ISL Person Identification Systems in the CLEAR Evaluations

Hazım Kemal Ekenel[1] and Qin Jin[2]

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

Abstract. In this paper, we presented three person identification systems that we have developed for the CLEAR evaluations. Two of the developed identification systems are based on single modalities- audio and video, whereas the third system uses both of these modalities. The visual identification system analyzes the face images of the individuals to determine the identity of the person. It processes multi-view, multi-frame information to provide the identity estimate. The speaker identification system processes the audio data from different channels and tries to determine the identity. The multi-modal identification system fuses the similarity scores obtained by the audio and video modalities to reach an identity estimate.

1 Introduction

Person identification in smart environments is very important in many aspects. For instance, customization of the environment according to the person's identity is one of the most useful applications. However, until now, person identification research has focused on security-oriented authentication applications and face recognition in smart rooms has been ignored in great extent.

In CHIL project [1], aiming to encourage the research efforts for person identification in smart environments, a data corpus and evaluation procedure has been provided. Following the two successful uni-modal identification evaluations [2], this year multi-modal identification is also included to the person identification task.

In this paper, the person identification systems that have developed at the Interactive Systems Labs for the CLEAR evaluations are presented. The organization of the paper is as follows. In Section 2, the algorithms used in each system are explained. Experimental results are presented and discussed in Section 3. Finally, in Section 4, conclusions are given.

2 Methodology

In this section, face recognition, speaker identification, and fusion algorithms that are used for the evaluations are presented.

R. Stiefelhagen and J. Garofolo (Eds.): CLEAR 2006, LNCS 4122, pp. 249 – 257, 2007.

2.1 Face Recognition

The face recognition system processes multi-view, multi-frame visual information to obtain an identity estimate. The system consists of the following building blocks:

- Image alignment
- Feature Extraction
- Camera-wise classification
- Score normalization
- Fusion over camera-views
- Fusion over image sequence

The system receives an input image and the eye-coordinates of the face in the input image. The face image is cropped and aligned according to the eye coordinates. If only one eye is visible, that image is not processed. The aligned image is, then, divided into non-overlapping 8x8 pixels resolution image blocks. Discrete cosine transform (DCT) is applied on each local block. The obtained DCT coefficients are ordered using zig-zag scan pattern. From the ordered coefficients, the first one is removed since it only represents the average value of the image block. The first M coefficients are selected from the remaining ones [3]. To remove the effect of intensity level variations among the corresponding blocks of the face images, the extracted coefficients are normalized to unit norm. For detailed information please see [4].

Classification is performed by comparing the extracted feature vectors of the test image, with the ones in the database. Each camera-view is handled separately. That is, the feature vectors that are extracted from the face images acquired by Camera 1 are compared with the ones that are also extracted from the face images acquired by Camera 1 during training. This approach speeds up the system significantly. That is, if we have N images from each camera for training, and if we have R images from each camera for testing, and if we have C cameras that do recording, it requires $(C*N)*(C*R)$ number of similarity calculations between the training and testing images. However, when we do camera-wise image comparison, then we only need to do $C*(N*R)$ comparisons between the training and testing images. Apparently, this reduces the amount of required computation by $1/C$. In addition to the improvement in system's speed, it also provides a kind of view-based approach that separates the comparison of different views, which was shown to perform better than doing matching between all the face images without taking into consideration their view angles [5].

Distance values obtained from each camera-view are normalized using Min-Max rule, which is defined as:

$$ns = 1 - \frac{s - \min(S)}{\max(S) - \min(S)},$$

where, s corresponds to a distance value of the test image to one of the training images in the database, and S corresponds to a vector that contains the distance values of the test image to all of the training images. The division is subtracted from one, since the lower the distance is, the higher the probability that the test image belongs to that identity class. 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". These scores are then normalized by dividing them to the sum of the confidence scores.

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

In this section, the building blocks of the speaker identification system are explained.

2.2.1 Reverberation Compensation

A distant-talking speech signal is degraded by additive background noise and reverberation. Considering room acoustics as a linear shift-invariant system, the receiving signal $y[t]$ can be written as,

$$y[t] = x[t] * h[t] + n[t], \tag{1}$$

where the source signal $x[t]$ is the clean speech, $h[t]$ is the impulse response of room reverberation, and $n[t]$ is recording noise. Cepstrum Mean Subtraction (CMS) has been used successfully to compensate the convolution distortion. In order for CMS to be effective, the length of the channel impulse response has to be shorter than the short-time spectral analysis window which is usually 16ms-32ms. Unfortunately, the duration of impulse response of reverberation usually has a much longer tail, as long as more than 50ms. Therefore traditional CMS will not be as effective under these conditions.

