ISL Person Identification Systems in the CLEAR 2007 Evaluations

Hazım Kemal Ekenel[1], Qin Jin[2], Mika Fischer[1], and Rainer Stiefelhagen[1]

[1] Interactive Systems Labs (ISL), Universität Karlsruhe (TH),
76131 Karlsruhe, Germany
{ekenel, mika.fischer}@ira.uka.de
[2] Interactive Systems Labs (ISL), Carnegie Mellon University,
15213 Pittsburgh, PA, USA
qjin@cs.cmu.edu

Abstract. In this paper, we present ISL person identification systems in the CLEAR 2007 evaluations. The identification systems consist of a face recognition system, a speaker identification system and a multi-modal identification system that combines the individual systems. The experimental results show that the face recognition system outperforms the speaker identification system significantly on the short duration test segments. They perform equally well on the longer duration test segments. Combination of the individual systems improves the performance further.

Keywords: Face recognition, Speaker Identification, Multimodal Person Identification, Person Identification in Smart Rooms.

1 Introduction

Person identification for smart environments has become an important application area [1,2,3]. Sample applications can be a smart car that can identify the driver; a smart lecture or meeting room, where the participants can be identified automatically. As can be expected, this group of applications requires identification of people naturally under uncontrolled conditions.

Among the biometric person identification methods, face recognition and speaker identification are known to be the most natural ones, since the face and voice modalities are the modalities we use to identify people in our daily lives. However, doing face recognition or speaker identification in a natural way poses many challenges. In terms of face recognition, there is no cooperation of the subjects being identified, there are no constraints on head-pose, illumination conditions, use of accessories, etc. Moreover, depending on the distance between the camera and the subject, the face resolution varies, and generally the face resolution is low. In terms of speaker identification, again, there is no cooperation, and the system should handle large variety of speech signals, corrupted by adverse environmental conditions such as noise, background, and channel. The only factors that can help to improve the person identification performance in smart rooms are the video data of the individuals from multiple views provided by several cameras and the multi-channel speech signal provided by

microphone arrays that are mounted in the smart room. Furthermore, with the fusion of these modalities, the correct identification rates can be improved further. Sample images from different smart rooms are shown in Fig. 1.

Fig. 1. Sample images from different smart rooms

The organization of the paper is as follows. In Section 2, the individual face recognition and speaker identification systems are explained briefly, and the utilized fusion approaches are described. Experimental results are presented and discussed in Section 3. Finally, in Section 4, conclusions are given.

2   Methodology

In this section, we present the single modality and multi-modal person identification systems that have been developed at the ISL.

2.1   Video-Based Face Recognition

The face recognition system is based on the local appearance-based models and it processes multi-view video data provided by four fixed cameras. In the training stage all the images from all the cameras are put together. Although the manual annotations of the images are available in the database, due to the low resolution of face images these manual labels might be imprecise. In order to prevent the registration errors that can be caused by these imprecise labels, 24 additional samples are also generated by modifying the manual face bounding box labels by moving the center of the bounding box by 1 pixel and changing the width or height by ±2 pixels.

The feature extraction step follows the approach in [7,8], which performs block-based discrete cosine transform (DCT) to non-overlapping blocks of size 8×8 pixels. The obtained DCT coefficients are then ordered according to the zig-zag scan pattern.

The first coefficient is discarded for illumination normalization as suggested in [7] and the remaining first ten coefficients in each block are selected in order to create compact local feature vectors. Furthermore, robustness against illumination variations is increased by normalizing the local feature vectors to unit norm [8]. The global feature vector is generated by concatenating the local feature vectors. Afterwards, these global feature vectors are clustered using k-means algorithm in order to realize real-time classification with a nearest neighbor classifier.

