IET Biometrics

Research Article

IET Journals
The Institution of Engineering and Technology

Automatic face recognition for home safety using video-based side-view face images

ISSN 2047-4938
Received on 28th September 2017
Revised 3rd April 2018
Accepted on 24th April 2018
E-First on 19th June 2018
doi: 10.1049/iet-bmt.2017.0203
www.ietdl.org

Pinar Santemiz[1] ✉, Luuk J. Spreeuwers[1], Raymond N.J. Veldhuis[1]

[1]*Faculty of EEMCS, University of Twente, Cybersecurity and Safety, P.O. Box 217 7500AE Enschede, The Netherlands*

✉ *E-mail: p.santemiz@utwente.nl*

Abstract: Face recognition from side-view positions is an essential task for recognition systems with real-world scenarios. Most of the existing face recognition methods rely on alignment of face images into some canonical form. However, alignment in side-view faces can be challenging due to lack of symmetry and a small number of reliable reference points. To the best of the author's knowledge, only a few of the existing methods deal with video-based face recognition from side-view images, and not many databases include sufficient video footage to study this task. Here, the authors propose an automatic side-view face recognition system designed for home safety applications. They first contribute a newly collected video face database, named UT-DOOR, where 98 subjects were recorded with four cameras attached at doorposts as they pass through doors. Secondly, they propose a face recognition system, where they automatically detect and recognise faces using side-view images in videos. One of the attractive properties of this system is that they use cameras with limited view angle to preserve the privacy of the people. They review several databases and test their system both on the CMU Multi-PIE database and the UT-DOOR database for comparison. Experimental results show that their system can successfully recognise side-view faces from videos.

1 Introduction

In real-life scenarios with an uncontrolled environment, face recognition is a challenging task due to occlusion, expression, or pose variations [1, 2]. Owing to the complex structure of the human face, face recognition up to side view is a challenging problem. In general, pose changes introduce projective deformations and self-occlusion. Moreover, in side-view faces, lack of symmetry and the small number of reliable reference points make tracking and alignment of faces difficult. With the availability of large-scale databases covering large pose variations, many recent methods are developed focusing on face recognition under varying poses [3]. To the best of our knowledge, there are not many methods that specifically focus on side-view face recognition. Many of the existing methods mainly focus on comparing faces from an unseen viewpoint to frontal poses and use only a few side-view images.

Initial attempts to recognise side-view face images mainly use profile curves or fiducial points on the profile curves [4–6]. Recent methods can be classified into two main categories, namely *multi-view face recognition* methods and *face recognition across pose* [7]. In multi-view face recognition systems, both enrollment and test set contain images of every subject at every pose. On the other hand, in face recognition across pose, the aim is to recognise a face from a viewpoint from which it has not been previously seen. In such systems, the appearance variations are handled or predicted using 2D techniques or 3D methods. In [8], existing methods of face recognition under pose variations are surveyed until 2009. More recent developments can be found in [3].

Systems using multi-view face recognition methods based on 2D face images rely on comparing face images in similar poses. These methods are based on similarity of one face image to multiple biometric templates of subjects having a number of face images under different poses. One such method is given in [9], where Schroff *et al.* sort the reference subjects according to their similarity to the input face image, and use this list to describe the input image. In [10], Li *et al.* propose to divide faces into several pose-specific subspaces and recognise faces across two different linear subspaces. In this study, partial least squares are used to learn the multi-view subspace for both poses, and recognition accuracy is improved by maximising the intra-individual correlations across pose. In a recent study [11], Ding *et al.* introduce a method where they apply a discrete wavelet transform to side-view face contours and apply the random forest method for recognition. On a database of 40 people with six side-view images, a recognition accuracy of 92.50% is achieved using one image in the probe set and the rest in the training set.

Rathore *et al.* propose a human authentication system in [12], where ear and side-view face images are used to construct fused templates from obtained SURF features. Using side-view face images, a recognition accuracy of 99.03% is achieved on the IIT Kanpur database with 190 subjects, and a recognition accuracy of 97.16% is reported on the University of Notre Dame database with 114 subjects.

In this study, we build a biometric system using side-view faces for recognition. Such a system can be used both as a complementary method to frontal face recognition to improve accuracy and as the main authentication method in the absence of the frontal view. It is an advancement of our previous works [13, 14]. In [13], we presented a side-view face recognition system using warped shape-free face images. We tested our system on side-view face images from the CMU Multi-PIE database [15] using manually labelled landmarks for registration and achieved 91.10% rank-1 identification accuracy. Later, in [14], we proposed a side-view face recognition method using automatically detected facial landmarks. In this method, we detect the face using skin colour and find three landmark points using the curvatures on the facial profile, and three SVMs trained using HOG for each landmark. The faces are represented using local binary patterns (LBP), and a recognition accuracy of 89.04% is achieved on the CMU Multi-PIE database [15] using 137 subjects.

We observed that only a few of the existing methods focus on recognising faces using mainly side-view images. Moreover, most of the existing databases are designed for applications that focus on face recognition across pose and contain few images with near side-view poses.

Our goal is to create a biometric system to be used in home safety applications for identifying people using side-view face images as they walk through doors. We explore possibilities of such a system to be used in a house or an elderly home with up to 50 people. Knowing the location of vulnerable people such as children or the elderly is essential for ensuring their safety. In our

scenario, we aim to estimate the location of the individuals in the house, and if someone enters a room that he/she is not allowed to enter unattended, or if a person stays in a room for a suspiciously long time, the system can alarm the carer in time. Thus, the system can help to increase the situational awareness and prevent factors that may cause further accidents or detect an emergency in time. One of the biggest concerns people have with such a system is that they are not comfortable when their behaviour is observed continuously. Therefore, in our system, the cameras have a limited view angle and only see the doorways. Hence they cannot record activities in adjacent spaces, and thus preserving the privacy of the people.

In this paper, we first review currently available databases in Section 2. Then, in Section 3, we introduce our UT-DOOR database, which consists of a large number of video recordings with a significant number of side-view images. As a second contribution, we introduce a novel method for side-view face recognition in Section 4, where we automatically find three landmark points and apply several descriptors to represent and recognise face images. We test our system on side-view face images from the CMU Multi-PIE database [15], and the UT-DOOR database with both face identification and face verification protocols. We analyse our experimental results in Section 5. Finally, we give our conclusion in Section 6 and discuss our future work.

2 State of the art in multi-pose face databases

In recent years, several large-scale databases have been collected that contain face images with varying poses [3]. In Table 1, we give a list of some of the most popular public databases that contain pose variation up to side view, with a brief summary of their characteristics.

