ARTICLE IN PRESS

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A video-based door monitoring system using local appearance-based face models [☆]

Hazım Kemal Ekenel [a,*], Johannes Stallkamp [a,1], Rainer Stiefelhagen [a,b]

[a] Karlsruhe Institute of Technology, Institute of Anthropomatics, Adenauerring 2, 76131 Karlsruhe, Germany
[b] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, 76131 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 14 January 2008
Accepted 4 June 2009
Available online xxxx

Keywords:
Face recognition from video
Face detection
Feature tracking
Door monitoring
Discrete cosine transform
Fusion

ABSTRACT

In this paper, we present a real-time video-based face recognition system. The developed system identifies subjects while they are entering a room. This application scenario poses many challenges. Continuous, uncontrolled variations of facial appearance due to illumination, pose, expression, and occlusion of non-cooperative subjects need to be handled to allow for successful recognition. In order to achieve this, the system first detects and tracks the eyes for proper registration. The registered faces are then individually classified by a local appearance-based face recognition algorithm. The obtained confidence scores from each classification are progressively combined to provide the identity estimate of the entire sequence. We introduce three different measures to weight the contribution of each individual frame to the overall classification decision. They are distance-to-model (DTM), distance-to-second-closest (DT2ND), and their combination. We have conducted closed-set and open-set identification experiments on a database of 41 subjects. The experimental results show that the proposed system is able to reach high correct recognition rates. Besides, it is able to perform facial feature and face detection, tracking, and recognition in real-time.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Building a robust face recognition system is one of the biggest challenges in computer vision research. A wide range of possible application areas, such as access control, surveillance, automatic photo tagging etc., have fueled significant amount of research efforts on this problem. However, most of the studies on face recognition have been conducted on data that was collected under controlled conditions [1]. This type of data contains changes in facial appearance that are generated by modifying a single or a combination of two variation sources in a controlled way. The main variation sources that have been mainly focused on are expression, illumination, occlusion, pose and time gap between training and testing data. Although the studies that have been conducted on this type of data show the tested algorithms' performance against a specific type of facial appearance variation and provide insights about face recognition under these specific conditions, they are not sufficient to imitate real-world conditions due to two main reasons. In real-world, the variations of facial appearance are caused by combinations of multiple sources in a continuous manner, that is, for example, one needs to deal with face images from any view angle, whereas in the databases collected under controlled conditions, there are only discrete head pose classes. Moreover, most of the face recognition algorithms are tested on already cropped and aligned face images that are registered according to manually labeled fiducial points on the faces. However, it is known that errors in the registration of the face deteriorate performance of face recognition significantly [2,3]. Therefore, a real-world system must be robust against registration errors that may occur due to imperfect fiducial point localization.

The need to determine or verify the identity of a person in a wide range of application areas has also led to many commercial face recognition systems. Most of these commercial systems are mainly focused on security related applications, such as access control or surveillance [4–8]. In addition, systems with multimedia focus are also available, such as searching celebrities in videos [9] or automatic photo tagging [10].

In this paper, extending our previous work [11], we present a robust real-world face recognition system for smart environments that identifies the individuals while they are entering a room. The main motivation to build a face recognition system to monitor the people entering a room is the wide range of applications in which it can be used. Both for surveillance of public areas, e.g., airports, and for people monitoring in smart environments, e.g., smart homes, one of the best instants to identify the persons is the moment they are entering the room. This provides face images with resolutions ranging from 45 × 45 pixels to 100 × 100 pixels. It facilitates face

[☆] A demo video of the system can be downloaded from http://cvhci.ira.uka.de/~ekenel/door_monitoring.mpg.

[*] Corresponding author.

E-mail addresses: ekenel@kit.edu (H.K. Ekenel), jstallkamp@ira.uka.de (J. Stallkamp), rainer.stiefelhagen@kit.edu (R. Stiefelhagen).

URLs: http://cvhci.ira.uka.de/~ekenel/ (H.K. Ekenel), http://cvhci.ira.uka.de/~stiefel/ (R. Stiefelhagen).

[1] Present address: Ruhr-Universität-Bochum, Institut für Neuroinformatik, 44801 Bochum, Germany.

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2009.06.009

detection by providing a smaller number of faces to detect and less cluttered background, e.g., compared to an airport hall where a face detection software is expected to detect many faces simultaneously in a very crowded and cluttered scene.

1.1. Previous work

Numerous approaches have been developed to recognize faces. While the main focus was on image-based methods in the beginning [12–14], it shifted more and more towards video-based approaches in the last years. These are developed in order to overcome shortcomings of image-based recognizers like sensitivity to low resolution, pose variations and partial occlusion.

Zhou et al. [15] use sequence importance sampling (SIS) to propagate a joint posterior probability distribution of identity and motion over time to do tracking and recognition of a person simultaneously. To overcome continous changes of head pose and facial expressions, Lee et al. [16] represent the appearance of a person by the means of pose manifolds which are connected by transition probabilities. In order to model person-specific appearance and dynamics, Liu and Chen [17] train individual hidden Markov models (HMM) on eigenface image sequences. In their approach, they use sequences which resulted in classification results with high confidences to adapt the models. The problem with models that are based on probability distributions is that they make strong assumptions about underlying distributions in the training set. A model may implicitly learn dependencies which are not characteristic for the data, if the training set turns out not to be representative. The counterpart are exemplar-based approaches which generally do not assume an underlying distribution and are, thus, less affected by non-representative training data.

A large variety of head pose and illumination variations, as well as occlusion, is encountered in feature films. Arandjelovic and Zisserman [18] built a system to retrieve all faces from a film that match one or multiple query images. The appearance-based approach uses a modified Euclidean distance for classification. Instead of doing frame-based retrieval, Sivic et al. [19] group all face views of a person within the same shot into a face-track, represented as a histogram. Given a query image in one of the scenes, the corresponding face-track is determined. All matching face-tracks are retrieved from the whole film by means of a chi-square goodness-of-fit test.

Face recognition systems that are to be deployed in a real-life scenario usually encounter the problem that they are confronted with unknown people. Li and Wechsler [20] make use of transduction to derive a rejection criterion. The k-nearest neighbors of a test sample are iteratively misclassifed to determine an error distribution. If classification of the test sample as any of the classes does not yield a credibility sufficiently different from this distribution, it is rejected, otherwise it is classified.

1.2. Motivation

The goal of this work is to build a real-time capable face recognition system (FRS) for smart environments. Sample application areas can be a smart lecture or meeting room, where the participants can be identified automatically; a smart home, identifying the family members while they are entering the rooms of the house or a smart store that can recognize its regular customers. While the number of subjects to be identified in these scenarios is limited, the central challenge arises from the aim of achieving unobtrusive recognition. The face recognition system is supposed to work in the background without the need of specific user interaction. The people to be recognized are not to be disturbed or interrupted in their actions by the presence of the computer vision system. This is essential to grant users the freedom to behave naturally. As a consequence of this freedom, difficulties arise from varying pose, like out-of-plane rotations, and different facial expressions. Accessories and facial hair can cause partial occlusions. Daylight leads to very different illumination depending on the time of day, time of year and weather conditions. In spite of these hardly controllable natural influences, even the artificial light sources are withdrawn from the system's control if unobtrusive recognition as postulated above is to be implemented. Since the users, i.e., the persons to be recognized, are not supposed to be restrained by the system, they are free to switch on and off any light sources that might be available. This leads to a wide variety of illumination configurations in terms of light intensity, direction and even color.

In the given scenario, the developed system is deployed at the entrance door to a seminar room. The camera is located opposite

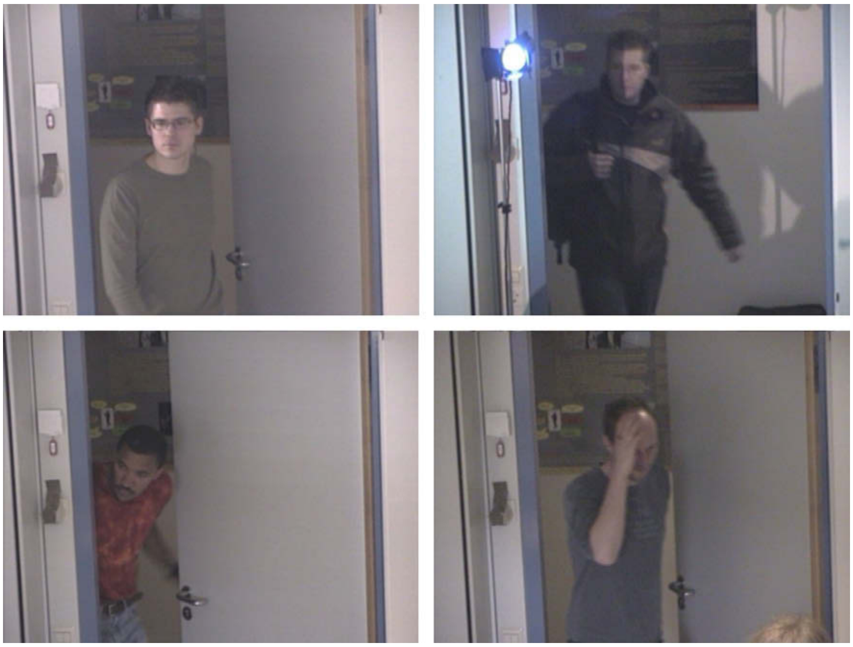

Fig. 1. Exemplary recognition situations showing a variety of different lighting, pose and occlusion conditions. No individual explicitly looks into the camera.

