Received: 15 January 2020 | Revised: 5 August 2020 | Accepted: 18 September 2020

DOI: 10.1049/bme2.12006

ORIGINAL RESEARCH PAPER

The Institution of Engineering and Technology WILEY

Hybrid face recognition under adverse conditions using appearance-based and dynamic features of smile expression

Murat Taskiran[1] | **Nihan Kahraman**[1] | **Cigdem Eroglu Erdem**[2]

[1]Department of Electronics and Communication Engineering, Yildiz Technical University, Istanbul, Turkey

[2]Department of Computer Engineering, Marmara University, Istanbul, Turkey

Correspondence

Murat Taskiran, Department of Electronics and Communication Engineering, Yildiz Technical University, Istanbul, Turkey.
Email: mrttskrn@yildiz.edu.tr

Funding information

The Scientific and Technological Research Council of Turkey (TUBITAK), Grant/Award Number: EEAG-116E088

Abstract

Although recent deep-learning-based face recognition methods give remarkable accuracies on large databases, their performance has been shown to degrade under adverse conditions (e.g. severe illumination and contrast variations; blur and noise). Under such conditions, soft-biometric features such as facial dynamics are expected to increase the performance if they are used together with appearance-based features. We propose a novel hybrid face recognition, which uses appearance-based features extracted using deep convolutional networks and statistical facial dynamics features extracted from facial landmark positions during smile expression. We evaluated the performances of three different state-of-the-art pre-trained deep convolutional neural networks (DCNNs) under a variety of severe image distortions with different parameters. The experimental results show that, although the face recognition performance using only DCNN-based features drops significantly under adverse conditions, the utilization of facial dynamics features together with DCNN-based features can compensate for the performance loss and increase the accuracy significantly. We believe the proposed system can be useful when face recognition is performed using videos obtained from systems, which may contain blurry and noisy images with a wide range of illumination variations.

1 | INTRODUCTION

The technological developments enabled us to perform many transactions in our daily lives using electronic devices. Therefore, the security of transactions in electronic environments has become an important problem. Hence, biometric systems have become more important for identity and recognition, which use static physiological or dynamic behavioural characteristics of a person.

Face has been one of the main biometric characteristic, and face recognition has many application areas including security, law enforcement, health, education, marketing, finance, entertainment and human–computer interaction. Face recognition systems have some advantages over other biometric modes since may operate remotely with minimal or no cooperation of the user. Facial biometric systems are mainly based on accessing the identity related information using physiological features obtained from face images or behavioural characteristics obtained from facial movements in a video. Using a facial image or video, not only the identity but also the age [1], gender [2] and race information can be determined. The emotional and mental state of the person can also be inferred [3–7] from changes in facial expressions over time.

Face recognition systems in the literature can be grouped under two main categories as image-based and video-based methods [8,9]. While image-based methods use facial appearance-based features, video-based methods can also exploit behavioural features, which can be considered as a soft biometric feature. Face recognition methods achieved over 90% recognition accuracy on face databases obtained in controlled environments in late 2000s [10–12]. After the introduction of deep learning methods to face recognition systems starting from the early 2010s, face recognition accuracies have exceeded 99% [13–15] on large-scale databases collected in the wild.

Video-based face recognition methods in the literature can be grouped as *set-based* and *sequence-based* methods. In *set-based* methods, frames of a video are treated as set of images, and temporal order is disregarded. An example is

frame-aggregation methods, which try to combine information from video frames effectively [16–19]. In Ref. [20], a deep learning based method is proposed for face recognition from videos, which tries to consider blur and occlusion effects. However, the temporal evolution of the frames is not taken into account. The frames of the video are passed through the network and average pooling is used to obtain the compact video representation. *Sequence-based* methods for face recognition from video can be grouped as temporal methods and spatio-temporal methods. Temporal methods use the facial dynamics information separately from the texture information [21], whereas spatio-temporal methods model the texture and the motion information together [22]. In Ref. [22], a deep neural network is trained using a loss function defined between two video streams to perform face verification from unlabelled videos.

There are comprehensive survey papers summarizing the recent developments on deep face recognition and verification [23–27]. In Ref. [27], an evaluation framework is also presented in order to measure how different aspects of deep-learning-based methods including network architecture, choice of loss function, data augmentation and training influence their performance. The reader is referred to the survey papers for an in-depth summary of face recognition literature.

Although deep-learning methods achieve high performance for both face identification and verification tasks under controlled environments, it was observed that their performance decreased significantly under adverse conditions (e.g. illumination, contrast and noise variations) [28,29], which are also referred to as semantic adversarial attacks [30]. Therefore, it can be expected that the use of soft-biometric features as well as appearance-based features obtained from deep learning networks can increase the performance of face recognition systems under adverse conditions. It is shown that the soft-biometric features obtained from facial dynamics carry information about the identity of the person, which is also supported by psychological studies [31,32]. However, they are not sufficient alone for identification of the individual with high accuracy [33–45]. In a recent work [41], histograms of facial action units detected during spontaneous facial expressions, when the subject was interacting with a quiz-game are shown to carry discriminative information. In Ref. [42], it is explained that individual differences in facial expressions can make a face recognition system more robust to spoofing attacks. The facial expression of pain is also used as a biometric feature [43]. The facial dynamics of smile expression were modelled using long short-term memory networks on the top of appearance features in [46]. In another recent work [44], changes in facial expression were shown to carry identity-related information. We would like to note that none of the approaches above aims to compensate the loss of accuracy under adverse conditions (i.e. severe image distortions or semantic adversarial attacks).