The FERET database [16] is one of the most widely used face databases with 1199 subjects and 9–20 pose variations. The available multi-pose subset of the database contains face images for 200 subjects across nine pose that span (−60° to +60°) in the horizontal direction. Each subject is captured once under a certain pose using ambient lighting, and expression remains the same across poses. Another database having a large number of subjects is the CAS-PEAL database [18] with 1040 subjects and 27 poses. The images are captured simultaneously using nine cameras spaced equally in a horizontal semi-circular shelf. Each subject is asked to look up and down to capture 18 additional images.

The FacePix database [17] was collected to cover a wide variety of pose variation using a precisely calibrated mechanism. It contains face images of 30 subjects covering pose variation between (−90° to +90°) at 1° increments. Recently, the CMU Multi-PIE database [15] was introduced to specifically address challenges due to pose and illumination variation. In total, the database contains 337 subjects captured under 15 viewpoints and 19 illumination conditions in up to four recording sessions. It is an extension of the CMU-PIE database [19], which contains 68 subjects and 13 poses. One other recent database is the Labeled Faces in the Wild (LFW) database [20] which is designed for studying the problem of unconstrained face recognition. It contains face images of 5749 people collected from the web, where 1680 of the people pictured have two or more distinct photos in the data set. Each face has been detected by the Viola–Jones face detector and labelled with the name of the person pictured.

In recent face recognition studies, there is an increasing interest in applications based on real-world scenarios, which is shifting the focus from image-based scenarios to videos. One of the largest video databases is the XM2VTS database [21], containing four recordings of 295 subjects taken over a period of four months. In each recording, there is a speaking headshot and a rotating headshot. The XM2VTS database includes high-quality colour images, sound files, video sequences, and a 3D model. It is an extension of the M2VTS database, which has voice and motion sequences of 37 people, who have been asked to count from 0 to 9 in their native language, and rotates their head from left to right. A newly published database covering the full range of pose variations is the IARPA Janus Benchmark A (IJB-A) database [22]. It contains 5712 face images and 2085 videos from 500 subjects. Images and videos are collected from the Internet, and both the bounding boxes for the face region and the ground truth eye and nose locations are manually annotated.

Although these databases are publicly available to researchers, we find that they are not sufficient to support our research where the main focus is on side-view face recognition. Most popular databases consist of still images collected in a controlled environment with uniform background and restricted pose variations. Moreover, despite the large number of pose variations and subjects, only a small percentage of the images and videos cover near side-view poses. Therefore, we present a new database that contains a large number of video recordings with side-view face images.

3 UT-DOOR database

In this study, we introduce the UT-DOOR database that consists of 4831 video recordings of 98 subjects. In our system, we aim to simulate a home safety application scenario, where we identify people using side-view face videos as they walk through doors. We capture video recordings from four cameras attached to doorposts under ambient illumination.

Each subject is recorded 15 times over two sessions. In the first session, one enrollment and six test recordings are made. In the second session, we repeated these recordings and capture one extra test recording, where the subject opens the door and then passes through. For enrollment, the subjects were asked to stand still and turn their head from one side to the other. For testing, we pre-designed three walking routes to include all possible head pose variations. The subjects trace a straight route in the first two test instances and diagonal routes in the third and the fourth instance. In the first four test instances, the subjects did not wear glasses and were asked to preserve a natural expression. In the fifth and sixth instances, the participants are asked to walk a straight route wearing glasses. In the sixth instance, they are also asked to make an expression. Some example images are shown in Fig. 1.

There are 98 subjects in the database, 66 males and 32 females. The ages of the subjects range from 22 to 63. The average age of the subjects is 30.4. In the database, 68 of the participants are Caucasian, 20 participants are Asian, and the remaining 10 subjects are of different races.

The videos are recorded in a room where we installed a stand-alone door as shown in Fig. 2. We attached four Sony X710CR

Table 1 Standard multi-pose face databases

Name	No. subjects	No. poses	Illumination	Expression	Image/video
FERET [16]	200	9	ambient lighting	neutral	image
FacePix [17]	30	181	ambient lighting	neutral	image
CAS-PEAL [18]	1040	27	15 conditions	six expressions	image
CMU-PIE [19]	68	13	43 conditions	four expressions	image
CMU Multi-PIE [15]	337	15	19 conditions	six expressions	image
LFW [20]	5749	not controlled	ambient lighting	not controlled	image
XM2VTS [21]	295	not controlled	ambient lighting	not controlled	video
IJB-A [22]	500	not controlled	not controlled	not controlled	image/video
Bosphorus [23]	105	14	ambient lighting	34	image

Fig. 1 *Examples of the recordings from our database. In the upper row, we see several images of the subject captured at the enrollment session. In the lower row, we see examples from each of the seven test recordings of the second session: walking a straight route (in the first two instances), diagonal to the left, diagonal to the right, wearing eyeglasses, wearing eyeglasses and making an expression, and opening the door*

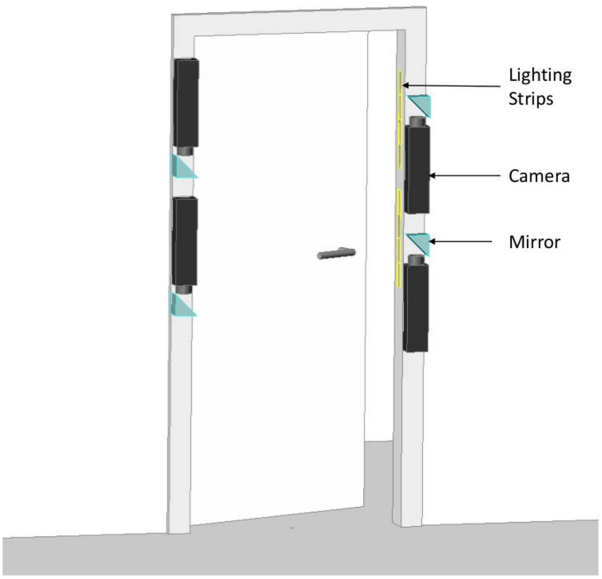

Fig. 2 *Video recording set-up. We attached four cameras to deal with several heights and use mirrors to direct the view. Additionally, we put four LED lighting strips to enhance illumination*

cameras in pairs symmetrically at two sides of the doorpost so that the cameras face each other. Moreover, we placed a LED lighting strip between each camera pair for illumination. In addition to the light that comes from the lighting strips, we use the ambient light in the room. Each video clip contains 100 frames where we use a resolution of 1024 × 768 pixels and a frame rate of 30 fps. Since the recordings are made from a close distance, we use lenses with 1.4 mm focal length.