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

the door with a distance of approximately six meters. Individuals are recognized when they enter the room. Depending on their intention, they turn sideways to get to the seminar area or collect a print-out, walk straight through to the next room or just stand in the door frame for some time before they leave. As outlined above, they are not explicitly cooperating, and recording conditions can vary largely. Some example views are shown in Fig. 1.

1.3. Our approach

In this paper, we propose a real-time video-based face recognition system for the mentioned real-world setting. The system consists of a robust eye tracking algorithm, that provides consistent eye locations to allow face registration, and a video-based face classification algorithm, that uses registered face images to derive an identity estimate.

The developed face classification system benefits from a local appearance-based face representation [21,22] and utilizes the video information in order to robustly handle strong variations in the data. Two main observations are exploited to derive two different schemes to weight the contribution of each individual frame to the overall classification result. The first, *distance-to-model* (DTM), takes into account how similar a test sample is to the representatives of the training set. The second, *distance-to-second-closest* (DT2ND), reduces the impact of frames which deliver ambiguous classification results. As a third measure, a combination of the two schemes is used.

The remainder of this paper is organized as follows. First, the face detection and video segmentation process is explained in Section 2. Face registration method is presented in Section 3. Details about the feature extraction and classification steps, including the introduction of the weighting schemes, are given in Section 4. Our approach is evaluated in Section 5. Section 6 finishes the paper with a conclusion and future directions.

2. Video segmentation

The face recognition system needs to first detect the instants at which someone is entering the room. This subsystem, that we named as *face recorder*, consists of three major parts: color-based

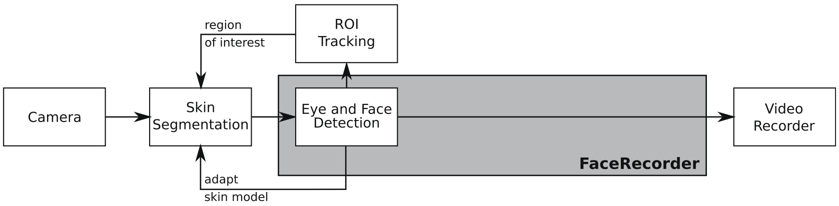

Fig. 2. Overview of the data collection system.

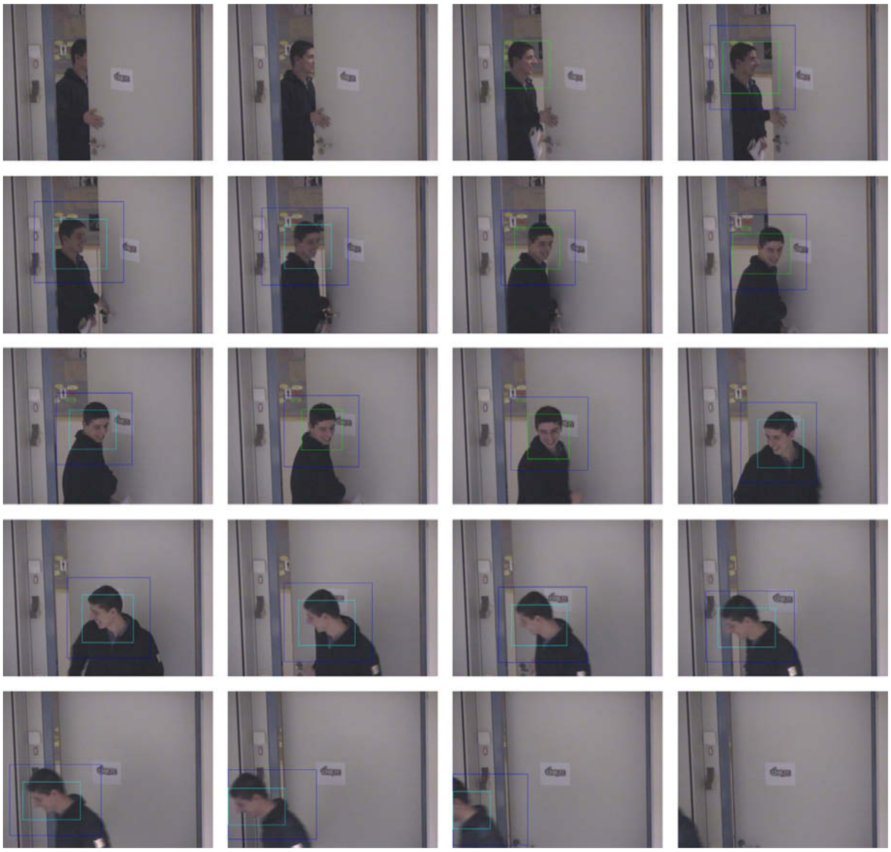

Fig. 3. Example of a recorded sequence. Blue box shows the search region and the green box shows the detected face. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

skin segmentation using ratio histogramming in order to select face candidates, feature-based face detection to confirm or discard them and a basic tracking method to ensure the complete entering sequence is recorded. Fig. 2 gives an overview of the system. A sample recorded sequence can be seen in Fig. 3.

The system's functional blocks are explained in the following subsections.

2.1. Skin color segmentation

In the given scenario, the face to be recognized is comparatively small with respect to the image dimensions of 640 × 480 pixels. In order to avoid unnecessary processing of the background, it is crucial to concentrate on meaningful areas of the image. To identify these regions, the image is searched for skin-like colors.

2.1.1. Skin color representation

In this study, a histogram-based model of 128 × 128 is used to represent the skin color distribution. It is learned from a representative training set of skin samples which are manually cropped from images by selecting large skin areas in faces in a set of input images. It is non-parametric and makes no prior assumption about the actual distribution of skin colors. The model utilized in this work is located in the normalized-rg color space.

The advantage of choosing a chrominance-based color space is a reduced sensitivity to illumination influences. At the same time, different skin tones, due to different ethnic backgrounds, get more similar to each other in this representation, forming a compact cluster in color space.

Based on a physical model for skin-reflectance, Störring et al. [23,24] show that the skin color of subjects with different backgrounds under illumination of varying color temperature generally forms an eye-brow-like shaped region in the chromaticity plane. This region is commonly referred to as skin-reflectance locus or, in short, *skin locus*.

Fig. 4a visualizes the skin model that we derived from 242 training samples or, to be more precise, 799,785 training pixels, captured with a Canon VC-C1 camera. The shape of our model is more elliptic, because the actual shape of the skin locus is camera-dependent [25]. Since in the given scenario, the encountered face sizes range from approximately 45 × 45 to 100 × 100 pixels, the model is scaled with respect to an average face size of 70 × 70 pixels. We will refer to this initial model as M_0.

During skin segmentation, we will use the skin locus to limit model adaptation. To do this, we describe its outline with two quadratic functions f_{min} and f_{max} that we fit to the boundary points of the skin distribution, i.e., to the outer-most histogram bins with non-zero count. The result can be seen in Fig. 4b.

A certain color (r, g) is part of the locus if

$$g > f_{min}(r) \quad \wedge \quad g < f_{max}(r) \tag{1a}$$

with

$$f_{min}(r) = 6.38r^2 - 4.79r + 1.15 \tag{1b}$$

$$f_{max}(r) = -3.51r^2 + 2.53r - 0.06. \tag{1c}$$

2.1.2. Image segmentation

The segmentation process is based on histogram backprojection, a technique that highlights colors in the image which are part of a histogram-based color model [26]. For a single image, the probability of a pixel being skin given a color vector (r, g) can be easily derived by application of Bayes' rule as described in detail in [27]. The result is a ratio histogram R, computed from the skin model histogram S and the image histogram I

$$p(skin|r, g) = R(r, g) = \frac{S(r, g)}{I(r, g)} \tag{2}$$

where r and g denote the histogram bin. Next, R is *backprojected* [26] onto the original image, which means, that each pixel $i(x, y)$ is replaced by $R(r_{x,y}, g_{x,y})$, where $r_{x,y}$ and $g_{x,y}$ denote the normalized color values of $i(x, y)$. In other words, R is used as lookup-table between pixel color and skin probability. This results in a gray scale image which can be interpreted as a probability map of skin presence. As elaborated in [27], application of Bayes' rule is only correct, if applied to the same image from which the histograms were originally computed. However, in practice, this works reasonably well for other images taken in a similar scenario. Backprojecting the ratio histogram instead of the model histogram itself emphasizes colors that are characteristic for the model. In turn, colors which are part of the model but which are also common in the background are weakened.