In this work, we propose a hybrid face recognition (HFR) system, which combines static appearance-based features and dynamic behavioural features extracted from facial landmarks of smile videos, in order to increase the recognition performance under adverse conditions. Experimental studies on two video databases have confirmed that the proposed hybrid model is beneficial for face recognition under adverse conditions.

To the best of our knowledge, this is the first work that aims to compensate the accuracy loss of face recognition systems under challenging image distortions using emotional facial dynamics information. In an earlier version of our work [45], we have shown that dynamic features extracted from facial landmarks of smile videos carry identity related information. In this paper, we extend our previous work mainly in three aspects:

- We show that using statistical dynamic features along with appearance-based features improve the robustness of face recognition under adverse conditions (i.e. severe image distortions)
- We make an extensive performance comparison of three recent DCNN based face recognition methods (VGGFace, VGGFace2 and ArcFace) under six different simulated adverse conditions with varying parameters
- We provide experimental results on two different smile databases (UvA-NEMO and FEEDTUM), to demonstrate that statistical facial dynamics features are useful under adverse conditions.

The outline of the paper is as follows. In Section 2, we give the details of the proposed method including the extraction of appearance-based and facial dynamics features from smile videos. In Section 3, the tested adverse conditions are described and experimental results are given on two databases. In Section 4, concluding remarks are given and future directions for research are indicated.

2 | HFR SYSTEM

The block diagram of the proposed HFR system is shown in Figure 1. First, the location of the face is detected at each frame in a given video that contains smile expression. Then, in order to capture the facial dynamics, facial landmarks are detected around the eyes, eyebrows, nose, lip and the chin. If the facial landmarks can be detected successfully, dynamics-based features extracted from the landmarks and appearance-based features obtained from a pre-trained deep convolutional neural network (DCNN) are used together for face recognition to improve the performance under adverse conditions. If the facial landmarks cannot be detected successfully, face recognition is performed using the appearance-based features only.

In the following, we provide detailed information about steps of the proposed face recognition system.

2.1 | Face detection

The first step of the proposed face recognition system is to detect the face locations in all frames of the video containing smile expression. Numerous methods have been proposed in the literature to perform face detection. A widely known is the

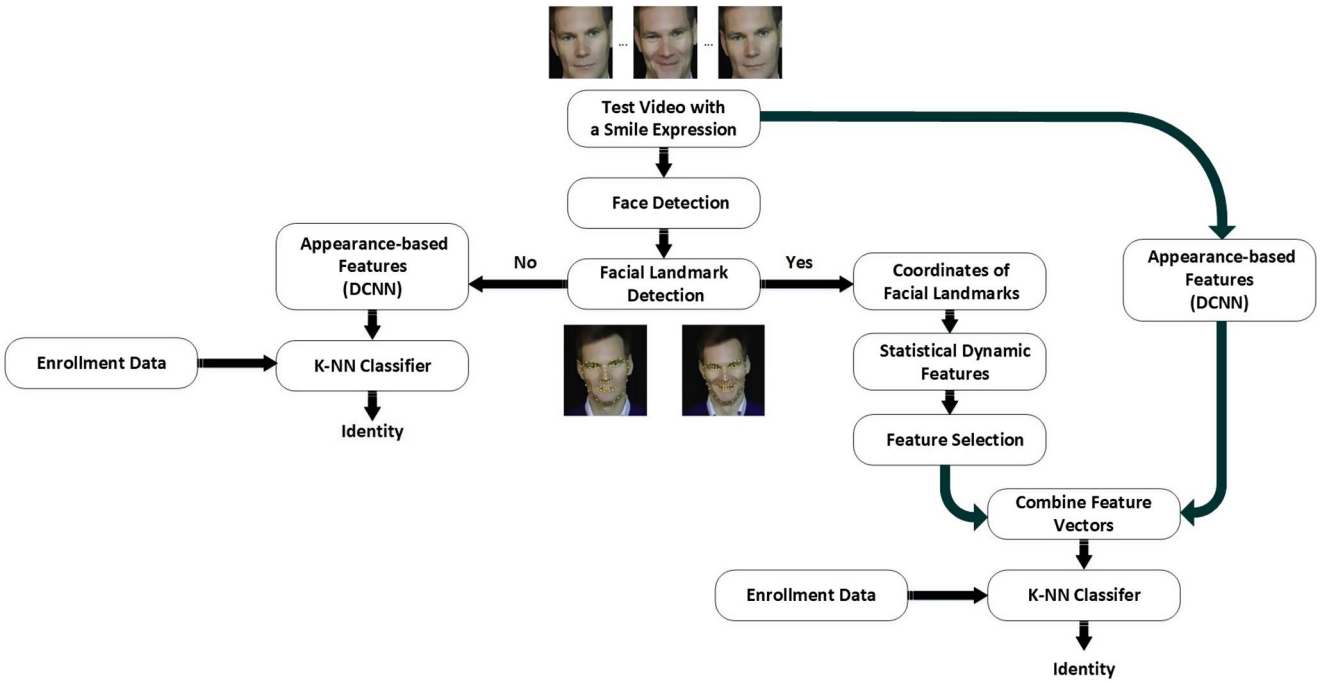

FIGURE 1 The block diagram of the proposed hybrid face recognition system. DCNN, deep convolutional neural network

