Video-Based Face Recognition Evaluation in the CHIL Project – Run 1

Hazım Kemal Ekenel
Universität Karlsruhe (TH)
Interactive Systems Labs
76131, Karlsruhe, Germany
ekenel@ira.uka.de

Aristodemos Pnevmatikakis
Athens Information Technology
Autonomic and Grid Computing Group
19002, Peania, Athens, Greece
apne@ait.edu.gr

Abstract

This paper describes the video-based face recognition evaluation performed under the CHIL project and the systems that participated to it, along with the obtained first year results. The evaluation methodology comprises a specially built database of videos and an evaluation protocol. Two complete automatic face detection and recognition systems from two academic institutions participated to the evaluation. For comparison purposes, a baseline system is also developed using well-known methods for face detection and recognition.

1. Introduction

Face recognition is one of the most important biometric recognition methods. Therefore many efforts have been carried out towards standardizing the way face recognition algorithms are tested [1-8]. One of the most well-known evaluation frameworks is the Facial Recognition Technology (FERET 1994, 1995, 1996) evaluation framework [1]. The FERET database used for evaluations contains high resolution (40 to 60 pixels distance between the centers of the eyes) still images having variations in expression, illumination, pose and time gap between the successive image acquisitions. Face Recognition Vendor Tests (FRVT 2000, 2002, 2005) have followed the FERET evaluations. They provide independent government evaluations of commercially available and prototype face recognition technologies [2,3]. Besides FERET and FRVT, Face Recognition Grand Challenge (FRGC 2004-2005) experiments have been carried out to promote and advance face recognition technology [4]. In FRGC data set higher resolution (~250 pixels distance between the centers of the eyes) images and 3D face data is provided to improve the face recognition algorithms' performance.

Face authentication systems has been tested with the XM2VTS database using the Lausanne protocol (2000, 2003) [5-7] and with the BANCA database using the BANCA protocol (2004) [8]. The XM2VTS database is a multi-modal database and contains also audio data to provide improvement in the correct recognition rate by fusing the facial features with the acoustic features.

All these evaluations of face recognition methods utilize databases that contain high resolution images, or image sequences taken under controlled conditions with a single camera. There have been no evaluations on image sequences that are acquired with multiple cameras in a natural smart environment or in a surveillance setting. The previous evaluation frameworks mainly focused on person authentication for access control. However, evolving identification requirements for smart environments [9] and surveillance, and corresponding increase in video-based face recognition research [10-18] impose the need for a different evaluation dataset and framework.

In this paper we present a video-based face recognition evaluation database and protocol that focus on recognizing people in a smart environment using the image sequences acquired from multiple cameras. This protocol is used in the Computers in the Human Interaction Loop (CHIL) project [19]. We also describe the two complete systems developed at the institutions that have participated to the first year evaluation using this evaluation methodology, and present the obtained results. The remainder of the paper is organized as follows. In Section 2 video-based face recognition evaluation framework in the CHIL project is explained. An overview of the two systems is given in Section 3. Results of the evaluations are presented and discussed in Section 4. Finally, in Section 5, conclusions and future planning are given.

Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)
0-7695-2503-2/06 $20.00 © 2006 IEEE

2. Video-based face recognition evaluation framework in the CHIL project – Run 1

In the CHIL Run 1 video-based face recognition evaluation, the seminars recorded by the CHIL partners were used. The smart-rooms used for capturing seminars are equipped with a variety of sensors, including multiple cameras, distant and close-talking microphones, as well as microphone arrays. For the face recognition task, four fixed cameras with a native resolution of 640 by 480 pixels are used. They are mounted on the corners of the rooms, as shown in Figure 1. They face the rooms in such a way, so that every point is monitored by at least two of them.

