FACE RECOGNITION FOR SMART INTERACTIONS

H.K. Ekenel, J. Stallkamp, H. Gao, M. Fischer, R. Stiefelhagen

interACT Research, Computer Science Department, Universität Karlsruhe (TH)
76131, Karlsruhe, Germany, e-mail: ekenel@ira.uka.de, web: http://isl.ira.uka.de/

ABSTRACT

In this paper an overview of face recognition research activities at the interACT Research Center is given. The face recognition efforts at the interACT Research Center consist of development of a fast and robust face recognition algorithm and fully automatic face recognition systems that can be deployed for real-life smart interaction applications. The face recognition algorithm is based on appearances of local facial regions that are represented with discrete cosine transform coefficients. Three fully automatic face recognition systems have been developed that are based on this algorithm. The first one is the "door monitoring system" that observes the entrance of a room and identifies the subjects while they are entering the room. The second one is the "portable face recognition system" that aims at environment-free face recognition and recognizes the user of a machine. The third system, "3D face recognition system", performs fully automatic face recognition on 3D range data.

1. INTRODUCTION

Person identification is one of the most crucial building blocks for smart interactions. Either as an assistant in human-human interactions, e.g. a memory aid that tells the person who he is talking to, or in human-machine interactions, e.g. a machine that recognizes its user and customizes the preferences accordingly, it provides the most important characteristic of natural interactions, *personalization*. Besides, the identity of a person can be used to improve the performances of the other perceptual technologies, such as expression analysis systems or appearance-based head pose estimation systems, by facilitating the use of person-specific models.

Among the person identification methods, face recognition and speaker identification are known to be the most natural ones, since the face and voice modalities are the modalities we use to identify people in our everyday lives. Although other methods, such as fingerprint identification, can provide better performance, they are not appropriate for natural smart interactions due to their intrusive nature. The advantage of speaker identification is its capability to perform identification over telephone lines where the person may not be visible to the identification system. In contrast, face recognition provides passive identification, that is the person to be identified does not need to cooperate or take any specific action. For example, a smart store can recognize its regular customers while they are entering the store. The customers do not need to talk or look directly to the camera to be recognized.

At the interACT Research Center, we classify the smart interaction applications we have developed into two groups:

1) Face recognition for smart environments: This application group comprises the identification tasks at a constant location [1]. For example, in a smart home, 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. This application group requires identification of people without any cooperation, and under uncontrolled conditions, without any constraints on head-pose, illumination, use of accessories, etc.
2) Face recognition for smart machines: In this application group, a machine identifies its user. For instance a car that identifies its driver [2], or a laptop that recognizes its user. In this application group an implicit cooperation exists between the person and the machine due to the standard actions the person performs, e.g. driver looking at the road, or computer user looking at the screen. Therefore, the head pose variations are limited in such systems. The difficulty in this group arises due to changing environmental conditions.

In this paper, we first explain our local appearance-based face recognition algorithm, then we introduce three fully automatic face recognition systems that are based on this algorithm. The first system, "door monitoring system", is a sample of the first group of smart interaction applications. The system identifies the individuals while they are entering a room. It operates completely in the background without requiring any cooperation or specific action. The second system, "portable face recognition system", is a sample of the second group of smart interaction applications. The system runs on a laptop computer and identifies its user at varying locations. The third system, "3D face recognition system", can be used for both smart interaction application groups. Recently, 3D face recognition has become popular, since the range data is not affected by changing illumination and environmental conditions. The developed system

utilizes the same face recognition algorithm, which we have used for the other two systems, for 3D range data. The experimental results show that the algorithm performs very well also in 3D domain and outperforms the well-known face recognition algorithms.

2. LOCAL APPEARANCE FACE RECOGNITION

Local appearance face recognition was first proposed in [3] as a generic, practical and robust face recognition algorithm. It is based on local representation of facial regions and fusion of local experts. The underlying ideas for preferring a local appearance-based approach over a holistic appearance-based approach are as follows:

- 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.
- 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.

