VIDEO-BASED DRIVER IDENTIFICATION USING LOCAL APPEARANCE FACE RECOGNITION

J. Stallkamp[1], H.K. Ekenel[1], H. Erdoğan[2], R. Stiefelhagen[1], A. Erçil[2]

[1]Universität Karlsruhe (TH), 76131 Karlsruhe, Germany
{jstallkamp,ekenel,stiefel}@ira.uka.de
[2]Sabanci University, 34956 Istanbul, Turkey
{haerdogan,aytulercil}@sabanciuniv.edu

ABSTRACT

In this paper, we present a person identification system for vehicular environments. The proposed system uses face images of the driver and utilizes local appearance-based face recognition over the video sequence. To perform local appearance-based face recognition, the input face image is decomposed into non-overlapping blocks and on each local block, discrete cosine transform is applied to extract the local features. The extracted local features are then combined to construct the overall feature vector. This process is repeated for each video frame. The distributions of the feature vectors over the video sequence are modeled using a Gaussian distribution function at the training stage. During testing, the feature vector extracted from each frame is compared to each person's distribution, and individual likelihood scores are generated. Finally, the person is identified as the one who has maximum joint-likelihood score over the whole video sequence. To assess the performance of the developed system, extensive experiments are conducted on different identification scenarios, such as closed-set identification, open-set identification and verification. For the experiments a subset of the CIAIR-HCC database, an in-vehicle data corpus that is collected at the Nagoya University, Japan is used. We show that, despite varying environment and illumination conditions, that commonly exist in vehicular environments, it is possible to identify individuals robustly from their face images.

Index Terms— Local appearance face recognition, vehicle environment, discrete cosine transform, fusion.

1. INTRODUCTION

Person identification is one of the most interesting signal processing problems in smart vehicles. It can facilitate many useful applications, such as automatic customization of the vehicle's environment or driver assistance according to the person's identity. Among the person identification techniques, face recognition is one of the most addressed techniques due to its naturality as a biometric trait. The intense research efforts on face recognition have provided significant improvements in face recognition performance under controlled laboratory conditions. However, face recognition in uncontrolled real-world environments is still a very difficult problem [1].

Face recognition algorithms suffer mainly from varying head pose and illumination conditions [1]. For face recognition in smart vehicles, the main problem stems from continuously changing environment while driving, hence continuously varying illumination conditions on the driver's face. On the other hand, the driver has a specific location and he/she moves his/her head within a limited range during driving. Thus, the pose variation is limited in a car which simplifies the problem.

Person identification in vehicles has recently attracted many research efforts [2,3,4,6]. In [2], acoustic data and lip images are used to determine the person's identity. In this study, lip images detected at each video frame are transformed onto an eigenspace that is constructed using the lip images in the training stage. The obtained eigenlip coefficients are fused with acoustic features, mel frequency cepstral coefficients (MFCC), of the corresponding speech signal. A hidden Markov model (HMM) is fed with the combined feature vector for classification. In [3], driving signals such as acceleration and brake pedal pressure readings, and vehicle speed variations, are investigated to find out whether they are useful for person identification. Modelling the distribution of these signals using mixtures of Gaussians and doing classification based on maximum-likelihood scores have been found to be a good way to identify the driver. Audio-visual information, as well as driving signals are utilized to perform multimodal identification in [4]. In this study, for speaker identification, mel-frequency cepstral coefficients are used, whereas eigenfaces [5] algorithm is employed for face recognition. Pressure readings of acceleration and brake pedals and their time-derivatives are used to model the driving behaviour. For each modality, a Gaussian mixture model (GMM) is trained to model each person's biometric data for classification. The modalities are combined using a weighted sum rule. Finally in [6], acoustic features and