Each session is recorded using four cameras simultaneously, which resulted in four different viewpoints. We assumed that when some of the cameras cannot capture a meaningful face image of the person due to its position, at least one of the other cameras can still capture the whole face. This can happen when the person traces a diagonal path and moves away from the cameras at one side. Moreover, we observed that in some instances where the person is walking too close to one side, or some of the cameras are mounted too high or too low compared to the person, the face can only be captured partially by these cameras. Also, occlusion may occur due to a person's hair. Some examples of these situations are given in Fig. 3.

As we can observe from the examples given in Fig. 3, while some recordings captured faces partially, in all recordings there is at least one image that contains the whole face. Also, the distance between the subject and the cameras may change in the recordings due to the walking route of the person. This results in a variation of the face size in different images.

For testing purposes, we manually inspected all face images and selected images where the whole face is visible. Then we manually annotated facial landmarks, namely, eye centres, nose tips, and mouth corners, whenever those landmarks were visible. We further annotated 15 head poses varying from frontal view (0°) to side view (±90°) with 15° increments. In most of the test images, the person is asked to trace a straight route. Therefore, face images are mostly side view or near side view. In the end, we obtained a database with a total of 4831 recordings with 62,192 images containing the whole face. There are in total 36,849 enrollment images, and 25,343 test images with an average of 376 enrollment images and 171 test images for each person. The enrollment and test videos contain 27.14 and 6.87 frames on the average, respectively. We observed that 60.84% of the whole set consists of near side-view images, where 50.44% of the enrollment images and 75.95% of the test images have pose variation between 75° and 90° in the horizontal direction. The distribution of the pose variations based on our annotation is given in Fig. 4.

4 Side-view face recognition system

In our system, we aim to recognise faces by comparing side view to side-view face images. This is a challenging problem due to the lack of symmetry and the small number of reliable reference points. To tackle this problem, we first detect the face and automatically find landmark points on the face, namely, the eye centre, the tip of the nose, and the corner of the mouth. We selected these landmarks as they are nearly always visible, whereas other possible landmark locations such as ear or face profile can be problematic due to occlusion and pose variation. We register the images to the average face using Procrustes analysis. Finally, we extract features to describe the face using LBP and histogram of oriented gradients and use them for recognition. Our framework can be seen in Fig. 5.

4.1 Registration

In our registration algorithm, we aim to find three landmark points, namely the eye centre, the nose tip, and the mouth corner. Using these landmarks, we register images to an average face using Procrustes analysis [24] and crop them to a fixed size of 180 × 90

Fig. 3 *Examples of four different viewpoints*
(a) The subject is opening the door and cameras on the right side cannot capture a meaningful face image, ***(b)*** The position of the upper cameras is too high, and only part of the face is visible in the upper images, ***(c)*** The position of the lower cameras is too low, ***(d)*** The left side of the face is occluded by the person's hair

pixels. In our algorithm, we first detect the face using the Viola–Jones algorithm [25]. To compute the average face, we select several training images from the CMU Multi-PIE database as described in Section 5. We register the training images to the first image using manually labelled landmarks and then compute their average to find the average face and its landmarks.

To begin, we first introduce the mathematical description of our method. Let $\boldsymbol{x}_{\mathrm{E}}, \boldsymbol{x}_{\mathrm{N}}, \boldsymbol{x}_{\mathrm{M}} \in \mathbb{R}^2$ be the x-, y-coordinates of the facial landmarks for the eye centre, the nose tip, and the mouth corner in an image I, respectively. Then the vector $\boldsymbol{S} = (\boldsymbol{x}_E, \boldsymbol{x}_N, \boldsymbol{x}_M) \in \mathbb{R}^6$ denotes the coordinates of all three facial landmarks in I. We will refer to the vector $\boldsymbol{S}$ as the shape.

We first find several landmark candidates within the detected face region using texture information. In our texture-based landmark-detection step, we use a support vector machine (SVM) classifier [26]. We train three separate SVMs for each landmark using histogram of oriented gradient (HOG) descriptors [27]. We define each location having positive SVM score as a candidate landmark. We further assume that around the correct landmark point, several candidates will form a cluster. Based on this assumption, we examine the number of candidates in each cluster and eliminate outliers. Then, from each cluster, we select the candidate having the maximum SVM score, and further eliminate remaining candidates to reach final candidate list for each landmark point:

$$\boldsymbol{x}_{\mathrm{E}}^{i}, \quad i = 1, \ldots, p \tag{1}$$

$$\boldsymbol{x}_{\mathrm{N}}^{j}, \quad j = 1, \ldots, q \tag{2}$$

$$\boldsymbol{x}_{\mathrm{M}}^{k}, \quad k = 1, \ldots, r \tag{3}$$

where the candidate lists for $\boldsymbol{x}_{\mathrm{E}}, \boldsymbol{x}_{\mathrm{N}}$, and $\boldsymbol{x}_{\mathrm{M}}$ contain p, q, and r candidates, respectively. At this stage, instead of treating each landmark separately, we compute all the landmark combinations from our candidate lists to find possible representations for the shape of the given face image. Each initial shape estimate for an image is

$$\boldsymbol{S}_t = (\boldsymbol{x}_{\mathrm{E}}^{i}, \boldsymbol{x}_{\mathrm{N}}^{j}, \boldsymbol{x}_{\mathrm{M}}^{k}) \tag{4}$$

with $t = 1, \ldots, N$, and $N = p \times q \times r$. To select the shape with the correct landmarks, we use a shape function f_T, and the SVM scores g_{SVM} to compute the SVM function g_S. We assume that, for a shape to describe a meaningful face, the vertical coordinates of the eye centre, the nose tip, and the mouth corner should preserve order, and the three landmarks should form a triangle. Therefore, as the shape function, we use a rule-based algorithm to eliminate the combinations that do not meet this assumption, and as the SVM function, we find the summation of the SVM scores for each shape estimate. Finally, to choose the combination with the correct landmarks, we compute the final score of each combination and select the one with the maximum score:

$$f_T(\boldsymbol{S}_t) = \begin{cases} 1 & \text{if shape meets the assumption} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$g_S(\boldsymbol{S}_t) = g_{\mathrm{SVM}}(\boldsymbol{x}_{\mathrm{E}}^{i}) + g_{\mathrm{SVM}}(\boldsymbol{x}_{\mathrm{N}}^{j}) + g_{\mathrm{SVM}}(\boldsymbol{x}_{\mathrm{M}}^{k}) \tag{6}$$

$$\boldsymbol{S}^{*} = \arg\max_{t} \left(f_T(\boldsymbol{S}_t) \cdot g_S(\boldsymbol{S}_t)\right) \tag{7}$$

In cases when we obtain the input from a video, we further refine the landmarks with the aid of the similarities through the image sequence. Here, we first register each image of the sequence to the average face and crop them to a fixed size. Then we compare these face images with each other using template matching to find incorrectly registered instances. To measure the similarities, we use normalised cross-correlation. For images that have low similarity scores during template matching process, we further investigate the remaining landmark combinations and select the registration resulting in a higher similarity score to improve registration in that frame.

