Engineering Journal issn: 0125-8281

Article

Face Synthesis and Partial Face Recognition from Multiple Videos

Warinthorn Nualtim[1,a], Watcharapan Suwansantisuk[2,b*], and Pinit Kumhom[2,c]

1 Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand
2 Department of Electronic and Telecommunication Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand
E-mail: [a]warinthorn.atom@mail.kmutt.ac.th, [b,*]watcharapan.suw@mail.kmutt.ac.th (Corresponding author), [c]pinit.kumhom@mail.kmutt.ac.th

Abstract. Surveillance videos provide rich information to identify people; however, they often contain partial facial images that make recognition of the person of interest difficult. The traditional method of partial face recognition uses a database that contains only full-frontal faces, resulting in a reduction in the performance of recognition models when partial face images are presented. In this study, we augmented the database of full-frontal face images and synthesized two- and three-dimensional facial images. We designed a method for partial face recognition from the augmented database. To synthesize the two-dimensional (2D) facial images, we divided the available video images into groups based on their similarity and chose a representative image from each group. Then, we fused each representative image with a full-frontal face image using the scale-invariant feature transform (SIFT) flow, and augmented the original database with the fused images. To design a partial face recognition algorithm, we carefully evaluated the similarity between a set of video images from cameras and an image from the augmented database by counting the number of keypoints given by the SIFT. Compared to competitive baselines, the proposed method of partial face recognition has the highest face recognition rates in four out of six test cases on the widely used ChokePoint dataset, using most subjects (so-called subject group B) in the gallery. The proposed method also has recognition rates of approximately 22% to 72% on the test cases. The 2D face synthesis was found to outperform the three-dimensional (3D) face synthesis on a large subject group, possibly because the method of 2D reconstruction retains important facial features. The methods of augmentation and partial-face recognition are simple and improve the face recognition rate of traditional methods.

Keywords: Multiple videos, partial face recognition, face synthesis.

ENGINEERING JOURNAL Volume 27 Issue 4
Received 12 September 2022
Accepted 20 April 2023
Published 30 April 2023
Online at https://engj.org/
DOI:10.4186/ej.2023.27.4.29

DOI:10.4186/ej.2023.27.4.29

1. Introduction

Among several methods to identify a person, face recognition stands out due to its multiple benefits. Unlike identity recognition through a smartphone application, fingerprint, key-card, or human-computer interface, face recognition avoids physical contact and could be beneficial during a pandemic [1]. Face recognition uses input images from video cameras, which are ubiquitous, and can be used for monitoring and surveillance purposes. Videos offer different views of a person, providing diversity for face recognition and potentially improving the face recognition rate. Face recognition from videos is widely used for identity verification.

The difficulties in face recognition arise from the environment in which the videos are recorded. The images captured by the video cameras may originate either from controlled or uncontrolled environments. In controlled environments [2], captured images have high resolution and contain full frontal faces of a person under suitable lighting conditions. In uncontrolled environments [3], captured images have low resolution and contain non-frontal, partial, or occluded faces. Face recognition in uncontrolled environments is challenging, but has significant applications [4–6] in surveillance and threat monitoring.

Existing research has provided several methods to improve partial face recognition in uncontrolled environments. A previously proposed model [7, 8] employed a super-recognition method by learning the relationship between the high-resolution space and very-low-resolution space, and mapping a very low-resolution image to a high-resolution image. For non-frontal facial input images, active appearance model (AAM) was proposed [9] that followed a multiresolution scheme. The first level of the multiresolution scheme initialized a generic AAM and enabled automatic estimation of the pose angle. The next level further refined the AAM model. A virtual frontal-face image was created before matching it to the images in the full-frontal face database [9]. An alignment-free face-recognition method was also proposed [10] that used multi-keypoint descriptor (MKD). It enabled advanced extraction of features, including both the position and direction of the facial image. Generally, MKD does not require facial alignment between the input and gallery image during the matching process. Instead, it uses sparse representation-based classification (SRC) that classifies partial and occluded faces [11].

The methods in existing research aim to enhance the rate of face recognition and mitigate the difficulties encountered in partial face recognition. They apply image fusion to extract and combine the representations of multiple video images. The image fusion method can be divided into two groups based on the fusion of pixels or features. A technique called pixel-fusion based stereo image retargeting was proposed that is suitable for fusing two distinct images [12]. The proposed approach used image pairs to evaluate performance. A single-pixel fusion method that combines two stereo images was also proposed [12]. In this method, once the input and reference information were fused, the input images were compared to determine whether they depicted the same face. Using feature fusion, the papers [13, 14] proposed a technique called scale invariant feature transform (SIFT) flow to combine the features of two images. SIFT flow is robust against the misalignment of two images. It is suitable for the fusion of still and dynamic video images taken in the same scene from different viewpoints. Image fusion combines numerous input images into a few images and simplifies the face recognition task.

In addition to image fusion, feature extraction is a common step in facial recognition. Feature extraction needs to be robust against illuminations, image scaling, and image rotation. The SIFT algorithm [15] extracts image features and can handle various illuminations. The SIFT algorithm finds matching keypoints between an input and database image. The SIFT algorithm was used to proposed a stereo matching method for partial face recognition using low-resolution face images from a surveillance video [16]. Additionally, a transformation matrix for facial features of low-resolution and high-resolution images was also developed [16]. The low-resolution image contains a face to be recognized and the high-resolution image was a gallery image. The stereo matching cost is obtained between the SIFT features of a low-resolution image and a high-resolution image. The approach proposed in [16] contains the training and testing states. In the training state, the model is trained using a transformation matrix between high-resolution images and low-resolution non-frontal face images. The training state computes the descriptor for various facial locations. Using the SIFT descriptor, the testing state reconstructs every point of the high-resolution gallery images from low-resolution images and applies a stereo matching cost to compute the distance between the two images. The features can be compared to measure the similarity between two images.

Augmentation is suitable for a database that contains a few full-frontal faces of each person, taken, for example, from an identification (ID) card or a passport. Several studies have improved face recognition by augmenting galleries using synthesized facial images. A technique called domain-specific face synthesis (DSFS) was proposed [17], which exploited the representative intraclass variation information in facial images that were captured in an uncontrolled environment and obtained from multiple video cameras. The paper [18] proposed a technique of 3D face reconstruction using more than one samples per person in the database to better recognize a pose face and an occluded face. A hybrid face recognition using face synthesis was proposed [19]. In this method, 2D face images were

transformed, and 3D techniques were combined by extracting the features obtained from the SIFT. A facial sketch synthesis system was developed [20] using 2D combined model-based face-specific Markov network (2DDCM) features. The face synthesis in the 2DDCM approach uses candidate patches that are derivatives of sample training patches. Face synthesis is performed by emphasizing strong lines or curves to enhance the shadow regions. A previous study [21] proposed a synthesized virtual frontal face from a pose image for face recognition. The first step is to build a model as the morphable displacement field (MDF) to encode the full-frontal-face database calculated from the 3D face model using the maximal likelihood correspondence estimation (MLCE) method. In the paper [21], a mask was generated from a 3D face model to aid face recognition. In the paper [22], 3D morphable model (3DMM) and generative adversarial network (GAN) were proposed to restore de-occluded facial images. The 3DMM framework removes face occlusions from other objects and synthesizes a face region. Synthesized images were added to the gallery to provide different views of the full-frontal face images.

Existing face recognition methods are fundamentally limited. Current 3D synthesis methods need to estimate the roll, pitch, and yaw of a face for face reconstruction. A synthesis of illumination and lightning conditions is also required [17, 18]. However, these steps are not straightforward. A simple method for augmenting the full-frontal face gallery, combined with a simple method for face recognition, will improve the face recognition rate.

In this study, we designed simple methods of face synthesis to augment the gallery. We also designed the corresponding method of partial face recognition to improve the performance of still-to-video face recognition. The proposed method takes videos of a person and the full-frontal face gallery as an input and identifies the person in the gallery that matches the person in the video. The video images were obtained from uncontrolled environments. The main aims of this study are as follows.

- design a simple method of 2D face synthesis, to augment the full-frontal face images in the gallery with the synthesized profile images.
- propose a method for partial face recognition.
- perform a comparison of the rate of face recognition of the proposed method with that of existing methods.

