4

Multimodal Person Identification

Aristodemos Pnevmatikakis[1], Hazım K. Ekenel[2], Claude Barras[3], Javier Hernando[4]

[1] Athens Information Technology, Peania, Attiki, Greece,
[2] Universität Karlsruhe (TH), Interactive Systems Labs, Fakultät für Informatik, Karlsruhe, Germany
[3] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), Orsay, France
[4] Universitat Politècnica de Catalunya, Barcelona, Spain

Person identification is of paramount importance in security, surveillance, human-computer interfaces, and smart spaces. All these applications attempt the recognition of people based on audiovisual data. The way the systems collect these data divides them into two categories:

- Near-field systems: Both the sensor and the person to be identified focus on each other.
- Far-field systems: The sensors monitor an entire space in which the person appears, occasionally collecting useful data (face and/or speech) about that person. Also, the person pays no attention to the sensors and is possibly unaware of their existence.

Near-field person identification systems require the person's attention. The person is aware of the sensors' location, approaches the sensors and offers the system samples of his or hers face and voice. Such systems are obtrusive, but the audiovisual streams thus collected have a high signal-to-noise ratio. The images depict faces of high resolution, approximately frontal in pose and neutral in expression, while the sound is almost free of the detrimental effects of the room's acoustics: reverberation and attenuation. The typical application of such systems is access control, where these systems are expected to offer close to perfect recognition rates. A typical example of such system using near infrared face recognition can be found in [12]

Far-field person identification systems [15] employ audiovisual sensors scattered in the space the person is expected to be in. The systems do not anticipate that the person will acknowledge the existence of the sensors, nor behave in any way that will facilitate the collection of noise-free audiovisual streams. Hence, far-field data streams are corrupted with noise: The video streams contain faces viewed from arbitrary angles, distances, under arbitrary illumination, and possibly, depending on the environment of the deployment, with arbitrary expressions. Similarly, the sound streams suffer from reverberations, large attenuations, and the coexistence of background sounds. The audiovisual environment changes dramatically as the person moves around the space. As a result, the faces collected are tiny (typically of 10 pixels between the eyes; see Fig. 4.1) and with gross variations in pose, expression, and illumination. The speech samples are also attenuated, corrupted with all sorts of background noises (occasionally entirely masked by them) and reverberations. Nevertheless, far-field systems have three features that allow them to offer usable recognition rates:

- the use of multiple sensors (many cameras and microphones),
- the abundance of training data that are audiovisual streams similar to those on which the system is expected to operate,
- the possibly long periods of time that the systems can collect data on which they are going to base their identity decision.

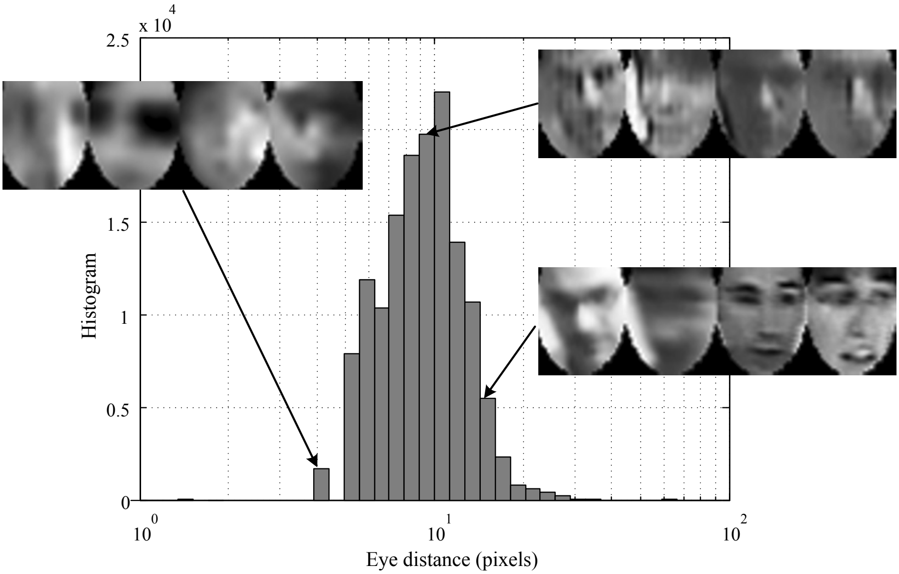

Fig. 4.1. Histogram of the distances between the eyes of the faces to be recognized in typical far-field scenarios. The faces are those of the CLEAR 2006 evaluation. Typical faces at three different eye distances are also shown.

