H}^{av}, \mathbf{H}^{va}$ can in fact be approximated reliably by the identity matrix $\mathbf{I}^K$ of size $K \times K$. This is an important advantage.

C. Model similarity measures - Person Identification

Let the training modality be audio and testing modality be visual. We term this as the (a-v) case. In the training phase, we generate a set of models $\{\mathbf{p}_\omega^a\}_{\omega=1}^{N_\Omega}$ in the audio feature space from the feature points in the train dataset (ref. Sec.III-A), each ω representing a different person. In the test phase, a model $\mathbf{p}_{\mathbf{X}}^v$ is generated in the visual feature space from feature points in the test dataset, extracted from visual data of an unknown person X. It is mapped to the audio feature space using eqn.7 to give the estimated PMF $\tilde{\mathbf{p}}_{\mathbf{X}}^a$ in the audio modality. This estimated PMF is matched with the set of PMFs $\{\mathbf{p}_\omega^a\}_{\omega=1}^{N_\Omega}$ using a suitable similarity measure Ψ and the person $\mathbf{X}$ is identified as the one whose model $\mathbf{p}_{\omega^*}^a$ has the highest similarity with the test model.

$$\omega^* = \arg\max_\omega \Psi(\mathbf{p}_\omega^a \tilde{\mathbf{p}}_{\mathbf{X}}^a) \tag{8}$$

The similarity measure Ψ can be the Bhattacharyya coefficient [18],

$$\Psi_B(\mathbf{p}_\omega^a, \tilde{\mathbf{p}}_{\mathbf{X}}^a) = \sum_{\forall k} \mathbf{p}_\omega^a(k)^{\frac{1}{2}} \tilde{\mathbf{p}}_{\mathbf{X}}^a(k)^{\frac{1}{2}} \tag{9}$$

the $\mathbf{L}^2$ inner product [20],

$$\Psi_{\mathbf{L}^2}(\mathbf{p}_\omega^a, \tilde{\mathbf{p}}_{\mathbf{X}}^a) = \sum_{\forall k} \mathbf{p}_\omega^a(k) \tilde{\mathbf{p}}_{\mathbf{X}}^a(k) \tag{10}$$

or a simplified form of the Kullback-Leibler Divergence [18],

$$\Psi_{KL}(\mathbf{p}_\omega^a, \tilde{\mathbf{p}}_{\mathbf{X}}^a) = \sum_{\forall k} \log(\mathbf{p}_\omega^a(k)) \tilde{\mathbf{p}}_{\mathbf{X}}^a(k) \tag{11}$$

In practice, all three gave comparable results, with $\Psi_{\mathbf{L}^2}$ performing slightly better than the others. For the reverse (v-a) case, where the training modality is visual while the testing modality is audio, a similar procedure was followed with the roles of the modalities interchanged.

IV. Experiments

A. Database and features

All experiments were performed on the standard M2VTS audio-visual database [21][5]. The database contains 10 female and 24 male subjects. For each subject, synchronized audio and visual data was recorded in a controlled environment across four sessions separated by one week intervals. In each session, the subjects counted from '0' to '9' in their native language. Lip annotations were obtained from `http://www.ee.surrey.ac.uk/Projects/M2VTS/experiments/lip_tracking/`. In this work, only the 24 male subjects were considered since the

[1] This is comparable to a Parzen window approach [18].

number of female subjects was too few to yield statistically significant results.

For the visual modality, we concentrated on lip appearance features since they have been shown to be efficient and robust to small errors in lip localization [17]. The video frame rate was 25fps. From each video frame, a 16×16 Region-Of-Interest (ROI) around the lips was extracted using available annotation, followed by geometric normalization and inter-frame alignment. Next, 2D-DCT features [17] were extracted and 3$^{\text{rd}}$ to 10$^{\text{th}}$ highest energy coefficients were retained to form the visual feature vectors. Mean normalization was performed for each video sequence [17]. For the audio modality, the audio data sampled at 8kHz was blocked into frames equal in duration to the video frames (corresponding to 320 samples per frame) and 16 Mel-Frequency Cepstral Coefficients (MFCC) [17][1] were extracted from each block, out of which 1$^{\text{st}}$ to 8$^{\text{th}}$ were retained to form the audio feature vectors.[2] For each audio sequence, Cepstral Mean Subtraction [17] was performed. It is to be noted that only voiced frames were used, both for audio and visual modalities.

