

# The AIT Multimodal Person Identification System for CLEAR 2007

Andreas Stergiou, Aristodemos Pnevmatikakis and Lazaros Polymenakos

Athens Information Technology, Autonomic and Grid Computing,  
Markopoulou Ave., 19002 Peania, Greece  
{aste, apne, lcp}@ait.edu.gr  
<http://www.ait.edu.gr/research/RG1/overview.asp>

**Abstract.** This paper presents the person identification system developed at Athens Information Technology and its performance in the CLEAR 2007 evaluations. The system operates on the audiovisual information (speech and faces) collected over the duration of gallery and probe videos. It comprises of an audio-only (speech), a video-only (face) and an audiovisual fusion subsystem. Audio recognition is based on the Gaussian Mixture modeling of the principal components of composite feature vectors, consisting of Mel-Frequency Cepstral Coefficients and Perceptual Linear Prediction coefficients of speech. Video recognition is based on combining three different classification algorithms: Principal Components Analysis with a modified Mahalanobis distance, sub-class Linear Discriminant Analysis (featuring automatic sub-class generation) with cosine distance and Bayesian classifier based on Gaussian modeling of intrapersonal differences. A nearest neighbor classification rule is applied. A decision fusion scheme across time and classifiers returns the video identity. The audiovisual subsystem fuses the unimodal identities into the multimodal one, using a suitable confidence metric.

## 1 Introduction

Person identification is of paramount importance in security, surveillance, human-computer interfaces and smart spaces. Hence, the evaluation of different recognition algorithms under common evaluation methodologies is very important. Even though the applications of person recognition vary, the evaluations have mostly focused on the security scenario, where training data are few but recorded under close-field conditions. An example of this for faces is the Face Recognition Grand Challenge [1], where facial images are of high resolution (about 250 pixels distance between the centers of the eyes).

The CLEAR 2007 person identification evaluations [2], following the CLEAR 2006 [3] and the Run-1 evaluations [4] of the CHIL project [5], focus on the surveillance and smart spaces applications, where training can be abundant, but on the other hand the recording conditions are far-field: wall-mounted microphone arrays record speech far from the speakers, and cameras mounted on room corners record

faces. These two modalities are used, either stand-alone or combined to recognize people in audiovisual streams.

The person identification system implemented in Athens Information Technology operates on short sequences of the two modalities of the far-field data, producing unimodal identities and confidences. The system is trained automatically, in the sense that there is no manual operation for the selection of the speech or the faces to be used in the training of the systems. The audio subsystem is analyzed in section 2, while the video one in section 3. The identities produced by the unimodal subsystems are then fused into a bimodal one by the audiovisual subsystem, detailed in section 4. The CLEAR 2007 experiments are presented in section 5. Finally, in section 6 the conclusions are drawn.

## 2 Speaker identification subsystem

In the training phase of our system the goal is to create a model for each one of the supported speakers and ensure that these models accentuate the specific speech characteristics of each person. To this end, we first break up the training segments into frames of appropriate size (i.e. duration), with successive frames having a predefined overlap percentage. The samples belonging to each frame are used to calculate a composite feature vector that represents the given frame during the model estimation process. Specifically, a set of Mel Frequency Cepstral Coefficients (MFCC) are extracted from each frame (as in [6]) and augmented by a corresponding set of Perceptual Linear Prediction (PLP) coefficients, in order to model the characteristics and structure of each individual's vocal tract. All composite feature vectors for a given person are collected and used to train a Gaussian Mixture Model (GMM), based on the Baum-Welch algorithm. A GMM is in essence a linear combination of multi-variant Gaussians that approximates the probability density function (PDF) of the MFCC+PLP features for the given speaker:

$$\lambda_k = \sum_{m=1}^M w_m N(o, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad k = 1, \dots, K \quad (1)$$

where  $K$  is the number of speakers (i.e. 28) and  $\lambda_k$  is the GMM for the  $k$ -th speaker. This model is characterized by the number of Gaussians ( $M$ ) that constitutes the mixture, each having its own weight ( $w_m$ ), mean vector ( $\boldsymbol{\mu}_m$ ) and covariance matrix ( $\boldsymbol{\Sigma}_m$ ).