In [28], it is stated that the background remains noisy in cluttered environments but this issue is successfully addressed with a two-stage thresholding algorithm based on region-growing. The first stage is a basic binary threshold at level T_{high}, which is set to 100. The second one is a hysteresis threshold similar to the one introduced by Canny [29] for edge detection. It uses a lower threshold value T_{low} than the initial one but it only adds those pixels to the previously created binary image which are 8-connected to already selected pixels. The thresholded image is less cluttered, if the backprojection is smoothed using a Gaussian kernel because this mitigates interlacing effects and noise. Morphological operators have been omitted for speed reasons. Possible *face candidates*

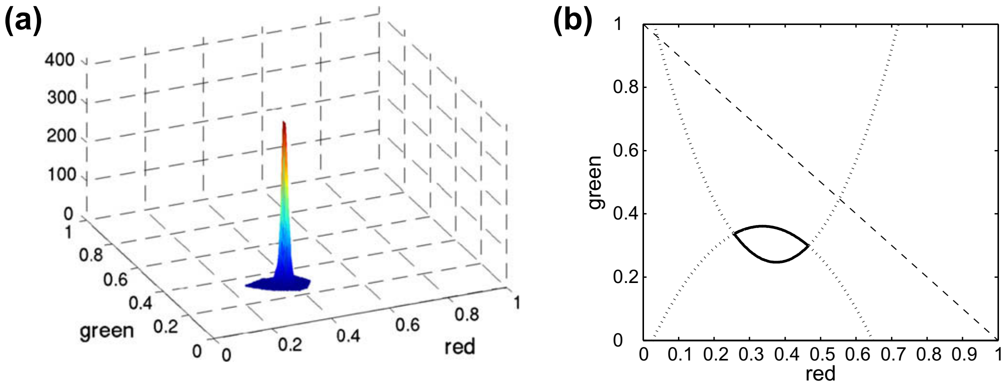

Fig. 4. (a) The skin color distribution as determined from a training set. (b) The skin locus in normalized-rg color space, described by two functions of quadratic order. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

ARTICLE IN PRESS

are extracted from the thresholded image using a connected components algorithm [30].

The lower threshold T_{low} is determined adaptively. It is chosen as the average gray level of the non-black pixels of the backprojection, i.e., as the mean probability of all skin-like colored pixels. This approach has a major advantage over a constant value of T_{low}. If the skin of an entering person is only poorly represented by the current model, due to color, size or both, only a small percentage of the skin pixels will be larger than $T_{\text{ high}}$ while the majority will have comparatively small values. If a constant T_{low} is chosen too large, these pixels will not be segmented. Choosing T_{low} small enough to successfully segment the badly modeled skin pixels, problems arise when a well-modeled face is encountered. The skin pixels of such a face will, to a large extent, get high probabilities of being skin. As a consequence, application of T_{high} already leads to reasonable segmentation. The small T_{low} from before will then add unnecessary clutter to the segmented image.

2.1.3. Model adaptation

The model generated from the skin samples, M_0, is only used for initial detection and is then adapted to the current illumination situation and the person's specific skin color. Whenever a face is successfully detected in a skin-colored area, the histogram $H_{\text{ face}}$ of this area is used to update the current model M_t.

$$M_{t+1}(r,g) = (1-\alpha)M_t(r,g) + \alpha H_{\text{face}}(r,g) \tag{3}$$

with update parameter α and bin indexes r and g. With $\alpha = 0.4$, this ensures fast adaptation to every specific case. Due to the Gaussian smoothing, the thresholding process described above leads to segmentation of non-skin pixels close to skin-colored ones, e.g., eyes, lips and hair. In order to avoid adaptation to these colors, only colors inside the skin locus are used to compute $H_{\text{ face}}$.

2.2. Feature-based face and eye detection

In order to detect the faces and the eyes we have used the approach proposed by Viola and Jones [31]. We use the implementation of their algorithm from the Open Computer Vision Library (OpenCV) [32]. We trained our own face and eye detection cascades. To account for in-plane face rotations to some extent, we rotated training face images up to 30°.

2.3. Region-of-interest tracking

A person's face is not necessarily detected in every frame because he or she might turn sideways or look down so that the face detector, which has been trained for quasi-frontal faces, fails. In order to be able to record the whole sequence until the person leaves the camera's field of view, a simple yet effective tracking algorithm has been employed. It is based on the fact that a profile view of a face still produces a face candidate during the skin segmentation step. This leads to the underlying assumption that a face candidate at or close to a position where a face was successfully detected in previous frames is likely to be this face. Basically, this face candidate is accepted as a face if its center lies within the bounding box of the previously detected face. To account for movement, the search region is enlarged by a certain amount. The processing of the next frame will then be restricted to this area which leads to an enormous speedup as image data is reduced to a fraction.

3. Face registration

As stable eye detections are crucial, eye locations are tracked over consecutive frames using Kalman filters. Both eyes are tracked separately. The state of each of the two Kalman filters covers the x- and y-position of one of the eyes, together with its speed of motion, v_x and v_y. The state estimates are supported by measurements of the (x,y) location of the eyes as determined by eye detectors.

The problem that arises with eye detection is, that an eye detector with a reasonable detection rate produces quite a few false positives. This is due to the fact that the intensity distribution of an eye, as captured by the classifier, is rather simple. Therefore, it can be observed in other parts of the processed area as well, e.g., on curly hair. This is especially true since the detector is trained with input data which is rotated up to 30°. In order to initialize the Kalman filters, it is necessary to decide on the "true" detection among all available ones. It is observed that the majority of false positives only show up in single frames or pairs of frames. Nevertheless, some of them are detected more consistently. In contrast, the *genuine* eye locations are not necessarily detected in every single frame.

To solve this problem, the approach depicted in Fig. 5 is implemented [33]. The detections of each eye cascade are used to generate track hypotheses over consecutive frames. Close detections in consecutive frames are associated to each other to form a track. Tracks that do not get updated with a new measurement are extrapolated based on previous observations. If several detections are associated with one track, it gets split into two. If two tracks overlap for several frames, one of them is discarded.

From the set of tracks, eye pairs are generated with the following constraints:

- Left eye is left of right eye.
- Eye distance is larger than a minimum.
- Left and right eye move into a similar direction.
- Left and right eye move at similar speed.

At this point, the number of possible eye candidates is already greatly reduced. To verify the eye pair hypotheses, the image is first rotated, so that the eye positions are on a horizontal line. Next, a face detector is used to finally confirm or discard the hypothesis. The rotation is necessary because the face detector is restricted to upright faces. Without that restriction, the false positive rate would strongly increase as in the eye detector case. If the face detector is successful, the Kalman filters are initialized accordingly. As a fallback solution, eye candidates trigger the Kalman filter initialization if they appear consistently over a long time. On the one hand, this is necessary because the face detector may still fail on an upright face. On the other hand, it is possible because normally only the true eye locations are consistently detected over a longer period of time. The face detector approach is able to succeed within three frames while the fallback solution is triggered after successful detection of a valid eye pair over 15 frames.

Despite the fact that the eye detector is trained to account for some amount of rotation, it still works best on horizontal eyes, i.e., upright faces. Therefore, the detection results can be greatly improved if subsequent face candidates are rotated based on the Kalman filter prediction prior to any detection/confirmation. If eye detection fails nevertheless, the prediction can be used as substitute.

For registration, the face image is rotated to bring the detected or predicted eye locations into horizontal position. Afterwards, the image is scaled and cropped to a size of 64 × 64 pixels, so that the eyes are located at certain coordinates in the resulting image. Fig. 6 shows some samples obtained with this method.