visual features extracted from lip and face images are combined using a sum rule and an adaptive cascade rule. The same feature extraction methods as in [2,4] are used in this study. In [4,6], it has been shown that face recognition's correct classification performance is relatively lower than the performance of speaker identification (89% vs. 98%, respectively). This result was expected due to continuously changing illumination conditions, occlusion, e.g., by the steering wheel or a close-talk microphone, and low resolution faces. Taking these problems into account, we aim to provide a robust face recognition system in this study that can reach the same performance levels as speaker identification. To overcome uncontrolled environmental conditions, we are utilizing local appearance-based face recognition [7,8] on multiple samples of the same face that are obtained from the video sequence. Using multiple evidences of the same biometric modality has been shown to improve the performance [9,10]. In addition, local appearance-based face recognition has been shown to be a superior approach over standard holistic based approaches [7,8]. The reason for the better performance is the ability to handle, to some extent, facial appearance variations caused by occlusion, illumination and expression, where a local change affects only the corresponding part of the representation and does not modify the representation vector as a whole. These variations can lead to modifications on the entire representation coefficients in a holistic representation scheme. The proposed local approach has been also tested in CLEAR and Face Recognition Grand Challenge (FRGC) evaluations and significant improvement has been observed over the baseline face recognition algorithm using holistic PCA/eigenfaces [11,12].

The organization of the paper is as follows. In Section 2, the proposed face recognition algorithm is explained. Experimental results are presented and discussed in Section 3. Finally, in Section 4, conclusions are given.

2. LOCAL APPEARANCE FACE RECOGNITION

Local appearance face recognition is based on the separate representation of local facial regions and their combination which preserves spatial relationships. In [7], discrete cosine transform (DCT) is proposed to be used to represent the local regions. DCT has been shown to be a better representation method for modelling the local facial appearance compared to principal component analysis (PCA) and discrete wavelet transform (DWT) in terms of face recognition performance [7].

Discrete cosine transform for 2D input f(x, y) is defined as

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right]$$

for *u, v = 0, 1, … , N - 1* where

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} & for\ u = 0 \\ \sqrt{\frac{2}{N}} & for\ u = 1,2,\ldots,N-1 \end{cases}$$

The corresponding bases are shown in Figure 1. The first coefficient, the one at position (0, 0), represents the average value of the input signal, i.e., in our case, the average gray level. Lower order coefficients represent lower frequencies, whereas higher order coefficients correspond to higher frequencies. An image can be represented just by using the lower frequencies without losing much of the total information content of the image. In order to benefit from this property and to provide compact representation, the DCT coefficients are ordered using a zig-zag scanning pattern [13], and only the first few coefficients that correspond to high energy content are selected to represent the entire local face region.

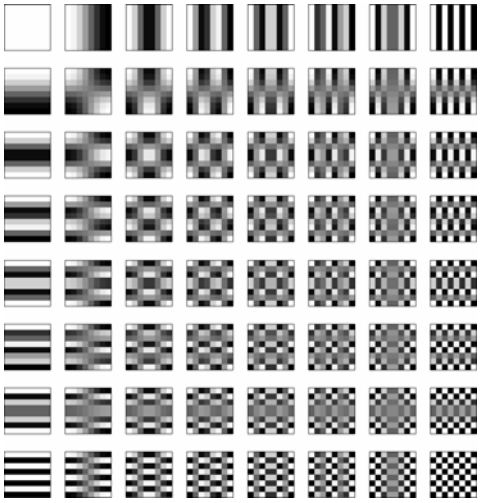

Figure 1. DCT basis functions for N = 8. (0, 0) is at the top left corner.

Feature extraction using local appearance-based face representation can be summarized as follows: A detected and normalized face image is divided into blocks of 8x8 pixels size. Each block is then represented by its DCT coefficients. The top-left DCT coefficient is removed from the representation in order to increase robustness against illumination variations, since it represents the average intensity value of the block. From the remaining DCT coefficients the ones containing the highest information are extracted via a zig-zag scan. These extracted local features are then concatenated to represent the entire face image and they are normalized to have unit norm. For each video frame this process is repeated. The distribution of the feature vectors over the video are modeled using a Gaussian distribution function at the training stage.

During testing, the feature vector extracted from each frame is compared to each person's distribution, and

individual likelihood scores are generated. Finally, the person is identified as the one who has maximum joint-likelihood score over the video.