B. Protocol

A complete experiment comprised of several independent runs. In each run, a certain fixed number of N_L subjects were chosen at random from the 24 to form the learning dataset, while the remaining $N_\Omega = 24 - N_L$ subjects formed the *match dataset*. All crossmodal mapping parameters were learnt using the learning dataset while all person identification tests were carried out on the match dataset. The match dataset itself was broken into train and test datasets (ref. Sec.II).

Hebbian projection matrices $\mathbf{H}^{av}, \mathbf{H}^{va}$ were estimated from the learning dataset (ref. Sec.III-B). Next, a 4-fold cross-validation was performed on the match dataset as follows: in each fold, a particular session out of the 4 was chosen to form the test dataset, while the other 3 formed the train dataset for each subject. Models were trained from both train and test datasets (Sec.III-A) and mapped to the same modality as required. Approximately 5 seconds of speech was used to create the test models, and about 15 seconds for the train models per subject. Finally, closed-set person identification was performed on all the N_Ω test recordings in the test dataset (one from each subject) against all the N_Ω models in the train dataset (ref. Sec.III-C). For *each* run, a total of $4 \times N_\Omega$ identification tests were performed. The correct identification rate for a run is measured as,

$$\tilde{P}^c = \frac{\text{No. of correct identifications}}{\text{Total no. of tests}} \times 100\% \qquad (12)$$

For each experiment, we report the mean identification rate P^c by averaging the correct identification rate $\tilde{P}^c$ for each run across 200 runs. A higher P^c indicates better performance. The total number of tests in an experiment is $N_T = 200 \times 4 \times N_\Omega$.

[2]For both the audio and visual modalities, the coefficients have been selected by trial-and-error to give best performance.

We performed separate experiments for different choices of N_Ω to analyse the effect of varying the number of subjects in the match dataset and learning dataset. Experiments were also repeated by varying the value of the smoothing parameter κ from a few tens to a few hundreds and the number of regions K from a few tens to $|\mathbf{D}_{\text{L}}^a|$, the size of the learning dataset, which was approximately 5000.

Two types of experimental conditions were investigated, (1) lexically matched condition and (2) lexically mismatched condition. For condition (1), speech content in test and train datasets were lexically matched: recordings from the database were used unchanged. For the second (more difficult) condition, the recordings were rearranged so that segments used for training were lexically mismatched with segments used for testing : if training data contained '0' to '4', testing data contained '5' to '9' and vice-versa.

Experiments were performed for both the (a-v) and (v-a) cases (ref. Sec.III-C). In all experiments, the Audio-Visual Mismatch criterion was strictly imposed. For reference, person identification experiments were also performed with both the train and test data from the *same* modality keeping the rest of the framework unchanged, i.e. both from audio modality or both from visual modality (conventional biometrics). We term these as the (a-a) and (v-v) cases respectively.

C. Results

We summarize the primary results of our experiments in Tables I and II, showing lexically matched and mismatched conditions respectively. Person identification performance is reported in terms of the mean identification rate P^c for 4 different choices of the number of subjects in the match set, $N_\Omega = \{4, 8, 12, 16\}$. In each case, the number of subjects in the corresponding learning dataset is $N_L = 24 - N_\Omega$. The performance of the proposed framework is given by the mean identification rates P^c_{a-v} and P^c_{v-a} for the (a-v) and (v-a) cases respectively. For comparison, the mean identification rate P^c_0 of a system based purely on random chance is also shown. It is calculated as $P^c_0 = \frac{1}{N_\Omega}$ since each test is a one-to-N_Ω matching problem. Additionally, mean identification rates of conventional biometric systems (a-a) and (v-v) are reported as P^c_{a-a} and P^c_{v-v} respectively (ref. Sec.IV-B).

