MULTIMODAL PERSON IDENTIFICATION IN A SMART ROOM

J.Luque, R.Morros, J.Anguita, M.Farrus, D.Macho, F.Marqués, C.Martínez, V.Vilaplana, J. Hernando

Technical University of Catalonia (UPC)
Jordi Girona, 1-3 D5, 08034 Barcelona, Spain
aluque@gps.tsc.upc.edu

ABSTRACT

In this paper we present a person identification system based on a combination of acoustic features and 2D face images. We address the modality integration issue on the example of a smart room environment. In order to improve the results of the individual modalities, the audio and video classifiers are integrated after a set of normalization and fusion techniques. First we introduce the monomodal acoustic and video identification approaches and then we present the use of combined input speech and face images for person identification. The various sensory modalities, speech and faces, are processed both individually and jointly. The result obtained in the CLEAR'06 Evaluation Campaign shows that the performance of the multimodal approach results in improved performance in the identification of the participants.

1. INTRODUCTION

Person identification consists in determining the identity of a person from a data segment, such as a speech, video segment, etc. Currently, there is a high interest in developing person identification applications in the framework of smart room environments. In a smart room, the typical situation is to have one or more cameras and several microphones. Perceptually aware interfaces can gather relevant information to recognize, model and interpret human activity, behaviour and actions. Such applications face an assortment of problems such a mismatched training and testing conditions or the limited amount of training data. In this work we present the audio, video and multimodal person identification techniques and the obtained results in the CLEAR'06 Evaluation Campaign inside the CHIL (Computers in the human interaction loop) project [1]. The CLEAR'06 Person Identification evaluation is a closed-set task, that is, all the possible speakers are known. Matched training and testing conditions and far-field data acquisition are assumed, as well as no a priori knowledge about room environment.

For acoustic speaker identification, the speech signals are parameterized using the Frequency Filtering (FF) [2] over the filter-bank energies, which is both computationally efficient and robust against noise. Next, in order to model the probability distribution of the parameters generated by each speaker, Gaussian Mixture Models (GMM) [3] with diagonal covariance are used.

In the case of visual identification, an appearance-based technique is used due to the low quality of the images. Face images of the same individual are gathered into groups. Frontal images within a group are jointly compared to the models for identification. These models are composed of several images representative of the individual. The joint recognition enhances the performance of a face recognition algorithm applied on single images. Individual decisions are based on a PCA [4] approach given that the variability of the users' appearance is assumed to be low and so are the lighting variations.

Multimodal recognition involves the combination of two or more human traits like voice, face, fingerprints, iris, hand geometry, etc. to achieve better performance than using monomodal recognition [5], [6]. In this work, a multimodal score fusion technique, Matcher Weighting with equalized scores, has been used. This technique has obtained an improvement for the correct identification rate on the closed-set $15/30$ seconds training and $1/5/10/20$ seconds testing conditions on the CLEAR'06 Evaluation task.

This paper is organized as follows: In sections 2 and 3 an overview of the audio and video algorithms and techniques is given. Section 4 presents the technique for multimodal fusion. Section 5 describes the evaluation scenario and the experimental results. Finally, section 6 is devoted to provide conclusions.

2. ACOUSTIC PERSON IDENTIFICATION

The speaker identification (SI) task consists in determining the identity of the speaker of a speech segment. In this task it is usually assumed that all the possible speakers are known. For this evaluation, recordings from 26 speakers using one microphone of an array have been provided. The first stage of current speaker recognition systems is a segmentation of the speech signal into regular segments. The speech signal is divided into frames of 30 ms at a rate of 10 ms. From each segment a vector of parameters

that characterizes the segment is calculated. In this work we have used the Frequency Filtering (FF) parameters [2]. These parameters are calculated as the widely used Mel-Frequency Cepstral Coefficients (MFFC) [7] but replacing the final Discrete Cosine Transform of the logarithmic filter-bank energies of the speech frame with the following filter:

$$H(z) = z - z^{-1} \tag{1}$$

These features have several interesting characteristics: they are uncorrelated, computationally simpler than MFCCs, have frequency meaning and they have generally shown an equal or better performance than MFCCs in both speech and speaker recognition. In order to capture the temporal evolution of the parameters the first and second time derivatives of the features are generally appended to the $\mathbf{o}_{FF}$, basic static feature vector. The so called delta coefficients [8] are computed using the following regression formula,