4. Face recognition

A local appearance-based face recognition algorithm is used [21,22] for face recognition. It is a generic face recognition ap-

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

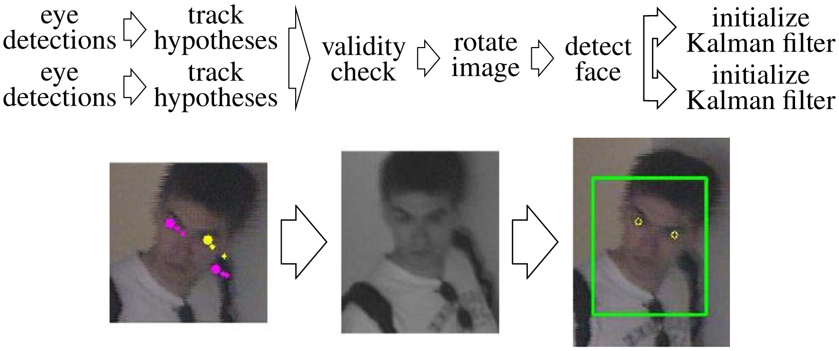

Fig. 5. Initialization of the Kalman filters for eye tracking.

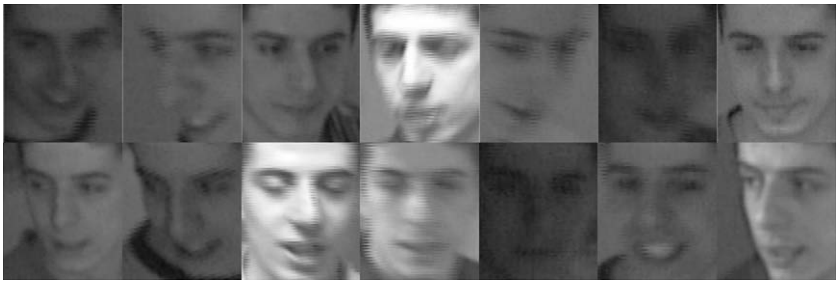

Fig. 6. Sample registered face images with the proposed system.

proach that has been found to be robust against expression, illumination, and occlusion variations as well as real-world conditions [34]. The algorithm has been evaluated on several benchmark face databases, such as AR [35], CMU PIE [36], FRGC [37], Yale B [14] and Extended Yale B [38] face databases, and found to be significantly superior to other generic face recognition algorithms, such as eigenfaces [12], Fisherfaces [13], embedded hidden Markov models [39] and Bayesian face recognition [40]. In addition, it achieved the best recognition rates in the CLEAR 2007 evaluations [41].

The approach utilizes representation of local facial regions and combines them at the feature level, which provides conservation of the spatial relationships. The algorithm uses discrete cosine transform (DCT) for local appearance representation. There are several advantages of using the DCT. Its data independent bases make it very practical to use. There is no need to prepare a representative set of training data to compute a subspace. In addition, it provides frequency information,which is very useful for handling changes in facial appearance. For instance, it is known that some frequency bands are useful for combating against illumination variations. Moreover, we have found that the DCT-based local appearance representation is better than representations based on the Karhunen-Loève, Fourier, Wavelet and Walsh–Hadamard transforms in terms of face recognition performance [34].

In the proposed approach, a detected and registered face image is divided into non-overlapping blocks of 8 × 8 pixels size. The reason for choosing a block size of 8 × 8 pixels is to have small enough blocks in which stationarity is provided and transform complexity is kept simple on one hand, and to have big enough blocks to provide sufficient compression on the other hand. Furthermore, the experiments conducted with different block sizes also showed that using a block size of 8 × 8 pixels is also beneficial for face recognition performance. Afterwards, on each 8 × 8 pixels block, the DCT is performed. The obtained DCT coefficients are ordered using zig–zag scanning. From the ordered coefficients, the first five AC coefficients are selected in order to create compact local feature vectors. The DC coefficient is discarded for illumination normalization as suggested in [21]. Furthermore, robustness against illumination variations is increased by normalizing the local feature vectors to unit norm [22]. This reduces illumination effects, especially illumination differences with a gradient pattern, while keeping the essential frequency information. Finally, the local feature vectors extracted from each block are concatenated to construct the overall feature vector. Both a discriminative and a generative approach are followed to classify the so-achieved feature vectors. With both approaches, individual models are derived for each person. The granularity of these models depends on the respective amount of available training data. This accounts for the fact that the real-life data collection setting leads to largely varying amounts of data among the different persons. Thus, the more often the system encounters a certain individual, the more detailed this individual's model will be, as more variation can be captured. The block diagram of the face recognition system is given in Fig. 7.

4.1. K-nearest neighbors model

A major advantage of discriminative approaches like K-nearest neighbors (KNN) is that they do not make an assumption about the distribution of the underlying data. This allows to build meaningful models with less data than would be necessary to train high-dimensional generative models like Gaussian mixtures. To determine the nearest neighbors, the L_1 norm is employed as distance measure $d(\cdot,\cdot)$, as it was shown to perform best among several popular distance metrics in [22]. The k closest neighbors S_i, $i = 1, 2, \ldots, k$ of a test vector x are selected with score $s_i = d(x, S_i)$. Because the distances and, thus, the resulting scores can differ largely between frames, they need to be normalized. This is achieved with linear *min-max normalization* [42],

$$s'_i = 1 - \frac{s_i - s_{min}}{s_{max} - s_{min}} \quad i = 1, 2, \ldots, k \tag{4}$$

which maps the scores to [0, 1]. To have equal contribution of each frame, these scores are re-normalized to $\sum_{i=1}^{k} s'_i = 1$. Of course, among the k closest representatives, there can be several ones from

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

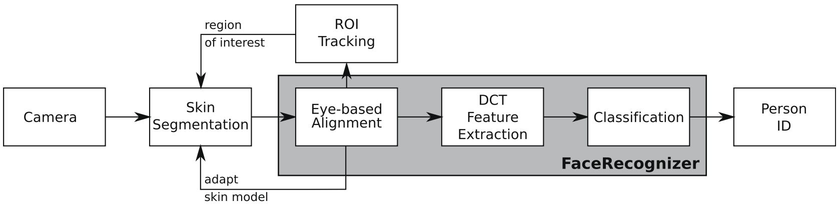

Fig. 7. Overview of the face recognition system.

the same class. Since some people have far fewer representatives than others, care must be taken that their scores are not dominated by those. Individual scores are selected by a simple max-rule [43], which only selects the maximum score for each class.

A sum-rule [43] decision fusion scheme is employed to take advantage of all frames in a video sequence to decide on the identity of a subject. Two baseline performances are determined. First, every single frame is evaluated individually to be able to evaluate the improvement contributed by video-based classification. Second, the baseline video-based recognition performance is determined by simply adding the scores of all frames.

4.2. Gaussian mixture model

Even though generative models, in our case Gaussian mixture models (GMM), usually require more training data than discriminative ones, they allow to model the data with probability density functions (pdf), and, as a consequence, the computation of conditional pdfs.

The Gaussian mixture model approach trains one GMM per class using an expectation-maximization algorithm [44,45]. Likewise the KNN model, the number of components per mixture depends on the number of training samples available for a person. At runtime, person x is classified as one of the N registered individuals in a maximum log-likelihood manner using

$$\arg\max_{i\in N}\log P(x|i) = \arg\max_{i\in N}\log\sum_{j=1}^{k_i}\alpha_{ij}\cdot\mathcal{N}(x;\mu_{ij},\Sigma_{ij}) \tag{5}$$

where k_i denotes the number of modes per person, α_{ij} the mixing parameters, and μ_{ij} and Σ_{ij} the mean and the variance of the jth component of person i's model, respectively. To keep the computational effort within reasonable bounds, only a diagonal rather than the full covariance matrix is used.

Three approaches are employed to evaluate the classification performance of the GMM setup on video input. Similar to the KNN model, frame-based evaluation determines the baseline performance of the model. Every frame is evaluated on its own based on a min-max normalization. For the video-based approaches, the classification of a sequence is made on the final score.

4.2.1. Bayesian inference

Using Bayes' rule, posterior probabilities are computed for each class. These posteriors are used as priors in the next frame. The posterior probability $P(i_t|x_{0:t})$ of person i at frame t given the all the previous observations $x_{0:t}$ is calculated as

$$P(i_t|x_{0:t}) = \frac{P(x_t|i_t)\cdot P(i_t|x_{0:t-1})}{P(x_t)} \tag{6}$$

The conditional observation likelihood $P(x_t|i_t)$ is computed by the GMM for person i, the unconditional one by

$$P(x_t) = \sum_{i=1}^{N}P(x_t|i_t)\cdot P(i_t|x_{0:t-1}) \tag{7}$$

with N being the number of individuals. The priors are initialized uniformly, i.e.,

$$P(i|x_0) = \frac{1}{N} \tag{8}$$

This approach takes into account the temporal dependency by computing the probability to observe a given sequence of input frames.