Although a wide range of values of κ and K were investigated, the performance did not vary considerably and we report identification rates corresponding only to their optimal values. It is to be noted that the optimal value for κ remained within 300, while optimal performance was obtained when $K \rightarrow |\mathbf{D}_{\text{L}}^a|$, the size of the learning dataset.

Although the primary notion of performance in a person identification task is given by the mean identification rate P^c, we also consider the case where it is required that the correct identity may not be the very first but should be at least within the first R^* identities selected according to decreasing similarity value Ψ (ref. Sec.III-C). We consider this relaxed scenario because this is intrinsically a difficult task. We report this performance in figs. 1 to 4 in terms of $P(R \leq R^*)$, i.e. the probability that the rank R of the correct

	N_Ω	N_L	Crossmodal biometric systems		Random chance	Conventional biometric systems	
			P^c_{a-v}	P^c_{v-a}	P^c_0	P^c_{a-a}	P^c_{v-v}
1.	4	20	43.0	44.1	25.0	99.8	96.8
2.	8	16	26.7	26.2	12.5	99.3	92.4
3.	12	12	17.5	17.8	8.3	98.4	87.8
4.	16	8	11.7	12.0	6.3	97.3	81.6

TABLE I

MEAN IDENTIFICATION RATES P^c (%) UNDER LEXICALLY MATCHED CONDITION. ROWS REPRESENT DIFFERENT CHOICES OF N_Ω AND N_L.

	N_Ω	N_L	Crossmodal biometric systems		Random chance	Conventional biometric systems	
			P^c_{a-v}	P^c_{v-a}	P^c_0	P^c_{a-a}	P^c_{v-v}
1.	4	20	30.0	31.1	25.0	69.9	80.3
2.	8	16	16.0	17.8	12.5	54.1	68.4
3.	12	12	11.0	12.4	8.3	44.4	58.0
4.	16	8	8.1	9.0	6.3	36.8	47.1

TABLE II

MEAN IDENTIFICATION RATES P^c (%) UNDER LEXICALLY MISMATCHED CONDITION. ROWS REPRESENT DIFFERENT CHOICES OF N_Ω AND N_L.

identity is less than or equal to R^*. R^* varies from 1 to N_Ω. A higher value of $P(R \leq R^*)$ indicates better performance. As before, we consider four choices, $N_\Omega = \{4, 8, 12, 16\}$. It is to be noted that the value of $P(R \leq R^*)$ at $R^* = 1$ is equal to P^c.

V. DISCUSSIONS

It is evident from tables I and II that our proposed crossmodal biometric systems ((a-v) and (v-a)) are able to perform consistently better than random chance in all the cases, although performance is degraded in the lexically mismatched condition (similar to the conventional systems). Given the high number of tests in each experiment (ranging from 3200 for $N_\Omega = 4$ to 12800 for $N_\Omega = 16$), it seems unlikely that our proposed system consistently performed better purely by chance. This suggests that crossmodal biometric information exists and our system is able to exploit this information.

As expected, our proposed crossmodal biometric systems are outperformed by the conventional biometric systems ((a-a) and (v-v)). This is not surprising since the crossmodal task is much more difficult than a direct matching within the same modality.

It is observed from tables I and II that although the absolute value of the mean identification rate P^c reduces as the number of subjects in the match set N_Ω increases, the ratio P^c/P^c_0 indicating the relative improvement of the proposed system compared to a purely random system remains fairly stable irrespective of the fall in the amount of data in the learning dataset (more than a 50% decrease from $N_\Omega = 4$ to $N_\Omega = 16$).

