Face Recognition in Smart Rooms

Hazım Kemal Ekenel, Mika Fischer, and Rainer Stiefelhagen

interACT Research, Computer Science Department, Universität Karlsruhe (TH)
Am Fasanengarten 5, Karlsruhe 76131, Germany
{ekenel,mika.fischer,stiefel}@ira.uka.de
http://isl.ira.uka.de/

Abstract. In this paper, we present a detailed analysis of the face recognition problem in smart room environment. We first examine the well-known face recognition algorithms in order to observe how they perform on the images collected under such environments. Afterwards, we investigate two aspects of doing face recognition in a smart room. These are: utilizing the images captured by multiple fixed cameras located in the room and handling possible registration errors due to the low resolution of the aquired face images. In addition, we also provide comparisons between frame-based and video-based face recognition and analyze the effect of frame weighting. Experimental results obtained on the CHIL database, which has been collected from different smart rooms, show that benefiting from multi-view video data and handling registration errors reduce the false identification rates significantly.

1 Introduction

Face recognition has attracted significant research efforts that are mainly fueled by security applications. Recently, face recognition for smart interactions has become another application area of significant interest [1]. There have been many papers published on the use of face recognition technology in human-robot interactions [2], smart cars [3], human-computer interfaces [4] as well as image and video retrieval applications [5], [6], [7].

One of the most interesting smart interaction applications is face recognition in smart rooms. Sample application areas can be a smart store that can recognize its regular customers while they are entering the store; a smart home, where family members can be identified while they are entering the rooms of the house and their location can be determined in order to automatically route incoming phone calls; a smart lecture or meeting room, where the participants can be identified automatically and their behaviours can be analyzed throughout the meeting or the lecture. This group of applications requires identification of people without any cooperation, and under uncontrolled conditions, without any constraints on head-pose, illumination, use of accessories, etc. Moreover, according to the distance between the camera and the subject the face resolution varies, and generally the face resolution is low. In these respects, face recognition in smart rooms is a very difficult task. The only factor that can help to improve

A. Popescu-Belis, S. Renals, and H. Bourlard (Eds.): MLMI 2007, LNCS 4892, pp. 120–131, 2007.

Fig. 1. Sample views of the smartrooms

the face recognition performance in smart rooms is the video data of the individuals from multiple views, provided by several cameras that are mounted in the smart room. Sample images from different smart rooms are shown in Figure 1.

Taking these facts into consideration, in this paper we present a detailed analysis of the face recognition problem in a smart room environment. We first compare the well-known face recognition algorithms in order to observe how they perform on the images collected in such environments. Afterwards, we investigate two typical aspects of doing face recognition in a smart room. These are: utilizing the images captured by multiple fixed cameras located in the room and handling possible registration errors due to the low resolution of the aquired face images. We propose a camera-weighting scheme in order to be able to give higher weights to the cameras that have a better view of the person. To be able to handle registration errors, we generate additional registered samples from the manually labelled training images by moving the manual eye label locations in the neighborhood and doing registration with respect to the newly obtained eye coordinates. Note that, even with manual labelling, due to the low resolution of the face images, there can be slight errors in the eye center coordinates. In addition, we also provide comparisons between frame-based and video-based face recognition and analyze the effect of frame weighting. We conduct the experiments on a data corpus that has been collected at different smart rooms. The experimental results indicate that utilizing video data and generating additional

samples reduces the false identification rates significantly. Camera and frame weighting have been found to improve the performance further.

The organization of the paper is as follows. In Section 2, local appearance-based face recognition using the discrete cosine transform is explained briefly. A baseline face recognition system is described in Section 3. In Section 4, experimental results are presented and discussed. Finally, in Section 5, conclusions are given.