$$\Delta\mathbf{o}_t(i) = \frac{\sum_{\theta=1}^{\Theta} \theta\big(\mathbf{o}_{t+\theta}(i) - \mathbf{o}_{t-\theta}(i)\big)}{2\sum_{\theta=1}^{\Theta}\theta^2} \tag{2}$$

where $\Delta\mathbf{o}_t(i)$ is the delta coefficient i at time t computed in terms of the corresponding static coefficients $\mathbf{o}_{t-\Theta}$ to $\mathbf{o}_{t+\Theta}$. The same formula is applied to the delta coefficients with another window size to obtain acceleration coefficients. For each speaker that the system has to recognize, a model of the probability density function of the parameter vectors is estimated. These models are known as Gaussian Mixture Models (GMM) [3], which is a weighted sum of Gaussian distributions:

$$\lambda_j = \sum_{m=1}^{M} w_m N\big(\mathbf{o}, \mu_m, \mathbf{\Sigma}_m\big) \tag{3}$$

where λ_j represents the model of the j^{th} speaker, $\mathbf{o}$ is the vector of parameters being modeled, M is the number of Gaussian mixtures, w_m is the weight of the Gaussian m, and N is a Gaussian function of mean vector μ_m and covariance matrix $\mathbf{\Sigma_m}$. The parameters of the models are estimated from speech samples of the speakers using the well-known Baum-Welch algorithm.

Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm. Twenty iterations are computed to estimate the client model. In the testing phase of a speaker identification system a set of parameter $\mathbf{O} = \{\mathbf{o}_i\}$ is computed from the testing speech signal. Next, the likelihood that each client model performs from the vector $\mathbf{O}$ is calculated and the speaker showing the largest likelihood is chosen,

$$s = \arg\max_j \big\{L\big(\mathbf{O}|\lambda_j\big)\big\} \tag{4}$$

where s is the recognized speaker and $L\big(\mathbf{O}|\lambda_j\big)$ is the likelihood that the vector $\mathbf{O}$ was generated by the speaker of the model λ_j.

3. VIDEO PERSON IDENTIFICATION

In this section, the Visual Person ID task is presented. We have developed for this task a technique for face recognition in smart environments. The technique takes advantage of the continuous monitoring of the scenario and combines the information of several images to perform the speaker recognition. Recognition is stand-alone, taking detection and tracking for granted. That is, the system is semi-automatic. Appearance based face recognition techniques are used given that the scenario does not ensure high quality images. As the visual identification evaluation is a close-set identification task, models for all individuals in the database are created off-line using two sets of video segments: the first one consists on one segment of 15 s per each individual in the database, while the second one consists on one segment of 30 s per individual.

The proposed system works with groups of face images of the same individual. For each test segment, face images of the same individual are gathered into a group. Then, for each group, the system compares such images with the model of the person.

We first describe the procedure for combining the information provided by a face recognition algorithm when it is applied to a group of images of the same person in order to, globally, improve the recognition results. Let $\{C\}_j = \{C_1, C_2, ..., C_S\}$ be the different models or classes stored in the (local or global) model database. S is the number of individual models. Let $\{x\}_i = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_M\}$ be a group of M probe images of the same person. Each model C_j contains N_j images, $\{y\}_n^j = \{\underline{y}_j^1, \underline{y}_j^2, ..., \underline{y}_{N_j}^j\}$ where N_j may be different for every class. We fix a decision threshold R_d so that $\underline{x}_i$ and $\underline{y}_n^j$ represent the same person if $d(\underline{x}_i, \underline{y}_n^j) < R_d$. If, for a given $\underline{x}_i$ the decision function is applied to every $\underline{y}_n^j \in C_j$, we can define the δ value of $\underline{x}_i$ relative to a class C_j, δ_{ij} as

$$\delta_{ij} = \#\big\{\underline{y}_n^i \in C_j \;;\; d(\underline{x}_i, \underline{y}_n^j) < R_d\big\} \tag{5}$$

That is, δ_{ij} counts the number of times that the face recognition algorithm matches $\underline{x}_i$ with an element of C_j. With this information, the δ-Table is built:

	$\mathbf{C_1}$	$\mathbf{C_2}$	...	$\mathbf{C_s}$
$\mathbf{X_1}$	δ_{11}	δ_{12}	...	δ_{1S}
$\mathbf{X_2}$	δ_{21}	δ_{22}	...	δ_{1S}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{X_M}$	δ_{M1}	δ_{M2}	...	δ_{MS}

Table 1. δ-Table

Relying on the table 1, the proposed technique allows to determine the class C_j which better represents the set of probe images $\{x_i\}$. Further information about this technique can be found in [9].