4.2.2. Bayesian inference with smoothing

Based on the previous approach, the idea of a consistent identity is introduced as suggested in [15]. The identity of an entering person does not change but depending on frame and model quality the classification of single frames can differ from previous ones. As a consequence, the influence of frames which are not consistent with the current sequence hypothesis, i.e., the current classification for a given sequence, is reduced. Extending Eq. (6), the smoothed posteriors are calculated as

$$P(i_t|x_{0:t}) = \frac{P(x_t|i)\cdot P(i_t|i_{t-1})\cdot P(i_t|x_{0:t-1})}{P(x_t)} \tag{9}$$

with

$$P(i_t|i_{t-1}) = \begin{cases}1-\epsilon & \text{if } i_t = i_{t-1}\\ \frac{\epsilon}{N} & \text{otherwise}\end{cases} \tag{10}$$

The amount of smoothing is determined by the smoothing parameter ϵ, where smaller values denote stronger smoothing. With a value of 0, the sequence is basically classified solely based on the first frame. Nevertheless, values *close* to 0 lead to a stabilization of the sequence hypothesis while still allowing a change to a different identity as the experiments in the next section will show.

4.3. Frame weighting

Due to the real-life quality of the data, not all frames are suitable to classify the subject. Low resolution, large occlusion and faulty alignment are examples of negative influences on frame quality. Besides, certain views of a person may simply not be captured by the model due to little training data or due to training data that contains too little variation. Two important observations have been made from the experiments conducted on a parameter estimation set, which are exploited in order to reduce the impact of ambiguous frames.

First, for wrong classifications, the distance to the closest representative is, on average, larger than for correct ones. Moreover, badly aligned frames result in larger distances as well. To account for this, we introduce the weighting scheme *distance-to-model* (DTM). The frames $f_i, i = 1, 2, \ldots$, are weighted with respect to the closest representative c with

$$w_{\mathrm{DTM}}(f_i) = \begin{cases}1 & \text{if } d(f_i,c) < \mu\\ e^{-\frac{d(f_i,c)-\mu}{2\sigma^2}} & \text{otherwise}\end{cases} \tag{11}$$

This weighting function is chosen according to the observed distribution of frame distances $d(f_i, c_{f_i,\mathrm{correct}})$, the distances of all frames f_i

to the closest representative $c_{f_i,\,\text{correct}}$ of the corresponding correct class. The distribution, determined on a parameter estimation set, resembles a normal distribution $\mathcal{N}(\cdot;\mu,\sigma^2)$. To increase robustness against outliers, μ is chosen as sample median and σ^2 as median absolute deviation (MAD)[46]. An example distribution and weight function is shown in Fig. 8. Using the weight function w_{DTM}, the influence of frames which are not sufficiently close to the model is reduced.

The second observation is that, in case of misclassification of frame f_i, the difference of the distances $\Delta(f_i)$ to the closest and the second closest representatives is generally smaller than in the correct case. The distribution of these distances follows approximately an exponential distribution

$$\varepsilon(x;\lambda) = 0.1\lambda e^{-\lambda x} \quad \text{with } \lambda = 0.5 \tag{12}$$

The weights are then computed as the cumulative distribution function of $\varepsilon(\cdot)$

$$w_{\text{DT2ND}}(f_i) = \mathscr{E}(\Delta(f_i)) = 1 - e^{-\lambda\Delta(f_i)} \tag{13}$$

An example distribution and weight function is shown in Fig. 9. This weighting scheme will be referred to as *distance-to-second-closest* (DT2ND).

DTM and DT2ND utilize different type of information. DTM takes into account how similar a test sample is to the representatives of the training set, whereas DT2ND takes into account how well the closest and second closest representatives are separated. For example, a badly aligned face image causes a large distance to the model. However, the best matches can be still well separated. Therefore, it is desirable to have both conditions satisfied. That is, having a small distance to the closest representative, and a good separation between the closest and second closest representatives. On account of this reason, in addition to individual weighting schemes, a joint weighting scheme is used that employs the product of w_{DTM} and w_{DT2ND} to weight the frames.

5. Experiments

In this section the evaluation results of the video segmentation and face recognition systems are presented and discussed.

5.1. Evaluation of the video segmentation

To assess the performance of the video segmentation algorithm, four continuous video streams were recorded on three different days and manually labeled for ground truth. They cover a time frame of 16.5 h and consist of approximately 1.5 million frames. Table 1 gives a detailed overview.

Looking at the share of less than one percent of relevant data within the recorded video, it is obvious that continuous recording is not an option for sensible data collection, not only concerning memory requirements but especially in terms of effort and time-consumption of tedious manual segmentation.

The results in Fig. 10 are given as *correct detection rate (CDR)* and *false detection rate (FDR)*. A correct detection is given if a detected sequence overlaps at least 50% of a labeled one. The total CDR for different overlap values can be seen from Fig. 11.

5.2. Evaluation of the face recognition system

To evaluate the face recognition system, we used a database that consists of 2292 video sequences (205,480 frames) of 41 subjects recorded during 6 months. It is chronologically divided into three sets for training, parameter estimation and testing as listed in Table 2. Face images are automatically extracted from training sequences using the registration process outlined in Section 3. Training data is augmented with virtual samples. These are generated during the extraction process by artificial perturbations of the originally detected eye locations by ±2 pixels in *x*- and *y*-direction. The face is then aligned according to these new coordinates and stored in the training database. Since nine locations per eye are evaluated, this increases the training set size by factor 81. As a consequence, the raw number of training samples is very large which would slow down the KNN approach. Since many samples from consecutive frames are very similar, k-means is applied to select representative exemplars. The clustering is performed for each person individually.

5.3. Closed-set identification

For closed-set identification, the system is only confronted with subjects that are registered in the database. The system has to classify each subject as one of the possible classes. The performance is measured as *correct classification rate (CCR)*, the percentage of correctly classified sequences in the test set. As baseline performance, every single frame is evaluated individually, that is, *CCR* is calculated as the percentage of correctly classified frames among all frames. The results are given in Table 3. *Uniform* denotes no weighting and therefore equal contribution of each frame, *combined* the combination of DTM and DT2ND by weight multiplication.

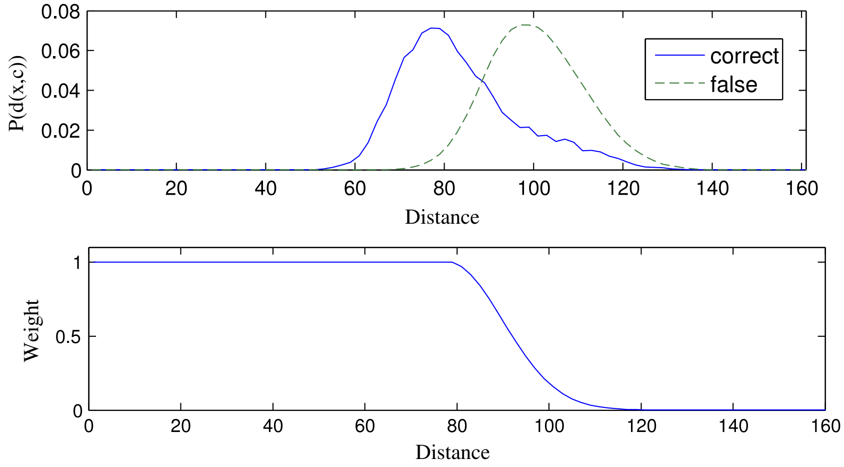

Fig. 8. DTM weight function. (top) Distribution of the distances to the closest representative of the correct class (blue, solid) and to all other classes (green, dashed) and (bottom) the actual weight function. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ARTICLE IN PRESS

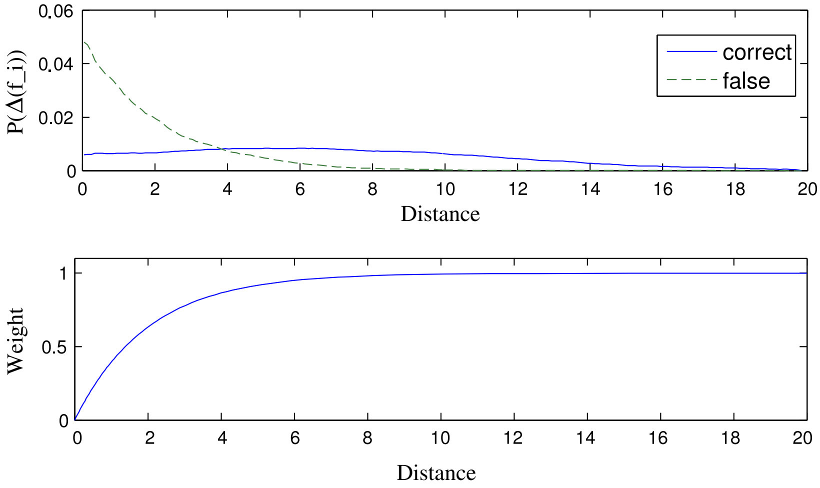

Fig. 9. DT2ND weight function. (top) Distribution of the distances between the closest and second closest representatives for correct (blue, solid) and false classifications (green, dashed) and (bottom) the actual weight function. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Overview of evaluation set. The number of sequences refers to situations in which somebody is actually entering the room.