2 Local Appearance-Based Face Recognition Using Discrete Cosine Transform

Local appearance-based face recognition was proposed as a fast and generic approach [8], [9] and does not require detection of any salient local regions, such as eyes, as in the modular or component based approaches [10], [11]. The underlying ideas for preferring a local appearance-based approach over a holistic appearance-based approach are as follows: (i) In a holistic appearance-based face recognition approach, a change in a local region can affect the entire feature representation, whereas in local appearance-based face recognition it affects only the features that are extracted from the corresponding block while the features that are extracted from the other blocks remain unaffected. This property provides robustness against both local registration imperfections and expression variations, (ii) a local appearance-based algorithm can facilitate weighting of local regions. It can put more weight to the regions which are found to be more discriminant.

In order to represent the local regions, the discrete cosine transform (DCT) is used. Its compact representation ability is superior to that of the other widely used input-independent transforms like the Walsh-Hadamard transform. Although the Karhunen-Loeve transform (KLT) is known to be the optimal transform in terms of information packing, its data dependent nature makes it infeasible for some practical tasks. Furthermore, DCT closely approximates the compact representation ability of the KLT, which makes it very useful for representation both in terms of information packing and in terms of computational complexity.

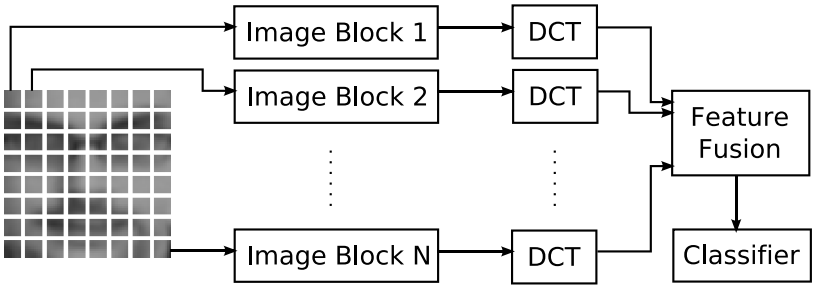

Fig. 2. System architecture of the face recognition system

Feature extraction using the local appearance-based face representation can be summarized as follows: A detected and normalized face image is divided into blocks of 8×8 pixels size. Then the DCT is applied on each block. The obtained DCT coefficients are ordered using zig-zag scanning. From the ordered coefficients, M are selected according to a feature selection strategy, and then normalized to unit norm, resulting in an M-dimensional local feature vector. These extracted local features are then concatenated to represent the entire face image (Figure 2). For details of the algorithm please see [8], [9].

3 Baseline Face Recognition System

In order to provide an identity estimate, the face recognition system processes multi-view, multi-frame visual information. The system components are: image registration, feature extraction, score normalization, fusion over camera-views and fusion over image sequence.

The baseline system receives an input image and the eye-coordinates of the face in the input image. The face image is cropped and registered according to the eye coordinates. The local appearance-based face recognition that is mentioned in Section 2 is used for feature extraction. The first DCT coefficient is removed since it only represents the average value of the image block. The first M coefficients are selected from the remaining ones. To remove the effect of intensity level variations among the corresponding blocks of the face images, the extracted coefficients are normalized to unit norm.

Classification is performed by comparing the extracted feature vectors of the test image with the ones in the training database. Each camera view is compared with all the others. Distance values of the 10-best matches obtained from each frame are normalized using the Min-Max rule, which is defined as:

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

where, s corresponds to the distance value of the test image to one of the ten closest training images in the database, and S corresponds to a vector that contains the distance values of the test image to the ten closest 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, we obtain a confidence score that is normalized to the value range of $[0, 1]$, closest match having the score '1', and the furthest match having the score '0'. These scores are then normalized by dividing them by the sum of the confidence scores. The obtained confidence scores are summed over camera-views and over the image sequence. The identity of the face image is assigned as the person who has the highest accumulated score at the end of a sequence.