In the testing stage, at an instant, all four camera views are compared to the representatives in the database. Their distances are converted to confidence scores using min-max normalization [4]. This way, the score is normalized to the value range of [0,1], closest match having the score "1", and the furthest match having the score "0". To have equal contribution of each frame, these scores are re-normalized by dividing them to the sum of their values. We weight each frame using the distance-to-second-closest (DT2ND) metric. In a previous study [9], it has been observed that the difference of the distances, x, between the closest and the second closest training samples is generally smaller in the case of a false classification than in the case of a correct classification. It has been found that the distribution of these distances resembles an exponential distribution:

$$\varepsilon(x; \lambda) = 0.1\lambda e^{-\lambda x} \text{ with } \lambda = 0.05.$$

The weights are then computed as the cumulative distribution function:

$$\omega_{DT2ND}(x; \lambda) = 1 - e^{-\lambda x}.$$

The obtained confidence scores are summed over camera-views and over image sequence. The identity of the face image is assigned as the person who has the highest accumulated score.

2.2 Speaker Identification

The speaker identification system is based on Gaussian mixture models (GMM) of Mel Frequency Cepstral Coefficients (MFCC) [10,11]. Feature warping and reverberation compensation are applied on MFCC in order to improve robustness against channel mismatch. Our reverberation compensation approach uses a different noise estimation compared to the standard spectrum subtraction approach [12]. The feature warping method warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval [12,13,14]. The identification decision is made as follows:

$$s = \arg\max_{i}\left\{L\left(Y|\Theta_i\right)\right\} \qquad Y = \left(y_1, y_2, \cdots, y_N\right)$$

where s is the identified speaker and $L\left(Y|\Theta_i\right)$ is the likelihood that the test feature set Y was generated by the GMM Θ_i of speaker i, which contains M weighted mixtures of Gaussian distributions

$$\Theta_i = \sum_{m=1}^{M} \lambda_m N(X, U_m, \Sigma_m) \qquad i = 1, 2, \cdots, S,$$

where X is the set of training feature vectors to be modeled, S is the total number of speakers, M is the number of Gaussian mixtures, λ_m is the weight of the Gaussian component m, and $N(X, U_m, \Sigma_m)$ is a Gaussian function with mean vector U_m and covariance matrix Σ_m. The parameters of a GMM are estimated from speech samples of a speaker using the EM algorithm.

As there are 64 channels for each speech recording, we train GMMs for each speaker on each of the 64 channels. We randomly select channel 7 as the test channel. We apply Frame-based Score Competition approach when computing the likelihood scores of test features given a speaker with 64 GMMs. The idea of the FSC approach is to use the set of multiple GMM models rather than a single GMM model. A multiple microphone setup emits speech samples from multiple channels. As a consequence, we can build multiple GMM models for each speaker k, one for each channel i and refer to it as $\Theta_{k,Chi}$. For a total number of 64 channels we get $\Theta_k = \{\Theta_{k,Ch1}, \cdots, \Theta_{k,Ch64}\}$ models for speaker k. In each frame we compare the incoming feature vector of channel $Ch7$ to all GMMs $\{\Theta_{k,Ch1}, \cdots, \Theta_{k,Ch64}\}$ of speaker k. The highest log likelihood score of all GMM models is chosen to be the frame score. Finally, the log likelihood score of the entire test feature vector set X from channel h is estimated as

$$LL(X \mid \Theta_k) = \sum_{n=1}^{N} LL(x_n \mid \Theta_k) = \sum_{n=1}^{N} \max \left\{ LL(x_n \mid \Theta_{k,Chj}) \right\}_{j=1}^{64}$$

32 Gaussians and 16 Gaussians are trained for each speaker for the training duration of 30-seconds and 15-seconds respectively. 13-dimensional MFCC is used as speaker features.

2.3 Fusion

The min-max normalization is used for score normalization. For modality weighting, we used a new adaptive modality weighting scheme based on the separation of the best two matches. It is named as cumulative ratio of correct matches (CRCM) and utilizes a non-parametric model of the distribution of the correct matches with respect to the confidence differences between the best two matches. It relies on the observation that the difference of the confidences between the closest and the second closest training samples is generally smaller in the case of a false classification than in the case of a correct classification. The greater the confidence difference between the best two matches is, the higher the weight the individual modality receives. This weighting model has been computed on a validation set by taking the cumulative sum of the number of correct matches achieved at a confidence difference between the best two matches. Finally, we combined the modalities using the sum rule [6].