Sequence	Duration (hh:mm:ss)	Total no. of frames	No. of sequences	No. of relevant frames
A	02:53:16	259,910	42	2929
B	04:04:13	366,318	12	3233
C	03:25:25	308,124	12	989
D	06:07:38	551,443	63	6220
Total	16:30:32	1,485,795	129	13,371

Sequence	CDR (%)	FDR (%)
A	92.9	9.3
B	83.3	0.0
C	100.0	0.0
D	95.2	9.1
Total	**93.8**	**7.6**

Fig. 10. Detection performance of face recorder. Results in the table are given as *correct detection rate* (CDR) and *false detection rate* (FDR). These measures are based on sequences rather than frames with an overlap of at least 50%.

Video-based evaluation outperforms frame-based evaluation since the increased amount of available data helps to resolve some ambiguities. Obviously, both weighting schemes improve the classification performance over uniform weighting. The combination takes advantage of both and performs better. The increase is slightly larger for DTM, as it assigns smaller weights to frames that are not similar enough to the representatives of the training set. DT2ND, in contrast, reduces the impact of ambiguous frames, i.e., frames which yield similar scores for the top two candidates, independent of how well the “face” is modeled. In fact, a badly aligned image can lead to a distinct score, but it is likely to have a large distance to the model. DTM is able to handle this case, DT2ND is not.

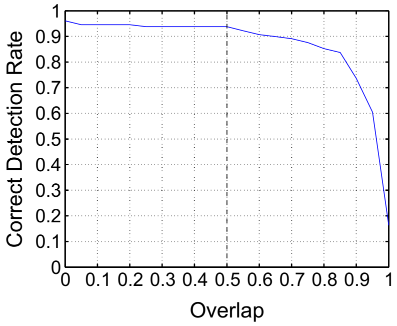

Fig. 11. Detection performance of face recorder with respect to overlap ratio.

Table 2
Sizes of the three subsets.

	Number of sequences
Training set	905
Parameter set	386
Test set	1001
Total	2292

Table 3
Closed-set correct classification results with different fusion schemes. Smooth GMM uses $\varepsilon = 10^{-5}$.

KNN	CCR (%)	GMM	CCR (%)
Frame-based	68.4	Frame-based	62.7
Uniform	90.0	Uniform	86.7
DTM	92.0	Smooth	87.8
DT2ND	91.3	DTM	90.6
Combined	**92.5**	DT2ND	89.1
		Combined	**91.8**

Nevertheless, reduction of ambiguity leads to better performance over uniform weighting as well.

As far as the different models are concerned, the discriminative approaches perform better than the generative ones. Since parametric models like GMMs need more training data with increasing dimensionality, this is possibly caused by insufficient training data for some individuals which can prevent derivation of meaningful

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

models. Besides, the number of mixture components might not be sufficient to approximate the underlying probability distribution. The discriminative models are less affected by little training data, as they classify new data only based on existing data, without making any assumptions about its distribution.

To investigate the robustness of the results, it is worth looking at the results including rank-2 and rank-3 classifications, i.e., cases in which the correct identity is among the best two or three hypotheses. As clearly depicted in Fig. 12 and Table 4, the frame-based approach often gets close to the correct decision. However, it has to decide on the identity even in the case that the single feature vector is of questionable quality. The approach lacks an opportunity to support or discard the hypothesis using additional data as done by the sequence-based methods. These are able to exploit the temporal dependency of consecutive frames and to promote the rank-2 and rank-3 classifications of the frame-based models to first place. Since many frames contribute to the decision, the overall performance improvement is larger than the difference between the correct and rank-3 classifications in the frame-based approach.

The more frames can be evaluated, the more likely it is to obtain a correct result. This gets confirmed by the observation that the average length of correctly classified sequences is larger – 39 frames – than that of misclassified ones with 28 frames as depicted in Fig. 13.

To justify the increased training efforts caused by the larger training set size, an experiment was conducted to compare the recognition performance using augmented and unaugmented training data. The comparison can only cover the KNN models as it is not possible to train appropriate GMMs due to the fact that many individuals have fewer images in the training set than the feature vector's dimensionality. As listed in Table 5, recognition performance increases significantly in all three KNN cases. This shows that the data augmentation is well worth the increased memory and time resources. Adding noise to detected eye locations leads to samples of different scale and rotation which increases the variation bandwidth and reduces the influence of possible registration errors. Since the data set size is increased by factor 81, even persons with few genuine training images can be modeled appropriately.

5.4. Open-set identification

This task extends the previous one by the difficulty that unknown people, i.e., persons which are not registered in the database, can be encountered. Therefore, prior to classification as one of the possible identities, the system has to decide whether a person is known or unknown. Impostors are to be rejected, while genuine members of the database need to be accepted and classified correctly. To model this task with the existing data set, the system is trained in a leave-one-out manner. One person at a time is removed from the database and is presented to the system as impostor during the subsequent evaluation on all sequences. This process is repeated N times, so that each person takes the impostor role once. The acceptance-rejection criterion is a threshold on the confidence of the classification, which is a value between 0 and 1. If the confidence is too low, the person is rejected.

Table 4
Correct recognition rate by rank for the KNN models.

	Frame-based	Uniform	Combined
Rank-1	68.4	90.9	92.5
Rank-2	76.5	94.8	95.6
Rank-3	81.1	96.2	96.7

A measure of confidence of the classification is derived by min-max normalization (see Eq. (4)) of the accumulated scores at the end of the sequence. The frame-based scores are already normalized and can serve as confidence measure without further processing.

Compared to closed-set identification, two more error types can occur. Additional to false classifications, the system can erroneously either reject genuine identities or accept impostors. All three errors have to be traded-off against each other as it is not possible to minimize them at the same time. For this reason, a different performance measure is necessary. The employed *equal error rate (EER)* denotes the minimum combined error rate. It is reached when

$$\mathrm{FAR} = \mathrm{FRR} + \mathrm{FCR} \qquad (14)$$

i.e., when the *false acceptance rate (FAR)* among the impostors is equal to the sum of the *false rejection rate (FRR)* and the *false classification rate (FCR)* among the registered persons. The rates are defined as

$$\mathrm{FAR} = \frac{n_{i,\mathrm{accepted}}}{n_i} \qquad (15)$$

$$\mathrm{FRR} = \frac{n_{g,\mathrm{rejected}}}{n_g} \qquad (16)$$

$$\mathrm{FCR} = \frac{n_{g,\mathrm{misclassified}}}{n_{g,\mathrm{accepted}}} \qquad (17)$$

where n denotes number of frames or sequences and the subscripts g and i denote genuine or impostor samples, respectively.

Looking at the results in Fig. 14, it can be seen that uniform and combined weighting form a lower and upper bound for the individual weighting schemes.

The two weighting schemes affect different parts of the ROC curve. The DTM scheme improves the recognition rate for high

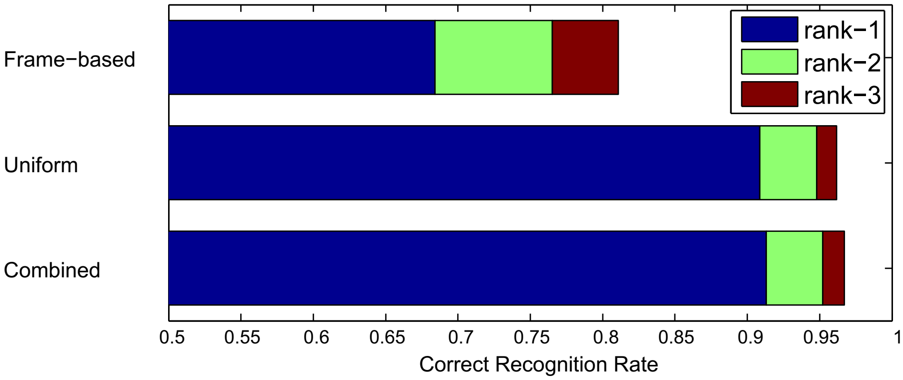

Fig. 12. Correct recognition rate by rank for the KNN models.