The proposed face recognition method is based on a simple but effective idea of scoring the difference between images in the gallery and input videos. The proposed face synthesis and recognition methods are easy to implement and result in a better face recognition rate than the baseline methods. They have practical utility and are suitable for recognizing partial faces in videos.

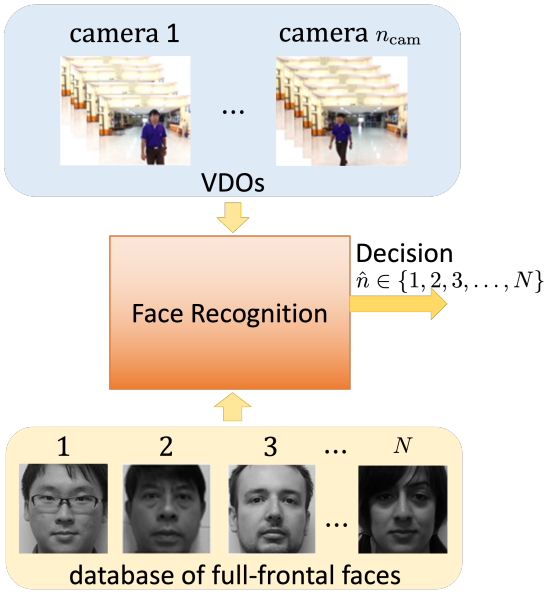


Fig. 1. Face recognition provides the index $\hat{n}$ of the gallery image that best matches the person in the videos.

The remainder of this paper is organized as follows. Section 2 describes this problem. Section 3 develops the methods of 2D and 3D face syntheses to augment gallery images, and develops a method for face recognition from augmented gallery images. Section 4 evaluates the proposed method, and compares the face recognition rate to the baselines. Section 5 concludes the paper and summarizes the important findings.

2. System Model

A face recognition system takes videos containing a person's face and gallery images of the full-frontal faces of different people as input, as shown in Fig. 1. Videos were recorded by n_{cam} cameras to provide different viewpoints of the person of interest, where $n_{\text{cam}} \geq 1$. Each video frame was a color image of the full or partial face of the person of interest. The gallery images consisted of N full-frontal face grayscale images of N distinct people, one of whom appears in the input videos. As per convention, N distinct people in the gallery were indexed and referred to as $1, 2, 3, \ldots, N$. In this study, we aimed to identify the person who appears in the videos so that the face recognition rate is as high as possible. The output of the face recognition system is the index $\hat{n}$ of the person in the gallery that best matches the person in the video, where $\hat{n} \in \{1, 2, 3, \ldots, N\}$.

3. Proposed Method

The proposed method for face recognition consists of two steps: face synthesis and face recognition. In the face synthesis stage, we augment each full-frontal face gallery image with images of the same person in different views.

DOI:10.4186/ej.2023.27.4.29

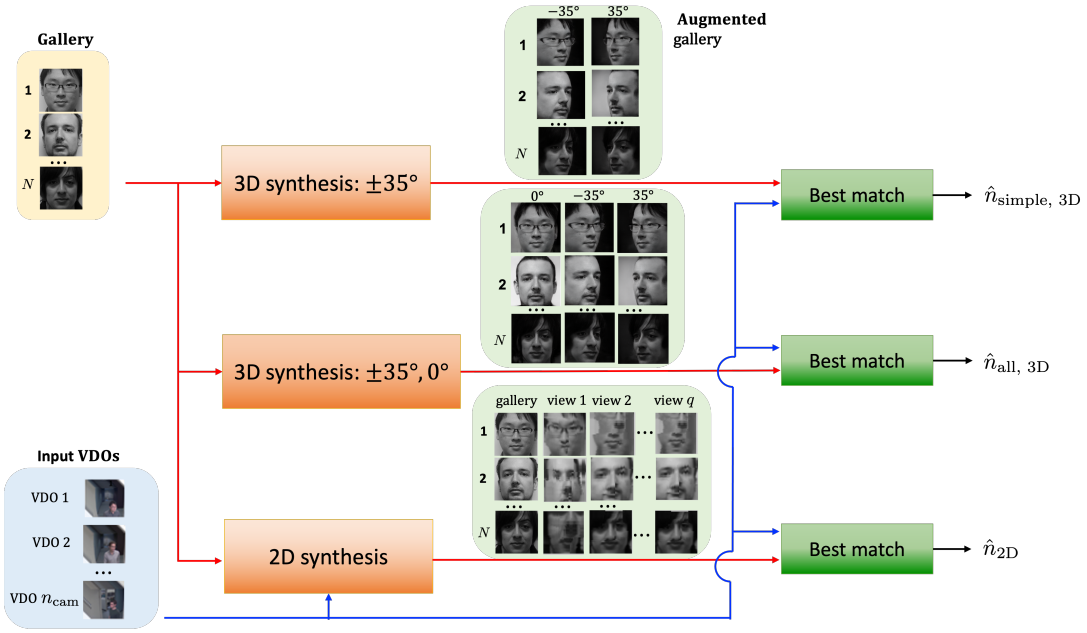


Fig. 2. The proposed 2D and 3D face syntheses extend the reference set.

We utilize two methods of face synthesis: 2D face synthesis and 3D face reconstruction. Both methods synthesize additional views for each full-frontal face image in a gallery. After face synthesis, we developed an algorithm to recognize the person in the videos from the gallery by matching the person in the videos to the most similar person in the augmented gallery.

Figure 2 depicts the overall workflow of the proposed method of face synthesis and face recognition. As described in the previous section, the inputs are videos and full-frontal gallery images, and the output is the index of the person: $\hat{n}_{2D}$ if the 2D face synthesis is in use; $\hat{n}_{\text{simple, 3D}}$ if the 3D face synthesis is in use, and if the full-frontal face images are excluded from the augmented gallery; and $\hat{n}_{\text{all, 3D}}$ if the 3D face synthesis is used, and if the full-frontal faces are included in the augmented gallery. The outputs from 2D and 3D face syntheses are augmented galleries, which contain various views of each full-frontal face image. The face recognition algorithm considers the augmented galleries as well as the input videos, and outputs the index of the person in the gallery that best matches the person in the videos.

Figure 3 depicts an example of the input and output of the proposed 2D face synthesis. The input consists of videos and a gallery of images. The output, through the process that we will soon describe, is a matrix of images. Each row represents a person who appears in the gallery. The columns show different views of the person, starting from full-frontal face images in the first column to other views in the remaining columns. The number of other views is denoted as q in the figure. The q is the number of groups of distinct images in the videos. For a given person, the image in a column is a fusion of the full-frontal face image and the representative image from each group. Figure 3 gives an overview of the 2D face synthesis procedure.

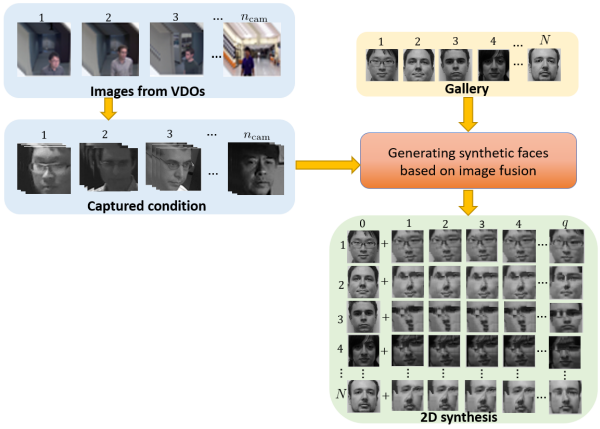


Fig. 3. The proposed 2D face synthesis fuses the gallery and input video images.