In this work, a PCA based approach [4] has been used. This way, the decision function is the Euclidean distance between the projections of x_i and y_n^j on the subspace spanned by the first eigenvectors of the training data covariance matrix:

$$d(\underline{x}_i, \underline{y}_n^j) = ||W^T \underline{x}_i - W^T \underline{y}_n^j|| \quad (6)$$

The XM2VTS database [10] has been used as training data for estimating the projection matrix and the first 400 eigenvectors have been preserved. Due to the images being recorded continuously using the corner cameras, face images can not be ensured to be all frontal. Mixing frontal and non-frontal faces in the same models can be quite a problem for face recognition systems. To avoid this situation, eye coordinates are used to determine the face pose for each image. Only frontal faces are used for identification. Note that, in our system, models per each person have been automatically generated, without human intervention. All images for a given individual in the training intervals are candidates to form part of the model. Candidate face bounding boxes are projected on the subspace spanned by the first eigenvectors of the training data covariance matrix WT. The resulting vector is added to the model only if different enough from the vectors already present in the model.

4. MULTIMODAL PERSON IDENTIFICATION

In a multimodal biometric system that uses several characteristics, fusion is possible at three different levels: feature extraction level, matching score level or decision level. Fusion at the feature extraction level combines different biometric features in the recognition process, while decision level fusion performs logical operations upon the monomodal system decisions to reach a final resolution. Score level fusion matches the individual scores of different recognition systems to obtain a multimodal score. Fusion at the matching score level is usually preferred by most of the systems.

Matching score level fusion is a two-step process: normalization and fusion itself [11], [12], [13], [14]. Since monomodal scores are usually non-homogeneous, the normalization process transforms the different scores of each monomodal system into a comparable range of values. One conventional affine normalization technique is z-score, that transforms the scores into a distribution with zero mean and unitary variance [12], [14].

After normalization, the converted scores are combined in the fusion process in order to obtain a single multimodal score. Product and sum are the most straightforward fusion methods. Other fusion methods are min-score and max-score that choose the minimum and the maximum of the monomodal scores as the multimodal score.

4.1. Normalization and Fusion Techniques

Scores must be normalized before being fused. One of the most conventional normalization methods is z-score (ZS), which normalizes the global mean and variance of the scores of a monomodal biometric. Denoting a raw matching score as *a* from the set *A* of all the original monomodal biometric scores, the z-score normalized biometric x_{zs} is calculated according to Eq. 7,

$$x_{zs} = \frac{a - \mu(A)}{\theta(A)} \quad (7)$$

where $\mu(A)$ is the statistical mean of A and $\theta(A)$ is the standard deviation.

Histogram equalization (HE) is a non linear transformation whose purpose is to equalize the variances of two monomodal biometrics in order to reduce the non linear effects typically introduced by speech systems. The HE technique matches the histogram obtained from the speaker verification scores and the histogram obtained from the face identification scores, both evaluated over the training data. The designed equalization takes as a reference the histogram of the scores with the best accuracy, which can be expected to have lower separate variances, in order to obtain a bigger variance reduction.

In Matcher Weighting (MW) fusion each monomodal score is weighted by a factor proportional to the recognition rate, so that the weights for more accurate matchers are higher than those of less accurate matchers. When using the Identification Error Rates (IER) the weighting factor for every biometric is proportional to the inverse of its IER. Denoting w^m and e^m the weigthing factor and the IER for the mth biometric x^m and M the number of biometrics, the fused score u is expressed as

$$u = \sum_{m=1}^{M} w^m x^m \quad (8)$$

where

$$w^m = \frac{\frac{1}{e^m}}{\sum_{m=1}^{M} \frac{1}{e^m}} \quad (9)$$

Before carrying out the fusion process, histogram equalization is applied over all the previously obtained monomodal scores. Since the best recognition results have been achieved in the acoustic recognition experiments, the histogram of the voice scores has been taken as a reference in the histogram equalization. After the equalization process, the weighting factors for both acoustic and face scores are calculated by using the corresponding Identification Error Rates, as in Eq. 9. Z-score normalization is also applied, and final fused scores are obtained by using Eq. 8.