4 Experiments

The experiments have been conducted on a database that has been collected by the CHIL consortium for the CLEAR 2006 evaluations [12]. The recordings are

Fig. 3. Views of the UKA smartroom taken at the same instant from the four cameras

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. Each smartroom has four cameras, one at each corner, which are labeled from 1 to 4. Sample images from the recordings can be seen in Figures 1 and 3. The evaluation data for the visual identification task consists of short video sequences taken from the database. The recording conditions are uncontrolled and lead to low resolution faces ranging between 10 to 50 pixels resolution, depending on the camera view and the position of the presenter/participant. The presenter to be recognized moves around the projection screen without facing the cameras. Shadows and the beam of the projector result in largely varying face illumination conditions. There are 26 subjects in the database. Two different training and four different testing durations are used in the experiments as presented in Table 1. The training sets contain one sequence for each subject, whereas the number of sequences for each subject in the testing set is varying. Identity estimates are provided at the end of each test segment using the available video data. In addition frame-based recognition — where an identity estimate is provided for every single frame — is also performed and the corresponding results are also presented.

In the database, eye center labels are available for every 200 ms. We only used the frames where both of the eyes are visible and are labelled at the same time. In total we processed 26494 images for the experiments, where 8689 of them belong

Table 1. Duration of Training and Testing Segments

Train/Test ID	Segment duration (sec)	No. of segments
Train A	15	26
Train B	30	26
Test 1	1	613
Test 2	5	411
Test 3	10	289
Test 4	20	178

Table 2. False Identification Results of the Baseline System

	A1	A5	A10	A20	B1	B5	B10	B20
Frame-based	50.1	50.9	50.7	50.4	43.5	43.6	43.7	43.7
Video-based	35.1	25.9	24.9	19.3	32.3	22.3	22.1	17.1

to the training set and the remaining 17805 belong to the testing set. The face images are aligned according to the labelled eye-center coordinates and scaled to 64×64 pixels resolution. The aligned images are then divided into 8×8 pixels resolution non-overlapping blocks making 64 local image blocks. From each image block five-dimensional DCT-based feature vectors are extracted and they are concatenated to construct the final 320-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. The distance values are converted to the matching scores and then the normalized matching scores are combined in order to provide the identity estimate. The identity candidate that has the highest score is assigned as the identity of the person.

The baseline results of both the frame-based and video-based identification are presented in Table 2. In this experiment, all the camera views are compared with each other and the cameras are weighted equally. No frame weighting is performed and no additional samples are used. Each column 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. For frame-based identification all the frames in the training-testing duration combination are used. For example in the combination 'A5', all the frames in the Train A set and all the frames in 5 seconds duration testing segments are used. Two main observations can be derived from the table. The first one is that using the video data improves the results significantly compared to the single

frame classification and the second one is that as the duration of training or testing increases the false identification rate decreases.

In the following experiments the baseline parameters — comparing all camera views with each other, no camera weighting, no frame weighting, no additional samples — will be kept and whenever a parameter is changed it will be indicated in the section.

4.1 Comparison of the Well-Known Face Recognition Algorithms

In the first part of the experiments, well-known face recognition algorithms have been tested on the smart room data. The experiments are conducted frame-based. That is, an identity estimate is provided for each frame. We used all the frames from the Train B set for training and all the frames in the 20 second segments for testing. We have compared our local appearance-based face recognition (LAFR) algorithm with Eigenfaces [13], [14], linear discriminant analysis (LDA) [15] and Bayesian face recognition [16] algorithms. In the Eigenfaces and Bayesian face recognition algorithms we kept the first 320 eigenvectors, in order to have the same dimensional feature vector that we used for the LAFR approach. For Bayesian face recognition we used 1000 intra-personal and extra-personal samples. For LDA, we used the LDA+PCA algorithm provided in the CSU face identification evaluation system [17]. This version of LDA uses a soft distance measure proposed by Zhao et al. [15]. We both used the L1 and MAHCOS [14] distance metrics in the Eigenfaces algorithm. The false identification rates are given in Table 3. As can be seen the local appearance-based face recognition approach outperforms the other well-known face recognition algorithms. The most interesting result that can be observed in this table is the very high false identification rate obtained by Bayesian face recognition which has been known to be one of the best performing algorithms in the FERET evaluations [18] and which has inspired many other algorithms that utilize intra-personal and extra-personal variations. The main reason for the bad performance on the smart room database is the multiple sources of variations that exist in the database. Varying pose and illumination changes, registration errors and low resolution make the intra-personal and extra-personal variations almost identical, therefore the approach loses its discriminative capability.