Viola-Jones (VJ) face detector, which uses Haar cascades [47]. Although the VJ face detector is successful in detecting frontal faces, it may not achieve the same performance for different head poses. In 2005, Dalal and Triggs proposed another face detection method based on Histogram of Oriented Gradient descriptors and Support Vector Machine classifier [48], which achieved more successful results as compared to the VJ face detector.

Recently, the use of deep learning networks has led to more robust face detectors. The Single Shot MultiBox Detector (SSMD) method [49] uses the ResNet-10 network as backbone architecture. Training is done using images collected from the web, that gives successful results for face detection at different angles and distances. Another DCNN-based method used for face detection is the Max-Margin Object Detector (MMOD) [50]. While the training of traditional DCNN architectures requires large-scale databases, training of the MMOD network was performed using around 7000 face images taken from different face databases.

In our work, we used the MMOD face detector since it gives accurate results with various head poses and works much faster than the other tested face recognition methods.

2.2 | Facial landmark detection

Facial landmarks are detected both in order to perform 2-D face alignment and extract statistical dynamic features during smile activity. Recently, many methods have been proposed in the literature for facial landmark detection [51–56]. The tree-structure based model proposed by Zhu and Ramanan in 2012 [55] has achieved good landmark localization results on faces in the wild database. The method has demonstrated similar performance as compared to a commercial software, which was trained with a large number of images. In the method known as CHEHRA [52], a general trained model is updated to be a specific model in order to perform facial landmark detection on face images in uncontrolled environments. In 2013, Xiong et al. [54] proposed a supervised descent method for optimization of a non-linear least squares function in order to solve the face alignment problem. In 2015, the supervised descent method was developed non-locally and successfully implemented to track facial landmarks on the face [51]. Kazemi et al. [53] proposed a method, which is based on gradient boosting for learning an ensemble of regression trees that optimizes the sum of square error loss and manages inherently missing data, for detecting facial landmarks.

In our work, we used the method in [53] for facial landmark detection since it was experimentally found to be more robust under tough conditions, which may be due to the fact that it has tolerance for missing data. In Figure 2 an example image is shown with 68 facial landmarks detected on a face image from the UvA-NEMO smile video database [57].

2.3 | Extraction of appearance-based features

The performance of face recognition systems has increased rapidly by the use of deep neural networks, yielding up to 99% face recognition accuracy on large-scale databases [13–15]. We evaluate three different pre-trained deep neural network architectures (VGG-Face [58], VGG-Face2 [59] and ArcFace [60]) to extract appearance-based feature vectors under adverse conditions, which are briefly explained below:

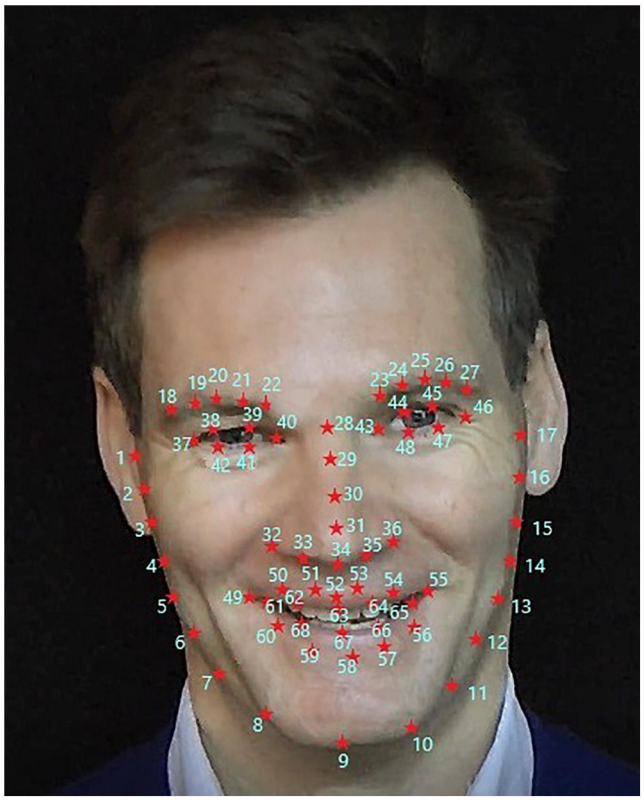

FIGURE 2 An example face image from the UvA-NEMO smile video database showing the 68 facial landmarks detected by Ref. [53]