The goal of the far-field multimodal person identification systems built within the CHIL Consortium is to equip smart spaces with a means of giving an identity to each person working in the space. Contrary to the more usual verification problems, where an identity claim is verified over hundreds of individuals, the CHIL systems perform identification of the people present from a rather limited set of possible identities. This is not a restriction in the intended application domain; the number of the people called to work in a given smartspace is indeed limited.

In order to measure progress in these two categories of video-to-video person identification systems, it is important to utilize formal evaluation protocols. This is a well-established procedure for near-field face recognition, but to the authors' knowledge, it is only through the CLEAR (2006 and 2007) person identification evaluations [18, 19] and their predecessor Run-1 evaluations [8] initiated by the CHIL project that a formal evaluation campaign focusing on far-field multimodal person identification has been launched.

The following sections outline, the different approaches of the CHIL Consortium for audio-only, video-only, and multimodal person identification systems, leading to the lessons learned and progress made in the field of far-field multimodal person identification over the course of the project.

4.1 Speaker Identification

All systems construct speaker models using Gaussian mixture models (GMM) of some short-term speech features. The features are derived from speech frames using the Mel frequency cepstral coefficients (MFCC) [11, 1, 2, 17, 16, 7, 6], the perceptually-weighted linear prediction coefficients (PLP) [17], or the frequency filtering (FF) [14, 13] approaches. Postprocessing of the features like cepstral mean normalization and feature warping are included in some systems [1, 2, 7, 6]. The models for the different speakers are trained either using the available training speech for each speaker independently, or by maximum a posteriori (MAP) adaptation of a universal background model (UBM) [1, 2]. The latter allows much larger GMMs by pooling all available speech and training a large UBM GMM.

During the recognition phase, a test segment is scored against all possible speakers and the speaker model with the highest log-likelihood is chosen.

The systems explore different options of utilizing the multiple audio streams. These fall into two categories: those that attempt to improve the quality of the signal prior to constructing or testing models, utilizing some sort of beamforming [2, 13], and those that attempt postdecision fusion of the classification outcomes derived from each standalone microphone [13].

Beamforming is not the only option for preprocessing the signals. Since the recordings are far-field and corrupted by all sorts of background noises, they can be preprocessed by a speech activity detector (SAD) to isolate speech. Also, principal components analysis (PCA) can be employed on the features prior to model construction or testing, efficiently combining different feature extraction approaches. The PCA transformation matrix can be derived globally from all speakers, or a per-speaker approach can be followed [17].

4.2 Face Identification

The face identification systems extract the faces from the video streams, aided by face labels. Only in the Run-1 evaluations were the faces automatically detected [8].

The face labels provide a face bounding box and eye coordinates every 200 ms. Some systems use only the marked faces [8], while others attempt to interpolate between the labels [17]. The labels are quite frequent in the CLEAR 2007 data set. That is not the case in the CLEAR 2006 data set, where face bounding boxes are provided every 1 sec [19]. For this data set, faces can be collected from the nonlabeled frames using a face detector in the neighborhood of the provided labels. Experiments with a boosted cascade of simple classifiers (Viola-Jones detector) [20] have shown enhanced performance over the simple linear interpolation [15].

The systems geometrically normalize the extracted faces, based either on the eyes [15, 16, 8] or on the face bounding box [15, 17, 7]. In the latter case, the faces are just scaled to some standard size. Some of the systems battle face registration errors by generating additional images, either by perturbing the assumed eye positions around the marked ones [8], or by modifying the face bounding box labels by moving the center of the bounding box by 1 pixel and changing the width or height by ± 2 pixels [7]. Geometric normalization is followed by further processing, aiming at intensity normalization. One approach is to normalize the integer values of the luminance to a mean of 128-mean and a standard deviation of 40 (luminance normalization) [17]. Other, more aggressive approaches like histogram equalization and edginess have been used in the past but have been abandoned as they degrade performance in the case of pose or expression variations [15, 8].

The processed faces obtained from the CLEAR 2007 data set lie on a highly non-linear manifold, making recognition difficult. This is shown in Fig. 4.2, where the first two PCA dimensions of the gallery faces for two different individuals are depicted. In this case, the manifold shows faces from person 1 ranging from left profile to frontal and finally to right profile, as the first PCA coefficient varies from -1,500 to +1,500. The same holds for person 2, only in this case there are not many frontal faces to occupy the range around the zero value of the first PCA coefficient. Hence, the first PCA coefficient is for pose and not person discrimination. Person 1 is discriminated from person 2 mainly by the second PCA coefficient, at the threshold value of 500, although there are many outliers from both people, since their projections fall into each Other's vicinity.