For the identification part, testing samples are again segmented into frames with the same characteristics as the ones created during the training process, and we subsequently extract MFCC and PLP coefficients from each frame and concatenate them into feature supervectors. To perform identification, each of the  $K$  GMM's is fed with an array of the coefficients (one row per sample), based on which we calculate two different metrics.

The first one, which is commonly used in MFCC+GMM speaker identification systems [6], measures the sum of log-likelihoods (across all test frames) that this set

of observations was produced by the given model. The GMM that produces the highest log-likelihood denotes the most probable speaker according to the system:

$$k_1 = \arg \max_k \{L(\mathbf{O} | \lambda_k)\}, \quad k = 1, \dots, K \quad (2)$$

where  $\mathbf{O}$  is the matrix of MFCC+PLP coefficients for the specific test segment and  $L(\mathbf{O} | \lambda_k)$  is the total log-likelihood that each model  $\lambda_k$  produces this set of observations. We have denoted this standard metric as Maximum Total Likelihood (MTL).

Although the MTL score proves adequate for tests of low duration (i.e., 1 second long), our experiments with this year's development data have shown that for longer durations a different approach is preferable. Specifically, we adopt the video modality identification metric from [6]: for each frame, we find the speaker with the highest log-likelihood, as well as compute a margin that is defined as the ratio of the second-best to best log-likelihood for this frame (since log-likelihoods are strictly negative, this ratio is always greater than one) raised to the sixth power. These margins are then used as weights in a weighted sum score which we denote as Maximum Total Margin – 6<sup>th</sup> power (MTM\_6). This second metric is used for tests with durations of 5, 10 and 20 seconds.

All samples are broken up into frames of length 1024 with 75% overlap. Since the data are sampled at 44.1 kHz, each frame has duration of a little over 23 msec. The size of the GMM, as well as the relative numbers of MFCC and PLP coefficients that make up the feature supervectors were determined after experiments conducted on this year's development data. Specifically, all GMM's consist of 32 Gaussians and the number of static coefficients (either MFCC or PLP) per frame has been set to 12, to which we concatenate the log-energy of the frame to create 13D vectors. Depending on the wealth of training data, we have deduced that either MFCC or PLP coefficients must be predominant in the composition of the supervectors. Consequently, for the 15 second training condition (TRAIN\_A), each feature vector consists of 39 MFCC (static + delta + delta-delta) and 26 PLP (static + delta) coefficients, whereas for the 30 second training condition (TRAIN\_B) we use a complementary setup: 26 MFCC and 39 PLP coefficients. In all cases, both MFCC and PLP coefficients were extracted using the HTK Toolbox [7].

A very crucial step in the creation of a successful GMM is the initialization of its parameters, which will be updated during the iterations of the EM training algorithm. The standard approach is to use the K-Means clustering algorithm to obtain some initial estimates for the Gaussian parameters; this strategy however suffers from the random characteristics of the outcome of K-Means, which in turn lead to a different GMM each time the same data are used for training. Moreover, the identification performance varies considerably across these different models. We have therefore utilized a deterministic initialization strategy for the EM algorithm, based on the statistics of the training data. Specifically, we compute a number of percentiles across all dimensions of the training data set and thus partition the data range in each dimension into as many subsets as the modes of the GMM. The K-Means algorithm is consequently run using the central values of each subset as initial cluster means, and the resulting clustered data are fed into the EM algorithm for parameter fine-tuning.

Our experiments have shown that this strategy gives on average lower error rates than the random K-Means initialization, although there are a few runs using the standard approach that lead to better identification performance.

Automatic identification systems are evaluated based on their response time and error rate. It is obviously important to minimize both these numbers, however in many cases it is not easy or even possible to do that and we must settle for a trade-off between speed and identification accuracy. We have addressed this issue by employing the standard Principal Components Analysis (PCA) as a pre-processing step. Specifically, we compute a transformation (projection matrix) for each speaker based on their training data and use that matrix to perform a mapping to the PCA coordinate system prior to GMM calculation. In the testing phase, we compute the log-likelihood of each speaker by first projecting the MFCC+PLP vectors to the respective PCA space.