- **VGG-Face** [58] was proposed by Parkhi et al. in 2015, which was based on VGG-Very-Deep-16 DNN architecture and was trained with 2.6 M face images from approximately 2.6 K subjects. The experimental results on the Labeled Faces in the Wild [61] and Youtube Faces Databases [62] achieved accuracies of 97.27% and 92.8%, respectively. The input image to this network has a size of 224 × 224 × 3 and the size of the extracted feature vector is 4096 × 1, which is obtained from layer fc7, just before the fully connected layers. We used this vector in our experiments
- **VGG-Face2** [59]: In 2018, a large-scale database named as VGG-Face2 was collected by Cao et al. for face recognition containing 3.31 M face images obtained from 9131 subjects with different lighting, and pose variations. ResNet-50 and Squeeze-and-Excitation blocks Resnet-50 architectures [63,64] were trained with the VGG-Face2 database and an increased recognition accuracy was observed. The input images to the network are of size 224 × 224 × 3 and a 2048 × 1 feature vector is obtained from layer fc7
- **ArcFace** [60]: In early 2019, Deng et al. proposed a method based on additive angular margin loss (ArcFace) in order to obtain high-level discriminative features for face recognition. Experimental results using 10 different wide-range face databases have shown that ArcFace achieves very high test performance for face recognition. In this work, we use ArcFace loss function with Inception-ResNet 50 v_1 deep neural network architecture. The training of the network was carried out using MS-Celeb-1M [65] and VGG-Face2 databases. After the training of the network was completed, ArcFace DCNN was used to obtain the 512 dimensional feature vector was obtained from layer fc1

2.4 | Extraction of statistical dynamic facial features

Statistical facial dynamics information extracted from landmark positions in smile videos were used for gender recognition in Ref. [2]. We adopt a similar approach for face recognition using facial dynamics features of smile expression.

First, the location of the face at each frame of the smile video was detected together with 68 facial landmarks locations. Then, 27 facial distances were calculated using these facial landmarks as described in Table 1, which were expected to change during the smile expression. In Table 1, ρ denotes the Euclidean distance between two facial landmarks and l_i denotes i^{th} facial landmark. Next, statistical dynamic features were calculated using these 27 distances after temporal segmentation.

The smile facial expression mainly consists of three temporal segments: onset, apex and offset. The onset refers to the time interval in which the facial expression changes from the neutral state to the expressive state with maximum intensity. Apex refers to the temporal segment during which the intensity of the emotion stays at maximum level. The offset refers to the time interval during which the facial expression changes from the maximum intensity of the expression to the neutral state. These three temporal segments must be estimated before statistical dynamic features are calculated. A low-pass filter of length 5 was used to filter the unwanted temporal variations of mouth length (D_5) before proceeding with the segmentation. The length of the mouth was then normalized with its maximum value and used to determine the onset, apex and offset segments of a smile facial expression. In Figure 3, an example plot of the normalized mouth length versus frame number is shown for a video in the UvA-Nemo smile database [57,66]. In order to detect the onset, apex and offset segments we consider the amplitude and the rate of change of the normalized mouth length. During onset, the mouth length is less than a specified threshold value (0.8) and it increases at a constant rate (ϵ) calculated based on the image resolution. The beginning of apex segment is marked when the normalized mouth length is above the threshold value and the rate of change of mouth length (ϵ) has slowed down. During offset, the mouth length should be below the threshold and decreasing at a constant rate.

After the smile expression is segmented into onset, apex and offset segments, 24 statistical dynamic features are calculated for each of the 27 facial-distances as described in Table 2. The superscripts $(^{+})$, $(^{a})$ and $(^{-})$ represent the onset, apex and offset segments of the smile expression, respectively. The velocity is calculated as $V = \frac{dD}{dt}$ and the acceleration is

TABLE 1 Facial distances used to extract dynamic features. $\rho(.,.)$ denotes the distance, and l_i denotes i^{th} facial landmark