4.2 Feature extraction

For describing the faces, we compare four different methods: LBP [28], histogram of oriented gradients (HOG) [27], principal component analysis (PCA) [29], and linear discriminant analysis (LDA) [30]. We implement PCA using the eigenface approach [29], and LDA using the Fisherface approach [30]. In our computations, we use the VLFeat library [31].

The most prominent advantages of LBP are its invariance against illumination changes and its computational simplicity. In our system, we divide the registered images into subregions of 10 × 10 pixels and compute uniform LBPs for each subregion. Then, we concatenate these histograms to form the feature vector.

HOG represents the shape via the distributions of local intensity gradients or edge directions. The main advantage of using HOG descriptors is that they offer some robustness to scene illumination changes while capturing characteristic edge or gradient structure. In our approach, we divide the registered image into non-overlapping cells with 10×10 pixels and form an orientation histogram having 32 bins for each cell. Each pixel in the cell calculates a weighted vote for the histogram based on the gradient orientations.

4.3 Recognition

In our recognition algorithm, we aim to compare faces that are similar in pose and in viewpoint. Therefore, we compare each test image to the images in the enrollment set, where each identity is represented with multiple enrollment images varying on pose or viewpoint. To handle pose variation, for each subject, we compute the distances between the test image and enrollment images and choose the image with least distance as the reference image.

In the UT-DOOR database, instead of single images we have frames of video sequences. Assuming at least one frame in each video sequence can adequately represent the corresponding identity, we compare all frames in the test video sequence with all enrollment images. Then we select the pair having the smallest distance.

The Chi-square distance is an accepted way to compare histograms, and it is reported to be effective compared to other dissimilarity measures for LBP and HOG [32, 33]. The cosine distance, on the other hand, is a popular and computationally efficient distance measure commonly used for PCA and LDA [34]. Therefore, we compute distance scores between images using Chi-square distance for LBP and HOG and cosine distance for PCA and LDA. For classification, we use one nearest neighbour classifier and test our system with both identification and verification protocols.

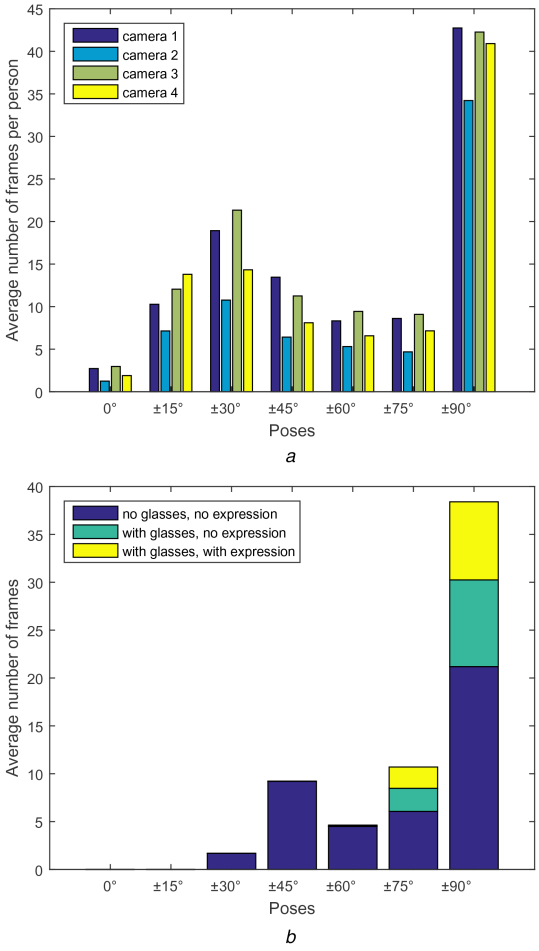

Fig. 4 *Average number of frames per person and pose distribution recorded during enrollment and test sessions. The head poses are varying between frontal view (0°) and side view (±90°)*
(a) Enrollment images, ***(b)*** Test images

5 Experimental results

We test our system on side-view images from the CMU Multi-PIE database and the newly proposed the UT-DOOR database. We use the CMU Multi-PIE database because it is one of the most widely used face databases containing a large number of poses and subjects. In the CMU Multi-PIE database, each subject is recorded under 15 poses in up to four sessions using 13 cameras that are spaced at 15° intervals. The images are acquired in a controlled environment with constant background and illumination and have a resolution of 640 × 480 pixels. We select the images acquired from the four cameras that have pose variation between 75° to 90° and −75° to −95° in the horizontal direction. In total, we use 3684 near side-view face images from all 337 subjects in our experiments. We divide these images into three subsets: a training set containing 292 images from 73 subjects that only attended one session, an enrollment/gallery set containing four images per person captured by different cameras, and a test/probe set containing the remaining images. In total, there are 264 subjects in the enrollment and test sets, where 1056 images were used for enrollment and 2336 images were used for testing.

In the UT-DOOR database, we used enrollment and test images of subjects without glasses and expression, and the face is fully visible. The images have a resolution of 1024 × 768 pixels. For comparison, we selected near side-view face images having −90° to −75° and 75° to 90° pose variation. In total, we used 685 enrollment videos with 18,588 frames and 1358 test videos with 9331 frames from 98 subjects.

In our experiments, we perform cross-database training to measure the generalisation capabilities of the system. We train our system using images from the CMU Multi-PIE, where we construct a Viola–Jones face detector, three SVMs for landmark detection, and learn the parameters of the average face, PCA, and LDA. For comparison purposes, we present recognition results using registration with both manually labelled landmarks and automatically detected landmarks.

In Section 5.1, we report our registration performance, which consists of face-detection and landmark-detection accuracies. Then, in Section 5.2, we present verification and identification results.