3.1. Two-Dimensional Face Synthesis

The proposed method of face synthesis consists of four steps: preprocessing, grouping, finding group representatives, and image fusion, as shown in Fig. 4. First, preprocessing is used to prepare the video images for enhanced face recognition. The preprocessing methods used in the proposed method are grayscale conversion, face detection, image resizing, and image equalization. Grayscale conversion converts the color image in each video frame into a grayscale image. Face detection uses the AdaBoost cascade classifier [23] to identify and crop a face in each video frame image. Image resizing scales the cropped facial image to di-

mensions of $m \times m$ pixel^2, where m is a design parameter. Image equalization [24] adjusts the contrast and enhances the quality of each cropped and resized image. Preprocessing converts a sequence of video frames into a sequence of grayscale images containing the face of the person of interest. Second, the incoming video images are partitioned into several groups, each containing similar images. The first image in the preprocessed video sequence is selected as a member of the first group. Subsequent video images similar to the first image of the current group are included as group members. We measure the similarity between two video images using the mean square error (MSE) of their difference. The two images are considered similar if their MSE is below the design threshold. A video image that is very different from the first image of the current group triggers the creation of a new group and becomes the first image of the new group. Upon completion of the comparison, the grouping step partitions the video images into a certain number, q, of groups. Third, to determine a representative of each group, we choose the first image in each group as the representative image. This approach is reasonable given that all images in the same group are similar in appearance, as measured by MSEs. Fourth, in image fusion, we used the SIFT flow [14] to fuse the full-frontal face image and the representative image from each group. The 2D face synthesis results in an augmented gallery, as illustrated in Fig. 5.

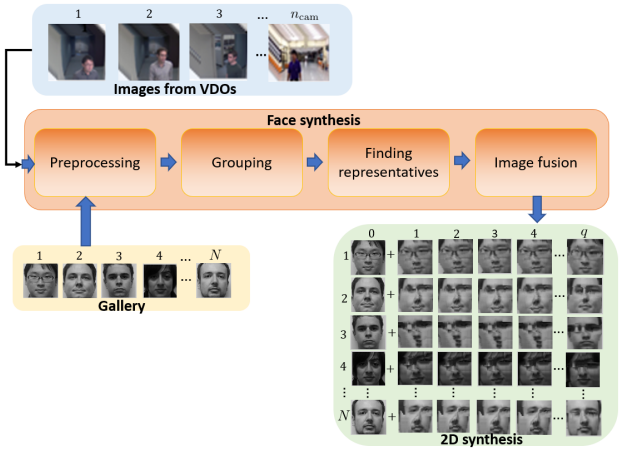


Fig. 4. The proposed 2D face synthesis method consists of four major steps.

Algorithm 1 shows the algorithm for 2D face synthesis. Lines 1–23 of the algorithm divide the frames in the video into groups. Each group consists of video images that are similar to one another. The variable `groupCount` denotes the current number of groups. The `groupCount` is initialized to zero. The algorithm then iterates through each frame of the video. If the number of groups is zero, the first image is immediately converted into a new, additional group, and is declared the representative of that group. A variable `Gimg` is a representative image of the current group. If the number of groups is not zero, the current image is compared to the representative image by the MSE in line 11. If the MSE exceeds a threshold, the current image is deemed to be different from the images in the current group, and the current image triggers the creation of a new group. The current image also becomes the representative image of the group, as seen in lines 16–19. Upon completion of the grouping procedure, the array `group`$[i]$ is a list of video images in the ith group, and the first element in the list is the representative image of the group. Finally, the algorithm fuses each representative image with each frontal facial image using the SIFT flow in line 31, resizes the fused image to the original dimension of $m \times m$ pixel^2, and augments the fused, resized image to the gallery. Subsequently, the algorithm of the 2D face synthesis is terminated.

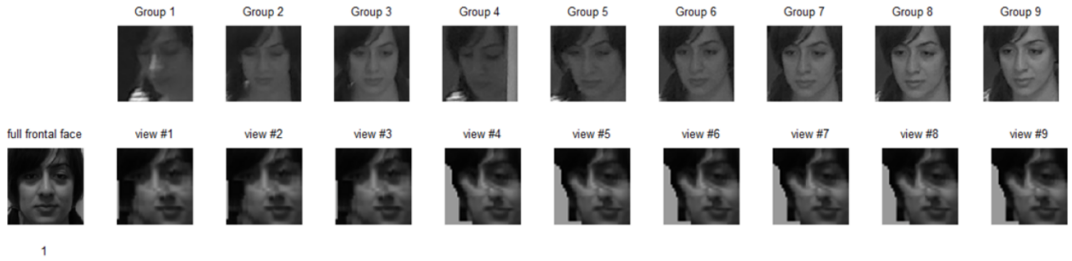
Algorithm 1 Two-dimensional Face Synthesis

Input: A sequence S_i of images from the ith camera, where $1 \leq i \leq n_{\text{cam}}$; A sequence D of N frontal-face square images from the database

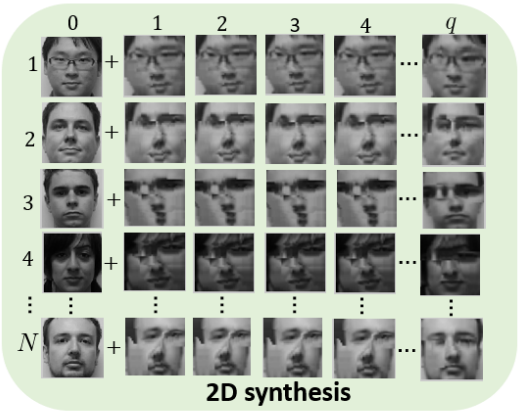
Output: an array $D_{\text{aug}}[1..N][0..q]$ of augmented gallery images

1: Initialize the total number of groups to zero: `groupCount` $= 0$
2: **for** $i = 1$ to n_{cam} **do**
3: **for** $j = 1$ to Number of images from the ith camera **do**
4: Let I denote the jth image from the ith camera: $I = S_i[j]$
5: **if** (`groupCount` == 0) **then**
6: `groupCount` = `groupCount` + 1
7: Initialize `group`[1] to be the empty array
8: Append image I to the array `group`[1], i.e., `Gimg` = `group`[1][0] = I
9: Initialize the number of images in the current group to one: `gsize` = 1
10: **else**
11: `mse`= MSE between I and `Gimg`
12: **if** (`mse`$\leq$ threshold) **then**
13: Append image I to group `group` [`groupCount`], i.e., `group`[`groupCount`] [`gsize`] = I
14: `gsize` = `gsize` + 1
15: **else**
16: `groupCount` = `groupCount` + 1
17: Initialize `group`[`groupCount`] to be the empty array
18: Append image I to array `group` [`groupCount`], i.e., `Gimg` = `group`[`groupCount`][0] = I
19: Initialize the number of images in the current group to one: `gsize` = 1
20: **end if**
21: **end if**

DOI:10.4186/ej.2023.27.4.29



(a) Input video images are converted into q groups, shown as an example for $q = 9$. The representative image of each group (top row) is fused with the full-frontal face image (left) to form a synthesized face.



(b) After the proposed algorithm partitions the video images into q groups, the algorithm fuses each one of N full-frontal face image with the representative image of each group.

Fig. 5. The 2D face synthesis produces the augmented gallery, which contains the original N full-frontal face images and Nq synthesized images.

end for
end for
Assign $q =$ `groupCount`
Initialize D_{aug} to be an array of N rows and $1 + q$ columns.
for $r = 1$ to N **do**
  `Fimg` $=$ the rth full-frontal face image, $D[r]$, in the gallery
  $D_{\text{aug}}[r][0] =$ `Fimg`
  for $k = 1$ to `groupCount` **do**
    `Gimg` $=$ the representative image, `group`$[k][0]$, of the kth group
    $D_{\text{aug}}[r][k] =$ `getSIFTFLOW(Gimg, Fimg)`
    Resize the augmented image D_{aug} to the size of $m \times m$ pixel[2]
  end for
end for

3.2. Three-Dimensional Face Synthesis

In 3D face synthesis, we use the render-and-rotate method [25] with a pre-trained GAN to produce profile images from each N full-frontal face image. The profile images are set such that their yaw angles are $\pm 35°$, which cover a reasonable range of angles, as shown in Fig. 6. The profile images provide different views of each person, improving the chance of correctly identifying the person of interest in the input videos.