Facial Distance	Description	Description by Facial Landmarks
D_1	Width of right eye	$\rho(\frac{l_{43}+l_{44}}{2}, \frac{l_{46}+l_{47}}{2})$
D_2	Width of left eye	$\rho(\frac{l_{37}+l_{38}}{2}, \frac{l_{40}+l_{41}}{2})$
D_3	Length of right eye	$\rho(l_{43}, l_{46})$
D_4	Length of left eye	$\rho(l_{37}, l_{40})$
D_5	Length of mouth	$\rho(l_{49}, l_{55})$
D_6	Width of mouth	$\rho(l_{52}, l_{58})$
D_7	Centre of mouth to left side upper lip	$\rho(\frac{l_{52}+l_{58}}{2}, l_{51})$
D_8	Centre of mouth to right side upper lip	$\rho(\frac{l_{52}+l_{58}}{2}, l_{53})$
D_9	Centre of mouth to left mouth corner	$\rho(\frac{l_{52}+l_{58}}{2}, l_{49})$
D_{10}	Centre of mouth to right mouth corner	$\rho(\frac{l_{52}+l_{58}}{2}, l_{55})$
D_{11}	Centre of mouth to upper lip	$\rho(\frac{l_{52}+l_{58}}{2}, l_{63})$
D_{12}	Centre of mouth to average distance of two mouth corners	$\rho(\frac{l_{52}+l_{58}}{2}, \frac{l_{49}+l_{55}}{2})$
D_{13}	Left side of right eyebrow to nose	$\rho(l_{23}, l_{31})$
D_{14}	Right side of left eyebrow to nose	$\rho(l_{22}, l_{31})$
D_{15}	Centre of right eyebrow to right side of right eyebrow	$\rho(l_{25}, l_{27})$
D_{16}	Centre of right eyebrow to left side of right eyebrow	$\rho(l_{23}, l_{25})$
D_{17}	Centre of left eyebrow to right side of left eyebrow	$\rho(l_{20}, l_{22})$
D_{18}	Centre of left eyebrow to left side of left eyebrow	$\rho(l_{18}, l_{20})$
D_{19}	Distance between eyebrows	$\rho(l_{22}, l_{23})$
D_{20}	Left corner of left eye to left mouth corner	$\rho(l_{37}, l_{49})$
D_{21}	Right corner of left eye to centre of mouth	$\rho(l_{40}, \frac{l_{63}+l_{67}}{2})$
D_{22}	Left corner of right eye to centre of mouth	$\rho(l_{43}, \frac{l_{63}+l_{67}}{2})$
D_{23}	Right corner of right eye to right mouth corner	$\rho(l_{46}, l_{55})$
D_{24}	Upper side of left eye to right corner of left eyebrow	$\rho(\frac{l_{38}+l_{39}}{2}, l_{22})$
D_{25}	Upper side of right eye to left corner of right eyebrow	$\rho(\frac{l_{44}+l_{45}}{2}, l_{23})$
D_{26}	Upper side of left eye to left corner of left eyebrow	$\rho(\frac{l_{38}+l_{39}}{2}, l_{18})$
D_{27}	Upper side of right eye to right corner of right eyebrow	$\rho(\frac{l_{44}+l_{45}}{2}, l_{27})$

calculated as $A = \frac{d^2D}{dt^2}$. The parameter η represents the number of frames and ω represents the frame rate of the video sequence. After calculating the 24 statistical dynamic features for each of the 27 facial-distances, a 648-dimensional feature vector was obtained for each smile video. Next, a feature selection algorithm is used in order to detect the features, which carry identity-related information of the person, the details of which are explained next.

2.5 | Feature selection and classification

We used the Extremely Randomized Trees Classifier (Extra Trees Classifier) [67] for feature selection from the 648 dimensional statistical facial dynamics vector. The Extremely Randomized Trees Classifier is a collective learning technique that uses the results of multiple correlative decision trees collected in a 'forest' to achieve a classification result. Each Decision Tree in the Extra Trees Forest uses a subset of original training examples.

At each node in the tree, a mathematical criterion (such as the Gini Index) is used to select the best features for classification from each set of features. This causes many interconnected decision trees to be created. In order to perform feature selection using this forest structure, the normalized total reduction of the mathematical criterion is calculated for each feature. If Gini Index is used as the mathematical criterion, this calculation is called as the Gini

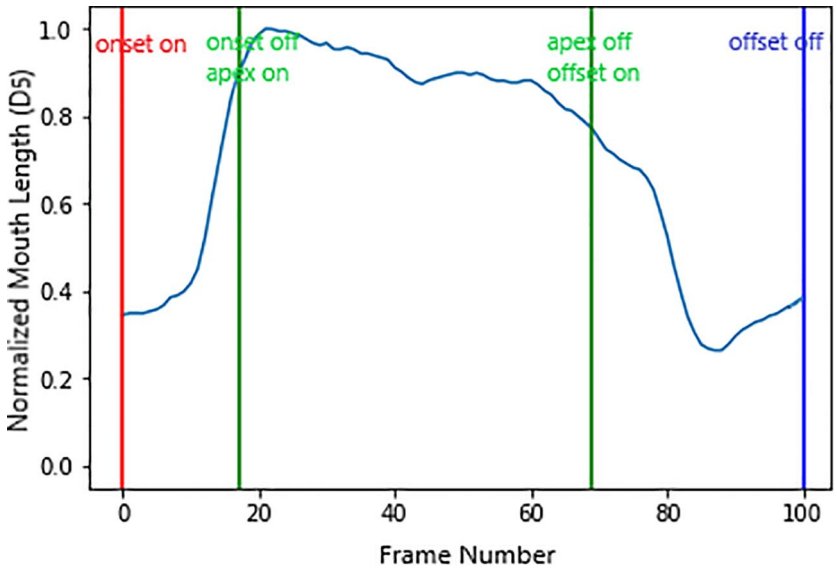

FIGURE 3 The plot of normalized mouth length versus frame number of a smile video. Onset, apex and offset segments of a smile video are marked