3   Experiments

The experiments have been conducted on a database that has been collected by the CHIL consortium [15] for the CLEAR 2007 evaluations [16]. The recordings are from lecture-like seminars and interactive small working group seminars that have been held at different CHIL sites: AIT, Athens, Greece, IBM, New York, USA, ITC-IRST, Trento, Italy, UKA, Karlsruhe, Germany and UPC, Barcelona, Spain. Sample images from the recordings can be seen in Figure 1. The used data for the identification task consists of short video sequences of 28 subjects, where the subject is both speaking and visible to the cameras at the same time. The recording conditions are uncontrolled, and depending on the camera view and the position of the presenter/participant, low resolution faces ranging between 10 to 50 pixels resolution are acquired. Two different training and four different validation/testing durations are used in the experiments as presented in Table 1. Identity estimates are provided at the end of each test sequence duration using the available audio-visual data.

Table 1. Duration and number of the training, validation and testing sequences

Sequence ID	Sequence Duration (sec)	No. of Sequences
Train A	15	28
Train B	30	28
Validation 1	1	560
Validation 2	5	112
Validation 3	10	56
Validation 4	20	28
Test 1	1	2240
Test 2	5	448
Test 3	10	224
Test 4	20	112

In the database, face bounding box labels are available every 200 ms. We only used these labeled frames for the experiments. The face images are cropped and scaled to 40x32 pixels resolution. They are then divided into 8x8 pixels resolution non-overlapping blocks making 20 local image blocks. From each image block ten-dimensional DCT-based feature vectors are extracted as described in Section 2.1 and they are concatenated to construct the final 200-dimensional feature vector. The classification is performed using a nearest neighbor classifier. The L1 norm is selected as the distance metric, since it has been observed that it consistently gives the best correct recognition rates when DCT-based feature vectors are used.

13-dimensional MFCC, with feature warping and reverberation compensation applied, is extracted from the speech signal as the speaker feature. We trained a GMM with 32 mixtures for each speaker using the expectation-maximization (EM) algorithm under the 30 seconds training condition and 16 mixtures for each speaker under the 15 seconds training condition. The classification is performed as described in Section 2.2.

3.1 Experiments on the Validation Set

The correct identification rates of the face recognition and speaker identification systems obtained on the validation set are presented in Table 2. In the table, each row shows the results for a different training-testing duration combination. The letter indicates whether the training is from set A or B which corresponds to 15 and 30 second training durations, respectively. The number indicates the duration of the testing segment in seconds. As expected, as the duration of training or testing increases the correct identification rate increases. Both systems achieve 100% correct identification when the systems are trained with 30 seconds of data and tested with the sequences of 20 seconds duration. Face recognition is found to be significantly superior to speaker identification at the other training-testing duration combinations.

Table 2. Correct identification rates of the individual modalities on the validation set

	Face Reco. (%)	Speaker Id. (%)
A1	91.4	56.4
A5	99.1	67.9
A10	100	89.3
A20	100	92.9
B1	94.3	61.1
B5	100	84.8
B10	100	98.2
B20	100	100

To assess the effect of amount of testing data used for face recognition, we linearly interpolate the manual face bounding box labels and use these interpolated labels for cropping the faces from unlabelled frames during testing. That is, we use not only the labelled frames, but also the frames in between the labelled frames for testing. The obtained results can be seen in Table 3. Compared to the face recognition results in Table 2, a minor increase in correct identification rate in training-testing duration combinations, A1 and A5, and a minor decrease in B1 and B5 are observed. This indicates that using frames every 200ms suffices for face recognition, which provides also a significant drop in processing requirements.