We separate the impulse response $h[t]$ into two parts $h_1[t]$ and $h_2[t]$, where,

$$h[t] = h_1[t] + \delta(t-T)h_2[t]$$

$$h_1[t] = \begin{cases} h[t] & t < T \\ 0 & otherwise \end{cases}$$

$$h_2[t] = \begin{cases} h[t+T] & t \geq 0 \\ 0 & otherwise \end{cases}$$

and rewrite formula (1) as

$$y[t] = x[t] * h_1[t] + x[t-T] * h_2[t] + n[t]$$

$h_1[t]$ is a much shorter impulse response, whose length is smaller than the DFT analysis window, thus it can be compensated by the conventional CMS. For $x[t-T] * h_2[t]$, we treat it the same as additive noise n[t], and apply the noise reduction technique based on spectrum subtraction. Assuming the noise $x[t-T] * h_2[t] + n[t]$ could be estimated from $y[t-T]$, and then the spectrum subtraction is performed as,

$$\hat{X}[t,w] = \max(Y[t,w] - a \cdot g(w)Y[t-T,w], b \cdot Y[t,w]),$$

where a is the noise overestimation factor, b is the spectral floor parameter to avoid negative or underflow values. We can empirically estimate the optimum *a, b* and *g(w)* on a development dataset. We found that the system performance is not sensitive to T. Within the range of 20-40 ms there is no significant difference on the effect of the spectra subtraction. However outside that range, there is obvious performance degradation. We found *a = 1.0*, *b = 0.1* and $g(w) = \left|1 - 0.9e^{jw}\right|$ optimal in most changing conditions based on development data as described in [6]. Standard CMS is applied after spectrum subtraction to eliminate the effect of $h_1[t]$.

2.2.2 Feature Warping

The feature warping method proposed in [7], which warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval, aims to make the features more robust to different channel and noise effects. The warping can be considered as a nonlinear transformation *T*, which transforms the original feature X to a warped feature $\hat{X}$, i.e,

$$\hat{X} = T(X)$$

This can be done by CDF matching, which warps a given feature so that its CDF matches a desired distribution, such as normal distribution. The method assumes that the dimensions of the MFCC vector are independent. So each dimension is processed as a separate stream. The CDF matching is performed over short time intervals by shifting a window. Only the central frame of the window is warped every time. The warping executes as follows, the same way as in [8].:

- for $i = 1, \cdots, d$ where d is the number of feature dimensions
- sorting features in dimension i in ascending order in a given window
- warping raw feature value x in dimension i of the central frame to its warped value $\hat{x}$ which statisfies:

$$\phi = \int_{-\infty}^{\infty} f(y)dy$$

Where $f(y)$ is the probability density function (PDF) of standard normal distribution, i.e.

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})$$

and ϕ is its corresponding CDF value. Suppose x has a rank r and the window size is N. Then the CDF value can be approximated as

$$\phi = \frac{(r - 1/2)}{N}$$

- $\hat{x}$ can be quickly found by lookup in a standard normal CDF table.

In our experiments, the window size is 300 frames and the window shifts one frame. Zeros are padded at the beginning and at the end of the raw feature stream.

2.2.3 Speaker Modeling

Over past decades, GMM has become the dominant approach for speaker modelling in speaker recognition systems which use untranscribed training data [9]. The recognition decision is made as follows

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

where s is the identified speaker and $L(Y|\Theta_i)$ 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 modelled, 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. In our system, 128 Gaussians and 32 Gaussians are trained for each speaker for the training duration of 30-seconds and 15-seconds respectively. We will show how we choose these numbers of Gaussians in the experimental results section.

2.3 Multimodal Identification

Multimodal identification is performed by fusing the match scores of each unimodality- audio and video. Since, different classifiers are used during classification in each modality (nearest neighbor vs. GMM), the confidence scores of each modality are normalized with a non-linear function to compensate this mis-match. Sigmoid function is used for this purpose. After normalizing the match scores, they are fused via sum rule. Since, there is no common validation set available to the evaluation participants, no prior information about the performances of audio-only and video-only testing is used. Therefore, audio and video modalities are equally weighted.

3 Experiments

In this section the evaluation data is described and the experimental results are presented.

3.1 Face Recognition Experiments

The evaluation data for visual identification task in CLEAR evaluations consists of short video sequences taken from the Seminar 2005 database recorded in the various CHIL sites. There are 26 individuals in the database.

In face recognition experiments, face images are aligned according to eye-center coordinates and scaled to 40x32 pixels resolution. Only every five frame that has the eye coordinate labels is used for training and testing. The aligned image is then

divided into 8x8 pixels resolution non-overlapping blocks making 20 local image blocks. From each image block 10 unit norm DCT-0 coefficients are extracted and they are concatenated to construct the 200-dimensional final feature vector. The classification is performed using nearest neighbor classifier. L1 norm is selected as the distance metric, since it has been observed that, it consistently gives the best correct recognition rates when unit norm DCT-0 coefficients are used. The distance values are converted to the matching scores by using the Min-Max rule. The normalized matching scores are accumulated over different camera views and over image sequence. The identity candidate that has the highest score is assigned as the identity of the person.

The false identification rates for different training and testing durations can be seen in Table 1. As can be observed from the table, the increase in the training segments' duration or in the testing segments' duration decreases the false identification rate.