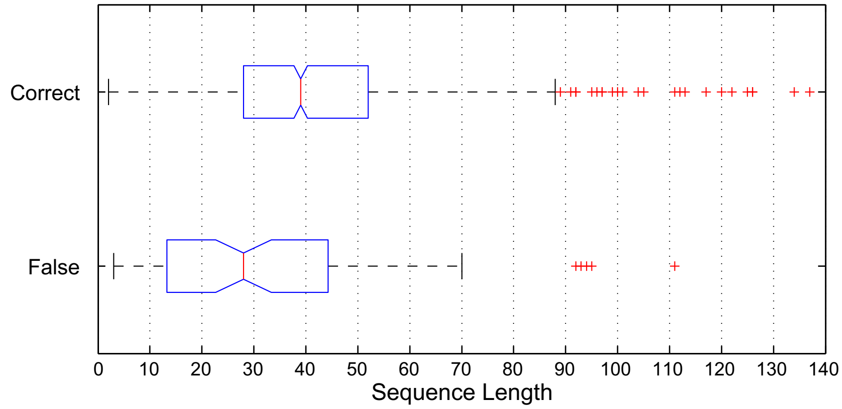

Fig. 13. Box plot showing the distribution of sequence lengths for correct and false classifications. The box denotes the median as well as the lower and upper quartiles whereas the "whiskers" denote the range of values. Outliers are marked by crosses. The plot is based on results achieved with the combined KNN approach.

Table 5
Influence of data set augmentation with virtual samples. All results improve significantly. Significance was computed with crosstabulation.

	Frame-based	Uniform	Combined
Unaugmented (%)	56.6	87.6	88.2
Augmented (%)	68.4	90.9	92.5
p-value	0.00	0.02	0.00
Significantly better	✓	✓	✓

Table 6
Open-set equal error rates. Smooth GMM uses $\varepsilon = 10^{-5}$.

KNN	EER (%)	GMM	EER (%)
Frame-based	50.0	Frame-based	43.7
Uniform	23.4	Uniform	23.0
DTM	23.3	Smooth	18.7
DT2ND	21.3	DTM	20.2
Combined	**21.0**	DT2ND	20.5
		Combined	**18.0**

false acceptance rates. A FAR of 100 percent is equivalent to closed-set identification in terms of the ROC curve because the CCR can only be computed over genuine samples. This confirms the results from the closed-set experiment.

DTM reduces the false classification rate, but it is not able to discriminate between known and unknown persons as the feature vector of an impostor can be indeed very similar to the training representatives. Therefore, as can be seen from Table 6, the EER is approximately the same than in the unweighted case.

Genuine identities, however, usually have smaller distances to one single class representative than to all other classes in the model, while impostors are similarly close to multiple classes (cf. Fig. 9). This ambiguity is exploited by the DT2ND weighting scheme to identify impostors. Their scores are reduced, leading to smaller confidence values, which in turn result in better rejection. The same threshold causes rejection of more impostors than in the unweighted case or, to put it the other way round, the threshold can be reduced causing fewer false rejections. As a consequence, the EER is reduced in open-set identification.

Looking at Fig. 15, the observations are similar if the GMM output is used as similarity measure. DTM and DT2ND improve the results over uniform weighting. DTM outperforms DT2ND for high FARs (i.e., getting closer to closed-set identification). The combination of both weighting approaches is able to join the performance improvements and leads to the lowest EER of all runs, as can be seen from Table 6.

To examine the effects of the constraint that a person's identity does not change within a sequence, as formulated in Section 4.2, the Smooth GMM model is evaluated at different levels of smoothing. The smaller the ϵ-value, the stronger is the constraint. As Fig. 16 shows, a moderate amount of smoothing improves the

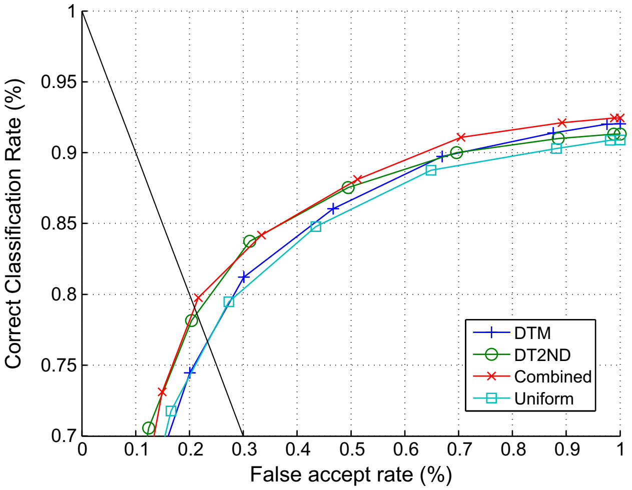

Fig. 14. ROC curve of open-set recognition results for the KNN approach. The black line denotes equal error.

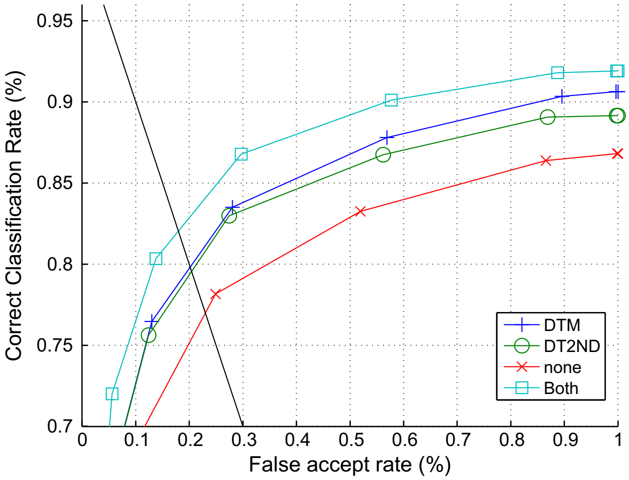

Fig. 15. ROC curve of of open-set recognition results for the GMM approach. The black line denotes equal error.

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

ARTICLE IN PRESS

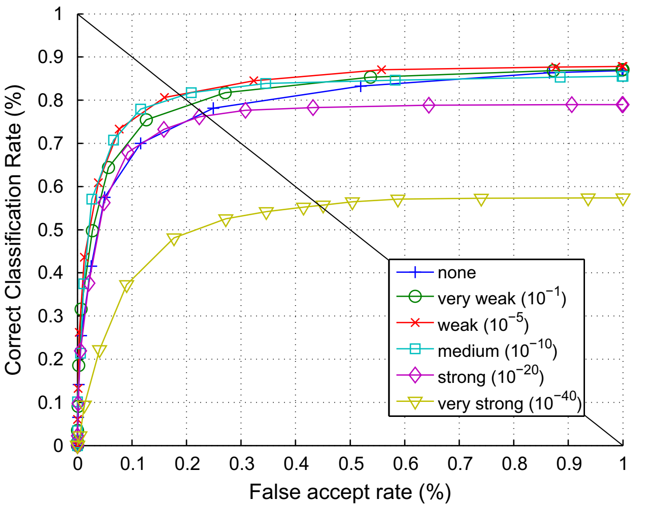

Fig. 16. Influence of identity smoothing. The values in parentheses are the ϵ-values used for smoothing in Eq. (9). The unsmoothed curve corresponds to the Uniform-GMM model.

open-set identification performance. Small numbers of ambiguous and inconsistent frames do not derogate the currently best score while many consistent frames increase the confidence of the decision. As a consequence, a smoothed classification result is more distinct than an unsmoothed one. Since smoothing generally favors sequences with consistent frame hypotheses over ones with inconsistent classifications, it does not necessarily reduce the number of false classifications but the augmented confidence leads to a reduction of the false rejection rate. In contrast to genuine identities, impostors often cause inconsistent frame scores, so that the resulting low confidence leads to a proper rejection.

However, if the smoothing factor is chosen too small, the system gets stuck on the decision of the first frames. Even if all subsequent frames are classified as a single different identity, this person's video-based score will grow only marginally because the frame-based scores are practically reduced to zero.

Based on the observation that the first frames are generally of low quality, especially due to low resolution, it would be possible to omit them in the classification process. This assumption is not included into the current system as this would restrict the system to this specific door scenario.