Table 3. Performance Comparison of Well-known Face Recognition Algorithms

Recognizer	FI rate (%)
LAFR	43.6
Eigenfaces L1	48.6
Eigenfaces MAHCOS	59.5
LDA	49.6
Bayesian	87.4

Table 4. False Identification Results of Camera-wise Classification

Recognizer	FI rate (%)
LAFR	37.8
PCA L1	45.8
PCA MAHCOS	60.8
Fisherfaces	46.5
Bayesian	82.5

Table 5. False Identification Results of Camera-wise and All Camera Classification for LAFR

	A1	A5	A10	A20	B1	B5	B10	B20
All cameras	50.1	50.9	50.7	50.4	43.5	43.6	43.7	43.7
Camera-wise	46.7	46.9	46.7	46.4	39.7	38.1	38.2	37.8

4.2 Camera-Wise vs. All Camera Classification

In the second part of the experiments, we compared camera-wise and all camera classification. In camera-wise classification, each camera-view is handled separately. That is, the testing image acquired by a camera is only compared with the training images acquired at each site by the camera with the same label. For example, if the testing image was acquired by a camera with label 1, we only compare it with training images also acquired by a camera with label 1. On the other hand, in all camera classification the testing image acquired by a camera is compared with the training images acquired by all the cameras. Camera-wise classification has many advantages. First of all, it speeds up the system significantly. That is, if we have N images from each camera for training, and if we have R images from each camera for testing, and if we have C cameras that do recording, $(C \cdot N) \cdot (C \cdot R)$ similarity calculations are performed between all the training and testing images. However, when we do camera-wise image comparison, then we only need to do $C \cdot (N \cdot R)$ comparisons between the training and testing images. Apparently, this reduces the amount of required computation by $1/C$. In addition to the improvement in the system's speed, it also provides a kind of view-based approach that separates the comparison of different views, which was shown to perform better than doing matching between all the face images without taking into consideration their view angles [10].

Table 4 shows the false identification results of the well-known face recognition algorithms. Again, all the frames from the Train B set and all the frames in the 20 second segments are used for training and testing, respectively. This time camera-wise classification is done instead of comparing all the camera views with each

Table 6. Effect of Camera Weighting

	A1	A5	A10	A20	B1	B5	B10	B20
Video-based	34.4	25.1	22.5	19.3	31.3	22.3	21.1	15.9

other. Compared to the results in Table 3, it can be noticed that the results have been improved for each recognizer except PCA MAHCOS.

In Table 5, camera-wise and all camera classification results are presented for the LAFR algorithm for different training-testing duration combinations. Both of the classifications are performed frame-based. As can be seen at each training-testing duration combination the results improved with camera-wise classification.

4.3 Camera Weighting

In the third part of the experiments, the effect of camera weighting is analyzed. The camera weighting is performed with respect to the distance between the eyes. The higher inter-eye distance implies either a high resolution face image or a lower resolution face image with a close to frontal pose. On the other hand, a small inter-eye distance implies either low resolution face image or a higher resolution face image with a close to profile head pose. Since we would like to weight the cameras that have better view of the subject more and since higher resolution or close to frontal face images are more desirable for face recognition, we did the weighting by taking into consideration the inter-eye distance. We put more weights to the camera views with high inter-eye distances, by using weights proportional to the inter-eye distance. The obtained results can be seen in Table 6. Compared to the results at the second row of Table 2 a slight decrease in the false identification rates can be observed.