Table 1. False visual identification rates

Test Duration (sec)	Segments	Train A (15 sec)	Train B (30 sec)
1	613	46.8%	40.1%
5	411	33.6%	23.1%
10	289	28.0%	20.4%
20	178	23.0%	16.3%

3.2 Speaker Identification Experiments

The evaluation data in CHIL 2005 Spring Evaluation was used as our development data set. This data set has been carried out on the union of the UKA-ISL_Seminar_2003 and UKA-ISL_Seminar_2004 databases. Non-speech segments have been manually removed both from the training and the testing segments. There are two microphone conditions: Closed-Talking-Microphone (CTM) and Microphone Array (ARR). The duration and number of segments selected for the training and testing as improving our system is described in Table 2.

Table 2. Description of development data

	Duration (sec)	CTM Segments	ARR Segments
Train A	30	11	11
Train B	15	11	11
Test	5	1100	682

In order to find an optimal number of Gaussians for a speaker model, we conducted several speaker identification experiments with different number of Gaussians in a speaker model.

Table 3. False identification rate with different number of Gaussians for 30-sec training

Number of Gaussians	64	128	256
Miss Classification Rate	0.36%	0.27%	0.36%

Table 4. False identification rate with different number of Gaussians for 15-sec training

Number of Gaussians	16	32	64
Miss Classification Rate	2.82%	2.00%	2.23%

According Table 3 and Table 4, we choose to use 128 Gaussians for the 30-second training condition and 32 Gaussians for the 15-second training condition.

Table 5 shows the system performance improvement by applying reverberation compensation and feature warping under the 30-seconds training condition. We can see from the table that signification improvement was achieved for both the CTM and ARR microphone conditions.

Table 5. Performance improvement by reverberation compensation and feature warping

	Baseline	RC+Warp	Relative Improvement
CTM	0.27%	0.18%	33.3%
ARR	6.74%	3.08%	54.3%

Table 6. False audio identification rate in clear 2006 evaluation

Test Duration (sec)	Segments	Train A (15 sec)	Train B (30 sec)
1	613	23.7%	14.4%
5	411	7.8%	2.2%
10	289	7.3%	1.4%
20	178	3.9%	0%

The overall system performances for different training and testing durations are given in Table 6. It is apparent that, as the duration of training or testing segments increases, the error rate decreases.

3.3 Multimodal Identification Experiments

The evaluation data for multimodal identification task in CLEAR evaluations consists of short audio-video sequences taken from the Seminar 2005 database recorded in the various CHIL sites. There are 26 individuals in the database.

To perform multimodal identification, the individual modality matching scores are fused as explained in Section 2.3. The experimental results can be seen in Table 7. Again, it can be observed that the increase in training segments' duration or in testing segments' duration decreases the false identification rate. Due to equal weighting of each modality, the multimodal identification results are higher than the visual-only results and lower than the audio-only results.

Table 7. False multi-modal identification rates

Test Duration (sec)	Segments	Train A (15 sec)	Train B (30 sec)
1	613	43.1%	35.7%
5	411	29.2%	19.7%
10	289	23.9%	16.6%
20	178	20.2%	12.4%

4 Conclusions

In this paper, we presented the person identification systems that have been developed at the Interactive Systems Labs for the CLEAR evaluations. The experimental results showed that, speaker identification performs better than face recognition for person identification in smart environments. The main reason for the performance difference is the low video quality. Multimodal identification results perform worse than speaker identification results. This result is expected, since the audio and video data are weighted equally, due to missing priori performance information which is caused by lack of a common validation set.

Acknowledgements

We would like to thank Kenichi Kumatani for his contributions to the ISL visual identification evaluation effort.

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909.

References

[1] Computers in the Human Interaction Loop –CHIL, http://chil.server.de/

[2] H.K. Ekenel, A. Pnevmatikakis, "Video-Based Face Recognition Evaluation in the CHIL Project – Run 1", 7th International Conference Automatic Face and Gesture Recognition (FG2006), Southampton, UK, April 2006.

[3] H.K.Ekenel, R. Stiefelhagen, "Local Appearance based Face Recognition Using Discrete Cosine Transform", 13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey, September 2005.

[4] H.K. Ekenel, R. Stiefelhagen, "Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization", CVPR Biometrics Workshop, New York, USA, June 2006.
[5] A. Pentland, B. Moghaddam, T. Starner and M. Turk, “View based and modular eigenspaces for face recognition”, Proceedings of IEEE CVPR, pp. 84-91, 1994.
[6] Q. Jin, Y. Pan and T. Schultz, "Far-field Speaker Recognition", International Conference on Acoustic, Speech, and Signal Processing (ICASSP) 2006.
[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification", Proc. Speaker Odyssey 2001 conference, June 2001.
[8] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy and R. Gopinath, "Short-time Gaussianization for Robust Speaker Verification", in Proc. ICASSP, 2002.
[9] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", Speech Communication, Vol. 17, No. 1-2, p. 91-108, August 1995.