TABLE 2 Statistical facial dynamics features

	Definition			
Feature	**General**	**Onset**	**Apex**	**Offset**
Duration		$\eta(\frac{D^+}{\omega})$	$\eta(\frac{D^a}{\omega})$	$\eta(\frac{D^-}{\omega})$
Duration ratio		$\frac{\eta(D^+)}{\eta(D)}$		$\frac{\eta(D^-)}{\eta(D)}$
Maximal amplitude	max(D)			
STD of amplitude	std(D)			
Mean amplitude		mean (D^+)	mean (D^a)	mean (D^-)
Total amplitude		$\sum(D^+)$		$\sum(D^-)$
Net amplitude	$\sum(D^+)$- $\sum(D^-)$			
Amplitude ratio		$\frac{\sum(D^+)}{\sum(D^+)+\sum(D^-)}$		$\frac{\sum(D^-)}{\sum(D^+)+\sum(D^-)}$
Maximal speed		max (V^+)		max (V^-)
Mean speed		mean (V^+)		mean (V^-)
Maximum acceleration speed		max (A^+)		max (A^-)
Mean acceleration speed		mean (A^+)		mean (A^-)
Net Amp., Duration ratio	$\frac{(\sum(D^+)-\sum(D^-))w}{\eta(D)}$			

Importance. Then, the Gini Importance calculated for each feature is sorted in descending order and the top k features are selected. In our work, we used the 128 features with the highest Gini Importance (i.e, larger than 0.002) among the 648 dimensional feature vector to be used in face recognition.

After the feature selection process, we fused the facial dynamics features with the appearance-based features obtained from the DCNN and used a K-Nearest-Neighbor (KNN) classifier for face identification. We tested the K values between 1 and 8 and the value, which achieved the highest test performance, was selected. It was observed experimentally that the best result was obtained for $K = 7$.

3 | EXPERIMENTAL RESULTS

In this section, we first described the used video databases. Then, the evaluation method and the tested distortions (adverse conditions) are described. We compared the three different DCNNs under the tested adverse conditions, and presented the experimental results of the proposed hybrid method on two databases.

3.1 | Video databases

In the literature, there are only a few publicly available databases, which contain multiple smile videos of the same subject.

TABLE 3 Description of the used video databases

Video Databases	Definitions						
	# of videos	# of subjects	Gender dist.	Age range	Resolution	Video length (s)	fps
UvA-NEMO database	1240	400	185 women, 215 men	From 8 to 76	1920 × 1080	Avg. 3.9	50
FEEDTUM database	378	18	9 women, 9 men	From 23 to 38	640 × 480	Avg. 3	25

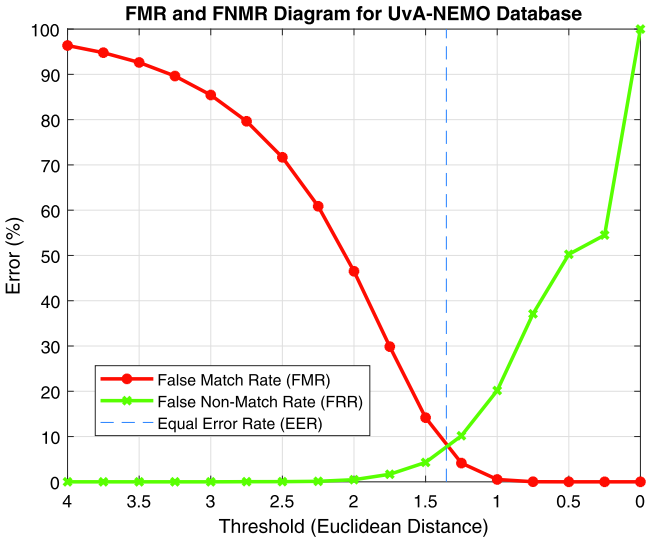

FIGURE 4 False match rate (FMR) and false non-match rate (FNMR) plots for face verification experiments on the UvA-NEMO smile database using only facial dynamics features. The equal error rate is 7.42%

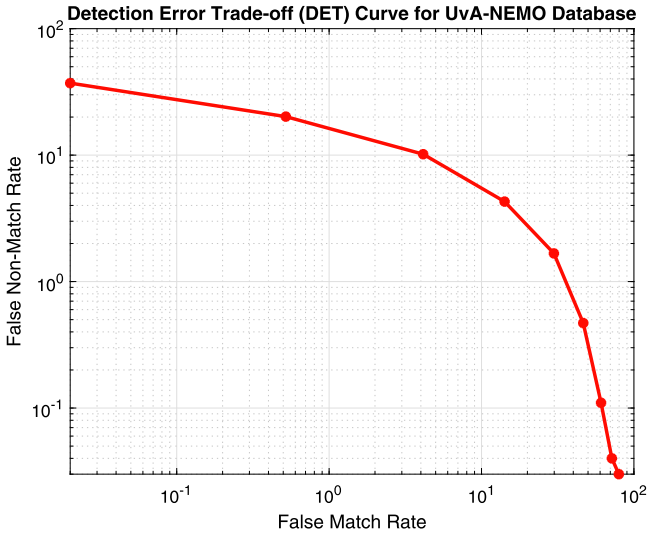

FIGURE 5 Detection error trade-off (DET) curve for face verification experiments on the UvA-NEMO smile database using only facial dynamics with no distortion

Databases containing expressive face videos are usually collected for the purpose of affecting recognition rate, and therefore contain only a single video for each emotion of the same subject. In this work, two different databases, which contain at least two smile videos for each subject were used to test the proposed HFR system.