We produce an augmented gallery in two slightly different manners. First, the augmented gallery consists of the profile images of $-35°$, $0°$, and $35°$ yaws from the render-and-rotate method.[1] Second, the augmented gallery contains only the $-35°$ and $35°$ profile images. Excluding the $0°$ yaw image is beneficial when the input video images are heavily occluded, because a full-frontal face image complicates the decision of the face recognition system. Three-dimensional face synthesis reconstructs the missing portion of the face and provides profile images that are similar to those appearing in the pre-trained dataset.

Algorithm 2 Face Recognition from Augmented Database

Input: A sequence S_i of images from the ith camera, where $1 \leq i \leq n_{\text{cam}}$; An array $D_{\text{aug}}[1..N][0..q]$ of

[1]The profile image of $0°$ yaw is slightly different from the input full-frontal face image to the render-and-rotate method.

(a) Example of full-frontal faces and rotated images

(b) Example of augmented gallery that contains only profile images

Fig. 6. Full-frontal face images from the ChokePoint dataset are rotated to $-35°$ and $35°$.

the augmented gallery images

Output: An index $\hat{n}$ of the gallery image that best matches the person in the input videos, where $1 \leq \hat{n} \leq N$

1: **for** $i = 1$ to n_{cam} **do**
2: **for** $j = 1$ to Number of images from the ith camera **do**
3: Detect a face in the image $S_i[j]$
4: Crop a square that contains a face in image $S_i[j]$
5: Resize the cropped facial image to $m \times m$ pixel^2 and equalize the image
6: Initialize $\tilde{S}_i[j]$ to be the resulting $m \times m$ equalized facial image
7: **end for**
8: **end for**
9: **for** $r = 1$ to N **do**
10: **for** $k = 0$ to q **do**
11: Resize the gallery image $D_{\text{aug}}[r][k]$ to the size of $m \times m$ pixel^2 and equalize it
12: Initialize $\tilde{D}_{\text{aug}}[r][k]$ to be the resized, equalized gallery image
13: **end for**
14: **end for**
15: `maxScore`$=-\infty$
16: **for** $r = 1$ to N **do**
17: **for** $i = 1$ to n_{cam} **do**
18: `selectedIndices` $=$ `imageSelection` $(\tilde{S}_i, \tilde{D}_{\text{aug}}[r][0])$
19: **for each** $j \in$ `selectedIndices` **do**
20: **for** $k = 0$ to q **do**
21: `curr` $=$ `Match`$(\tilde{S}_i[j], \tilde{D}_{\text{aug}}[r][k])$
22: **if** (`curr` $>$ `maxScore`) **then**
23: `maxScore` $=$ `curr`
24: $\hat{n} = r$
25: **end if**
26: **end for**
27: **end for**
28: **end for**
29: **end for**
30: **return** $\hat{n}$

3.3. Face Recognition

The proposed method for face recognition uses video images and augmented gallery images as the input and identifies the person in the gallery image that best matches the person in the videos. The overall face recognition method consists of three main steps: pre-processing, image selection, and image identification, as shown in Fig. 7. Pre-processing detects a face from each video image and adjusts the image contrast. Image selection selects a subset of video images that are suitable for face recognition and discards poor quality images that do not provide sufficient features for face recognition. Image identification finds the features of each selected video image and measures the similarity between each selected video image and gallery image. The output of the face recognition method is the index $\hat{n}$ of the augmented gallery.

The proposed face recognition method is presented in Algorithm 2. The input is the video image S_i from the ith camera, where $1 \leq i \leq n_{\text{cam}}$, and array $D_{\text{aug}}[1..N][0..q]$ of augmented gallery images, which come from either the 2D face synthesis, the 3D face synthesis with full-frontal faces and the profile faces, or the 3D face synthesis with the profile faces only, where q denotes the number of views. The output of the algorithm is the index $\hat{n}$ of the gallery image that best matches the video images, where $1 \leq \hat{n} \leq N$. Lines 1–14 represent the preprocessing steps. The video images are cropped to contain the face, resized to a square of size $m \times m$ pixels^2, and equalized to adjust the contrast. Similarly, gallery images are resized to the same size and equalized. Line 15 initializes the variable `maxScore` to the smallest number. The variable `maxScore` is the level of similarity between the video and gallery images. Lines 16–29 iterate through each N gallery image and n_{cam} camera. Line 18 selects a subset of the video images (from the current camera) that look similar to the reference image, which is considered to be the first available view in the augmented

DOI:10.4186/ej.2023.27.4.29

gallery. The selection method appears in our previous contribution [26, Algorithm 4] and will not be repeated here. The rationale for image selection is to discard low-quality images from the pool to enhance the face recognition rate.[2] Lines 19–20 iterate through each of the selected video images and view in the augmented image. Line 21 finds the similarity between the current video image and the current gallery image. The score is computed using SIFT [15], which returns the number of matching keypoints between two images. Lines 22–25 update `maxScore` and the best matching index $\hat{n}$, if the number of keypoints is greater than the previously recorded maximum. Line 30 returns the index $\hat{n}$ that best matches video images. Subsequently, the algorithm terminates.

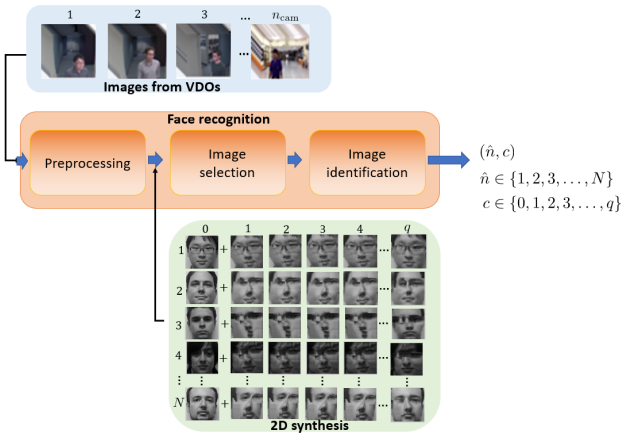


Fig. 7. The proposed face recognition method uses an extended reference set.

4. Performance Evaluation

To evaluate the performance of our proposed method, we applied the proposed face recognition algorithm, given by Algorithm 2, to a benchmark dataset and compared the face recognition rate to those of the baseline algorithms [26,27]. The benchmark dataset was the Chokepoint dataset [28], which is widely used. The dataset provided a database of 25 high-resolution full-frontal-face gallery images, consisting of 19 males and six females. The baselines are unified face image (UFI) [27] and multiple-camera algorithms [26], which are suitable for face recognition from multiple videos. All the algorithms were evaluated in the same setting to ensure fairness of the comparison. A higher face recognition rate indicates better performance.

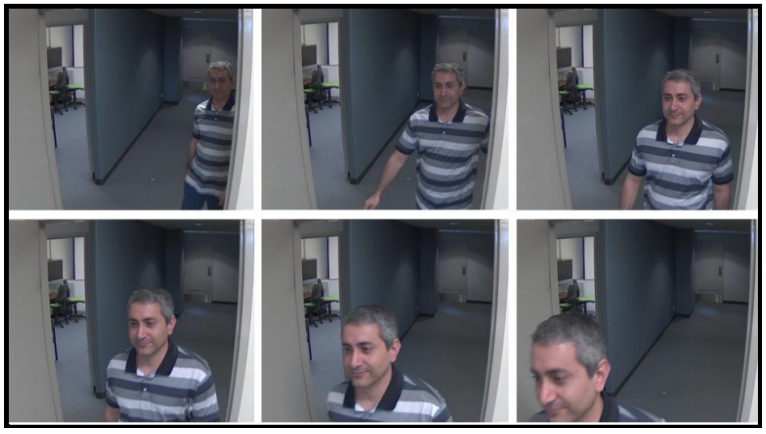
We selected suitable parameters for performance evaluation. We determined the threshold that appears in line 12 of Algorithm 2 as 65. A small threshold yields a large number q of synthesized faces per person and a large augmented gallery, which may lead to ambiguity in face detection and a low face recognition rate. A large threshold yields a small number of synthesized faces and a small augmented gallery, which could be too small for the detection of occluded faces in the input video. The threshold of 65 is derived from trial-and-error, and suits the benchmark dataset. Twelve subjects (denoted subject group A) and 22 subjects (subject group B), out of the 25 available in the dataset, were selected to reduce the running time of the evaluation test and to mimic a typical situation for which the number of persons in the gallery is larger than the number of persons being recognized through the videos. The subjects served as fair representatives for performance evaluation. In the Chokepoint dataset, we selected six test cases, called ES1, ES2, ES3, LS1, LS2, and LS3, to be consistent and to facilitate comparison with the baseline methods [26, 27]. Each test case provided images captured by two video cameras in two different views of a person. See Fig. 8 as an example of an input video image. In each test case, the video frames from the same camera were sufficiently similar. We took the first thirty frames as an input to the face synthesis algorithm (Algorithm 1) in subject group A, and took all frames in subject group B, to demonstrate that, if large enough, the number of frames for face synthesis is insignificant. However, when we performed face recognition (Algorithm 2), we used all available video frames to enhance the face recognition rate. In the UFI algorithm, we averaged $k = 2$ images [27]. The chosen parameters were suitable and enhanced the performance of both the proposed and the baseline algorithms.