5.1 Landmark detection and registration experiments

In our registration algorithm, we first detect the face using the Viola–Jones algorithm [25] and find several landmark candidates within this region. Then we use a texture-based landmark-detection approach, where we use three separate SVMs for each landmark. We then select the correct landmarks among these candidates and register images using these landmarks. For face detection and texture-based landmark-detection steps, we train our system using the training images of the CMU Multi-PIE database. We use these

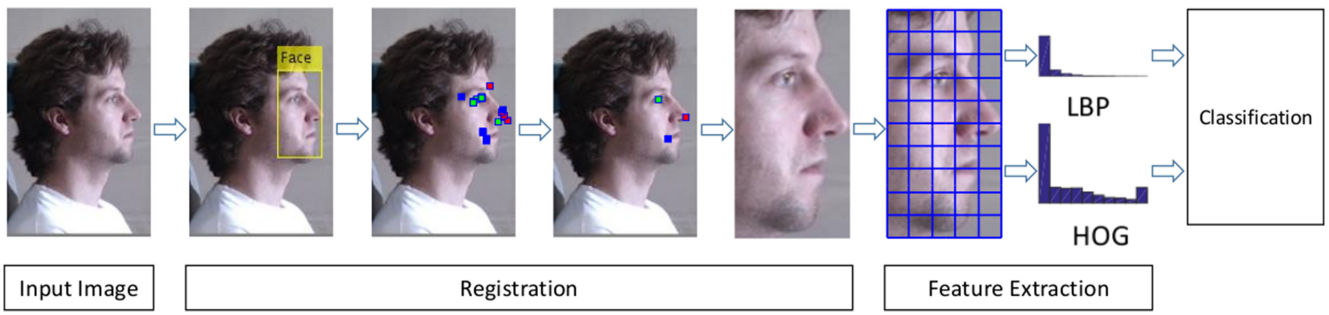

Fig. 5 *Framework of our proposed system*

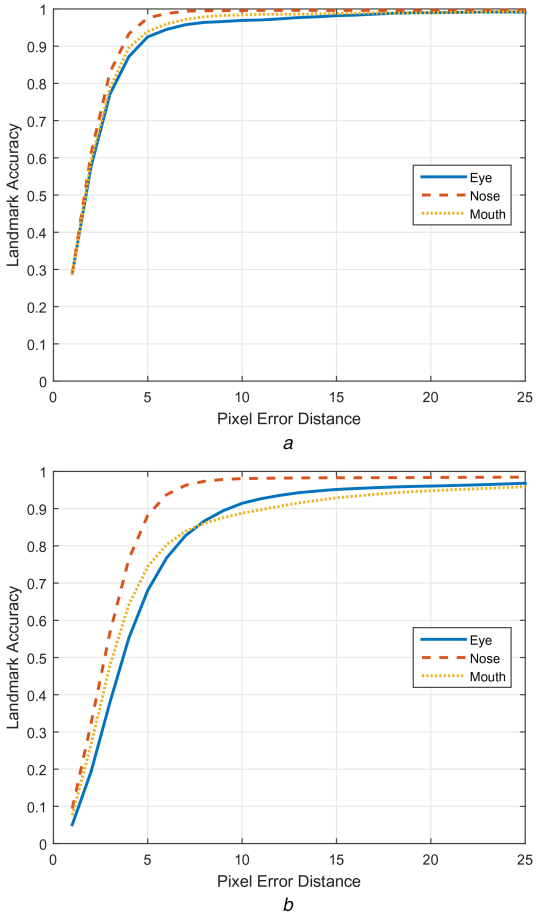

Fig. 6 *Landmark accuracies*
(a) CMU Multi-PIE, ***(b)*** UT-DOOR

classifiers for tests on both the CMU Multi-PIE and the UT-DOOR databases.

In our experiments, in 99.82% of the 3392 images in the CMU Multi-PIE database, and in 98.74% of the 27,919 images in the UT-DOOR database, the face is detected correctly. When we analyse videos in the UT-DOOR database, in 99.61% of the 2043 videos, the face is correctly detected in at least one frame. In some images, our system cannot find three landmarks in the given face region due to a face-detection error, motion blur, or poor illumination. We can detect three landmarks in 99.82% of the images of the CMU Multi-PIE database, and in 98.06% of the images of the UT-DOOR database. Since registration cannot be performed on images where no face or landmark is detected, these images are excluded in the remaining experiments.

The average distance between the eye centre and the mouth corner in the CMU Multi-PIE database is 79.91 pixels. An automatically detected point found within 10-pixels distance from the ground truth is accepted as a correct detection. Using this threshold, the correct detection on the CMU Multi-PIE database for the eye centre, the tip of the nose, and the mouth corner are 96.92, 99.56, and 98.40% in 3382 images, respectively.

The images in the UT-DOOR database have a higher resolution, where the average distance between the eye centre and the mouth corner is 190.75 pixels. For comparison with the CMU Multi-PIE database, we scale the landmarks to the same resolution and investigate detection for the UT-DOOR database within the same distance. We detect the eye centre, the tip of the nose, and the mouth corner with 91.91, 98.43, and 89.15% accuracy in 26,839 images of the UT-DOOR database, respectively. We plot the localisation accuracies for different thresholds in Fig. 6.

In our experiments on the CMU Multi-PIE database, we see that both our face-detection and landmark-detection approaches perform sufficiently well. We use 291 images for training and performed our test on 3382 images. Our performance shows that even with a small training set, we can achieve high detection accuracies.

When we compare the accuracies from both databases, we see a lower registration performance on the UT-DOOR database. A decrease in the performance is expected due to challenges in the UT-DOOR database such as pose variation, motion blur, complex background, and poor illumination. On the other hand, as we use videos rather than single images, the system does not need to detect landmarks in each frame. We can further eliminate falsely registered images in the recognition step. Therefore, we investigate if at least one frame in each video sequence has correct registration, and see that in 98.03% of the 2031 videos, we can register at least one frame correctly.

Fig. 7 illustrates registration of a test video where the person is correctly recognised despite some erroneously registered images. In this example, we see the image sequence before and after registration, and the identified frames from the enrollment set. On the first, the third, and the fifth frames landmark detection is erroneous, and in the last frame, no landmark could be detected. When we look at the recognition results (third row), we see that the first and the third frames are wrongly identified due to false registration, whereas in the fifth frame, the identification is not affected by the landmark-detection error. Finally, correct match has the highest similarity, and the test video is identified correctly.

5.2 Recognition experiments

In our experiments, we compare recognition performances using registration with manually labelled landmarks and automatically detected landmarks. Our results can be seen in Table 2. We test our system with both identification and verification protocols. For identification, we use K-fold cross-validation to get multiple enrollment/test set pairs. We calculate the rank-1 recognition accuracies and the standard deviation of multiple runs. For verification, we use non-parametric bootstrapping method [35] for computing confidence intervals. We provide the average EER and the 95% confidence intervals for comparison. The receiver operating characteristic (ROC) curves and the cumulative matching characteristic (CMC) curves are shown in Figs. 8 and 9, respectively.