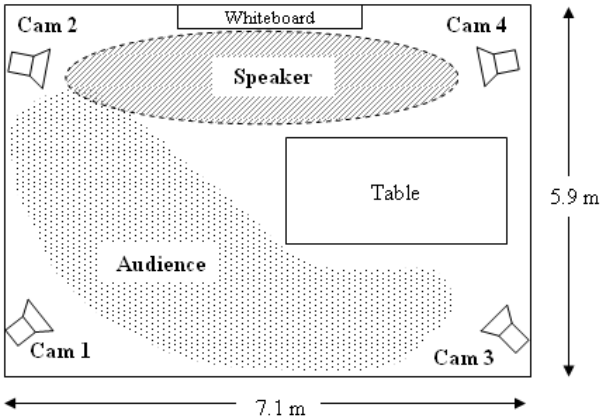

Figure 1. The smart-room is equipped with 4 fixed cameras at a height of approximately 2.7m. The cameras' joint field-of-view covers the entire room

The recording conditions are unconstrained and lead to very small faces with distances between the eyes of typically 10 to 20 pixels, depending on the camera view and the position of the presenter. The presenter to be recognized moves around the projection screen without facing the cameras. Shadows and the beam of the projector result into largely varying face illumination conditions. As an example, the views from all four cameras at the same instant are shown in Figure 2.

The Run 1 evaluation database consists of 7 seminars recorded in the Interactive Systems Labs of the Universität Karlsruhe (TH), Germany in 2003. Each seminar is given by a different presenter. These seminars are recorded in five different dates. The time gap between the different recording dates ranges from one week to one month. The lectures in the same day are recorded consecutively. Each seminar is divided into four sparse, non-overlapping segments of 5 minutes duration. From these, segments 3 and 4 are used for training and segments 1 and 2 for testing.

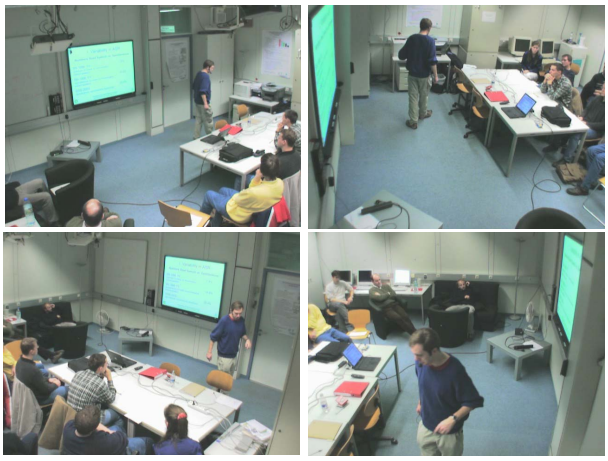

Figure 2. Example images of the presenter, taken at the same instant from the four cameras

In the Run 1 evaluation framework, the systems are trained using pre-selected still images. Five frames that contain frontal face images of each of the presenters are manually selected from segments 3 and 4 for training. Some of the selected faces are shown in Figure 3.

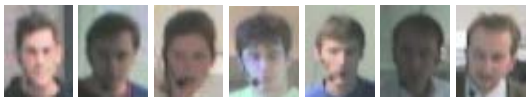

Figure 3. Sample training face images

Segments 1 and 2 are used for testing: 20 uniformly sampled, non-overlapping sequences of 100 frames from each camera and each seminar are selected. Example faces in these frames are shown in Figures 4 and 5. The classification is performed over these 100 frames to yield a single identity. Manual operations are allowed only at the training stage, while testing is done fully automatically.

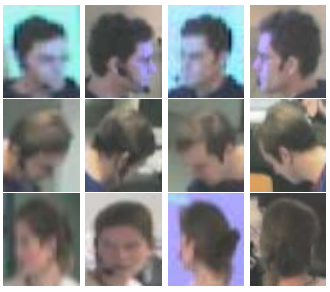

Figure 4. Test samples at the same instant from cameras 1 to 4 (left to right)

The rationale behind the proposed frontal still image-to-video recognition scheme is the assumption that frontal pictures of lecturers are readily available. Using these as training images, we want to automatically identify the lecturers during the lecture.