Tables 1 and 2 show the face recognition rates of the various methods for the subjects in group A and group B, respectively. The existing methods consist of the UFI [27] and the multiple-camera method [26]. The proposed method includes three different variations to augment the gallery: 2D face synthesis, 3D face synthesis with $\pm 35°$ profile images, and 3D face synthesis with $-35°, 0°, +35°$ profile images. Different test cases appear in different columns, which are labelled as ES1, ES2, ES3, LS1, LS2, and LS3. Videos in test cases ES1–ES3 contained large yaw angles of the people, while videos in test cases LS1–LS3 contained small yaw angles, approximately 10°–20°. In general, test cases ES1–ES3 were more difficult for face recognition than test cases LS1–LS3. Entries in the table denote the face recognition rate of each method under each test case. The highest face recognition rate for each test case is emphasized with an underline.

Several conclusions can be drawn from the results in Table 1. Our method has a higher recognition rate than existing methods for the test cases ES2, ES3, and LS3. The highest face recognition rates were mostly due to the method of 3D face synthesis, combined with the proposed face recognition, and were 41.7%, 41.7%, and 66.7% for test cases ES2, ES3, and LS3, respectively. Test cases ES1, LS1, and LS2 were the ones in which the proposed method were not the best performer. For ES1, the method of fusion

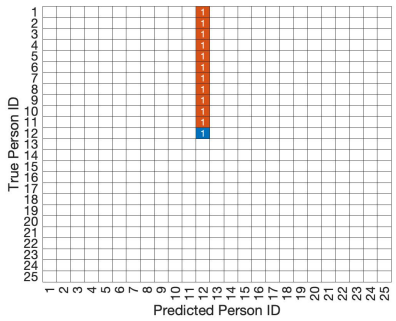
[2]The selection process occurs for a 2D-augmented gallery only. For a 3D-augmented gallery, all available video images are in use.

(a) Example input images at every eighth frame of a video from portal camera 1

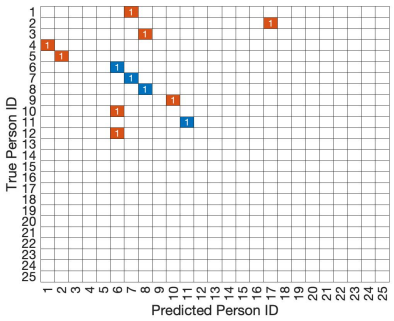


(b) Example input images at every eighth frame of a video from portal camera 2

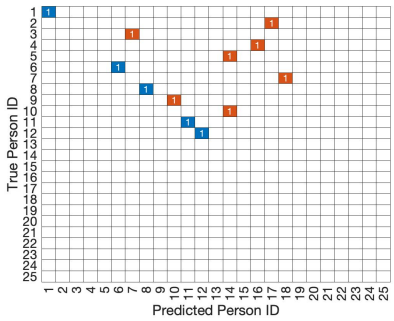
Fig. 8. Input images of ChokePoint dataset are captured from camera 1 and camera 2.



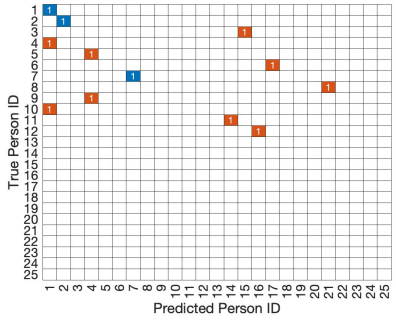
(a) UFI



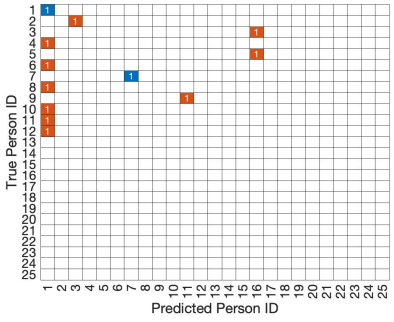
(b) No Fusion



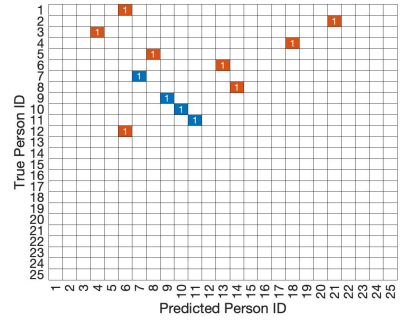
(c) Fusion



(d) Proposed: $\pm 35°$

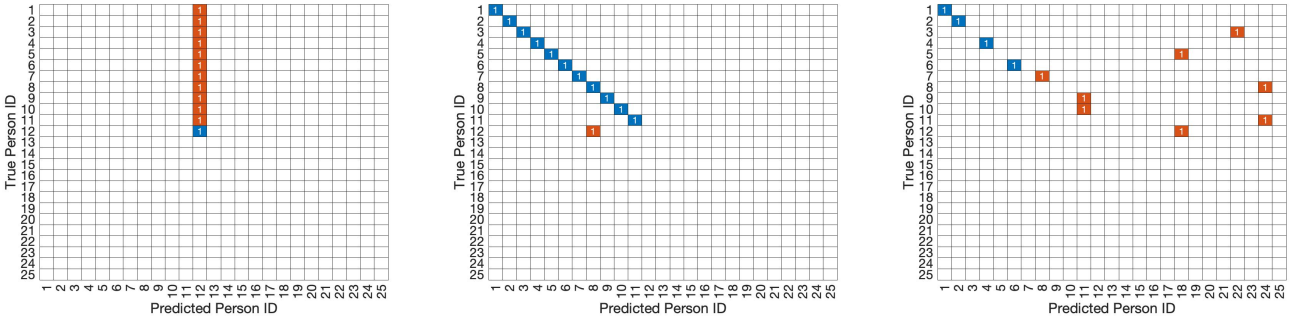


(e) Proposed: $\pm 35°, 0°$

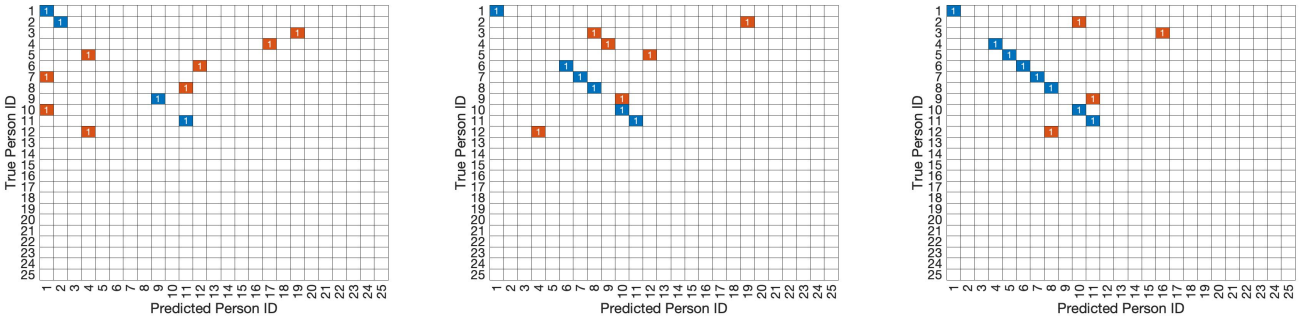


(f) Proposed: 2D

Fig. 9. Confusion matrices are shown for the most difficult test case ES1, where most video frames consist of occluded faces (Subject group A).