4.4 Additional Samples

In the fourth part of the experiments, we analyze the contribution of additional training sample generation to the face recognition performance on the smart room data. Note that, even with manual labelling, due to the low resolution of the face images, there can be slight errors in the eye center coordinates. To be able to handle registration errors, we generate additional registered samples from the manually labelled training images. In order to do this, we move the left and right eye center labels in their 4-neighborhood, $(x+1, y), (x-1, y), (x, y-1), (x, y+1)$. This gives 5 locations for each eye and 25 combinations of eye positions (including the original eye coordinates). The face image is then registratered using each of these 25 eye coordinates. This way, we generated 24 additional training samples per original training sample. Table 7 shows the results. Both the frame-based and video-based results improved significantly, around 10% absolute decrease is achieved in the false identification rates.

24 additional training samples implies 24 times more processing time that must be spent in a nearest neighbor classification scheme which is not desirable. To

Table 7. Effect of Using Additional Samples

	A1	A5	A10	A20	B1	B5	B10	B20
Frame-based	38.7	39.0	38.8	38.3	31.9	32.1	32.1	31.8
Video-based	28.2	17.5	17.2	11.9	22.8	13.2	11.9	9.1

Table 8. Effect of Using Additional Samples with Clustering

	A1	A5	A10	A20	B1	B5	B10	B20
Frame-based	40.5	41.9	41.9	41.7	35.0	34.3	34.2	34.1
Video-based	25.3	15.7	17.9	13.6	20.7	12.2	12.6	9.1

reduce the number of training samples we used k-means clustering. We chose k to be the number of original samples and used the resulting cluster centers as representatives. This way the processing time for classifying new images remains the same. The resulting false identification rates are shown in Table 8. The results are very close to the ones that were obtained without clustering. Even, at some cases the false identification rates decrease. These indicate that there is no need to sacrifice from the processing time in order to obtain better results using the additional samples.

4.5 Frame Weighting

In the fifth part of the experiments, we investigated the effect of frame weighting. It has been observed that the distance 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 [19]. It has been found that the distribution of these distances resembles an exponential distribution:

$$\varepsilon(x;\lambda) = 0.1\lambda \mathrm{e}^{-\lambda x} \quad \text{with } \lambda = 0.05 \tag{2}$$

The weights are then computed as the cumulative distribution function:

$$\mathcal{E}(x;\lambda) = 1 - \mathrm{e}^{-\lambda x} \tag{3}$$

Note that this distribution is extracted completely on a different database and is not specific to the mentioned smart room scenario [19]. We weighted each frame using this formula. The results are given in Table 9. Again an improvement over the baseline system is achieved.

4.6 Combining All the Parameters

In the last experiment, we combined all the parameters we have analyzed so far. We used additional samples with clustering, camera weighting and frame

Table 9. Effect of Frame Weighting

	A1	A5	A10	A20	B1	B5	B10	B20
Video-based	34.9	25.4	22.1	17.6	31.5	19.3	18.3	14.8

Table 10. Effect of Combining all the Parameters

	A1	A5	A10	A20	B1	B5	B10	B20
Video-based	26.3	17.0	16.1	11.9	21.5	12.9	11.2	9.1

weighting for this experiment. Interestingly, the improvements observed in the previous experiments do not sum up in the combined experiment. We noticed that the largest impact comes from using additional samples with clustering.