The use of PCA introduces one further degree of freedom in the system, namely the dimensionality of the projection space. It is obvious that by keeping an increasingly smaller number of eigenvalues from the PCA scatter matrix we can reduce this dimensionality accordingly, therefore achieving a significant execution speed increase. The choice of the number of discarded eigenvalues will be ultimately dictated by the truncation error introduced due to the reduction of the projection space dimension. Specifically, if the initial space dimension is  $d$  and we discard the  $q$  smallest eigenvalues, the truncation error will be equal to

$$e = 1 - \frac{\sum_{i=d-q+1}^d \lambda_i}{\sum_{i=1}^d \lambda_i} \quad (3)$$

In [6], we had employed an automatic decision process that determines the number of retained eigenvalues in a way that ensures that the average truncation error across all speakers is no more than 0.2%. The maximum value of  $q$  that satisfies this condition is chosen, so that we achieve the greatest speed increase possible while retaining (mostly) optimal identification accuracies. Extensive experimentation with this year's development data indicated that the augmentation of the feature vectors with the PLP coefficients forces us to be stricter with this limit, which was ultimately set to 0.1%.

Our experiments indicate that this selection strategy gives a value for  $q$  that is at most one above or below the number of eigenvalues that minimizes the error rates. Even if our choice of  $q$  leads to slightly sub-optimal solutions, the achieved error rates are still superior to using the standard GMM algorithm approach without PCA pre-processing. We have therefore achieved faster response times as well as enhanced identification performance.

### 3 Face identification subsystem

Face recognition on still images has been extensively studied. Given sufficient training data (many gallery stills of each person) and/or high resolution images, the 90% recognition barrier can be exceeded, even for hundreds of different people to be recognized [1]. Face recognition on video streams has only recently begun to receive attention [6, 8-13]. Video-to-video face recognition refers to the problem of training and testing face recognition systems using video streams. Usually these video streams are near-field, where the person to be recognized occupies most of the frame. They are also constrained in the sense that the person looks mainly at the camera. Typical such video streams originate from video-calls and news narration, where a person's head and upper torso is visible.

The CLEAR series of person identification evaluations address a much more interesting application domain: that of the far-field unconstrained video streams. In such streams the people are far from the camera, which is typically mounted on a room corner near the ceiling. VGA-resolution cameras in such a setup can easily lead to quite small faces – down to less than ten pixels between the eyes [3], contrasted to over two hundred pixels in many of the latest face recognition evaluations [1]. Also, the people go about their business, almost never facing the camera directly. As a result, faces undergo large pose, expression and lighting variations. Part of the problem is alleviated by the use of multiple cameras; getting approximately frontal faces is more probable with four cameras at the corners of a room than with a single one. The problem is further alleviated by the fact that the goal is not to derive a person's identity from a single frame, but rather from some video duration. Faces to be recognized are collected from a number of frames; the person identity is then established based on that collection of faces. Hence, far-field unconstrained video-to-video face recognition needs to address the following challenges:

- Detection, tracking, segmentation and normalization of the faces from the video streams, both for system training and recognition.
- The face recognition algorithm needs to cope with very small faces, with unconstrained pose, expression and illumination, and also with inaccurate face framing.
- Fusion of the individual decisions on faces, to provide the identity of the person given some time interval.

The video subsystem for person identification utilizes all four camera streams to extract frontal and profile faces for training and testing of the system. The faces are extracted employing the provided face bounding box label files. These are sampled at 200 ms intervals. Cubic interpolation is used between those labels, to get approximate face locations in all frames. Contrary to the CLEAR 2006 AIT system, no geometric face normalization, neither selection of the most frontal faces is applied. The face patches are scaled to a standard size of 48 tall by 32 wide pixels. Intensity normalization to a mean value of 127 and a standard deviation of 40 is applied on every face patch thus obtained. Right profile faces are flipped to become left profile. All these pre-processing steps are based on the analysis of the CLEAR 2006 results, presented in [14].

To cope with different impairments present in the task (pose, expression and illumination changes), three face recognition algorithms are applied and then suitably combined to yield the identity of the person. These algorithms are based on Principal Components Analysis (PCA), subclass Linear Discriminant Analysis (LDA) and intrapersonal differences. They are presented in the next subsystems.