Figure 5. Sub-sampled testing sequence from a single camera

Although a database of 7 individuals may seem too small for an identification task, considering the identification requirement of smart environments, i.e. recognizing members of a laboratory or family members at home, this number is reasonable.

3. Overview of algorithms

The face recognition task described in the previous section is very difficult due to three reasons: (a) The faces need to be detected in a cluttered scene and under different illumination conditions. (b) The video is unconstrained, so the suitable views need to be selected form the four cameras, and (c) The faces are very small, with typical eye distance of 10 to 20 pixels. Next, two systems that attempt to deal with these difficulties are presented.

3.1. Athens Information Technology

The face recognition system developed at Athens Information Technology (AIT) comprises a face detection, a recognition and an identity fusion stage. It operates on the video frames of all four cameras obtained in a given time interval, and returns the estimated identity for that interval. The block diagram of the end-to-end face recognition system is given in Figure 6.

The face detector extracts foreground segments from the video by processing the difference of the current frame from a background estimate. Skin is detected in the foreground segments by processing the skin versus non-skin likelihood map of the foreground pixels. Large and round enough such segments are considered possible faces. The last operation of the detector is to normalize the faces based on the position of the eyes. Eye detection is a two-stage process. Initially, the zone of the eyes is detected by manipulating the horizontal and vertical averages of the grayscale face region. Then the eyes themselves are detected using k-means clustering of particles allocated to the pixels inside the eye zone. More particles are assigned to darker pixels. The eye zone is also utilized to obtain the best linear estimate of illumination variation across the face. The faces are normalized for this illumination change and also their intensity is normalized to zero mean and unit variance. Finally, the faces are fit in a 48 by 34 pixels template by rotating and scaling the segments based on the estimated eye positions.

The face recognizer returns the identity of every candidate face segment and a confidence associated with it. The identity of every candidate face segment is produced by a Linear Discriminant Analysis (LDA) face recognizer. Since the eye detection can be inaccurate, especially for the smallest face segments, the normalization of the segments is not perfect. Given the sensitivity of LDA to the geometric normalization [20], some performance loss occurs. Nevertheless, most of the lost performance can be regained if the inaccurate eye detection is taken into account in the training phase of the LDA classifier. The eyes in each of the five training images per person are deliberately perturbed to the eight neighboring pixels around their hand-annotated positions. Thus, 81 training images are artificially produced out of each of the original training image. The deliberate introduction of face normalization errors in the training stage allows for matched training and testing conditions and enhances performance.

The confidence associated with each one of the recognitions is related to the degree of confidence that the candidate face segment is indeed a frontal face. The frontal face verification of the normalized faces is again a two-stage process.

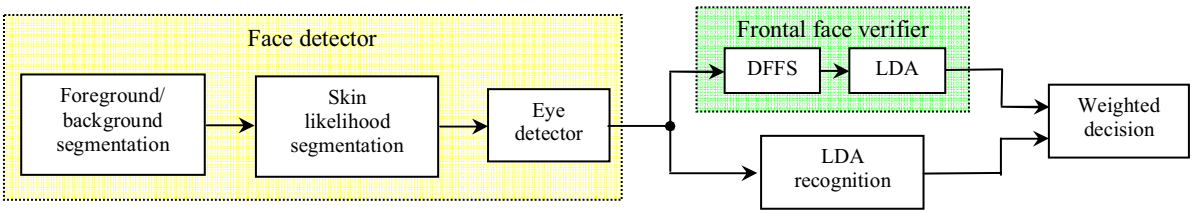

Figure 6. Block diagram of the AIT end-to-end face recognition system

First the Distance From Face Space (DFFS) [21] of every normalized segment is calculated. Small DFFS values indicate that the segment is a frontal face. Due to the small resolution of the faces, a class of profile faces also results to small DFFS values. These images are discarded by a second stage which comprises an LDA classifier trained to tell frontal faces and that class of profile faces apart. The DFFS of the segments that are classified as profile is artificially increased. That DFFS value is used as a measure of 'frontalness' of the segment. Employing this second stage of frontal verification significantly boosts recognition performance.