With manual landmarks, we achieve rank-1 identification accuracies of 94.2% using LBP on the CMU Multi-PIE, and 99.5% using HOG on the UT-DOOR database. Using automatically detected landmarks, rank-1 identification rates are 92.7% for the CMU Multi-PIE using LBP, and 96.7% for the UT-DOOR databases using HOG. When we investigate our verification performances, we observe that the EERs for recognition using manual landmarks are 6.1% using HOG on the CMU Multi-PIE database, and 4.3% using LBP on the UT-DOOR database. When we use automatically detected landmarks, we achieve EER of 6.9% on the CMU Multi-PIE database using HOG, and 5.3% on the UT-DOOR database using LBP. Based on these recognition accuracies, we observe that the system would already be useful for a house or an elderly home environment. In the future, information acquired over time at different doors could be combined to get better accuracy.

It should be noted that considering our automatic landmark-detection accuracy on the UT-DOOR database, we achieve relatively high performance on recognition. When we test our localisation on single images in the UT-DOOR database, we achieve 91.91, 98.43, and 89.15% detection accuracies for the mouth corner, the tip of the nose, and the mouth corner, respectively. On the other hand, when we test our method on videos, we see that in 98.03% of the videos we can register at least one frame correctly. In our recognition approach, we compare videos rather than single images and achieve rank-1 identification accuracy of 96.7% using HOG. Therefore, we can conclude that our system can handle few falsely registered image due to the high accuracies of the correctly registered images in videos.

We further use several methods to describe the face, where we compare LBP, HOG, and as baseline methods PCA and LDA. It has been shown that compared to holistic methods, LBP is less sensitive to variations that occur due to illumination, expression, or

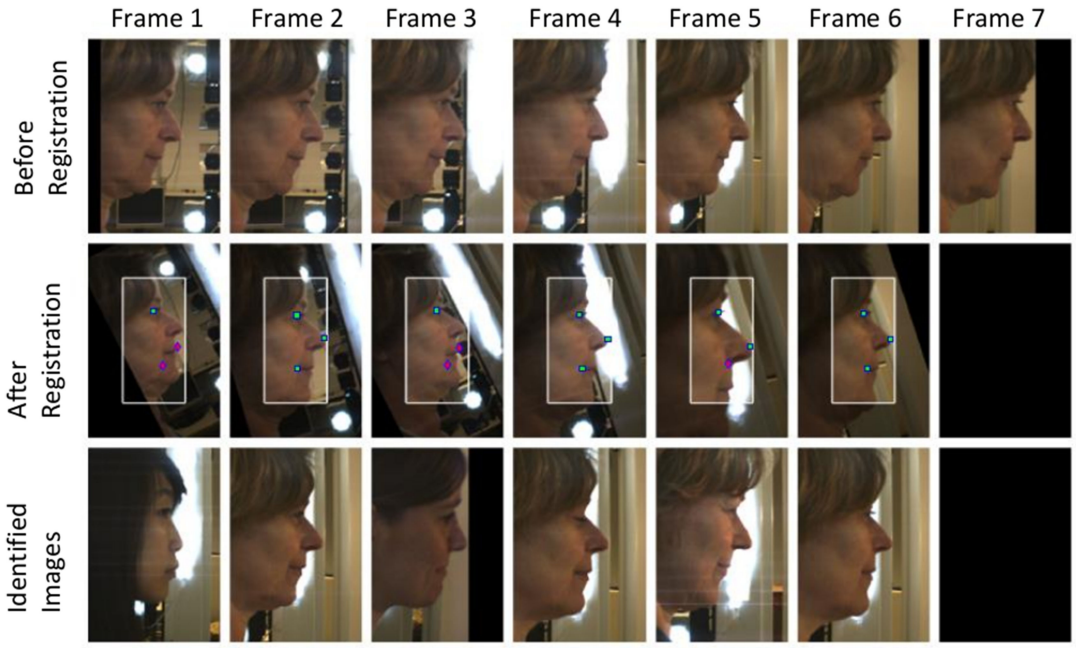

Fig. 7 *Test example. Upper row: test video, middle row: video after registration, lower row: identified reference images. In the first, the third, and the fifth frame, landmark detection is erroneous, and in the last frame, no landmark could be detected*

Table 2 Recognition performances. We provide rank-1 accuracy (%) and standard deviations to demonstrate the identification rates. For the verification tests, we provide equal error rate (EER) and 95% confidence intervals

	LBP	HOG	PCA	LDA
CMU Multi-PIE				
manual landmarks				
rank-1 accuracy (%)	**94.2 ± 0.4**	93.9 ± 0.4	76.1 ± 1.6	58.0 ± 3.1
EER (%)	6.5[5.9, 7.6]	**6.1[5.4, 7.2]**	6.6 [5.9,7.8]	9.5 [8.6,10.8]
automatic landmarks				
rank-1 accuracy (%)	**92.7 ± 0.6**	92.6 ± 0.6	77.9 ± 1.7	63.2 ± 2.8
EER (%)	7.4 [6.7,8.3]	6.9[6.2,7.7]	7.2 [6.3,8.4]	9.2 [8.5,10.5]
UT-DOOR				
manual landmarks				
rank-1 accuracy (%)	99.2 ± 0.1	**99.5 ± 0.1**	91.4 ± 1.0	86.1 ± 1.6
EER (%)	4.3[3.8, 4.6]	5.0 [4.3,5.7]	8.6 [8.0,9.6]	9.1 [8.5,9.6]
automatic landmarks				
rank-1 accuracy (%)	93.8 ± 0.7	**96.7 ± 0.5**	83.5 ± 2.2	79.7 ± 2.9
EER (%)	5.3[4.6, 5.9]	6.5 [5.7,7.1]	9.5 [8.6,9.9]	10.3 [9.3,11.2]

Bold values indicate the best results.

pose [28]. Both HOG and LBP describe the image by dividing it into local regions, extracting histograms of texture features for each region independently, and then combining these histograms to form a description of the face. Consequently, they are not affected by small local changes compared to methods like PCA or LDA.

When we analyse the results of PCA and LDA, we see that PCA outperforms LDA in identification tests. In verification tests, we see that the difference is not statistically significant, but we see that PCA performs slightly better than LDA. This can occur when the number of samples describing each class is not sufficiently large, or when the training data non-uniformly sample the underlying distribution [36]. Our training set contains 292 images from the CMU Multi-PIE database, where for each person there are four images captured by different cameras. Therefore, we can conclude that our training data is not sufficiently representative for LDA. Moreover, we see that on the CMU Multi-PIE, we achieve better results using automatic landmarks compared to manual landmarks. This shows that due to small training data PCA and LDA do not learn the underlying class distributions optimally.

When we compare the identification results that we achieve using LBP and HOG, we observe that the difference is not statistically significant in most cases. On the UT-DOOR database, we see that HOG performs better than LBP when we use automatically detected landmarks. On the other hand, when we look at the verification results, we see that the confidence intervals for LBP overlap with the confidence intervals for HOG. Therefore, we cannot make a statistically significant claim that HOG is better than LBP.