Different features are extracted by the different systems. They are all classified using a nearest-neighbor classifier, although the distance from class centers has been used by some systems in the past [15, 16, 8]. Obviously from Fig. 4.2, this is not a good choice, since the projected faces of a class are quite spread out and many times form disjoint clusters.

One approach performs block-based discrete cosine transform (DCT) to non-overlapping blocks of size 8×8 pixels [5, 4, 7, 9]. The obtained DCT coefficients are then ordered according to the zigzag scan pattern. The first coefficient is discarded for illumination normalization and the remaining first 10 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. The global feature vector is generated by concatenating the local feature vectors.

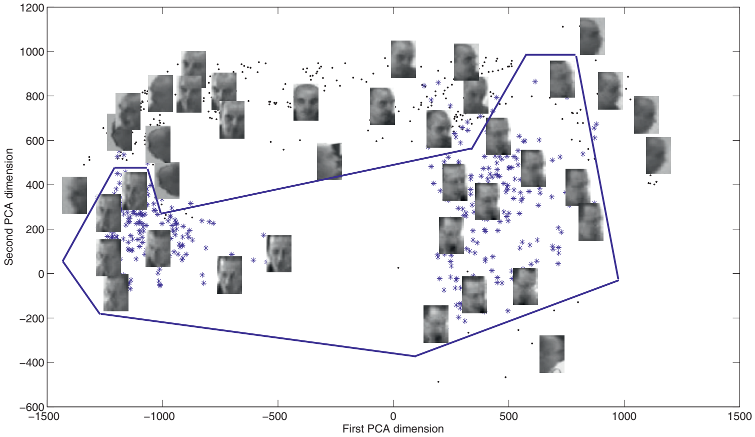

Fig. 4.2. 2D face manifold for people 1 and 2 of the CLEAR 2007 data set and representative faces centered on their respective projections. The faces of the two people lie on arcs, with the first PCA coefficient characterizing the pose and not the person. The projected faces of person 1 are depicted with black dots, while those of person 2 with blue stars. The characteristic faces belonging to person 2 are enclosed by a blue polygon.

Alternatively, PCA can be used to obtain features in a face subspace [15, 17, 16, 8]. The distance employed in the classifier is a modification of the weighted Euclidean, with the weights of each dimension depending on the neighborhood of the particular gallery point. The closest neighbors (using Euclidean distance) to the given gallery point are used to estimate the scatter matrix of gallery points in the vicinity of that gallery point [17]. The gallery point-dependent weights are the eigenvalues of that scatter matrix. Although this estimation of point-dependent weights is computationally expensive, it is performed once at system training and the weights for each of the projected gallery images are stored to be used in the recognition phase.

Linear discriminant analysis (LDA) has also been used in the past [15, 16, 8], without much success, as the face manifolds obtained under unconstrained conditions are nonseparable in a linear way (see the example of Fig. 4.2). Subclass LDA can address this problem [17]. Hierarchical clustering trees are used to automatically generate subclasses corresponding to face clusters belonging to the same class, by pruning the tree at some distance value.

Gaussian modeling of intrapersonal variations is also used to evaluate the probability that the difference of a gallery face from a probe face is indeed intrapersonal [17]. Forming all difference images is not computationally feasible; hence, a selection of the faces to be used is performed by grouping the gallery images of any person using hierarchical clustering trees. The trees are constructed using the projected gallery images onto a PCA subspace. For every person, some projected images are selected as the median of every cluster obtained by the trees. The intrapersonal differences are formed and modeled at the reduced dimension of the PCA subspace.

The fusion of the decisions obtained from all camera views within the testing period is performed using the weighted-sum rule [15, 17, 16, 7, 8, 6]. The weights are calculated using various metrics based on the distance of the best and second-best matches. Some systems also fuse decisions from different feature extraction methods [15, 17, 16], as each of these is best suited for different types of impairments in the face images.

4.3 Multimodal Person Identification

All multimodal person identification systems perform postdecision fusion. The weighted-sum rule is again utilized, with the individual modality confidences being used for the calculation of the weights.