3. EXPERIMENTS

The experimental data set consists of a subset of CIAIR-HCC database, an in-vehicle data corpus that was collected at the Nagoya University, Japan [14]. It contains image sequences of ten male and ten female subjects while they are driving. Each person has twenty image sequences and each of these image sequences contains 25 frames. The sequences are sparsely sample from all available video data. The recordings are done with a camera that is mounted in the front left part of the car facing the driver. A sample shot from the camera is shown in Figure 2.

Figure 2. A sample image from the recording

Most of the images have been automatically extracted from the video stream with a Haar features-based face detector [15] using OpenCV library [16]. Sample automatically cropped face images can be seen in Figure 3. On parts of the videos where this procedure failed (due to occlusion by the steering wheel or the driver's hand, for example), faces have been selected manually. The cropped faces are then scaled to 64x64 pixels resolution. In local appearance face recognition, five DCT coefficients are used to represent each local region. To provide a baseline system, PCA/eigenfaces algorithm is implemented [5]. In the PCA approach, the first 20 eigenvectors are used to represent the face images.

Figure 3. Sample cropped face images

In the experiments, 20-fold cross-validation is performed. The feature vectors have been modeled with a single Gaussian per person for face recognition. A generic face model is constructed, with a Gaussian mixture model using three Gaussians, for face verification. The experiments consist of the following three tasks:

Closed-set identification: This case corresponds to the identification scenario we envision to be used in a vehicle as one of the components of smart human-car interfaces, which can be defined as follows: From a known set of drivers, such as family members, find out who drives the car. After determining the identity, the vehicle can customize itself automatically according to the pre-learned preferences of the current driver. In this experiment, the system is trained on all 20 individuals and every test sequence is assigned to an identity in the training set. Performance is measured as correct classification rate (CCR).

Verification: This case corresponds to a more security-oriented application. In this scenario, the driver tries to convince the vehicle that he/she is one of the genuine drivers of the car. In this experiment, the system is trained on all 20 individuals and each test sequence is compared against each individual's model. This results in 1 genuine and 19 impostor tests to see whether they get correctly or falsely accepted or rejected. With the cross-validation approach mentioned above, this results in 7600 impostor test (19 per set × 20 individuals × 20 sets). The performance can be modified by a threshold on the likelihood, in order to accept or reject the identity. Performance is measured as equal error rate (EER), which is defined as the point on the receiver operating characteristics (ROC) curve where the false accept rate (FAR) and the false reject rate (FRR) are equal.

Open-set identification: This case corresponds to a scenario, in which the vehicle should determine whether the driver is one of the genuine drivers of the car and if yes, find out who he/she is. In this test, one person at a time is ex-cluded from training and afterwards presented to the system as an impostor (leave-one-out). This is repeated for every person in the set. In this scenario, three types of error exist:

1. False accept: The impostor is accepted as one of the individuals in the database.
2. False reject: An individual is rejected even though he/she is present in the database.
3. False classify: An individual in the database is correctly accepted but misclassified as one of the other individuals in the training data.

Since there are three error types, we have to redefine the EER from above to take this into account. Thus, EERO is defined as the error rate where the false accept rate is equal to the sum of the false reject and false classification rate.

Table 1 shows the obtained experimental results. Results are given as correct classification rate for closed-set identification, equal error rate (EER) for verification and redefined equal error rate (EERO) for open-set identification. It is noteworthy, that the correct classification rate

increased to 96.3% compared to the 89% achieved by PCA. That is a decrease in error rate by approximately two thirds. A performance improvement can be also observed for the verification and open-set identification tasks.

	PCA (%)	DCT (%)
Closed Set	89.0	96.3
Verification	7.3	5.5
Open Set	18.3	17.6

Table 1. Experimental results for video-based identification.

In order to assess how much the video-based approach adds to face recognition results, both techniques have been evaluated on single frames for the closed-set identification task. The results are shown in Table 2. As can be seen, the CCRs of both the PCA and DCT-based approaches increase significantly when they utilize image sequences to determine the identity instead of just a single frame. Again, also on single frames, the local appearance-based face recognition approach performs superior to the eigenfaces algorithm.