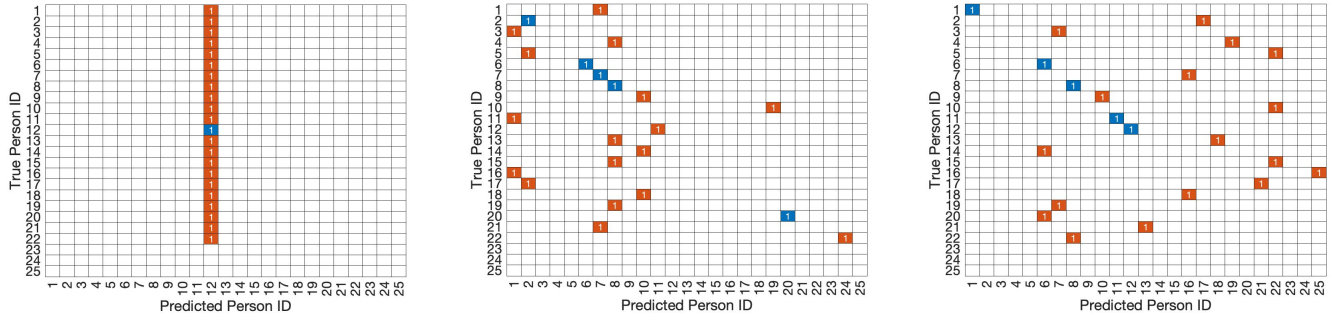


(a) UFI (b) No Fusion (c) Fusion

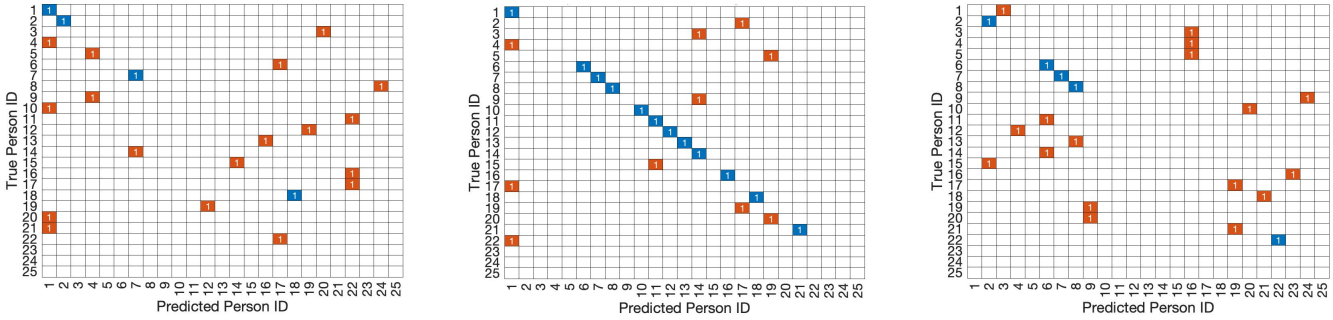


(d) Proposed: ±35° (e) Proposed: ±35°, 0° (f) Proposed: 2D

Fig. 10. Confusion matrices are shown for the simplest test case LS1, where most video frames consist of full-frontal faces (Subject group A).



(a) UFI (b) No Fusion (c) Fusion



(d) Proposed: ±35° (e) Proposed: ±35°, 0° (f) Proposed: 2D

Fig. 11. Confusion matrices are shown for the most difficult test case ES1, where most video frames consist of occluded faces (Subject group B).

Table 1. Face-recognition rates of the proposed methods outperform face-recognition rates of a baseline for subject group A.

Method	Test cases on subject group A					
	ES1	**ES2**	**ES3**	**LS1**	**LS2**	**LS3**
UFI [27]	8.3%	8.3%	8.3%	8.3%	8.3%	8.3%
No Fusion [26]	33.3%	25.0%	25.0%	91.7%	66.7%	41.7%
Fusion [26]	41.7%	25.0%	33.3%	33.3%	25.0%	41.7%
Proposed: 3D Rotated Face $-35°, 35°$	25.0%	25.0%	16.7 %	33.3%	41.7%	25.0%
Proposed: 3D Rotated Face $-35°, 0°, 35°$	16.7 %	41.7 %	41.7%	50.0%	41.7%	66.7%
Proposed: 2D Face Synthesis	33.3%	25.0%	25.0%	66.7%	50.0%	58.3%

Table 2. Face-recognition rates of the proposed methods outperform face-recognition rates of a baseline for subject group B.

Method	Test cases on subject group B					
	ES1	**ES2**	**ES3**	**LS1**	**LS2**	**LS3**
UFI [27]	4.5%	4.5%	4.5%	4.5%	4.5%	4.5%
No Fusion [26]	22.7%	22.7%	27.3%	77.3%	63.6%	40.9%
Fusion [26]	22.7%	22.7%	31.8%	36.4%	40.9%	31.8%
Proposed: 3D Rotated Face $-35°, 35°$	18.2%	27.3%	18.2 %	36.4%	36.4%	36.4%
Proposed: 3D Rotated Face $-35°, 0°, 35°$	13.6 %	18.2 %	27.3 %	54.5%	45.5%	50.0%
Proposed: 2D Face Synthesis	22.7%	27.3%	31.8%	72.7%	63.6%	36.4%

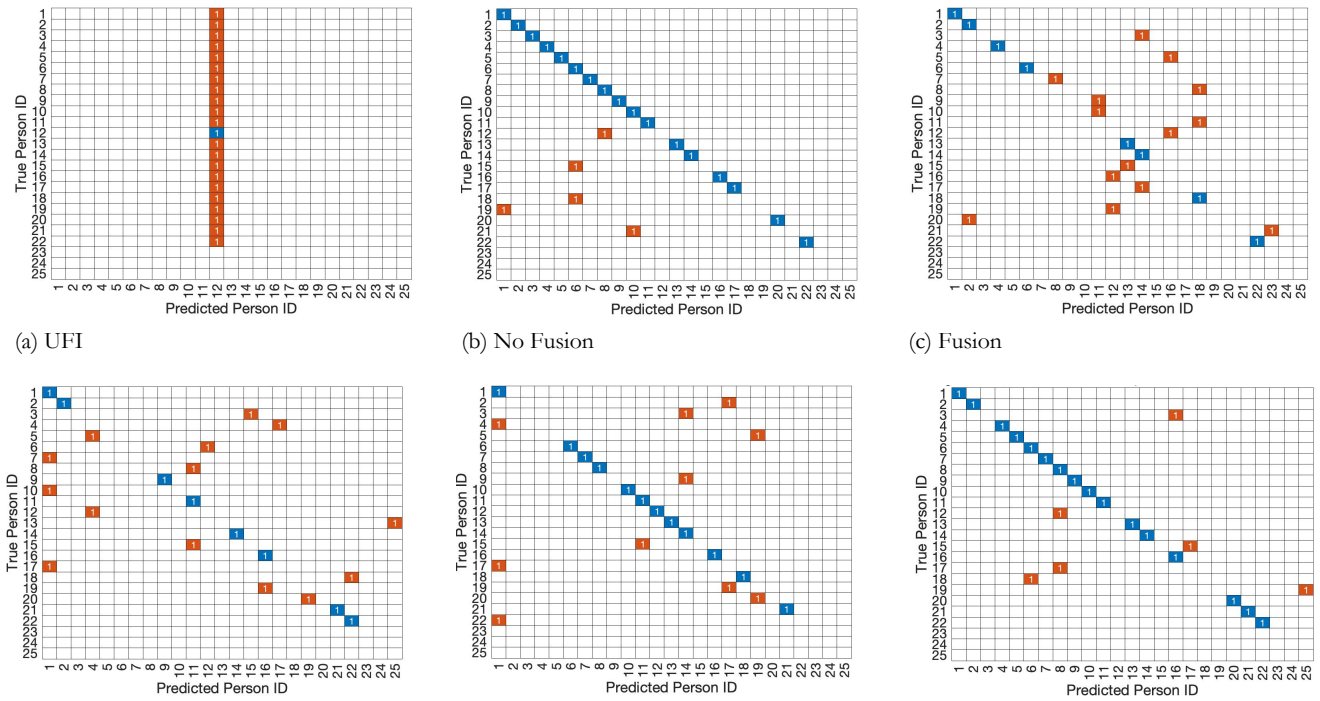
achieved the highest face recognition rate of 41.7%, which is better than the face recognition rate of 16.7% achieved by the proposed 3D face synthesis with $-35°, 0°, +35°$ profile images. In test cases LS1 and LS2, the method of no fusion achieved the highest face recognition of 91.7% and 66.7%, which are better than the face recognition rates of 50.0.% and 41.7% achieved by the proposed 3D face synthesis with $-35°, 0°, +35°$ profile images. The proposed method of 3D face synthesis with $-35°, 0°, +35°$ profile images achieves the highest face recognition rates in three out of six test cases and outperforms the other methods in subject group A.