### 3.1 PCA face recognition

The first algorithm uses Principal Components Analysis (PCA) to find from the gallery faces a recognition subspace of dimension  $N_{PCA}$ . Classification is done by projecting each probe face in this subspace and then use a nearest neighbor classifier. 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  $N_{neib}$  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. The gallery point-dependant weights are the eigenvalues of that scatter matrix. Although this estimation of point-dependant weights is computationally expensive, it is performed once at system training and the  $N_{PCA}$  weights for each of the projected gallery images are stored to be used in the recognition phase.

### 3.2 Subclass LDA face recognition

LDA can prove close to useless for class discrimination problems that are non-separable in a linear way [15]. The face manifolds obtained under unconstraint conditions can easily fall in this problem category. For such problems [15, p. 453] suggests sub-class LDA, where training samples belonging to some class are allocated to different subclasses. Even though the resulting classification space is of larger dimension and the training images per subclass are fewer than those per original class, the problem is turned to a linearly separable one that the resulting classifier can cope with. For automatic training of subclass LDA systems, the choice of the split from a class to the subclasses has to be automatic.

Hence an automatic subclass selection process is needed that subdivides each class into subclasses, making no assumption for the number of subclasses or the number of samples in any subclass. These requirements are met using hierarchical clustering trees. During the training phase, PCA is employed to project the faces into a subspace of dimension  $N$ . The Euclidean distance between all projected faces of each person is calculated, to build a hierarchical clustering tree. The tree is built bottom-up, using a measure of the distance between two clusters of projected faces: At each step up the tree, the two clusters with minimum distance between them are merged into a higher-level cluster. The chosen distance measure is the increase in the total within-cluster sum of squares as a result of joining the two clusters  $c_1$  and  $c_2$ , comprising  $n_1$  and  $n_2$  projected faces respectively. The within-cluster sum of squares is defined as the

sum of the squares of the Euclidean distances between all projected faces in the cluster and the centroid of the cluster. The centroids of the two clusters are:

$$\bar{x}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} x_k^{(i)}, \quad i = 1, 2 \quad (4)$$

where  $x_k^{(i)}$  is the  $k$ -th projected faces of the  $i$ -th cluster. Then, the cluster distance measure is given by:

$$d(c_1, c_2) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \|\bar{x}_1 - \bar{x}_2\|_2 \quad (5)$$

where  $\|\bar{x}_1 - \bar{x}_2\|_2$  is the Euclidean distance of the two centroids. The hierarchical tree thus obtained is used to cluster the samples of the class in subclasses, by cutting the tree at any distance value  $D$ . Small cutoff distances result to many subclasses, whereas large cutoff distances result to few subclasses. The PCA subspace dimension  $N_{PCA}$  and the tree cutoff distance  $D$  are the two parameters of the algorithm.

### 3.3 Bayesian face recognition

The images of the face of a person can vary substantially due to pose, expression and illumination variations. Such variations are intrapersonal and should not lead to a decision that a particular face depicted in a probe image is different than those of the same person depicted in the gallery images. A Bayesian face recognition framework can be build by modeling the intrapersonal variations in the gallery images by a Gaussian distribution, and evaluating the probability that the difference of a gallery face from a probe face is indeed intrapersonal [16].

The problem with this approach is the computational complexity of modeling the intrapersonal differences, when galleries have many images, as is the case with the training conditions of the CLEAR evaluations. To overcome this difficulty, only some of the gallery faces are utilized. The automatic selection of the faces to use 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 of dimension  $N_{PCA}$ . For every person,  $N$  projected images are selected as the median of every of the  $N$  clusters obtained by the trees. The intrapersonal differences are formed and modeled at the reduced dimension  $N_{PCA}$  of the PCA subspace.

Classification is again based on the nearest neighbor classifier, with the inverse of the intrapersonal probability serving as a distance.