Figures 1 to 4 show that the proposed systems are consistently better than random chance in terms of the probability $P(R \leq R^*)$ across different values of R^*. Furthermore, it can

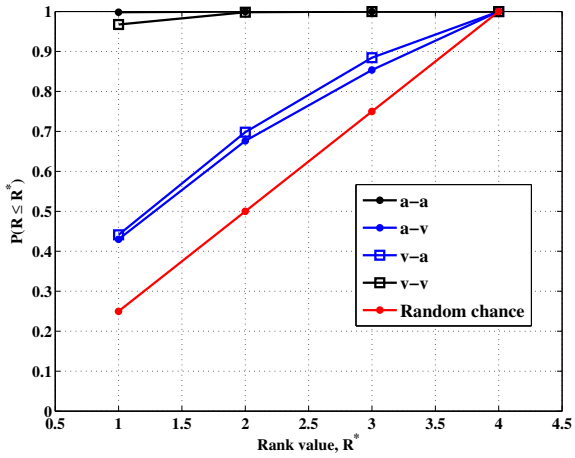

Fig. 1. Expected probability that the rank of the correct identity R is lower than or equal to R^* at different values of R^*, for the case $N_\Omega = 4$, $N_L = 20$ under lexically matched condition.

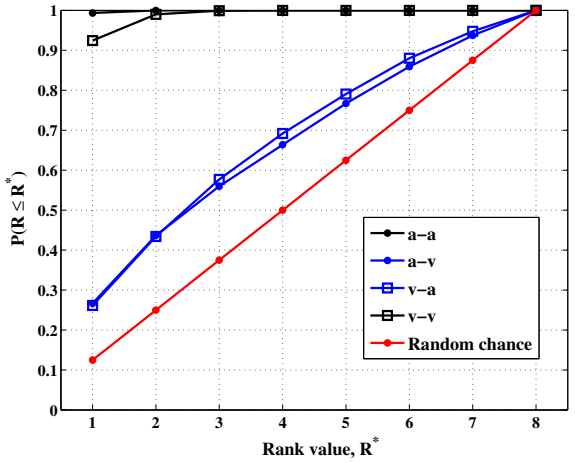

Fig. 2. Expected probability that the rank of the correct identity R is lower than or equal to R^* at different values of R^*, for the case $N_\Omega = 8$, $N_L = 16$ under lexically matched condition.

be shown that the value of $P(R \leq R^*)$ at $R^* = N_\Omega/2$ in figs. 1 to 4 provides a rough estimate of the performance of such a framework in an XAB matching task (ref. Sec.I and [10][14]). It is interesting to note that the value is close to 0.65 which is correlated by the values obtained by human observers as reported in these studies although these are not directly comparable since the databases used were different.

VI. CONCLUSION AND FUTURE WORKS

A. Conclusions

In this work, we formally introduced the concept of crossmodal biometrics, person-specific information which is embedded "crossmodally" across two modalities. We investigated approaches to extract and exploit such information for a crossmodal person identification task. In particular, we considered audio-visual crossmodal biometrics, i.e. matching test data from the visual modality with training data from

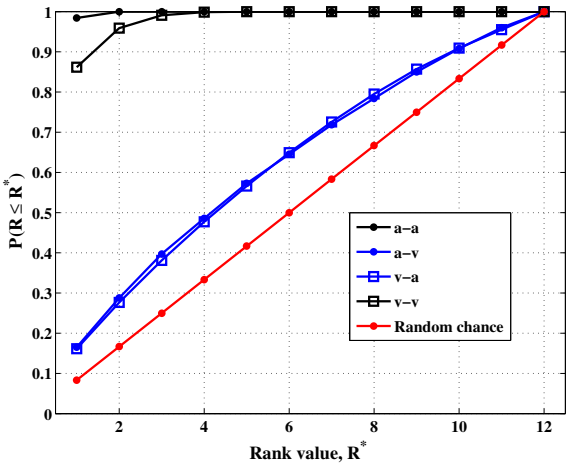

Fig. 3. Expected probability that the rank of the correct identity R is lower than or equal to R^* at different values of R^*, for the case $N_\Omega = 12$, $N_L = 12$ under lexically matched condition.