For a larger subject group B in Table 2, the proposed face recognition method of 2D face synthesis is the best performer, achieving the highest face recognition rates at 22.7%, 27.3%, 31.8%, and 63.6% in four (ES1, ES2, ES3, LS2, respectively) out of six test cases. In test case LS1, the method of no fusion has the highest face recognition rate of 77.3%, while the 2D face synthesis method is the runner-up having the face recognition rate of 72.7%. In test case LS3, the best performer is the proposed 3D face synthesis with $-35°, 0°, +35°$ profile images. For the subject in group B, the best performer is the proposed 2D face synthesis.

Figures 9–12 show the confusion matrices for the most difficult test case ES1 and the simplest test case LS1, to give further details on the face recognition rates in Tables 1 and 2. The video frames for the LS1 test case contain a full frontal face. In contrast, the video frames for the ES1 test case contain an occluded face. Each row of the confusion matrix is the ID number of the actual person, while each column is the ID number of the predicted person by the face recognizer. The red entries in the confusion matrices are misclassifications, while the blue ones are correct classifications. In these test cases, the UFI predicts every person in the test videos to be person number 12, hence the face recognition rate of $\frac{1}{12} = 8.3\%$ for subject group A and $\frac{1}{22} = 4.5\%$ for subject group B throughout both ES1 and LS1, and indeed for all test cases. Another systematic error appears in the face recognition from the 3D face synthesis with $-35°, 0°, +35°$ profile gallery, in Fig. 9(e). There, the predicted person tends to be person ID 1, reducing the face recognition rate of the method in the difficult ES1 test case. The confusion matrices offer the insight into the prediction performance of the face recognition methods.

Figure 13 depicts the face recognition rates of different methods across test cases and subject groups. The largest face recognition rate is at 91.7%, due to the method of no fusion on test case LS1 and subject group A. Except the UFI, all methods tend to have a larger face recognition rates on the LS1–LS3 test cases than on the ES1–ES3. Although face recognition rates are affected by the subject group, their face recognition rates do not change much. For ex-

DOI:10.4186/ej.2023.27.4.29



(a) UFI (b) No Fusion (c) Fusion

(d) Proposed: ±35° (e) Proposed: ±35°, 0° (f) Proposed: 2D

Fig. 12. Confusion matrices are shown for the simplest test case LS1, where most video frames consist of full-frontal faces (Subject group B).

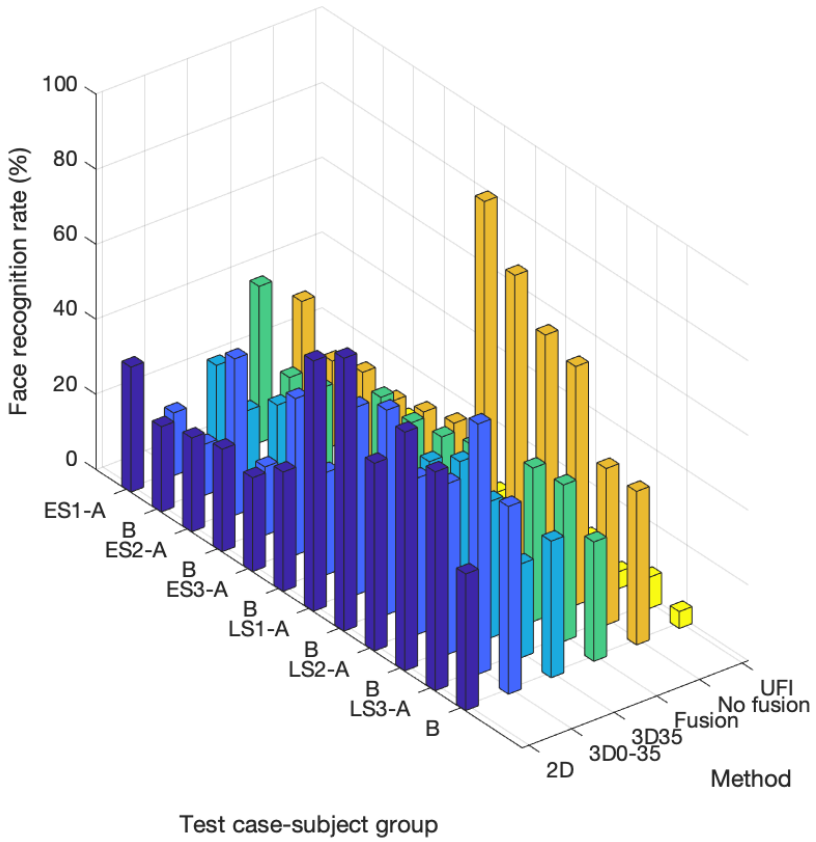


Fig. 13. While the method of no fusion achieves the highest face recognition rate in a certain test case, the 2D method has the most consistent performance.

ample, the method of 2D face synthesis performs relatively well on the test cases for both subject groups A and B, and so is the method of no fusion. The UFI consistently underperforms comparing to the other methods. Given its consistent performance, the method of 2D face synthesis is a recommended method.

The 2D face synthesis has several appealing properties that enhance a face recognition rate and lead to a better recognition performance than other methods in a large subject group B. Both 2D and 3D face syntheses extend the full-frontal-face galleries, and improve the face recognition rates of the UFI, fusion method, and non-fusion method, which do not augment the full-frontal-face gallery. The main difference between the 2D and 3D face syntheses is that the 2D synthesis uses the input video frames in constructing the augmented gallery. The 3D face synthesis uses only the full-frontal face images, together with a neural network that is previously trained with generic images of persons. The videos of the person to be recognized helps the 2D face synthesis capture important facial features of the person. These properties of the 2D face synthesis lead to the face recognition rate that is generally larger than the face recognition rates of other methods observed in a large subject group B.

The proposed face recognition method has several practical applications. Overall, the proposed methods can recognize faces well in input videos with small to moderate facial occlusions. The proposed methods adjust the contrast of the input video frames, making them more robust to different lightning conditions. The proposed method works with a limited gallery, which consists of a full-frontal face for each person and augments the gallery images to improve the face recognition rate. The proposed method is suitable for partial face recognition in uncontrolled environments.

Although the proposed method has higher face recognition rates than the baselines in most test cases, several future studies can be applied to enhance its performance. The first future work is to enhance the input video images to have a higher resolution. Future work will help improve the face recognition rate for input videos that are poor in quality. The second future work will be the exploration of different techniques for face recognition. The techniques in the proposed method evaluate the similarity between the input videos and each gallery image based on the number of matching keypoints. Other promising techniques include face synthesis that uses machine learning, such as the face augmentation generative adversarial network (FA-GAN) [29] and classification with sparse representation-based classification (SRC) [4, 17]. The third future work is performance evaluation on a larger dataset, such as the COX Face Database [30], to add more test cases to the evaluation scenario. Using this work as a basis, several directions can lead to future extensions of this study.