### 3.4 Fusion across time and face recognition algorithms

The individual decisions of the three different face recognition algorithms are fused using the sum rule [17]. According to it, each decision  $ID_i$  in a testing segment casts a vote that carries a weight  $w_i$ . The weights  $w_i$  of every decision such as  $ID_i = k$  are summed to yield the weights  $W_k$  of each class:

$$W_k = \sum_{i:ID_i=k} w_i \quad (6)$$

where  $k = 1, \dots, K$  and  $K$  is the number of classes. Then the fused decision based on the  $N$  individual identities is:

$$ID^{(N)} = \arg \max_k (W_k) \quad (7)$$

The weight  $w_i$  in the sum rule for the  $i$ -th decision is the sixth power of the ratio of the second-minimum distance  $d_i^{(2)}$  over the minimum distance  $d_i^{(1)}$ :

$$w_i = \left[ \frac{d_i^{(2)}}{d_i^{(1)}} \right]^6 \quad (8)$$

This choice for weight reflects the classification confidence: If the two smallest distances from the class centers are approximately equal, then the selection of the identity leading to the smallest distance is unreliable. In this case the weight is close to unity, weighting down the particular decision. If on the other hand the minimum distance is much smaller than the second-minimum, the decision is heavily weighted as the selection of the identity is reliable. The sixth power allows for a few very confident decisions to be weighted more than many less confident ones.

The decisions  $ID^{(PCA)}$ ,  $ID^{(LDA)}$  and  $ID^{(LDA)}$  of the PCA, the subclass LDA and the Bayesian face recognition algorithms are again fused using the sum rule to yield the reported identity. For this fusion, the class weights  $W_k$  of equation (6) are used instead of the distances in equation (8). Setting:

$$\begin{aligned} k_1 &\equiv [\text{best matching class}] = ID^{(N)} \\ k_2 &\equiv [\text{second-best matching class}] \end{aligned} \quad (9)$$

the weights of the PCA, the subclass LDA and the Bayesian decisions become:

$$w_i = \frac{W_{k_1}^{(i)}}{W_{k_2}^{(i)}}, \quad i \in \{PCA, LDA\} \quad (10)$$

Then the combined decision of the three algorithms to be reported by the visual subsystem is:

$$ID^{(\text{visual})} = \begin{cases} ID^{(\text{PCA})} & \text{if } w_{\text{PCA}} \geq w_{\text{LDA}} \\ ID^{(\text{LDA})} & \text{if } w_{\text{PCA}} < w_{\text{LDA}} \end{cases} \quad (11)$$

#### 4 Audiovisual person identification

The audiovisual system is again based on post-decision fusion using the sum rule. In this case the decision is:

$$ID^{(\text{A/V})} = \begin{cases} ID^{(\text{audio})} & \text{if } w_{\text{audio}} \geq \min(\{w_{\text{thr}}, w_{\text{visual}}\}) \\ ID^{(\text{visual})} & \text{if } w_{\text{audio}} < w_{\text{visual}} \end{cases} \quad (12)$$

where the audio weight is the ratio of the log-likelihood  $L(\mathbf{O} | \lambda_{k_1})$  that the best matching model  $\lambda_{k_1}$  produces the set of observations  $\mathbf{O}$ , over the log-likelihood  $L(\mathbf{O} | \lambda_{k_2})$  that the second-best matching model  $\lambda_{k_2}$  produces  $\mathbf{O}$ :

$$w_{\text{audio}} = \frac{L(\mathbf{O} | \lambda_{k_1})}{L(\mathbf{O} | \lambda_{k_2})} \quad (13)$$

The visual weights are the maximum of the PCA and LDA weights of (10), transformed by a factor  $c$  so that they have the same mean value as the audio weights and remain greater than or equal to unity:

$$w_{\text{visual}} = c \left[ \max(\{w_{\text{PCA}}, w_{\text{LDA}}\}) - 1 \right] + 1 \quad (14)$$