Table 3. Correct identification rates of face recognition using every frame in test segments

	A1	A5	A10	A20	B1	B5	B10	B20
Performance (%)	92.1	100	100	100	93.9	98.2	100	100

Table 4 compares system performances when both training and testing are on channel 7 (CH7-CH7) vs. training using all 64 channels and testing on channel 7 (All

Table 4. Comparison of speaker identification performance on validation set

Test Duration (sec)	Train A (15 sec)		Train B (30 sec)	
	CH7-CH7 (%)	All CHs-CH7 (%)	CH7-CH7 (%)	All CHs-CH7 (%)
1	53.6	56.4	50.9	61.1
5	66.1	67.9	82.1	84.8
10	83.9	89.3	92.9	98.2
20	92.9	92.9	100	100

Table 5. Comparison of modality weighting schemes on validation set

	DPC (%)	IPF (%)	CRCM (%)	DPC+CRCM (%)	IPF+CRCM (%)
A1	91.8	91.6	92.0	91.8	92.7
A5	100	100	98.2	99.1	100
A10	100	100	100	100	100
A20	100	100	100	100	100
B1	94.5	94.5	94.5	94.3	94.5
B5	100	100	100	100	100
B10	100	100	100	100	100
B20	100	100	100	100	100

CHs-CH7) using Frame-based Score Competition (FSC). We can see from the table that combining information from multiple channels provides significant system improvement especially when test duration is short.

Table 5 compares different modality weighting schemes. The correct identification rates in Table 2 are used to determine the fixed weights that each modality receives at each training-testing duration combination. It is done in two different ways. The first way, which is named as DPC, is by determining the weights directly proportional to the correct identification rates. For example, if the face recognition system has 100% and the speaker identification system has 85% correct identification rates, then they are weighted by 1 and 0.85 respectively for that training-testing duration combination. The second way, which is named as IPF, is by determining the weights inversely proportional to the false identification rates. For instance, if the face recognition system has 5% and the speaker identification system has 10% false identification rates, then the face recognition system receives twice as much weight than the speaker identification system. In addition to fixed modality weighting schemes, we also utilized CRCM which is the adaptive weighting scheme explained in Section 2.3. DPC+CRCM and IPF+CRCM are the combinations of fixed and adaptive weighting schemes. According to the obtained results on the validation set, IPF+CRCM has been selected as the modality weighting scheme to be used on the testing set.

3.2   Experiments on the Test Set

The correct identification rates of the face recognition and speaker identification systems obtained on the test set are given in Table 6. Similar to the obtained results on the validation set, as the duration of training or testing increases the correct identification rate increases. As can be noticed, on the test set the speaker identification performs as well as or even better than the face recognition at longer duration test segments. In the case of fixed modality weighting, this implies that the validation set is misleading, since on the validation set face recognition has been found to be more successful at these segments. The other observation that can be derived by comparing Tables 2 and 6 is the higher false identification rates obtained on the testing set. The main reason is that the time gap between the training set and test set is greater than the time gap between the training and validation set.

Table 6. Correct identification rates of the individual modalities on the test set

	Face Reco. (%)	Speaker Id. (%)
A1	84.6	41.9
A5	90.8	69.6
A10	93.3	92.0
A20	94.6	96.4
B1	89.3	41.2
B5	94.4	78.3
B10	94.6	96.4
B20	96.4	99.1

3.3   Fusion Experiments

In the fusion experiments, we analyzed the modality weighting schemes. The results are presented in Table 7. Even with the primitive fixed weights, in most of the training-testing duration combinations the correct identification rates are higher than the ones obtained by the individual modalities. The results with the more sophisticated adaptive modality weighting scheme, CRCM, show a significant increase in correct identification rates. Note that, in terms of performance of each modality, the validation set was not quite representative. As we have seen, under some training-testing duration combinations, face recognition was found superior than speaker identification on the validation set, but on the test set, it was the opposite. Therefore, performance based fixed weighting can be misleading. On the other hand, the results obtained by CRCM indicates that confidence differences are more robust cues for modality weighting. There is no significant performance difference between the CRCM and DPC+CRCM results. The performance degrades with IPF+CRCM, which was chosen as the primary modality weighting scheme for official results submission due to its best performance on the validation set. The