The individual modality confidences are calculated based on the observation that the difference of the confidences between the closest and second-closest matches is generally smaller in the case of a false classification than in the case of a correct classification. A method named the *cumulative ratio of correct matches* (CRCM), which uses a nonparametric model of the distribution of the correct matches with respect to the confidence differences between the best two matches, is used. This way, the classification results with a greater confidence difference between the two best matches receive higher weights [5, 4, 7, 9]. Alternatively, the ratio between the closest and the second-closest matches [15, 17, 16], or a histogram equalization of the monomodal confidence scores [10, 3], can be utilized.

4.4 Lessons Learned

As seen in the previous sections, the unimodal person identification systems evolved in the CHIL Consortium have explored different options regarding their three parts: data extraction and preprocessing, feature extraction, and classification. Following many experiments and four evaluations (two internal to CHIL and two evaluations conducted with the CLEAR workshops 2006 and 2007, which were open to the scientific community; see also Chapter 15), many lessons have been learned.

The following can be concluded regarding audio person recognition:

- Speech extraction and preprocessing: The experiments on the use of SAD to preprocess the audio for speech extraction are not conclusive; most systems just extract features from the whole audio duration. The preprocessing of the multiple microphone channels to produce a single, hopefully cleaner, audio signal using beamforming degrades performance. Postdecision fusion of the recognition outcomes from the different channels is the way to utilize the multiple audio sources.

- Feature extraction: The FF features are very promising. The standalone PLP features, or those obtained by the combination of PLP and MFCC into a single feature vector using PCA, are better than standalone MFCC. Feature postprocessing is counterproductive in the context of the CHIL seminars, where the speaker is generally recorded in a stable acoustic context.
- Classifier: All systems employ a Bayesian classifier based on the GMMs of the features. The use of a UBM with MAP adaptation is better than estimating speaker-specific models directly. Training the UBM by pooling all the training data outperforms using other training data or a direct MLE training of the target models.

The conclusions for face recognition are as follows:

- Face extraction and normalization: It is better to extract faces from only the labeled frames, without any interpolation. Selecting just the frontal faces (using the provided labels) is detrimental to performance. There are not enough experiments with the different geometric normalization approaches to be conclusive.
- Feature extraction: The feature extraction methods can be compared only under the same face extraction and normalization methods. This makes comparison very difficult. The best system employs local appearance-based face recognition using DCT. Experiments using the same face extraction and normalization methods led to the following ranking of feature extraction methods:
  - Intrapersonal modeling (Bayesian) > Subclass LDA > PCA > LDA
  - LDA with Kernel PCA combination > Kernel PCA > LDA > PCA
- Classifier: Since CLEAR 2006,we have known that the nearest-neighbor classifier, albeit slower, greatly outperforms the distance from the class centers one. Also, there is lot to gain by exploiting the optimum distance metric for each feature extraction method. While some systems only used Manhattan or Euclidean distance metrics, experiments have shown performance gain when using modifications of the Mahalanobis distance metric for PCA-based methods, or the cosine for LDA-based methods.

Finally, regarding multimodal systems, the one with the best audio subsystem always outperforms the rest. The results show that multimodal fusion provides an improvement in the recognition rate over the unimodal systems in all train/test conditions.

Recognition performance greatly increased in the last two years of the project, where results are somewhat comparable. In CLEAR 2006, the task involved 26 people and the segments had been selected so that speech was present in them, paying no attention to the faces. In CLEAR 2007, the people to be recognized increased to 28 and the choice of the segments was more balanced across the two modalities. In both evaluations, two training conditions (15 and 30 seconds) and four testing conditions (1, 5, 10, and 20 seconds) were selected. The best performance obtained in the two evaluations is summarized in Fig. 4.3.

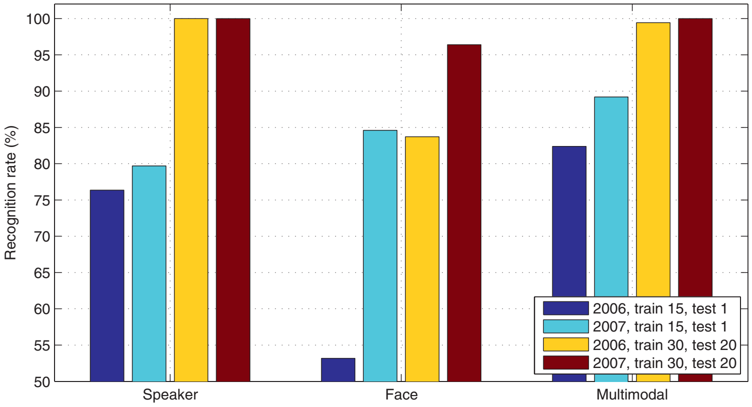

Fig. 4.3. Performance evolution of the person identification systems in the CLEAR 2006 and 2007 evaluations. Two out of the eight conditions are shown per evaluation; the shortest training and testing (15 and 1 seconds, respectively) and the longest training and testing (30 and 20 seconds, respectively).