For every sequence of 100 frames the identities and the confidences of all segments detected from the two best cameras are collected. Every identity is weighted by two to the power of the negative of the associated confidence value. The weights for every person are added and the person identity with the largest sum is selected. The selection of the two best cameras is based on the average frontal face confidence.

3.2. Universität Karlsruhe (TH)

The system developed by the Universität Karlsruhe (UKA) consists of an automatic face detector and a face recognizer. The face detector has two steps. The first step utilizes prior location distribution to separate the lecturer's presentation area from the audience area. The second step consists of haar-like features based multi-view face detection [22, 23, 24].

The developed face recognition system consists of three parts: Block-based discrete cosine transform (DCT) based representation, two-class LDA-based classification per frame and assignment of probabilities to each frame using Bayesian formulation, and accumulation of them over the video sequence. The face appearance is modeled locally, that is, the detected and resized face is divided into 8x8 pixels blocks and each block is represented with DCT coefficients [25, 26]. To provide robustness against detection and alignment errors, artificial samples are generated from the original training face images by translating and scaling them. Two-class LDA is used for classification [27]. In this approach a single M-class LDA classifier is divided into M two-class LDA classifiers. The training data for genuine class consists of samples from the true candidate, whereas the training data for impostor class consists of the other peoples' samples plus random background samples. In this method, each class has its own projection vector. When a test image arrives, it is projected onto each individual's decision space, using the corresponding projection vector. The distribution of projected genuine and impostor data in one-dimensional space is modeled with univariate Gaussians. The decision is taken by applying Bayes rule

$$P(C_{k,1} \mid x) = \frac{P(x \mid C_{k,1})P(C_{k,1})}{\sum_{i=1}^{2} P(x \mid C_{k,i})P(C_{k,i})} \qquad (1)$$

where $C_{k,1}$ denotes genuine class and $C_{k,2}$ denotes impostor class of the k^{th} individual. $P(C_{k,1})$ and $P(C_{k,2})$ are taken 0.5. From the equation above, there may be three cases observed. In the first case, for every "k", $P(C_{k,1} \mid x)$ may be smaller than 0.5. This can occur either from a background sample detected as a face or from an unknown face. In the second case, there may be more than one "k" such that $P(C_{k,1} \mid x)$ is bigger than 0.5. In this case the most probable candidate can be selected. In the third case, which is the ideal case, only one of the individuals has $P(C_{k,1} \mid x)$ bigger than 0.5. The extension of this system to multiple cameras and video is straightforward – accumulate the $P(C_{k,1} \mid x)$s that are bigger than 0.5 over multiple camera views and video sequence. By doing this, the video data and multiple camera information is utilized naturally, and good frames, those containing properly detected, high resolution images are inherently weighted more.

3.3. Baseline System

To evaluate and compare the performance of the proposed face recognition systems, a baseline system is also developed by the Universität Karlsruhe (TH). This system consists of an automatic face detector mainly based on skin color and a face recognizer mainly based on principal component analysis. The face detector utilizes prior location distribution to determine the lecturer's presentation area. It then performs skin color segmentation in the RGB color space. Background subtraction, median filtering and morphological analysis are used to process the output of the skin segmentation. Finally, it uses simple heuristics such as the ratio between face width and height, to determine the face region. In face recognition training, the eye centers of each face are labeled manually and the face images are cropped and aligned according to these labels. For testing, the incoming test face image is projected onto the face space, and its DFFS value is calculated. The 10 face images that have lower DFFS are selected out of 400 face images. According to their DFFS score, these 10 face images are assigned with separate weights, the lower its DFFS score, the higher its weight. Finally, the decisions of 10 images are fused by weighted voting.