The algorithm does not rely on detection of any salient facial features, such as eyes. It just partitions an aligned face image into 8x8 pixels resolution non-overlapping blocks. Discrete cosine transform (DCT) is used to represent the local regions. Its compact representation ability is superior to that of the other widely used input-independent transforms like Walsh-Hadamard transforms [4]. Although 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.

Feature extraction using local appearance-based face representation can be summarized as follows: A detected and normalized face image is divided into 8x8 pixels resolution blocks. Then, DCT is applied on each block. The obtained DCT coefficients are ordered using the zig-zag scanning pattern [4]. From the ordered coefficients, M of them are selected according to a feature selection strategy, and then normalized resulting in an M-dimensional local feature vector. These extracted local features are then concatenated to represent the entire face image. For details of the algorithm please see [3,5].

3. DOOR MONITORING SYSTEM

The real-time face recognition system presented in this section monitors the entrance door of a seminar room. Individuals are recognized when they enter the room. They behave naturally, since they are not required to interact with the recording system in any special way, e.g., holding their head in a certain position. As a consequence, the system is confronted with real-life facial appearance variations that are caused by changing illumination and head pose (Fig. 1).

Faces are detected in a two-stage process. First, regions of interest are determined by skin color segmentation, and then the eyes are detected with a classifier cascade of Haar-like features [6]. The eye positions are used to register the faces to a fixed orientation and scale (Fig. 2). Please note the variations in expression, illumination, pose and resolution as well as blurring effects from movement.

Fig. 1. Sample images from the door monitoring system

To evaluate the recognition performance, both a k-nearest-neighbors (k-NN) and a Gaussian mixture model (GMM) approach are used. In the k-NN case, video-based classification is achieved by summing up the normalized individual frame scores. In the Gaussian approach, this is done with Bayesian inference. As not all frames are of the same quality, a *weighting* scheme consisting of two sub-schemes is introduced into the k-NN approach to modify each frame's influence on the final decision. The scheme *distance-to-model* identifies frames that are inconsistent with the training data, therefore modeled inappropriately, and assigns them a lower weight. *Distance-to-second-closest* compares the top-2 matches and reduces a frame's weight if the classification is ambiguous, that is, if the top-2 matches in a frame are very close. A *smoothed* version of the GMM approach is also developed with the underlying idea that the identity of a person does not change over time. Consequently, frames which are inconsistent with the current hypothesis get a small weight. This approach still allows a change of identity if there is strong enough evidence, but it avoids rough sudden jumps between different classifications.

Fig. 2. Sample aligned face images.

In order to show the robustness of the local appearance-based face recognition approach under real-life conditions, we first compare it with several well-known face recognition algorithms, such as eigenfaces [7,8], Fisherfaces [9] and Bayesian face recognition [10] on the collected door database. We conduct this experiment frame-based, that is, the classification is performed based on single frames. A database of 2294 video sequences of 41 individuals, which have been automatically recorded during seven months with the proposed system, is used for the experiments. The data is divided into training and testing sets according to the recording date. The sequences recorded earlier are used for training and the ones recorded later are used for testing. Five-dimensional local DCT-based features are used for the local appearance-based face recognition algorithm, making a 320-dimensional combined feature vector. The feature vectors are classified using a nearest neighbor classifier. The L1 norm is used as a distance metric. The same feature dimensionality is used for the other face recognition approaches as well. For the eigenfaces, Mahalanobis cosine (MAHCOS) is also used as a distance metric in nearest neighbor classification. The correct identification rates are given in Table 1. The local appearance-based face recognition approach outperforms the well-known algorithms. The most interesting result that can be observed from this table is the very low correct identification rate obtained by Bayesian face recognition which has been known to be the one of the best performing face recognition algorithms. Varying pose, illumination changes, registration errors make the intra-personal and extra-personal variations almost identical, which causes this low performance.

Table 1. Correct recognition rates obtained on the door database.