Table 7. Correct identification rates of fixed weighting schemes

	DPC (%)	IPF (%)	CRCM (%)	DPC+CRCM (%)	IPF+CRCM (%)
A1	84.8	84.6	86.3	86.7	86.7
A5	91.1	90.8	93.5	93.5	91.7
A10	94.2	93.3	98.2	98.2	93.3
A20	94.6	94.6	99.1	99.1	94.6
B1	89.8	89.4	89.6	89.9	89.9
B5	94.9	94.4	97.3	97.3	94.4
B10	95.5	94.6	98.7	98.7	94.6
B20	97.3	97.3	99.1	99.1	99.1

reason is the hard modality weighting in IPF. Since, on the validation set at some train-ing-testing duration combinations, face recognition reached 0% false identification rate, at these combinations only the face recognition system's decision is trusted.

4   Conclusion

In this paper, we presented ISL person identification systems in the CLEAR 2007 evaluations. In Table 8, the false identification rates of the individual modalities and the multi-modal system are listed. The multi-modal system included in the table uses CRCM modality weighting scheme. From the table, it is clear that multi-modal fusion significantly improves the performance compared to each of the single modalities. This also indicates that the face and voice modalities are complementary biometric traits.

Table 8. Correct identification rates of individual modalities and the multi-modal system

	Face Reco. (%)	Speaker Id. (%)	Fusion (%)
A1	84.6	41.9	86.3
A5	90.8	69.6	93.5
A10	93.3	92.0	98.2
A20	94.6	96.4	99.1
B1	89.3	41.2	89.6
B5	94.4	78.3	97.3
B10	94.6	96.4	98.7
B20	96.4	99.1	99.1

References

1. Vendrig, J., Worring, M.: Multimodal Person Identification in Movies. In: Proceedings of the Intl. Conf. on Image and Video Retrieval, pp. 175–185 (2002)
2. Erzin, E., et al.: Multimodal Person Recognition for Human-Vehicle Interaction. IEEE Multimedia 13(2), 18–31 (2006)
3. Hazen, T.J., et al.: Multi-Modal Face and Speaker Identication for Mobile Devices. In: Hammoud, R.I., Abidi, B., Abidi, M. (eds.) Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems, Springer, Heidelberg (2007)
4. Snelick, R., et al.: Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(3), 450–455 (2005)
5. Brunelli, R., Falavigna, D.: Person Identification Using Multiple Cues. IEEE Trans. on Pattern Analysis and Machine Intelligence 17(10), 955–966 (1995)
6. Kittler, J., et al.: On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
7. Ekenel, H.K., Stiefelhagen, R.: Local Appearance based Face Recognition Using Discrete Cosine Transform. In: 13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey (September 2005)
8. Ekenel, H.K., Stiefelhagen, R.: Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization. In: IEEE CVPR Biometrics Workshop, New York, USA (June 2006)
9. Stallkamp, J.: Video-based Face Recognition Using Local Appearance-based Models., Thesis report, Universität Karlsruhe (TH) (November 2006)
10. Furui, S.: Recent Advances in Speaker Recognition. Pattern Recognition Letters 18, 859–872 (1997)
11. Reynolds, D.: Speaker Identification and Verification Using Gaussian Mixture Speaker Models. Speech Communication 17(1-2), 91–108 (1995)
12. Jin, Q., Pan, Y., Schultz, T.: Far-field Speaker Recognition. In: International Conference on Acoustic, Speech, and Signal Processing (ICASSP) (2006)
13. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. Speaker Odyssey 2001 Conference (June 2001)
14. Xiang, B., Chaudhari, U., Navratil, J., Ramaswamy, G., Gopinath, R.: Short-time Gaussi-anization for Robust Speaker Verification. In: Proc. ICASSP (2002)
15. Computers in the Human Interaction Loop –CHIL, http://chil.server.de/
16. CLEAR 2007, Evaluation, http://www.clear-evaluation.org/