	PCA (%)	DCT (%)
Single Frame	77.1	88.5
Video	89.0	96.3

Table 2. Results for PCA and DCT face recognition on single frames vs. video sequences.

4. CONCLUSIONS

In this paper, a robust person identification system for vehicular environments is presented. Local appearance-based face recognition is performed on image sequences to overcome uncontrolled environmental conditions, and occlusions due to e.g. steering wheel or driver's hand. 67.4% decrease is obtained in false classification rate using multiple samples of the individual's face images instead of a single frame to determine the identity. Using local appearance-based face recognition instead of eigenfaces on image sequences decreased the false classification rate by 65.9%. Overall, 83.6% decrease in false classification rate is achieved compared to performing eigenfaces algorithm on single images.

	Decrease in the error rate
Single frame vs. video	67.4%
PCA vs. DCT	65.9%
Overall	83.6%

Table 3. Decrease in the error rate due to, first row: using video instead of single frame, second row: using local appearance face recognition instead of eigenfaces on image sequences. Third row: Overall improvement compared to the single frame eigenfaces performance.

With these performance improvements, we obtained a robust face identification system, that performs almost as well as the speaker identification system (96.3% vs. 98%, respectively) [4,6]. This result shows that without expecting the collaboration of the driver, i.e. the driver speaking to the car, the driver can be identified implicitly within one second (25 frames) while he/she is in the car.

ACKNOWLEDGEMENTS

We would like to acknowledge Professor Kazuya Takeda of Nagoya University and his laboratory for providing the CIAIR database.

This work is sponsored by the European Union under the FP6-2004-ACC-SSA-2016684 SPICE project.

5. REFERENCES

[1] W. Zhao et al., "Face Recognition: A Literature Survey", ACM Computing Surveys, Vol. 35, No. 4, pp. 399-458, 2003.

[2] A. Kanak et al., "Joint Audio-Video Processing for Robust Biometric Speaker Identification in Car", Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan, April 2003.

[3] K. Igarashi et al., "Biometric Identification Using Driving Behavior", IEEE ICME 2004, Taipei, Taiwan, June 2004.

[4] H. Erdogan et al., "Multimodal Person Recognition for Vehicular Applications", 6th Intl. Workshop on Multiple Classifier Systems (MCS 2005), California, USA, June 2005.

[5] M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal Cognitive Neuroscience, Vol. 3, pp. 71-86, 1991.

[6] E. Erzin et al., "Multimodal Person Recognition for Human-Vehicle Interaction", IEEE MultiMedia, Vol. 13, No. 2, pp.18-31, April-June 2006.

[7] H.K. Ekenel, R. Stiefelhagen, "Local Appearance based Face Recognition Using Discrete Cosine Transform", EUSIPCO 2005, Antalya, Turkey, September 2005.

[8] H.K. Ekenel, R. Stiefelhagen, "Analysis of Local Appearance-based Face Recognition: Effects of Feature Selection and Feature Normalization", CVPR Biometrics Workshop, New York, USA, June 2006.

[9] P.J. Phillips et al., "Overview of the face recognition grand challenge", CVPR 2005, San Diego, USA, June 2005.

[10] Classification of Events, Activities and Relationships (CLEAR'06) Evaluation Workshop, www.clear-evaluation.org.

[11] H.K. Ekenel, R. Stiefelhagen, "A Generic Face Representation Approach for Local Appearance based Face Verification", CVPR Workshop on FRGC Experiments, San Diego, USA, June 2005.

[12] H.K. Ekenel, Q. Jin, "ISL Person Identification Systems in the CLEAR Evaluations", CLEAR Evaluation Workshop, Southampton, UK, April 2006.

[13] Gonzales, R.C., Woods, R.E., Digital Image Processing, Prentice Hall, 2001.

[14] N. Kawaguchi et al., "Multimedia Data Collection of In-Car Speech Communication", EUROSPEECH, 2001.

[15] P. Viola, M. Jones, "Robust Real-Time Face Detection", Intl. J. of Computer Vision, Vol. 57, No. 2, pp. 137-154, May 2004.

[16] The Open Computer Vision Library (OpenCV), http://sourceforge.net/projects/opencvlibrary/.