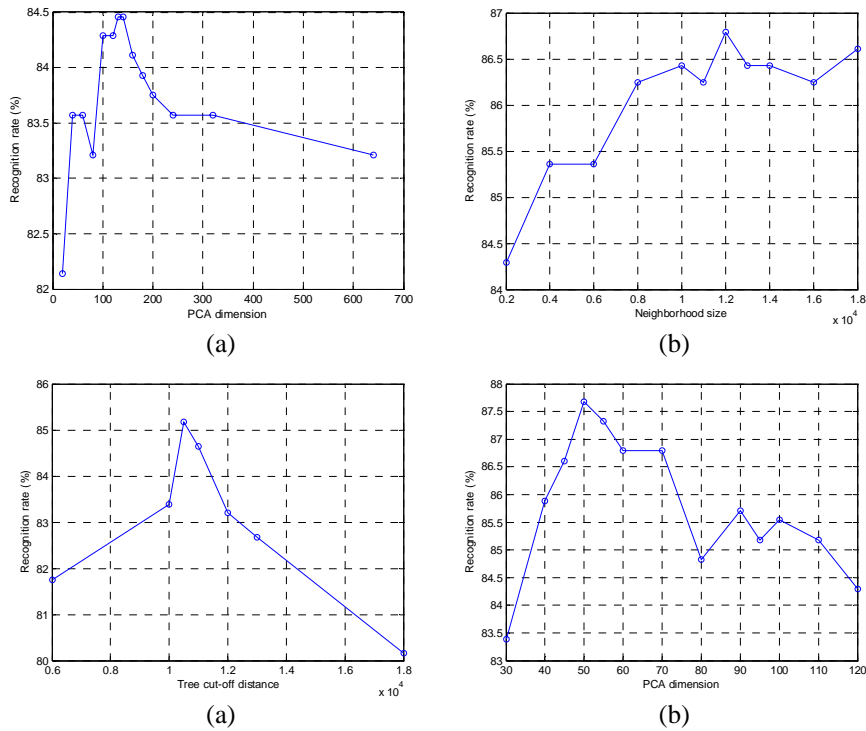
$w_{\text{thr}}$  is an audio weight threshold above which the audio decision is absolutely trusted. This reflects the confidence on the adequately weighted audio decisions, no matter the video ones. This is needed as the performance of video is not expected to be as good as the audio, due to the adverse effect of resolution, label interpolation and pose variation. The choice of this threshold is 1.016 for 15 seconds training, and 1.008 for 30 seconds, where experiments show that audio recognition should be error-free.

#### 5 Experiments

In CLEAR 2007, training, validation and testing segments have been defined. The two training conditions are 15 and 30 seconds long. Four durations have been defined both for validation and testing: 1, 5, 10 and 20 seconds long. All these segments contain mostly speech, so a speech activity detection algorithm [18] has not been used. The following subsections detail the experiments conducted using these datasets.

### 5.1 Parameter tuning using the development set

The development set is used to establish the values of the various parameters of the algorithms. An example of this parameter tuning is shown in Figure 1, for the three algorithms of the face recognition system, 15 sec training and duration of the validation tests of 1 sec. Note that for the longer testing durations, there are not enough validation tests to be performed for trusted analysis. Hence, only the 1 and 5 sec durations of the validation sets are used.



**Fig. 1.** Use of validation set for parameter tuning of the three algorithms involved in the face recognition system. The effect of the various parameters on the fused recognition rate of the whole 1 sec sequence is shown. (a) PCA subspace dimension and (b) neighborhood size for the PCA-based algorithm. (c) Tree cut-off distance for the automatic selection of subclasses in the subclass LDA algorithm. (d) PCA subspace dimension for the projection of the images to the space where representative images for the intrapersonal difference modeling are selected for the Bayesian algorithm.

### 5.2 Evaluation results

The results of the audio, visual and multimodal person recognition are shown in Table 1 per training and testing duration.

## 4 Conclusions

The results of the person identification system of Athens Information Technology at the development and test sets of the CLEAR 2007 evaluations are quite different. Overfitting to the development set is not a likely cause for the reduced performance on the test set. This is because most parameters have a value region leading to similar recognition performance. The most plausible explanation is the temporal separation of the training, validation and testing sets. Training and testing are chosen as far apart in time as possible, with the validation set being in-between. Also, the small number of segments of the longer durations, render any parameter tuning for these durations very risky from a statistical point of view.

**Table 1.** Audio, video and audiovisual recognition performance in CLEAR 2007.

Training duration	Testing duration	Recognition rate (%)		
		Audio	Video	Audiovisual
15	1	79.69	79.38	89.20
	5	90.40	86.16	94.42
	10	94.64	89.29	96.88
	20	95.54	91.07	96.43
30	1	84.51	86.12	92.90
	5	94.87	91.29	95.98
	10	96.88	94.20	96.88
	20	99.11	94.64	97.32

It is not straightforward to judge the relative difficulty of the CLEAR 2007 and CLEAR 2006 person identification evaluations. Although the number of people is slightly increased (28 compared to 26), the sequences are selected keeping a balance between the two modalities (in CLEAR 2006, attention had been paid to the existence of speech) and the video labels are more accurate and are all provided every 200 ms (compared to every 1 sec for the face bounding boxes of CLEAR 2006).