The first database is the **UvA-NEMO** smile videos database [57,66], which contains 1240 smile videos collected from 400 people (185 women and 215 men), ranging in age from 8 to 76 years. The videos are collected using 1920 × 1080 high definition format and have an average length of 3.9 s. The videos contain either spontaneous or deliberate. We use all smiles videos of a subject regardless of the smile type. All the smile videos start with a neutral expression, reach the apex and return back to the neutral state, hence contains the onset, apex and offset segments.

The second database used in the experiments is the **FEEDTUM** facial expression database, which contains 378 videos from 18 subjects with an age range of 23–38 [68]. Each subject has 21 videos including six different facial expressions and the neutral state. Since we only need the smile expression to test the proposed method, 54 smile videos obtained from 18 subjects were utilized. Description of the used video the databases is summarized in Table 3.

3.2 | Evaluation method and tested distortions

The main objective of this work is to show that the statistical dynamic features obtained from the smile expression have the potential to improve the performance of deep-learning-based face recognition systems, which deteriorates under adverse conditions. Therefore, we first evaluated the face recognition performance of pre-trained deep neural networks under adverse conditions. Then, we investigated the effect of combining appearance-based features obtained from pre-trained DCNNs and statistical dynamic features on face recognition performance under adverse conditions.

We applied various image distortions to the test images, which are described in detail below.

3.2.1 | Gaussian blur

Surveillance cameras may capture blurry videos due to out-of-focus or motion blur. In order to simulate out-of-focus blur, Gaussian low-pass filters with different standard deviations were applied to test images with the goal of investigating the effect of blur degradations on deep-learning-based face recognition methods. The standard deviation (σ) value in this work was ranged from 2.5 to 15.

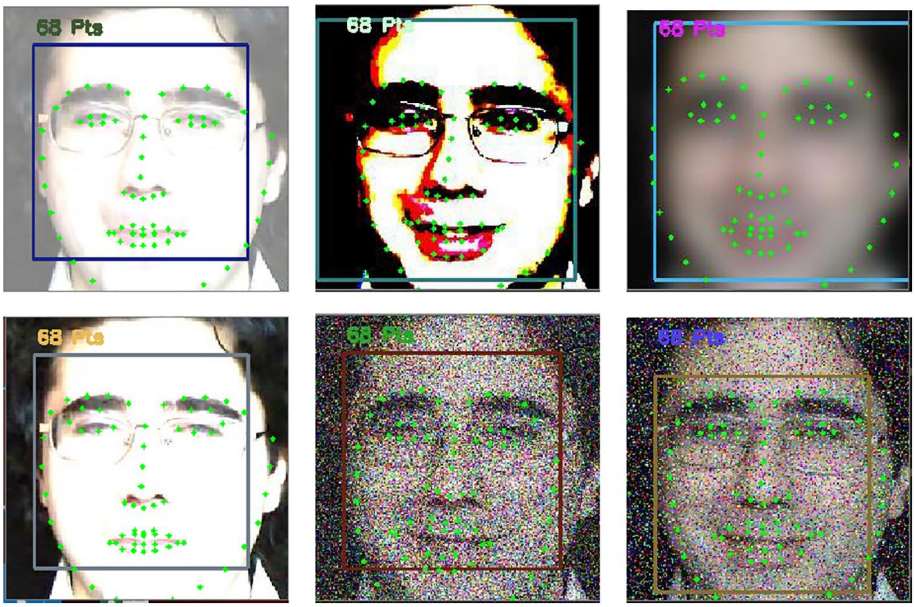

FIGURE 6 Detected facial landmarks obtained on sample face images from the UvA-NEMO database with different image distortions are shown. The parameters are indicated in parenthesis. From top to bottom and left to right: Additive illumination (125), contrast variation (0.49–0.51), Gaussian blur (10), Multiplicative illumination (2.5), Gaussian noise (0.1), salt and pepper noise (0.3) are shown

3.2.2 | Illumination variations

Two different methods were applied to investigate the effects of illumination changes on the face recognition performance. In the first method, pixel values in the images were modified by adding a constant value between 5 and 200. In the second method, the pixel intensities are multiplied by a constant value changing between 0.5 and 5. The modified pixel values were truncated to the range (0–255), if necessary.

3.2.3 | Gaussian noise

Gaussian noise was added to the test images, with zero mean and standard deviation between 0.1 and 0.5.

3.2.4 | Salt and pepper noise

The pixel intensities in the test images were replaced with a value of 0 or 255 with a given probability. The probability value is changed between 0.1 and 0.5.

3.2.5 | Contrast variations

In order to change the contrast of the test images, intensities between a low and high value were linearly mapped to the full range. The widest input range was selected as [0.1–0.9] and the narrowest range was selected as [0.49–0.51].

3.3 | Results on UvA-NEMO database

We performed both face verification and face identification experiments of the UvA-NEMO smile database, which are presented below.

