Multi-modal Person Identification in a Smart Environment

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

[1]Interactive Systems Labs (ISL), Universität Karlsruhe (TH), 76131 Karlsruhe, Germany

{ekenel,mika.fischer,stiefel}@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 a detailed analysis of multi-modal fusion for person identification in a smart environment. The multi-modal system consists of a video-based face recognition system and a speaker identification system. We investigated different score normalization, modality weighting and modality combination schemes during the fusion of the individual modalities. We introduced two new modality weighting schemes, namely, the cumulative ratio of correct matches (CRCM) and distance-to-second-closest (DT2ND) measures. In addition, we also assessed the effects of the well-known score normalization and classifier combination methods on the identification performance. Experimental results obtained on the CLEAR 2007 evaluation corpus, which contains audio-visual recordings from different smart rooms, show that CRCM-based modality weighting improves the correct identification rates significantly.

1. Introduction

Biometric person identification problem has attracted significant research efforts that have been mainly fueled by security applications. Recently, person identification for smart environments has become another application area of significant interest [1,2,3]. Sample application areas can be a smart video-conferencing system that can recognize the speaker; a smart lecture or meeting room, where the participants can be identified automatically and their behaviors can be analyzed throughout the meeting or the lecture. 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 smart room 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 a 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. A sample smart room layout and sample images from different smart rooms are shown in Figures 1 and 2.
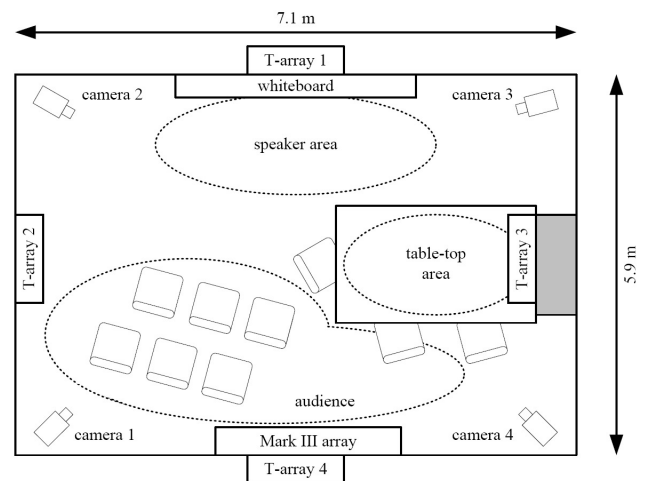

Figure 1. A smart room layout.

In this paper, in order to obtain a robust person identification system, we investigate in detail the multi-modal fusion strategies for person identification in a smart room environment. For score normalization, we compare the min-max [4] and hyperbolic tangent methods [4,5]. For modality weighting, we introduce two new schemes, namely, the cumulative ratio of correct matches (CRCM)

Figure 2. Sample images from different smart rooms.

and distance-to-second-closest (DT2ND) measures which are based on the separability of the best two matches. The classifiers are combined using the sum rule, product rule and max rule approaches [6]. We conduct experiments on a data corpus that has been collected at different smart rooms. The experimental results indicate that the CRCM modality weighting approach provides a significant increase in the correct identification rate.

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 briefly explain the single modality person identification systems and list the investigated fusion strategies.

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 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],

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

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 the ten best matches among 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". 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:

$$\varepsilon(x; \lambda) = 0.1\lambda e^{-\lambda x} \text{ with } \lambda = 0.05 . \qquad (2)$$

The weights are then computed as the cumulative distribution function:

$$\omega_{DT2ND}(x; \lambda) = 1 - e^{-\lambda x} . \qquad (3)$$

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:

$$s = \arg\max_i \{L(Y|\Theta_i)\} \qquad Y = (y_1, y_2, \cdots, y_N), \qquad (4)$$

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

$$\Theta_i = \sum_{m=1}^{M} \lambda_m N(X, U_m, \Sigma_m) \qquad i = 1, 2, \cdots, S, \qquad (5)$$

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, U_m, and Σ_m are the weight, mean, and diagonal covariance matrix of the mth Gaussian distribution.

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 the "frame-based score competition (FSC)" 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:

$$LL(X|\Theta_k) = \sum_{n=1}^{N} LL(x_n|\Theta_k) = \sum_{n=1}^{N} \max\{LL(x_n|\Theta_{k,Chj})\}_{j=1}^{64} . \quad (6)$$

This competition process based on multiple channels differs from the standard scoring process based on one channel in that the per-frame log likelihood scores are not necessarily derived from the same microphone.