Algorithm	Performance
Local DCT	**80.6%**
LDA	75.9%
PCA, L1	68.7%
PCA, MAHCOS	66.1%
Bayesian	28.0%

In video-based evaluations, the same database is used, but this time the classification is performed using the entire sequence. As can be seen from Table 2, the system successfully extends the frame-based approach proposed in [3,5] to video-based data. Correct recognition rates are significantly higher if the sequences of images are evaluated as the increased amount of input data compensates for low-quality frames. Results improve further if bad frames can be identified and their influence be reduced. Note that, the frame-based result with k-NN in Table 2 is lower than the one in Table 1. The reason is that the training samples are clustered in video-based face recognition to make the system real-time. A more detailed view of this system is beyond the scope of this overview paper. Please refer to [11] for a complete description and analysis.

Table 2. Correct recognition rates of the door monitoring system. *Weighted* and *Smooth* are only available for k-NN and GMM, respectively (see text).

	Frame-based	Video-based	Weighted	Smooth
KNN	68.4%	90.9%	92.5%	*N/A*
GMM	62.7%	86.7%	*N/A*	87.8%

4. PORTABLE FACE RECOGNITION SYSTEM

The portable face recognition system is deployed on a laptop computer and uses a standard webcam for image acquisition. The system is similar to the "door monitoring system" but faces the additional difficulties that it has to run on a less powerful mobile system and that the changes in environmental conditions can be much larger because it can be used at completely different locations. The system works as follows. New persons are learned by the system capturing a short sequence of length 150 frames of the subject. Face and eyes of the subject are detected automatically using a similar approach to the one in the "door monitoring system". Using the eye coordinates the image is then aligned and feature vectors are extracted from it using the local appearance based face recognition approach. Subjects are identified when they sit in front of the laptop. There is no required amount of time the person has to stay in front of the camera, however the confidence of the identification result increases with longer video sequences.

For the evaluation of the system, we have recorded sequences of 39 different subjects at various locations. We then train the system using the first available recording of every person. The testing set consists of 42 sequences from 14 different subjects, from which more than one sequence is available.

We applied two techniques to improve the results of the system. First we generate additional training samples by varying both eye coordinates by +2 or –2 pixels in both directions. This provides 80 additional images for each input image. To keep the classification fast we use k-means clustering to reduce the number of training vectors to the same number as without the additional samples. Second we preprocess the input images using the local binary patterns (LBP) operator [12]. This operator replaces the image intensities by a label describing the local image texture. This method reduces the effect of varying illumnation conditions on the system. The results of our evaluation can be seen in Table 3. As can be seen, both the additional steps improve the recognition rate significantly, with the system reaching around 79 % correct recognition rate.

Table 3. Correct recognition rates of the portable system.

Algorithm	Performance
Local DCT	52.4%
Extended Local DCT	64.3%
Extended Local DCT with LBP	78.6%

5. 3D FACE RECOGNITION SYSTEM

The proposed 3D face recognition system utilizes depth map images to extract 2D local features. The block diagram in Fig. 3 shows the processing steps to generate a depth image from 3D range data. A test face is first preprocessed to fill holes and to remove noise on the face surface. Eleven landmarks representing salient facial feature points are automatically estimated according to mean landmarks. With these landmarks, the face surface is warped to a base mesh using thin plate spline (TPS) deformation. This deformation transforms all test faces to a common framework, so that they have the same size, position and pose. The expression variations are also mitigated with the TPS process. After re-sampling the warped face surface with the base mesh, the corresponding depth map is created by picking the z-coordinate of each vertex.