Regarding the performance of the AIT system in the test set, the CLEAR 2007 results are superior to those of CLEAR 2006. The main reason for this is the enhanced algorithms employed, as indicated by the results of the system on the CLEAR 2006 test set.

## Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909. The authors wish to thank the people involved in data collection, annotation and overall organization of the CLEAR 2007 evaluations for providing such a rich test-bed for the presented algorithms.

## References

- [1] [www.clear-evaluation.org](http://www.clear-evaluation.org)
- [2] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa and P. Soundararajan: The CLEAR 2006 Evaluation, in R. Stiefelhagen and J. Garofolo (eds.), “*CLEAR 2006, Lecture Notes in Computer Science 4122*”, Springer-Verlag (2007), 1-44.
- [3] H. Ekenel and A. Pnevmatikakis: Video-Based Face Recognition Evaluation in the CHIL Project – Run 1, *Face and Gesture Recognition 2006*, Southampton, UK, (Apr. 2006), 85-90.
- [4] A. Waibel, H. Steusloff, R. Stiefelhagen, et. al: CHIL: Computers in the Human Interaction Loop, *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Lisbon, Portugal, (Apr. 2004).
- [5] A. Stergiou, A. Pnevmatikakis and L. Polymenakos: A Decision Fusion System across Time and Classifiers for Audio-visual Person Identification, in R. Stiefelhagen and J. Garofolo (eds.), “*CLEAR 2006, Lecture Notes in Computer Science 4122*”, Springer-Verlag (2007), 218-229.
- [6] HTK (Hidden Markov Toolkit), <http://htk.eng.cam.ac.uk/>
- [7] J. Phillips, P. Flynn, T. Scruggs, K. Boyer and W. Worek: Preliminary Face Recognition Grand Challenge Results, *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, Southampton, UK (2006) 15-21.
- [8] J. Weng, C.H. Evans, and W.-S. Hwang: An Incremental Learning Method for Face Recognition under Continuous Video Stream, *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, Grenoble, France (2006) 251-256.
- [9] K.-C. Lee, J. Ho, M.-H. Yang and D. Kriegman: Video-based face recognition using probabilistic appearance manifolds, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA (2003) 313-320.
- [10] X. Liu and T. Chen: Video-based face recognition using adaptive hidden markov models, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA (2003) 340-345.
- [11] B. Raytchev and H. Murase: Unsupervised recognition of multi-view face sequences based on pairwise clustering with attraction and repulsion, *Computer Vision and Image Understanding*, 91 (2003) 22-52.
- [12] G. Aggarwal, A.K. Roy-Chowdhury and R. Chellappa: A System Identification Approach for Video-based Face Recognition, *Proceedings of International Conference on Pattern Recognition*, Cambridge, UK (2004).
- [13] C. Xie, B.V.K. Vijaya Kumar, S. Palanivel and B. Yegnanarayana: A Still-to-Video Face Verification System Using Advanced Correlation Filters, in D. Zhang and A.K. Jain (eds.) “*ICBA 2004, Lecture Notes in Computer Science 3072*” Springer-Verlag (2004) 102-108.
- [14] A. Pnevmatikakis and L. Polymenakos: Far-Field Multi-Camera Video-to-Video Face Recognition, in V. Kordic (ed.), “*Face Recognition*”, *Advanced Robotics Systems*, accepted.
- [15] K. Fukunaga: *Statistical Pattern Recognition*, Academic Press (1990)
- [16] B. Moghaddam: Principal Manifolds and Probabilistic Subspaces for Visual Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, 20, 3 (March 1998), 226–239. J. Kittler, M. Hatef, R.P.W. Duin and J. Matas: On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.*, 24, 6 (2002).
- [17] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas: On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.*, 20, 3 (March 1998), 226–239.
- [18] J. Sohn, N.S. Kim and W. Sung: A Statistical Model Based Voice Activity Detection, *IEEE Sig. Proc. Letters*, 6, 1 (1999).