5. Conclusion

The objective of this study was to design a method for face synthesis and partial face recognition from the input of multiple videos. We propose two methods for face synthesis: 2D and 3D. In 2D face synthesis, we partition the input video frames into several groups according to the similarity of the frame images, choose a representative image from each group, and fuse the representative image with each full-frontal face image in the gallery using SIFT flow. In 3D face synthesis, we used a pre-trained neural network to reconstruct the profile images of each person in the gallery. The proposed partial face recognition method measures similarity by counting the number of SIFT keypoints between the video frames and each image in the augmented gallery. The SIFT algorithm is robust to partial and occluded face recognition. The augmented gallery image with the maximum matching keypoints was chosen to be the person who appears in the video frames. Face synthesis enhances the face recognition rate by adding different viewpoints to a full-frontal face image.

To evaluate the performance, we compared the face recognition rate of the proposed method with those of the baselines. The baselines were UFI, fusion, and no-fusion methods. The benchmark dataset was the Chokepoint dataset, which is widely used and provides several test cases. Compared with the baseline methods, the proposed method had the highest face recognition rates in four out of six test cases, in a large subject group (so-called subject group B). The face recognition rates of the proposed method are generally high when the videos contain faces with small yaw angles of approximately $\pm 20^{\circ}$. The 2D face synthesis method has a better face recognition rate than the 3D method in a large subject group, perhaps because it constructs the augmented gallery using the facial features. Overall, the proposed face recognition method outperformed baseline methods.

The proposed face synthesis and partial face recognition methods are useful in several situations. When the gallery contains only full-frontal face images, the proposed face synthesis method does not require knowledge of the rolls, pitches, or yaws of the input video images, making it easy to use. The proposed partial face recognition method works seamlessly with an augmented gallery, potentially increasing the face recognition rate. The proposed method of face synthesis and partial face recognition has practical utility and is suitable for recognizing partial faces in uncontrolled environments.

Acknowledgment

This research was supported in part by the Research Strengthening Project of the Faculty of Engineering, King Mongkut's University of Technology Thonburi.

References

[1] F. Scarpina, "Detection and recognition of fearful facial expressions during the coronavirus disease (COVID-19) pandemic in an Italian sample: An online experiment," *Frontiers in Psychology*, vol. 11, pp. 1–10, Sep. 2020, Article 2252.

[2] Z. Xu, H. R. Wu, X. Yu, K. Horadam, and B. Qiu, "Robust shape-feature-vector-based face recognition system," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 12, pp. 3781–3791, Dec. 2011.

[3] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 980–993, Mar. 2015.

[4] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[5] H. T. Ho and R. Chellappa, "Pose-invariant face recognition using Markov random fields," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1573–1584, Apr. 2013.

[6] Y. Zhu, Y. Li, G. Mu, S. Shan, and G. Guo, "Still-to-video face matching using multiple geodesic flows," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2866–2875, Dec. 2016.

[7] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[8] N. Hao, H. Liao, Y. Qiu, and J. Yang, "Face super-resolution reconstruction and recognition using non-local similarity dictionary learning based algorithm," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 213–224, Apr. 2016.

[9] L. Teijeiro-Mosquera and J. L. Alba-Castro, "Performance of active appearance model-based pose-robust face recognition," *IET Computer Vision*, vol. 5, no. 6, pp. 348–357, Nov. 2011.

[10] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[11] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[12] J. Lei, M. Wu, C. Zhang, F. Wu, N. Ling, and C. Hou, "Depth-preserving stereo image retargeting based on pixel fusion," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1442–1453, Jul. 2017.

[13] Y. Luo, P. Xue, and Q. Tian, "Scene alignment by SIFT flow for video summarization," in *International Conference on Information, Communications and Signal Processing*, Macau, China, Dec. 2009, pp. 1–5.

[14] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, May 2011.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[16] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 1034–1040, May 2016.

[17] F. Mokhayeri, E. Granger, and G.-A. Bilodeau, "Domain-specific face synthesis for video face recognition from a single sample per person," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 757–772, Mar. 2019.

[18] M. Abdelmaksoud, E. Nabil, I. Farag, and H. A. Hameed, "A novel neural network method for face recognition with a single sample per person," *IEEE Access*, vol. 8, pp. 102 212–102 221, Jun. 2020.

[19] M. Cadoni, A. Lagorio, and E. Grosso, "Augmenting SIFT with 3D joint differential invariants for multi-modal, hybrid face recognition," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, USA, Sep. 2013, pp. 1–6.

[20] C.-T. Tu, Y.-H. Chan, and Y.-C. Chen, "Facial sketch synthesis using 2D direct combined model-based face-specific Markov network," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3546–3561, Aug. 2016.

[21] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Maximal likelihood correspondence estimation for face recognition across pose," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4587–4600, Oct. 2014.

[22] X. Yuan and I. K. Park, "Face de-occlusion using 3D morphable model and generative adversarial network," in *IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, Oct. 2019, pp. 10 061–10 070.

[23] K. Dang and S. Sharma, "Review and comparison of face detection algorithms," in *International Conference on Cloud Computing, Data Science Engineering - Confluence*, Noida, India, Jan. 2017, pp. 629–633.

[24] H. Ibrahim and N. S. Pik Kong, "Image sharpening using sub-regions histogram equalization," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 891–895, May 2009.

[25] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation

from single-view images," in *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, Jun. 2020.

[26] W. Nualtim, W. Suwansantisuk, and P. Kumhom, "Face recognition based on multiple video cameras," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, Phuket, Thailand, Jun. 2020, pp. 324–330.

[27] L. An, B. Bhanu, and S. Yang, "Face recognition in multi-camera surveillance videos," in *International Conference on Pattern Recognition*, Tsukuba, Japan, Nov. 2012, pp. 2885–2888.

[28] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Colorado Springs, CO, USA, Jun. 2011, pp. 74–81.

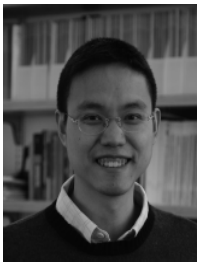
[29] M. Luo, J. Cao, X. Ma, X. Zhang, and R. He, "FA-GAN: Face augmentation GAN for deformation-invariant face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2341–2355, Jan. 2021.

[30] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on COX face database," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5967–5981, Dec. 2015.

DOI:10.4186/ej.2023.27.4.29

Warinthorn Nualtim received the diploma in electronics from Thai-Austrian Technical College, Chonburi, Thailand (1999); the B.Tec. degree in electronics technology from Rajabhat Rajanagarindra University, Chachoengsao, Thailand (2002); and the M.S. degree in robotics and automation from Institute of Field Robotics (FIBO), King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand (2007).

He is currently pursuing a doctoral degree in electrical and computer engineering at the Department of Computer Engineering, KMUTT. His main research interests are automation systems, robotics, image processing, computer vision, and machine learning.



Watcharapan Suwansantisuk is an Assistant Professor at King Mongkut's University of Technology Thonburi (KMUTT), Thailand. He received B.S. degrees in electrical and computer engineering and in computer science from Carnegie Mellon University, Pennsylvania (2002); and the M.S. and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology (2004 and 2012).

Before joining KMUTT, he spent summers at the University of Bologna in Italy as a visiting research scholar and at Alcatel-Lucent Bells Laboratory, New Jersey, as a research intern. His main research interests are wireless communications, synchronization, and statistical signal processing.

Dr. Suwansantisuk serves on the technical program committees for various international conferences and serves as a symposium co-chair of the IEEE Global Communications Conference (2015). He received the Leonard G. Abraham Prize in the Field of Communications Systems from the IEEE Communications Society (2011), jointly with Professor Marco Chiani and Professor Moe Win; and the Best Paper Award from IEEE RIVF International Conference on Computing and Communication Technologies (2016), jointly with Nasiroh Chedoloh.

Pinit Kumhom is an Assistant Professor at King Mongkut's University of Technology Thonburi (KMUTT), Thailand. He received a B.Eng. degree in electrical engineering from King Mongkut's Institute of Technology Thonburi (1988), a former name of KMUTT, and a Ph.D. degree in electrical and computer engineering from Drexel University, Pennsylvania (2000). His research interests include Internet of Things and their applications, digital system design and implementation, and signal and image processing.