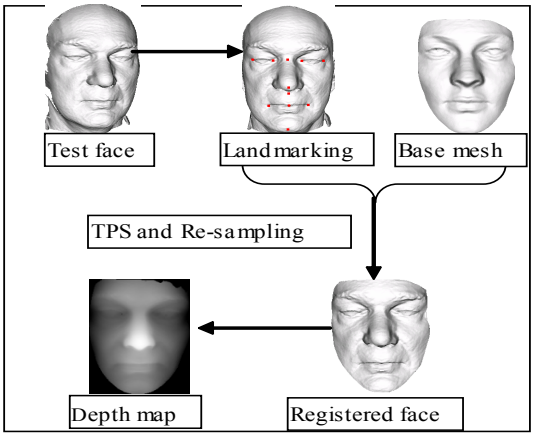

Fig. 3. Block diagram of the 3D face recognition system

The system was tested on the FRGC version 2 data set [13]. The 3D data corpus of FRGC database was collected by imaging subjects using a range scanner. We used the range images that were acquired in spring 2003 for training and the ones recorded in spring 2004 for testing. The training data contains neutral expressions, whereas the testing data contains different expressions, such as frowning, smiling, etc. In total, we used 218 range images of 109 subjects for training, where each individual has two samples, and 758 range images for testing, where individuals have different numbers of samples ranging from two to twelve. Table 4 shows the results of our approach and the other well known face recognition approaches. Our algorithm is found to be superior also in the 3D domain. For details please see [14].

Table 4. Correct recognition rates of 3D face recognition.

Methods	Performance
Local DCT	**93.1%**
Eigenfaces	86.5%
LDA	88.5%
Bayesian	89.7%

6. CONCLUSIONS

In this paper, we gave an overview of the face recognition research activities at the interACT Research Center. We first explained our face recognition algorithm. In addition, three fully automatic face recognition systems are also presented. Both of the presented 2D face recognition systems attain high correct recognition rates and they both run in real-time. The algorithm is also shown to perform well in the 3D domain. Videos of the systems can be downloaded from http://isl.ira.uka.de/~ekenel/demovideos.zip.

7. ACKNOWLEDGEMENTS

This work was sponsored by the European Union under the integrated project CHIL, *Computers in the Human Interaction Loop*, contract number 506909 and by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 *Humanoid Robots – Learning and Cooperating Multimodal Robots*.

8. REFERENCES

[1] R. Stiefelhagen et al., “Audio-Visual Perception of a Lecturer in a Smart Seminar Room”, Signal Processing, Vol.86(12), 2006.
[2] E. Erzin et al., “Multimodal Person Recognition for Human-Vehicle Interaction”, IEEE Multimedia, Vol.13(2), p.18-31, 2006.
[3] H.K. Ekenel, R. Stiefelhagen, "Local Appearance based Face Recognition Using Discrete Cosine Transform", 13th European Signal Processing Conference (EUSIPCO 2005), Turkey, 2005.
[4] R.C. Gonzales, R.E.Woods. Digital Image Processing. Prentice Hall, 2001.
[5] H.K. Ekenel, R. Stiefelhagen, "Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization", CVPR Biometrics Workshop, NYC, USA, 2006.
[6] P. Viola, M. Jones, “Robust Real-Time Face Detection”, Intl. Journal of Computer Vision, Vol. 57 (2), pp. 137-154 2004.
[7] M. Turk, A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Science, pp. 71–86, 1991.
[8] B.A. Draper et al., “Analyzing PCA-based Face Recognition Algorithms: Eigenvector Selection and Distance Measures”, Empirical Evaluation Methods in Computer Vision, 2002.
[9] W. Zhao et al., "Sub- space Linear Discriminant Analysis for Face Recognition", UMD, 1999.
[10] B. Moghaddam et al., "Bayesian Face Recognition". Pattern Recognition, Vol. 33 (11), pp. 1771-1782, 2000.
[11] J. Stallkamp, “Video-based Face Recognition using Local Appearance-based Models”, Thesis report, Universität Karlsruhe (TH), Nov. 2006
[12] T. Ahonen, A. Hadid, M. Pietikainen, “Face description with local binary patterns: application to face recognition”, IEEE Trans. on PAMI, Vol. 28 (12), pp. 2037-2041, 2006.
[13] P.J. Phillips et al., “Overview of the Face Recognition Grand Challenge”, CVPR 2005, San Diego, USA, June 2005.
[14] H. Gao, “Local Appearance-based 3D Face Recognition”, Thesis report, Universität Karlsruhe (TH), Nov. 2006.