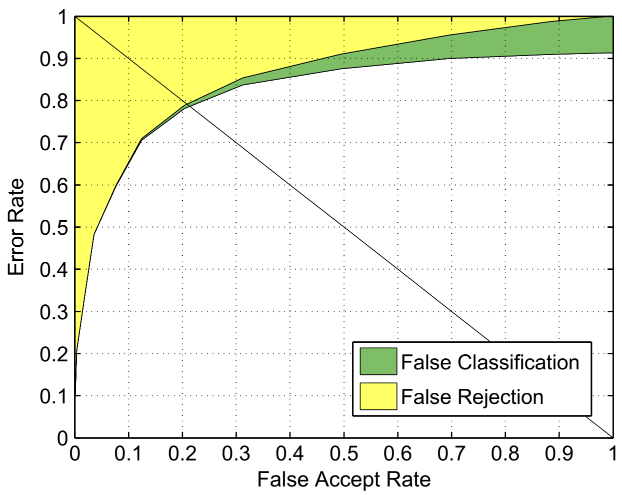

Fig. 17. Analysis of the contribution of FRR and FCR to the overall error rate of Combined-KNN. The black line denotes equal error.

Since the ROC curves show the CCR depending on the FAR, the question arises to which extent each of the other two types of error, false rejection and false classification, impair classification performance. To investigate this, Fig. 17 plots FRR and FCR separately for Combined-KNN. The lower bound corresponds to the CCR as depicted in Fig. 14. At the point of equal error, there remains only a minimal FCR of about 1 percent while the major part of about 20 percent is caused by false rejection of genuine identities. This reflects the difficulty of separating impostors and genuines as they represent arbitrary subsets of all possible faces.

6. Conclusion

In this study, a real-time video-based face recognition system is presented. The system is developed for smart environment applications, such as for smart homes or for smart stores, in which the main challenge is to perform unobtrusive identification without requiring the cooperation of the person to be identified. The key to this unaware recognition is a system that is able to process data under everyday conditions and to accomplish this in real-time. Violations of these requirements necessarily interrupt the persons to be identified or restrict them in their actions. An example of such a restriction is the necessity to stop for a moment and watch straight at the camera while not being allowed to wear any accessories in order to be recognized. As shown in this work, the proposed system is able to fulfill both requirements.

A large set of segmented data was automatically collected under real-life conditions including extreme variations in illumination, expression, pose, appearance or partial occlusion. In this system, robust registration is achieved by eye tracking which compensates for eye detection failures. These failures are mainly caused by pose variations and changing illumination.

Three weighting schemes have been introduced to weight the contribution of individual frames in order to improve the classification performance. The first, *distance-to-model*, DTM, takes into account how similar a test sample is to the representatives of the training set and therefore reduces the negative impact of unfamiliar data (e.g. due to bad recording conditions or faulty face registration). The second, *distance-to-second-closest*, DT2ND, reduces the influence of frames which deliver ambiguous classification results. Finally, the combination of DTM and DT2ND was shown to join the benefits of both.

Extensive experiments on a video database of 41 non-cooperative subjects have been conducted for both closed-set and open-set identification tasks. The results show that the combination of video-based face recognition with a local appearance-based approach is able to handle a large number of variations of facial appearance caused by illumination, pose, expression, and occlusion.

The proposed face recognition system requires on average 37 ms per frame on a Pentium 4 with 3 GHz and 1 GB RAM. Thus, it is able to process 25 frames per second which is the frame rate of the used camera.

In the future, detection of impostors in the open-set identification task needs further investigation. We plan to incorporate a measure which considers the consistency of individual frame-based results over the whole sequence. However, it has to be kept in mind that impostor detection is a non-trivial task because it requires the separation of an arbitrary subset of all possible faces from the rest.

References

[1] R. Gross, Face databases, in: Handbook of Face Recognition, Springer, 2005.
[2] A. Lemieux, M. Parizeau, Experiments on eigenfaces robustness, in: Proceedings of the International Conference on Pattern Recognition, vol. 1, 2002, pp. 421–424.

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009

[3] E. Rentzeperis, A. Stergiou, A. Pnevmatikakis, L. Polymenakos, Impact of face registration errors on recognition, in: Artificial Intelligence Applications and Innovations (AIAI06), Springer, 2006, pp. 187–194.
[4] Cognitec Systems GmbH. <http://www.cognitec-systems.de/>.
[5] Cross Match Technologies. <http://www.crossmatch.com/>.
[6] L-1 Identity Solutions. <http://www.l1id.com/>.
[7] Ex-Sight. <http://www.ex-sight.com/>.
[8] XID Technologies. http://www.xidtech.com/.
[9] Viewdle. <http://www.viewdle.com/>.
[10] Riya. <http://www.riya.com/>.
[11] J. Stallkamp, H.K. Ekenel, R. Stiefelhagen, Video-based face recognition on real-world data, in: Proceedings of the International Conference on Computer Vision, 2007.
[12] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[13] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class-specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[14] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 643–660.
[15] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (1–2) (2003) 214–245.
[16] K.C. Lee, J. Ho, M.H. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Proceedings of the 2003 IEEE Computer Society Computer Vision and Pattern Recognition, IEEE Computer Society, 2003, pp. 313–320.
[17] X. Liu, T. Chen, Video-based face recognition using adaptive hidden markov models, in: Proceedings of the 2003 IEEE Computer Society International Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2003, pp. 340–345.
[18] O. Arandjelovic, A. Zisserman, Automatic face recognition for film character retrieval in feature-length films, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2005, pp. 860–867.
[19] J. Sivic, M. Everingham, A. Zisserman, Person spotting: video shot retrieval for face sets, in: W.K. Leow, M.S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, E.M. Bakker (Eds.), Proceedings of the International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 3568, Springer, 2005, pp. 226–236.
[20] F. Li, H. Wechsler, Open set face recognition using transduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1686–1697.
[21] H.K. Ekenel, R. Stiefelhagen, Local appearance-based face recognition using discrete cosine transform, in: 13th European Signal Processing Conference (EUSIPCO 2005), 2005.
[22] H.K. Ekenel, R. Stiefelhagen, Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization, in: Conference on Computer Vision and Pattern Recognition Workshop 2006, 2006, p. 34.
[23] M. Störring, H.J. Andersen, E. Granum, Skin colour detection under changing lighting conditions, in: H. Araujo, J. Dias (Eds.), 7th International Symposium on Intelligent Robotic Systems, 1999, pp. 187–195.
[24] M. Störring, Computer vision and human skin colour, Ph.D. Thesis, Aalborg University, Denmark, August 2004.
[25] B. Martinkauppi, M.N. Soriano, M.H. Laaksonen, Behavior of skin color under varying illumination seen by different cameras at different color spaces, in: M.A. Hunt (Ed.), SPIE Machine Vision in Industrial Inspection IX, vol. 4301, 2001.
[26] M.J. Swain, D.H. Ballard, Color indexing, International Journal of Computer Vision 7 (1) (1991) 11–32.
[27] K. Schwerdt, J.L. Crowley, Robust face tracking using color, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2000, p. 90.
[28] T.-W. Yoo, I.-S. Oh, A fast algorithm for tracking human faces based on chromatic histograms, Pattern Recognition Letters 20 (1999) 967–978.
[29] J. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (6) (1986) 679–698.
[30] A. Rosenfeld, J.L. Pfaltz, Sequential operations in digital picture processing, Journal of the ACM 13 (4) (1966) 471–494.
[31] P. Viola, M.J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[32] Intel Corporation, Open source computer vision library (OpenCV), last visit: November 2006, 2006. <http://www.intel.com/technology/computing/opencv/index.htm>.
[33] Y. Bar-Shalom, T.E. Fortmann, Tracking and Data Association, Academic Press, New York, 1988.
[34] H.K. Ekenel, A robust face recognition algorithm for real-world applications, Ph.D. Thesis, Universität Karlsruhe (TH), Germany, 2009.
[35] A.M. Martinez, R. Benavente, The AR face database, Technical Report 24, CVC, 1998.
[36] T. Sim, S. Baker, M. Bsat, The cmu pose, illumination, and expression database, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1615–1618.
[37] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, Los Alamitos, CA, USA, 2005, pp. 947–954.
[38] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.
[39] A. Nefian, A hidden markov model-based approach for face detection and recognition, Ph.D. Thesis, Georgia Institute of Technology, USA, 1999.
[40] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, Pattern Recognition 33 (11) (2000) 1771–1782.
[41] R. Stiefelhagen, K. Bernardin, R.T.R.R. Bowers, M. Michel, J. Garofolo, The CLEAR 2007 evaluation, in: Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, 2007, pp. 3–34.
[42] R. Snelick, U. Uludag, A. Mink, M. Indovina, A.K. Jain, Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 450–455.
[43] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.
[44] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from incomplete data via EM algorithm, Journal Royal Statistical Society, Series B 39 (1977) 1–38.
[45] G. Welch, G. Bishop, An introduction to the Kalman filter, in: SIGGRAPH, Course 08, 2001.
[46] P.J. Huber, Robust Statistics, Wiley, 1981.

Please cite this article in press as: H.K. Ekenel et al., A video-based door monitoring system using local appearance-based face models, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2009.06.009