4. Results and discussion

In the experiments detailed in this section, the focus is on the two presented end-to-end systems. However it is also interesting to try to separate the impact of automatic face detection and camera selection. To do so, 15 testing frames per person depicting frontal faces from the cameras have been selected. The manually extracted and normalized faces have been recognized using the Eigenfaces method [21] yielding a correct recognition rate of 75.9% and the PCA+LDA combination [28] yielding 79.5%. It is expected that the performance of the end-to-end systems will be lower.

The average recognition rate of the baseline system on the database is just 43.9%. This increases to 66.1% for the Athens Information Technology and to 66.2% for the Universität Karlsruhe (TH) systems (see Table 1). Compared to the 79.5% correct recognition rate of the semi-automatic system, the fully-automatic systems perform very well, having in mind that on top of the recognition they have to perform detection and decision on the best views to trust. Given that the number of classes is just seven, these results seem low. But this is not the case since: (a) In training, only frontal face images are used. However, testing sequences include every kind of views; there even may be 100 frames sequences that do not contain any frontal face image. (b) Significant illumination variations are caused by the projector's beam and illumination sources in the room, and (c) The face resolution is very poor as the eye distance is typically 10 to 20 pixels.

Table 1. Experimental Results

Algorithm	Performance
Manual Baseline (Eigenfaces)	75.9%
Fully Automatic Baseline	43.9%
AIT System	66.1%
UKA System	66.2%

For the Athens Information Technology system, it is interesting to identify the effect on that performance of two extensions to the standard techniques used in the literature. Without the second (LDA) stage of the frontal face verifier that follows the standard DFFS stage, performance drops to 49.9%. Also, without the deliberate perturbation of the eyes in the training faces performance drops to 63.5%. Overall, performance without any of the two extensions is just 40.2%, i.e. very close to the baseline system. The performance of the AIT system as a function of the duration of the testing segments is shown in Figure 7.

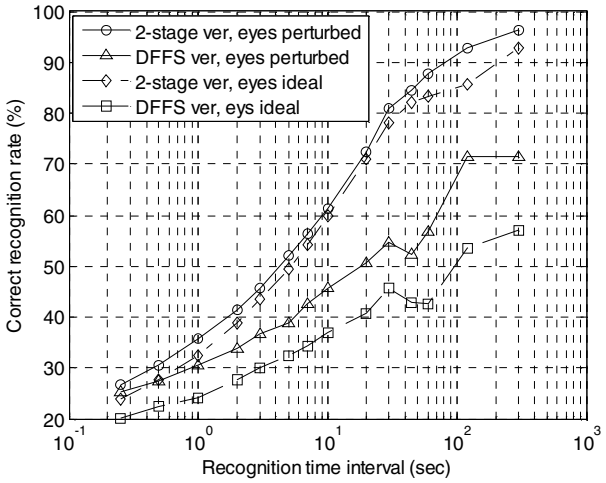

Figure 7. Performance of the AIT system as a function of the duration of the testing segments

5. Conclusions

This paper presents first-year results on the Run 1 video-based face recognition evaluation framework under the CHIL project. The Run 1 evaluation comprises a challenging database for the face recognition task and an evaluation protocol. Currently, the evaluation scenario contains only still-to-video based face recognition and the evaluation database contains 7 seminars. In the second year evaluations (which will be open to the external participants) video-to-video based face recognition scenario will be included and the database will contain more seminars. In addition, there will be a joint audio-visual identification evaluation using the same database.

6. Acknowledgements

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

7. References

[1] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms", IEEE Trans. on PAMI, Vol. 22, No. 10, October 2000.

[2] P. Phillips, P. Grother, R. Michaels, D. Blackburn, E. Tabassi and M. Bone, "Face recognition vendor test 2002: Evaluation report", NISTIR 6965, NIST, 2003.

[3] D. M. Blackburn, M. Bone, P.J. Phillips, "Face recognition vendor test 2000: Evaluation report", NIST, 2001.