2.3. Fusion

We investigated the effects of three main steps of the fusion process. They are: Score normalization, modality weighting and modality combination.

Score normalization is the first step in the process of modality fusion. Due to the different ways of feature extraction and classification, the distribution of the resulting scores may differ between the modalities. In our case, the face recognition system generates accumulated min-max normalized scores, whereas the speaker identification system provides likelihood scores. In order to combine these scores, we utilized and compared two well-known normalization methods, namely the min-max and hyperbolic tangent normalization methods. The min-max normalization is the normalization method we also used to transform the distance values to the normalized confidence scores in the face recognition system and can be calculated as in Equation 1 without the need of subtracting the obtained division value from one, since the modality scores are directly proportional to the modality confidences. Hyperbolic tangent normalization non-linearly maps the confidence scores to the (0,1) range and is calculated as,

$$nc = \frac{1}{2}\left[\tanh\left(0.01\frac{(c - mean(C))}{std(C)}\right) + 1\right], \qquad (7)$$

where, nc denotes normalized confidence score, c denotes confidence score of an identity candidate in the database, and C denotes the vector that contains the confidence scores of all the identity candidates in the database.

Modality weighting is the second step in the fusion process. In this study, we introduced 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 to 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. Figures 3 and 4 show the obtained correct match distribution over the confidence differences and the corresponding weighting model for the face recognition system, respectively. 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.
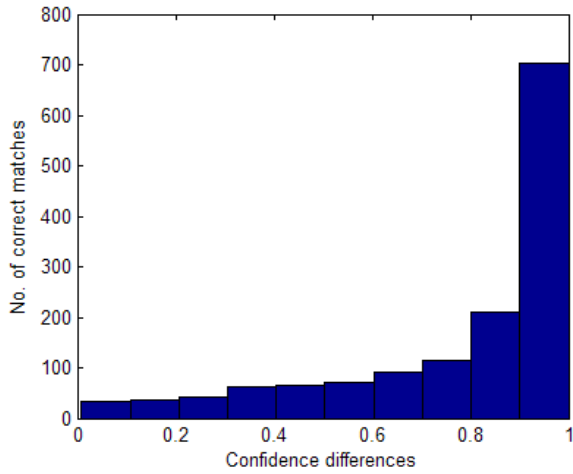

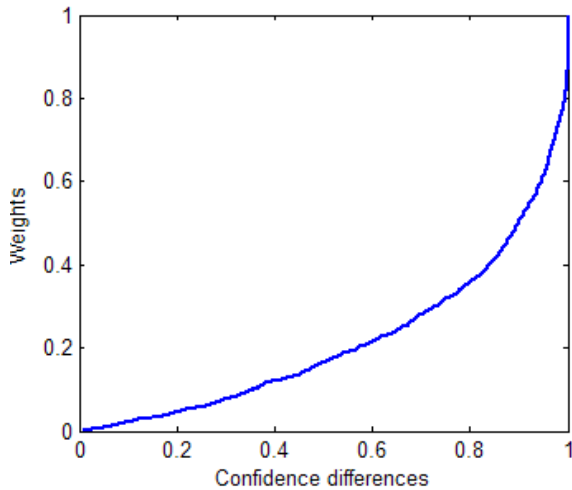**Figure 3.** The distribution of the correct matches.

Figure 4. The weighting model.

As a second way of modality weighting, we adapted the DT2ND to modality fusion, by simply using the confidence differences instead of using the differences between the normalized distance scores in Equation 3. In addition to the adaptive modality weighting schemes, we also took into account the individual correct identification rate of each modality that is obtained on the validation set and assigned fixed weights to the modalities according to their performance.

Finally, we combined the modalities using the well-known classifier combination methods: Sum rule, product rule and max 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 2. 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 Sequeces
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 labelled 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 false 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 false identification rate decreases. 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.

These results are used to determine the fixed weights that each modality receives. It is done in two different ways. The first way 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 is by determining the weights indirectly 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.