5. EXPERIMENTS AND DISCUSSION

5.1. Experimental set-up

A set of audiovisual recordings of seminars and of highly-interactive small working-group seminars have been used. These recordings were collected by the CHIL consortium for the CLEAR 06 Evaluation. The recordings were done according to the "CHIL Room Setup" specification [1]. A complete description of the different recordings can be found in [15]. The figure 1 depicts a brief description of the UPC smart-room sensors and space conditions.

Data segments are short video sequences and matching far-field audio recordings taken from the above seminars. In order to evaluate how the duration of the training signals affects the performance of the system two training durations have been considered: 15 and 30 seconds. Test segments of different durations ($1, 2, 5, 10$ and 20 seconds) have been used during the algorithm development and testing phases. A total of 26 personal identities have been used in the recognition experiments.

	Number of segments	
Segment Duration	Development	Evaluation
1 sec	390	613
2 sec	182	0
5 sec	78	411
20 sec	26	178

Table 2. Number of segments for each test condition

Each seminar has one audio signal from the microphone number 4 of the Mark III array. Each audio signal has been divided into segments which contain information of a unique speaker. These segments have been merged to form the final testing segments of $1, 5, 10$ and 20 seconds (see Table 2) and training segments of 15 and 30 seconds. Video is recorded in compressed JPEG format, with different frame-rates and resolutions for the various recordings. Far-field conditions have been used for both modalities, i.e. corner cameras for video and Mark III microphone array for audio. In the audio task only one array microphone has been considered for both development and testing phases. In the video task, we have four fixed position cameras that are continuously monitoring the scene. All frames in the 1/5/10/20 seconds segments and all synchronous camera views can be used and the information can be fused to find the identity of the concerned person. To find the faces to be identified, a set of labels is available with the position of the bounding box for each person's face in the scene. These labels are provided each 1s. The face bounding boxes are linearly interpolated to estimate their position in intermediate frames. To help this process, an extra set of labels is provided, giving the position of both eyes of each individual each 200 ms. The metric used to benchmark the quality of the algorithms is the percentage of correctly recognized people from test segments.

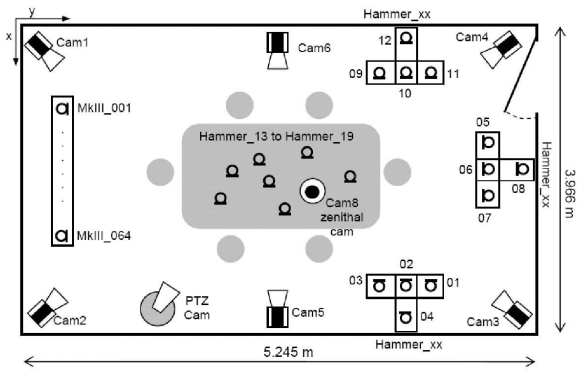

Fig. 1. *The UPC smart room setup*

5.2. Results

In this section we summarize the results for the evaluation of different modalities and the result improvement with the multimodal technique. Table 3 shows the correct identification rate for both audio and video modalities and the fusion identification rate obtained depending on the length of the used test files. Some improvements have been performed on the system since the CLEAR Evaluation, leading to better results than the ones presented in that. Related to acoustic identification task, it can be seen that the results, in general, are better as the segments length increases. Table 3 reports that for the different test segment lengths the recognition rate increases when more data is used to test the speaker models. Overall, using the 30 seconds training segments, an improvement of up to 6% in the recognition rate is obtained with respect to the case where 15 seconds segments are used. For the face identification evaluation, in general, these results show a low performance of the system. Results for the training set B (using a segment of 30 sec. to generate the models) show only a slight increase of performance with respect to training set A. It can also be seen that the results improve slowly as the segments length increases. The reasons for this low performance are manifold: First of all, the system uses only frontal faces to generate the models and for recognition. However, most of the face views found in the recordings are non frontal. Another reason for the low percentage of correctly identified persons is the low quality of the images. The need to cover all the space in the room with four cameras results in small images, were the person's faces are tiny. In the worst cases, face sizes are only 13x13 pixels.