[4] P. Phillips et al., "Overview of the Face Recognition Grand Challenge", CVPR 2005.
[5] K Messer et al., “XM2VTSDB: The Extended M2VTS Database”, AVBPA 1999.
[6] J Matas et al., “Comparison and face verification results on the XM2VTS database”, ICPR 2000.
[7] K. Messer et al., “Face verification competition on the XM2VTS database”, AVBPA, 2003.
[8] K. Messer et al., “Face authentication competition on the BANCA database”, ICBA, 2004.
[9] A. Waibel et al., “CHIL: Computers in the Human Interaction Loop”, WIAMIS 2004.
[10] J. Weng, C.H. Evans, W.S. Hwang, “An Incremental Learning Method for Face Recognition under Continuous Video Stream”, AFGR 2000.
[11] V. Krüger, S. Zhou, “Exemplar-Based Face Recognition from Video”, AFGR 2002.
[12] X. Liu, T. Chen, “Video-Based Face Recognition Using Adaptive Hidden Markov Models”, CVPR 2003.
[13] S. Zhou, V. Krueger and R. Chellappa, "Probabilistic Recognition of Human Face from Video", Computer Vision and Image Understanding, Vol. 91, pp. 214-245, July 2003.
[14] B. Raytchev, H. Murase, “Unsupervised recognition of multi-view face sequences based on pairwise clustering with attraction and repulsion”, Computer Vision and Image Understanding, Vol. 91, pp. 22-52, 2003.
[15] S. Zhou, R. Chellappa, B. Moghaddam, “Visual tracking and recognition using appearance-adaptive models in particle filters”, IEEE Trans. on Image Processing, Vol. 13, No. 11, pp. 1491-1506, 2004.
[16] G. Aggarwal, A.K.R Chowdhury, R. Chellappa, “A system identification approach for video-based face recognition”, ICPR 2004.
[17] C. Xie et al, “A Still-to-Video Face Verification System Using Advanced Correlation Filters”, ICBA 2004.
[18] K.C. Lee, D. Kriegman, “Online Learning of Probabilistc Appearance Manifolds for Video-based Recognition and Tracking“, CVPR 2005.
[19] CHIL Project; http://chil.server.de.
[20] A. Pnevmatikakis and L. Polymenakos, “A Testing Methodology for Face Recognition Algorithms”, in S. Renals and S. Bengio (Eds.): MLMI 2005, LNCS 3869, pp. 218-229, Springer-Verlag, Berlin, Heidelberg, 2006.
[21] M. Turk and A. Pentland, “Eigenfaces for Recognition”, J. Cognitive Neuroscience, Vol. 3, pp. 71-86, 1991.
[22] R. Lienhart, J. Maydt, “An Extended Set of Haarlike Features for Rapid Object Detection”, IEEE ICIP, 2002.
[23] M. Jones, P. Viola, “Fast Multi-view Face Detection”, CVPR 2003.
[24] H.K. Ekenel, K. Nickel, R. Stiefelhagen, “Locating and Identifying the Lecturer in a Smart Room”, Second Workshop on Face Processing in Video (FPiV'05), Canada, May 2005.
[25] H.K. Ekenel, R. Stiefelhagen, "A Generic Face Representation Approach for Local Appearance based Face Verification", CVPR IEEE Workshop on FRGC Experiments, 2005.
[26] H.K. Ekenel, R. Stiefelhagen, “Local Appearance based Face Recognition Using Discrete Cosine Transform”, 13th European Signal Processing Conference (EUSIPCO), Antalya, Turkey, September 2005.
[27] H.K. Ekenel, R. Stiefelhagen, “Two-class linear discriminant analysis for face recognition”, Tech. Report, Universität Karlsruhe (TH), 2005.
[28] A. Pnevmatikakis and L. Polymenakos, ‘Comparison of Eigenface-Based Feature Vectors under Different Impairments’, Int. Conf. Pattern Recognition 2004, vol. 1, pp. 296-300, 2004.