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

	Face Reco. (%)	Speaker Id. (%)
A1	8.6	43.6
A5	0.9	32.1
A10	0.0	10.7
A20	0.0	7.1
B1	5.7	38.9
B5	0.0	15.2
B10	0.0	1.8
B20	0.0	0.0

3.2. Experiments on the Test Set

The false identification rates of the face recognition and speaker identification systems obtained on the test set are given in Table 3. Similar to the obtained results on the validation set, as the duration of training or testing increases the false identification rate decreases. 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 3 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 3. False identification rates of the individual modalities on the test set.

	Face Reco. (%)	Speaker Id. (%)
A1	15.4	58.1
A5	9.2	30.4
A10	6.7	8.0
A20	5.4	3.6
B1	10.7	58.8
B5	5.6	21.7
B10	5.4	3.6
B20	3.6	0.9

3.3. Fusion Experiments

In the following subsections, we examine the steps of the fusion process.

3.3.1 Comparison of Score Normalization Methods

In the first fusion experiment, we compared the min-max and hyperbolic tangent score normalization methods. For modality weighting, we used the fixed weights which are directly proportional to the correct identification rates obtained on the validation set. We used the sum rule for classifier combination. The resulting false identification rates are shown in Table 4. As can be observed, there is no significant performance difference between using min-max or hyperbolic tangent methods for score normalization. Although, we used a very primitive fixed weighting scheme for this experiment, in most of the training-testing duration combinations the false identification rates are lower than the ones obtained by the individual modalities.

Table 4. Results of min-max and hyperbolic tangent score normalization methods.

	min-max	tanh
A1	15.2	14.2
A5	8.9	8.3
A10	5.8	4.9
A20	5.4	5.4
B1	10.2	9.6
B5	5.1	4.0
B10	4.5	3.1
B20	2.7	1.8

3.3.2   Comparison of Modality Weighting Methods

In the second fusion experiment, we analyzed the modality weighting schemes. We first compared the fixed weighting schemes at which the weights are either directly proportional to the correct classification rate or indirectly proportional to the false identification rate obtained on the validation set. These fixed modality weighting schemes are named as DPC and IPF respectively. The results are presented in Table 5. As already mentioned in Section 3.3.1, even with these primitive weights, in most of the training-testing duration combinations the false identification rates are lower than the ones obtained by the individual modalities.

Table 5. Results of fixed weighting schemes.

	DPC		IPF	
	min-max	tanh	min-max	Tanh
A1	15.2	14.2	15.4	15.2
A5	8.9	8.3	9.2	9.2
A10	5.8	4.9	6.7	6.7
A20	5.4	5.4	5.4	5.4
B1	10.2	9.6	10.6	10.1
B5	5.1	4.0	5.6	5.6
B10	4.5	3.1	5.4	5.4
B20	2.7	1.8	2.7	1.8

The results with the more sophisticated adaptive modality weighting schemes are given in Table 6. Compared to the Table 5, CRCM provides a significant drop in false identification rates. Although DT2ND has not been specifically designed for modality weighting, it still provides slightly better results. 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.

Table 6. Results of adaptive weighting schemes.

	CRCM		DT2ND	
	min-max	tanh	min-max	tanh
A1	13.7	19.0	15.3	14.2
A5	6.5	8.3	8.9	7.4
A10	1.8	2.7	6.7	5.4
A20	0.9	1.8	5.4	5.4
B1	10.4	15.3	10.4	10.0
B5	2.7	3.1	5.4	5.1
B10	1.3	0.9	5.4	4.0
B20	0.9	0.0	2.7	1.8

We also tried combining the fixed weights with the adaptive weights. However, as can be observed from Table 7, there is no significant performance difference between the CRCM and DPC+CRCM results. The performance degrades with IPF + CRCM. The reason is the hard modality weighting in IPF. Since, on the validation set at some training-testing duration combinations, face recognition reached 0% false identification rate, at these combinations only the face recognition system's decision is trusted.

Table 7. Results of combined CRCM and fixed weighting schemes.

	DPC + CRCM		IPF + CRCM	
	min-max	tanh	min-max	tanh
A1	13.3	16.8	13.3	13.5
A5	6.5	7.1	8.3	7.1
A10	1.8	2.2	6.7	6.7
A20	0.9	1.8	5.4	5.4
B1	10.1	12.8	10.1	10.3
B5	2.7	3.1	5.6	5.6
B10	1.3	0.9	5.4	5.4
B20	0.9	0.0	0.9	0.0

Table 8. Results of combined DT2ND and fixed weighting schemes.