5 Conclusions

In this paper we provided a detailed analysis of face recognition in smart rooms. We first compared the well-known face recognition algorithms in order to observe how they perform on the images collected under such environments. We found the local appearance-based face recognition algorithm to be superior to the other well-known face recognition algorithms. We also observed that the Bayesian face recognition approach, which is based on intra- and extra-personal variations, does not work well on this kind of uncontrolled data. Afterwards, we investigated two typical aspects of doing face recognition in a smart room. The first one is utilizing the video data captured from multiple fixed cameras located in the room. The obtained results show that benefiting from video data provided by multiple cameras decreases the false identification rates significantly compared to the frame-based results. The second aspect is handling possible registration errors due to the low resolution of the aquired face images. We generated additional registered samples from the manually labelled training images by moving the manual eye label locations in the neighborhood and did registration with respect to the newly obtained eye coordinate pairs. We also clustered the newly generated additional samples in order to have the same number of representative training samples as we had original training samples. In both cases — without and with clustering — the false identification rates decreased significantly, which indicates that registration errors are one of the most important problems in low resolution face recognition. In addition, we also analyzed the effect of camera and frame weighting. Camera and frame weighting have been found to improve the performance further.

Acknowledgement

This work is sponsored by the European Union under the integrated project CHIL—Computers in the Human Interaction Loop, contract number 506909.

References

1. Pentland, A., Choudhury, T.: Face recognition for smart environments. Computer 33(2), 50–55 (2000)
2. Nickel, K., Ekenel, H.K., Voit, M., Stiefelhagen, R.: Audio-Visual Perception of Humans for a Humanoid Robot. In: 2nd Intl. Workshop on Human-Centered Robotic Systems, Munich, Germany (2006)
3. Erzin, E., Yemez, Y., Tekalp, A.M., Ercil, A., Erdogan, H., Abut, H.: Multimodal Person Recognition for Human-Vehicle Interaction. IEEE Multimedia 13(2), 18–31 (2006)
4. Stiefelhagen, R., Bernardin, K., Ekenel, H.K., McDonough, J., Nickel, K., Voit, M., Wölfel, M.: Audio-Visual Perception of a Lecturer in a Smart Seminar Room. Signal Processing 86(12) (2006)
5. Berg, T., Berg, A.C., Edwards, J., Forsyth, D.: Who is in the picture. In: Neural Information Processing Systems(NIPS) (2004)
6. Arandjelovic, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: CVPR 2005, Washington, DC, USA, pp. 860–867 (2005)
7. Sivic, J., Everingham, M., Zisserman, A.: Person spotting: Video shot retrieval for face sets. In: Proc. Intl. Conf. on Image and Video Retrieval, Springer, Heidelberg (July 2005)
8. Ekenel, H.K., Stiefelhagen, R.: Local appearance-based face recognition using discrete cosine transform. In: 13th European Signal Processing Conference, Antalya, Turkey (2005)
9. Ekenel, H.K., Stiefelhagen, R.: Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization. In: CVPR Biometrics Workshop, New York, USA (June 2006)
10. Pentland, A., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: CVPR 1994 (1994)
11. Heisele, B., Ho, P., Poggio, T.: Face recognition with support vector machines: Global versus component-based approach. In: ICCV 2001, pp. 688–694 (2001)
12. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The CLEAR 2006 Evaluation. In: Stiefelhagen, R., Garofolo, J. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 1–45. Springer, Heidelberg (2007)
13. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Science, 71–86 (1991)
14. Draper, B.A., Yambor, W.S., Beveridge, J.R.: Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures. In: Empirical Evaluation Methods in Computer Vision, Singapore (2002)
15. Zhao, W., Chellappa, R., Phillips, P.J.: Subspace linear discriminant analysis for face recognition, Technical Report, UMD (1999)
16. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian Face Recognition. Pattern Recognition 33(11), 1771–1782 (2000)
17. The, C.S.U.: Face Identification Evaluation System: `http://www.cs.colostate.edu/evalfacerec/`
18. Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. on PAMI 22(10) (October 2000)
19. Stallkamp, J.: Video-based Face Recognition Using Local Appearance-based Models, Thesis report, Universität Karlsruhe (TH) (November 2006)