The most impressive performance boost has been achieved in face identification. Significant improvements have also been achieved in speaker and multimodal identification for the short training and testing segments.

References

1. C. Barras, X. Zhu, J.-L. Gauvain, and L. Lamel. The CLEAR'06 LIMSI acoustic speaker identification system for CHIL seminars. In *Multimodal Technologies for Perception of Humans. First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*, LNCS 4122, pages 233–240, 2006.
2. C. Barras, X. Zhu, C.-C. Leung, J.-L. Gauvain, and L. Lamel. Acoustic speaker identification: The LIMSI CLEAR'07 system. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 233–239, Baltimore, MD, May 8-11 2007.
3. P. Ejarque, A. Garde, J. Anguita, and J. Hernando. On the use of genuine-impostor statistical information for score fusion in multimodal biometrics. *Annals of Telecommunications, Special Issue on Multimodal Biometrics*, 62(1-2):109–129, Apr. 2007.
4. H. K. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen. Multi-modal person identification in a smart environment. In *CVPR Biometrics Workshop 2007, IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Jun. 2007.
5. H. K. Ekenel, M. Fischer, and R. Stiefelhagen. Face recognition in smart rooms. In *Machine Learning for Multimodal Interaction, Fourth International Workshop, MLMI 2007*, Brno, Czech Republic, Jun. 2007.
6. H. K. Ekenel and Q. Jin. ISL person identification systems in the CLEAR evaluations. In *Multimodal Technologies for Perception of Humans. First International Evaluation*

Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, LNCS 4122, pages 249–257, Southampton, UK, Apr. 6-7 2007.

7. H. K. Ekenel, Q. Jin, and M. Fischer. ISL person identification systems in the CLEAR 2007 evaluations. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 256–265, Baltimore, MD, May 8-11 2007.
8. H. K. Ekenel and A. Pnevmatikakis. Video-based face recognition evaluation in the CHIL project – run 1. In *7th IEEE International Conference on Automatic Face and Gesture Recognition, FG06*, pages 85–90, 2006.
9. H. K. Ekenel and R. Stiefelhagen. Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In *CVPR Biometrics Workshop*, New York, Jun. 2006.
10. M. Farrús, P. Ejarque, A. Temko, and J. Hernando. Histogram Equalization in SVM Multimodal Person Verification. In *ICB*, pages 819–827, 2007.
11. J.-L. Gauvain and C. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, Apr. 1994.
12. S. Z. Li, L. Zhang, S. Liao, X. Zhu, R. Chu, M. Ao, and R. He. A near-infrared image based face recognition system. In *7th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2006)*, pages 455–460, Southampton, UK, April 2006.
13. J. Luque and J. Hernando. Robust speaker identification for meetings: UPC CLEAR07 meeting room evaluation system. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 266–275, Baltimore, MD, May 8-11 2007.
14. J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrus, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando. Audio, video and multimodal person identification in a smart room. In *Multimodal Technologies for Perception of Humans. First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*, LNCS 4122, pages 258–269, Southampton, UK, Apr. 6-7 2006. Springer-Verlag.
15. A. Pnevmatikakis and L. Polymenakos. *Far-Field Multi-Camera Video-to-Video Face Recognition*. I-Tech Education and Publishing, 2007.
16. A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. A decision fusion system across time and classifiers for audio-visual person identification. In *Multimodal Technologies for Perception of Humans, Proceedings of the first International CLEAR evaluation workshop, CLEAR 2006*, LNCS 4122, pages 223–232, Southampton, UK, Apr. 6-7 2006.
17. A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. The AIT multimodal person identification system for CLEAR 2007. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 221–232, Baltimore, MD, May 8-11 2007.
18. R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. The CLEAR 2006 evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop, CLEAR 2006*, LNCS 4122, pages 1–45, Southampton, UK, Apr. 6-7 2006.
19. R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 3–34, Baltimore, MD, May 8-11 2007.
20. P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 511–518, Kauai, HI, Dec. 2001.