	DPC + DT2ND		IPC + DT2ND	
	min-max	tanh	min-max	tanh
A1	15.4	14.7	15.4	15.1
A5	8.9	8.0	9.2	8.9
A10	6.7	5.4	6.7	6.7
A20	5.4	5.4	5.4	5.4
B1	10.5	9.8	10.6	10.1
B5	5.4	5.1	5.6	5.6
B10	5.4	4.0	5.4	5.4
B20	2.7	1.8	2.7	1.8

3.3.3 Comparison of Classifier Combination Methods

Finally, we combined the weighted scores using the sum rule, product rule and max. rule. From Table 9, it can be seen that the max. rule operates better on the min-max normalized confidence scores. Sum rule and max rule are found to perform slightly better than the product rule. However, no big difference is observed in the false identification rates.

Table 9. Results of classifier combination methods.

	Sum Rule		Product Rule		Max. Rule	
	min-max	tanh	min-max	tanh	min-max	tanh
A1	13.3	16.8	13.3	13.2	15.2	31.2
A5	6.5	7.1	7.4	7.4	7.8	16.5
A10	1.8	2.2	4.5	4.5	2.2	6.2
A20	0.9	1.8	4.5	4.5	0.9	2.7
B1	10.1	12.8	9.6	9.3	11.8	27.3
B5	2.7	3.1	3.8	3.8	3.8	10.9
B10	1.3	0.9	2.7	3.1	0.9	2.2
B20	0.9	0.0	1.8	1.8	0.0	0.0

4. Conclusions

In this paper, we presented a detailed analysis of multi-modal fusion for person identification in a smart environment. We found no significant performance difference between well-known score normalization or classifier combination approaches. We observed that the modality weighting has the major impact on the correct identification rate. We proposed an adaptive modality weighting model that is derived from the confidence differences between the best two matches. It is named as cumulative ratio of correct matches (CRCM) and the weighting model is computed by taking the cumulative sum of the number of correct matches achieved at a confidence difference between the best two matches. In Table 10, the false identification rates of the individual modalities and the multi-modal system are listed. The multi-modal system included in the table uses min-max normalized confidence scores, CRCM modality weighting and the sum rule. 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 10. Results of individual modalities and the multi-modal system.

	Face Reco. (%)	Speaker Id. (%)	Fusion (%)
A1	15.4	58.1	13.7
A5	9.2	30.4	6.5
A10	6.7	8.0	1.8
A20	5.4	3.6	0.9
B1	10.7	58.8	10.4
B5	5.6	21.7	2.7
B10	5.4	3.6	1.3
B20	3.6	0.9	0.9

References

[1] J. Vendrig, M. Worring, "Multimodal Person Identification in Movies", in Proceedings of the Intl. Conf. on Image and Video Retrieval, pp. 175-185, 2002.

[2] E. Erzin et al., "Multimodal Person Recognition for Human-Vehicle Interaction", IEEE Multimedia, Vol.13(2), pp.18–31, 2006.

[3] T. J. Hazen et al., "Multi-Modal Face and Speaker Identication for Mobile Devices", in R. I. Hammoud, B. Abidi, and M. Abidi, eds., *Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems*, Springer-Verlag, Heidelberg, Germany, April 2007.

[4] R. Snelick 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] R. Brunelli, D. Falavigna, "Person Identification Using Multiple Cues", IEEE Trans. on Pattern Analysis and Machine Intelligence, 17(10):955–966, 1995.

[6] J. Kittler et al., "On combining classifiers", IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(3):226–239, 1998.

[7] 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.

[8] H.K. Ekenel, R. Stiefelhagen, "Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization", IEEE CVPR Biometrics Workshop, New York, USA, June 2006.

[9] J. Stallkamp, "Video-based Face Recognition Using Local Appearance-based Models", Thesis report, Universität Karlsruhe (TH), Nov. 2006.

[10] S. Furui. Recent Advances in Speaker Recognition. *Pattern Recognition Letters*, 18:859–872, 1997.

[11] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," Speech Communication, Vol. 17, No. 1-2, p. 91-108, August 1995.

[12] Q. Jin, Y. Pan and T. Schultz, "Far-field Speaker Recognition", International Conference on Acoustic, Speech, and Signal Processing (ICASSP) 2006.

[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification", Proc. Speaker Odyssey 2001 conference, June 2001.

[14] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy and R. Gopinath, "Short-time Gaussianization 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/.