6 Conclusion and future work

In this work, we propose a face recognition system designed for home safety applications that use side-view faces. We first contribute a newly collected video face database, named UT-DOOR, where 98 subjects were recorded with four cameras attached to a doorpost as they pass through a door. Compared with other available face databases, UT-DOOR consists of video recordings with a significant number of side-view images. It is a large-scale database with a total of 4831 recordings and 62,192 images containing the whole face.

Secondly, we propose a face recognition method, where we automatically detect and recognise faces using side-view images in videos. We present recognition results for side-view face images from the CMU Multi-PIE and the UT-DOOR databases. In our system, we automatically detect the eye centre, nose tip, and mouth corner landmarks with detection accuracies of 96.92, 99.56, and 98.40% in the CMU Multi-PIE database, and 91.91, 98.43, and 89.15% in the UT-DOOR database, respectively. We use these facial points for registration, and test our system using PCA, LDA, LBP, and HOG feature extraction methods. For the fully automatic approach, the rank-1 identification accuracies are 92.7% for the CMU Multi-PIE database and 96.7% for the UT-DOOR database.

In our database, some cameras captured the face partially due to the camera positions or some natural occlusions. Moreover, we have captured videos, where the person is wearing eyeglasses and making expressions. In the future, we aim to improve our registration algorithm to handle instances where some landmarks cannot be detected. We also aim to include a higher pose variation in our experiments to achieve a more robust recognition system.

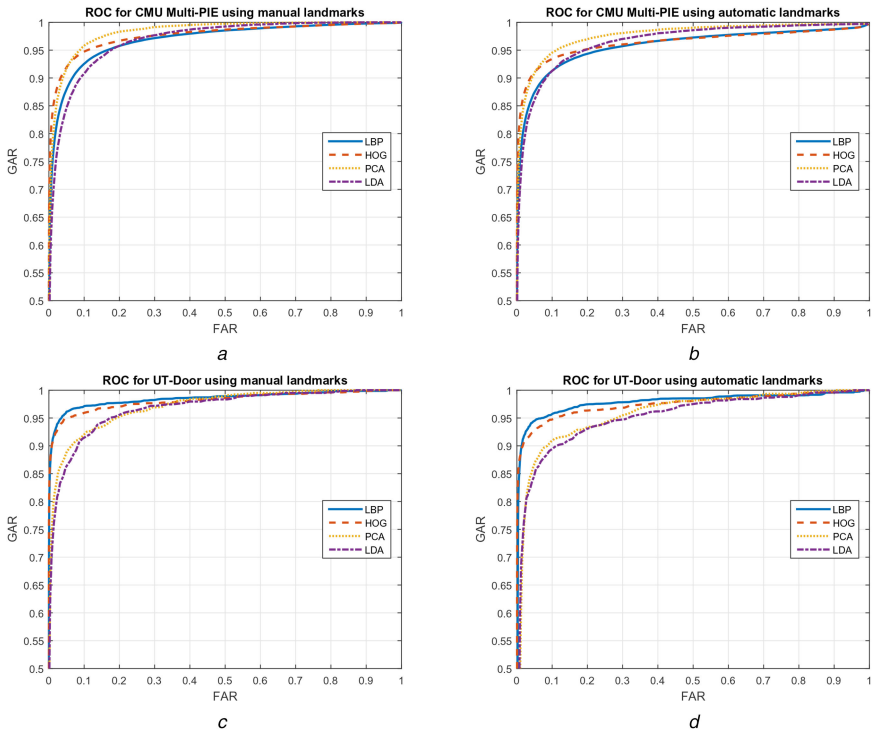

Fig. 8 *ROC curves*

(a) ROC curves for the CMU Multi-PIE images registered using manually labelled landmarks, ***(b)*** automatically detected landmarks, ***(c)*** ROC curves for the UT-DOOR images registered using manually labelled landmarks, ***(d)*** automatically detected landmarks

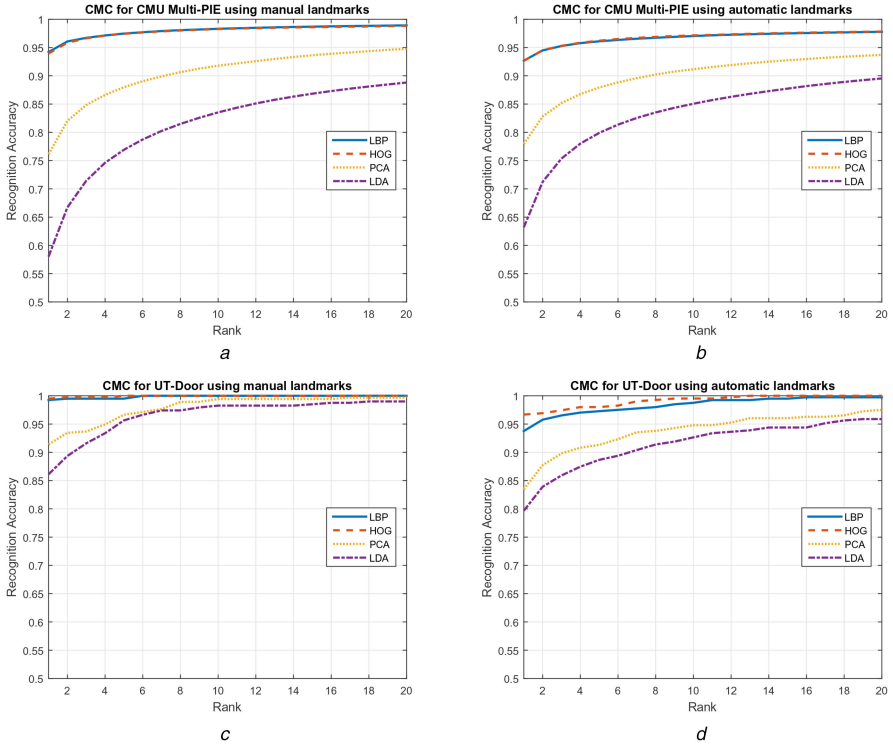

Fig. 9 *CMC curves*

(a) CMC curves for the CMU Multi-PIE images registered using manually labelled landmarks, ***(b)*** Automatically detected landmarks, ***(c)*** CMC curves for the UT-DOOR images registered using manually labelled landmarks, ***(d)*** Automatically detected landmarks

7 Acknowledgment

This work was supported by the GUARANTEE (ITEA 2) 08018 project.