The determination of the weighting factors for the multimodal fusion has been done by using the training signals of 30 seconds as a development set. The first 15 seconds have been used for training and the other 15 seconds for testing. The recognition results obtained in the evaluation for multimodal identification can also be seen in Table 3. Fusion results of both systems are

		Train A			Train B		
Duration	Segments	Speech	Video	Fusion	Speech	Video	Fusion
1	613	75.0 %	20.2 %	77.3 %	84.0 %	19.6 %	87.8 %
5	411	89.3 %	21.4 %	92.0 %	97.1 %	22.9 %	97.3 %
10	289	88.2 %	22.5 %	93.4 %	97.6 %	25.6 %	98.6 %
20	178	92.1 %	23.6 %	97.7 %	98.8 %	27.0 %	100.0 %

Table 3. Percentage of correct identification for both audio and video unimodal modalities and multimodal fusion. The first column shows the duration of test segments in seconds. The second one shows the number of tested segments. Train A and B are the training sets of 15 seconds and 30 respectively

also shown for the different lengths. Fusion correct identification rates are higher than the monomodal rates. The obtained fusion results outperform those obtained with both monomodal systems.

6. CONCLUSIONS

In this paper we have described two techniques for visual and acoustic person identification in smart room environments. A Gaussian Mixture Model of the distribution of the Frequency Filtering coefficients has been used to perform speaker recognition. For video, an approach based on joint identification over groups of images of a same individual using a PCA approach has been followed.

For the acoustic identification task, the results show that the presented approach is well adapted to the conditions of the experiments. For the visual identification task, the low quality of the images results in a low performance of the system. In this case, results suggest that identification should be performed combining more features other than frontal face bounding-boxes. To improve the obtained results, a multimodal score fusion technique has been used. Matcher Weighting with histogram equalized scores is applied to the scores of the two monomodal tasks. The results show that this technique can provide an improvement of the recognition rate in all train/test conditions.

Acknowledgements

This work has been partially sponsored by the EC-funded project CHIL (IST-2002 – 506909) and by the Spanish Government-funded project ACESCA (TIN2005 – 08852).

7. REFERENCES

[1] R. Stiefelhagen J. Casas, "Multi-camera/multi-microphone system design for continuous room monitoring,," in *CHIL Consortium Deliverable D4.1*, 2005.

[2] C. Nadeu, D. Macho, and J. Hernando, "Time and Frequency Filtering of Filter-Bank Energies for Robust Speech Recognition," in *Speech Communication*, 2001, vol. 34, pp. 93–114.

[3] D. A. Reynolds, "Robust text-independent speaker identifi cation using Gaussian mixture speaker models," in *IEEE Transactions ASSP*, 1995, vol. 3 N^o 1, pp. pp. 72–83.

[4] L.Sirovich M. Kirby, "Application of the Karhunen-Loeve procedure for the characterization of human faces," in *IEEE Trans. PAMI*, 1990, vol. 12, no. 1, pp. pp. 103–108.

[5] R.M. Bolle et al., "Guide to Biometrics," in *Springer*, 2004.

[6] R.Brunelli and D. Falavigna, "Person Identifi cation Using Multiple Cues," in *IEEE on PAMI*, 1995, vol. 17, No. 10, pp. 955–966.

[7] Mermelstein P. Davis S. B., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions ASSP*, 1980, vol. Vol. 28, pp. pp. 357–366.

[8] Furui S., "Speaker independent isolated word recognition using dynamic features of speech spectrum," in *IEEE Transactions ASSP*, 1986, vol. No. 34, pp. pp.52–59.

[9] J.Cruz V.Vilaplana, C.Martinez and F.Marques, "Face Recognition Using Groups of Images in Smart Room Scenarios," in *ICIP*, 2006.

[10] K. Messer et al., "XM2VTSDB: The extended M2VTS Database," in *AVBPA*, 1999.

[11] N.A Fox et al., "Person Identifi cation Using Automatic Integration of Speech, Lip and Face Experts," in *ACM SIGMM WBMA*, 2003, pp. pp. 25–32.

[12] M. Indovina et al., "Multimodal Biometric Authentication Methods: A COTS Approach," in *MMUA*, 2003, pp. pp. 99–106.

[13] T. Chen S. Lucey, "Improved Audio-visual Speaker Recognition via the Use of a Hybrid Combination Strategy," in *The 4th International Conference on Audio- and Video- Based Biometric Person Authentication*, 2003.

[14] T. Tan Wang Yuan, Wangm Yunhong, "Combining Fingerprint and Voiceprint Biometrics for Identity Verifi cation: an Experimental Comparison," in *ICBA*, 2004, pp. pp. 663–670.

[15] Djamel Mostefa et al., "CLEAR Evaluation Plan v1.1," in `http://isl.ira.uka.de/~nickel/clear/downloads/chil-clear-v1.1-2006-02-2%1.pdf`, 2006.