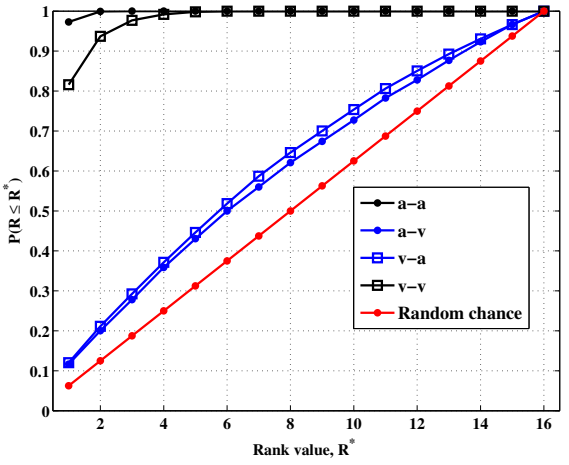

Fig. 4. Expected probability that the rank of the correct identity R is lower than or equal to R^* at different values of R^*, for the case $N_\Omega = 16$, $N_L = 8$ under lexically matched condition.

the audio modality, and vice-versa. We proposed a framework to perform this task and evaluated its performance using a standard audio-visual database under a variety of test cases and experimental conditions. Results from these experiments seem to suggest that such crossmodal person-specific information exists, and it is possible to exploit it for person identification when conventional biometric systems are not practical.

B. Future works

Although our framework performed consistently better than chance, identification rates are still too low. In future, we will address this issue and aim to improve our framework so that it is suitable for practical deployment in a real scenario. One possibility is to take into account dynamic audio and visual information in addition to the static feature vectors considered in this work. More sophisticated features and modelling techniques optimized for the task will also be considered. Furthermore, we aim to use a larger database with the possibility of a richer and more varied learning dataset, which could have a positive impact on the performance of the system.

VII. ACKNOWLEDGMENTS

The authors would like to thank Mathew M.-Doss and Francesco Orabona for their helpful comments and advice.

REFERENCES

[1] F. Bimbot et al., "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing*, no. 4, pp. 431–451, 2004.

[2] F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM classifiers for face verification on XM2VTS," in *Proc. International Conference on Audio- and Video-based Biometric Person Authentication*, 2003, pp. 1058–1059.

[3] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. 855–861.

[4] A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*. Springer Verlag, 2006.

[5] S. Bengio, "Multimodal Authentication using Asynchronous HMMs," in *Proc. of 4th Intl. Conf. on Audio- and Video- based Biometric Person Authentication*, vol. 4. Springer, 2003.

[6] K. Kumar and et al., "Audio-Visual Speech Synchronization Detection Using a Bimodal Linear Prediction Model," in *CVPR*, 2009.

[7] G. Chetty and M. Wagner, "Audio visual speaker verification based on hybrid fusion of cross modal features," vol. 4815/2007, pp. 469–478, 2007.

[8] D. Li, C. Taskiran, N. Dimitrova, W. Wang, M. Li, and I. Sethi, "Cross-modal analysis of audio-visual programs for speaker detection," in *MMSP*, 2005.

[9] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviours," 2004.

[10] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'Putting the Face to the Voice': Matching Identity across Modality," *Current Biology*, vol. 13, pp. 1709–1714, 2003.

[11] L. Lachs and P. Pisoni, "Crossmodal source identification in speech perception," *Ecological Psychology*, vol. 16, no. 3, pp. 159–187, 2004.

[12] S. Campanella and P. Bellin, "Integrating face and voice in person perception," *Trends in Cognitive Science*, vol. 11, no. 12, 2007.

[13] R. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, pp. 618–625, 2002.

[14] L. Rosenblum, N. Smith, S. Nichols, S. Hale, and J. Lee, "Hearing a face: Cross-modal speaker matching using isolated visible speech," vol. 68(1), pp. 84–93, 2006.

[15] A. Roy and S. Marcel, "Crossmodal matching of speakers using lip and voice features in temporally non-overlapping audio and video streams," in *20th International Conference on Pattern Recognition*, 2010.

[16] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. Asif, "The natural statistics of audiovisual speech," 2009.

[17] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. MIT Press, 2004.

[18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. John Wiley and Sons, 2000.

[19] M. Coen, "Multimodal Dynamics : Self-Supervised Learning in Perceptual and Motor Systems," Massachusetts Institute of Technology, PhD Thesis, 2006.

[20] W. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, 2006.

[21] "M2VTS Multimodal Face Database, Release 1.00," http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html.