8 References

[1] Zou, X., Kittler, J., Messer, K.: 'Illumination invariant face recognition: a survey'. IEEE Conf. BTAS, Crystal City, VA, USA, 2007, pp. 1–8

[2] Zhao, W., Chellappa, R., Phillips, P.J., *et al.*: 'Face recognition: a literature survey', *ACM Comput. Surv.*, 2003, **35**, (4), pp. 399–458

[3] Ding, C., Tao, D.: 'A comprehensive survey on pose-invariant face recognition', *ACM Trans. Intell. Syst. Technol.*, 2016, **7**, (3), pp. 37:1–37:42

[4] Gao, Y., Leung, M.: 'Line segment Hausdorff distance on face matching', *Pattern Recognit.*, 2002, **35**, (2), pp. 361–371

[5] Bhanu, B., Zhou, X.: 'Face recognition from face profile using dynamic time warping'. Int. Conf. Pattern Recognition, Cambridge, UK, vol. 4, 2004, pp. 499–502

[6] Zhou, X., Bhanu, B.: 'Human recognition based on face profiles in video'. IEEE Conf. CVPR Workshops, San Diego, CA, USA, 2005, p. 15

[7] Abate, A.F., Nappi, M., Riccio, D., *et al.*: '2D and 3D face recognition: a survey', *Pattern Recognit. Lett.*, 2007, **28**, (14), pp. 1885–1906
[8] Zhang, X., Gao, Y.: 'Face recognition across pose: a review', *Pattern Recognit.*, 2009, **42**, (11), pp. 2876–2896
[9] Schroff, F., Treibitz, T., Kriegman, D., *et al.*: 'Pose, illumination and expression invariant pairwise face-similarity measure via Doppelgänger list comparison'. IEEE Conf. Computer Vision, Barcelona, Spain, 2011, pp. 2494–2501
[10] Li, A., Shan, S., Chen, X., *et al.*: 'Cross-pose face recognition based on partial least squares', *Pattern Recognit. Lett.*, 2011, **32**, (15), pp. 1948–1955
[11] Ding, S., Zhai, Q., Zheng, Y.F., *et al.*: 'Side-view face authentication based on wavelet and random forest with subsets'. IEEE Conf. Intelligence and Security Informatics, Seattle, WA, USA, 2013, pp. 76–81
[12] Rathore, R., Prakash, S., Gupta, P.: 'Efficient human recognition system using ear and profile face'. IEEE Conf. BTAS, Arlington, VA, USA, 2013, pp. 1–6
[13] Santemiz, P., Spreeuwers, L.J., Veldhuis, R.N.J.: 'Side-view face recognition'. WIC Symp. Information Theory, Brussels, Belgium, 2011, pp. 305–312
[14] Santemiz, P., Spreeuwers, L.J., Veldhuis, R.N.J.: 'Automatic landmark detection and face recognition for side-view face images'. Int. Conf. Biometrics Special Interest Group, Darmstadt, Germany, 2013, pp. 337–344
[15] Gross, R., Matthews, I., Cohn, J., *et al.*: 'Multi-PIE', *Image Vis. Comput.*, 2010, **28**, (5), pp. 807–813
[16] Phillips, P.J., Moon, H., Rizvi, S.A., *et al.*: 'The FERET evaluation methodology for face-recognition algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (10), pp. 1090–1104
[17] Little, G., Krishna, S., Black, J., *et al.*: 'A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle'. IEEE Conf. Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, vol. 2, 2005, pp. 89–92
[18] Gao, W., Cao, B., Shan, S., *et al.*: 'The CASPEAL large-scale Chinese face database and baseline evaluations', *IEEE Trans. Syst. Man Cybern. A*, 2008, **38**, (1), pp. 149–161
[19] Sim, T., Baker, S., Bsat, M.: 'The CMU pose, illumination, and expression database', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2003, **25**, (12), pp. 1615–1618
[20] Huang, G.B., Ramesh, M., Berg, T., *et al.*: 'Labeled faces in the wild: a database for studying face recognition in unconstrained environments'. University of Massachusetts, Amherst, Technical Report, 2007. pp. 7–49
[21] Messer, K., Matas, J., Kittler, J., *et al.*: 'XM2VTSDB: the extended M2VTS database'. Int. Conf. Audio and Video-Based Biometric Person Authentication, Washington D.C., USA, 1999, pp. 72–77
[22] Klare, B.F., Klein, B., Taborsky, E., *et al.*: 'Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A'. IEEE Conf. CVPR, Boston, MA, USA, 2015, pp. 1931–1939
[23] Savran, A., Alyuz, N., Dibeklioglu, H., *et al.*: 'Bosphorus database for 3D face analysis', *Biometrics Identity Manag.*, 2008, **5372**, pp. 47–56
[24] Goodall, C.: 'Procrustes methods in the statistical analysis of shape', *R. Stat. Soc. Ser. B, Methodol.*, 1991, **53**, (2), pp. 285–339
[25] Viola, P., Jones, M.J.: 'Rapid object detection using a boosted cascade of simple features'. IEEE Conf. CVPR, Kauai, HI, USA, vol. 1, 2001, pp. 511–518
[26] Cortes, C., Vapnik, V.: 'Support-vector networks', *Mach. Learn.*, 1995, **20**, pp. 273–297
[27] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. IEEE Conf. CVPR, San Diego, CA, USA, vol. 2, 2005, pp. 886–893
[28] Ojala, T., Pietikäinen, M., Mäenpää, T.: 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (7), pp. 971–987
[29] Turk, M., Pentland, A.: 'Eigenfaces for recognition', *J. Cogn. Neurosci.*, 1991, **3**, (1), pp. 71–86
[30] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: 'Eigenfaces vs. Fisherfaces: recognition using class specific linear projection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (7), pp. 711–720
[31] Vedaldi, A., Fulkerson, B.: 'VLFeat: an open and portable library of computer vision algorithms'. Proc. ACM Int. Conf. Multimedia, New York, NY, USA, 2010, pp. 1469–1472
[32] Ahonen, T., Hadid, A., Pietikainen, M.: 'Face recognition with local binary patterns'. European Conf. Computer Vision, Prague, Czech Republic, 2004, pp. 469–481
[33] Mandal, B., Wang, Z., Li, L., *et al.*: 'Performance evaluation of local descriptors and distance measures on benchmarks and first-person-view videos for face identification', *Neurocomputing*, 2016, **184**, pp. 107–116
[34] Ruiz-del Solar, J., Navarrete, P.: 'Eigenspace-based face recognition: a comparative study of different approaches', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2005, **35**, pp. 315–325
[35] Dunstone, T., Yager, N.: '*Biometric system and data analysis design, evaluation, and data mining*' (Springer-Verlag, New York, NY, 2008)
[36] Martinez, A.M., Kak, A.C.: 'PCA versus LDA', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (2), pp. 228–233