3.3.1 | Face verification results using facial dynamics

In our earlier work [45], it was shown that face verification can be performed using statistical facial dynamics features obtained from smile videos. We improved our previous results by using the feature selection algorithm described above to estimate which dynamic features contribute the most to facial recognition, which reduced the number of dynamic features from 648 to 128. Statistical facial dynamics were extracted from each of the 1215 videos, and a Euclidean distance matrix was calculated containing the Euclidean distance between the feature vectors of each pair of videos. Then, false match rates (FMR) and false non-match rates (FNMR) were calculated using this matrix. All the videos were split into almost equal parts considering the total number of videos for each subject and making sure that the train and test videos for each subject were distinct.

In Figure 4, FMR versus FNMR plot for the UvA-NEMO smile database is given. The equal error rate (EER) is also marked with a dashed line, where FMR is equal to FNMR. The EER without using feature selection was 31.20 %, where it reduces to 7.42% with feature selection, which indicates a significant increase in face verification performance using only facial dynamics information of smile expression. The detection error trade-off (DET) curve is also given in Figure 5.

TABLE 4 Face recognition accuracies (%) on UvA-NEMO smile database using only statistical facial dynamics for various distortions

Image Degradations	Parameters/Accuracy (%)				
No degradation	81.02 ± 0.16				
Gaussian blur	**$\sigma = 2.5$**	**$\sigma = 5$**	**$\sigma = 7.5$**	**$\sigma = 10$**	
	76.20 ± 0.57	52.14 ± 0.36	24.33 ± 0.46	17.65 ± 0.74	
Additive illumination	**5**	**10**	**15**	**20**	**25**
	81.02 ± 0.16	81.02 ± 0.16	81.02 ± 0.16	81.02 ± 0.16	81.02 ± 0.16
Additive illumination	**50**	**70**	**100**	**125**	
	81.02 ± 0.16	81.02 ± 0.16	78.07 ± 0.80	62.84 ± 0.28	
Multiplicative illumination	**1**	**1.5**	**2**	**2.5**	
	81.02 ± 0.16	76.20 ± 0.20	69.79 ± 0.29	42.25 ± 0.25	
Gaussian noise	**$\sigma = 0.01$**	**$\sigma = 0.05$**	**$\sigma = 0.1$**		
	51.24 ± 0.34	4.54 ± 0.98	2.14 ± 0.14		
Contrast variations	**[0.1–0.9]**	**[0.2–0.8]**	**[0.3–0.7]**	**[0.4–0.6]**	**[0.49–0.51]**
	81.02 ± 0.16	81.02 ± 0.16	81.02 ± 0.16	81.02 ± 0.16	78.07 ± 0.27
Salt and pepper noise	**0.1**	**0.2**	**0.3**		
	36.90 ± 0.40	7.22 ± 0.32	2.14 ± 0.04		

TABLE 5 Face recognition accuracies (%) on UvA-NEMO smile database for the three compared DCNNs (VGGFace, VGGFace2 and ArcFace) under various image degradations. HFR stands for 'Hybrid Face Recognition', which is the proposed method. Cases where the proposed method gives higher recognition rates are indicated with bold letters

Image degragations	VGGFace	HFR-VGGFace	VGGFace2	HFR-VGGFace2	ArcFace	HFR-ArcFace
No image degragation	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Gaussian blur ($\sigma = 2.5$)	99.98	99.98	100.00	100.00	99.98	**100.00**
Gaussian blur ($\sigma = 5$)	99.91 ± 0.01	**99.93 ± 0.01**	99.66 ± 0.02	**99.82 ± 0.02**	99.78 ± 0.03	99.78 ± 0.03
Gaussian blur ($\sigma = 7.5$)	73.99 ± 1.52	**79.16 ± 0.38**	89.29 ± 0.61	**92.72 ± 0.09**	85.69 ± 2.38	**91.72 ± 0.20**
Gaussian blur ($\sigma = 10$)	19.48 ± 1.41	**24.13 ± 0.41**	48.74 ± 1.36	**56.07 ± 0.21**	36.79 ± 1.05	**50.90 ± 0.23**
Gaussian blur ($\sigma = 12.5$)	5.10 ± 0.99	5.10 ± 0.99	18.01 ± 0.65	18.01 ± 0.65	14.08 ± 0.42	14.08 ± 0.42
Gaussian blur ($\sigma = 15$)	2.31 ± 0.14	2.31 ± 0.14	5.33 ± 0.33	5.33 ± 0.33	6.83 ± 0.20	6.83 ± 0.20
Additive illumination (5)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (10)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (15)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (20)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (25)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (30)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (50)	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (70)	99.91 ± 0.01	**100.00**	100.00	100.00	99.98	**100.00**
Additive illumination (100)	99.57 ± 0.02	**99.73 ± 0.03**	99.91 ± 0.01	**99.93 ± 0.01**	99.98	**100.00**
Additive illumination (125)	97.21 ± 0.08	**97.85 ± 0.05**	99.55 ± 0.05	**99.80 ± 0.02**	99.91 ± 0.01	**100.00**
Additive illumination (150)	82.40 ± 0.57	82.40 ± 0.57	95.19 ± 0.06	95.19 ± 0.06	98.30 ± 0.05	98.30 ± 0.05
Additive illumination (175)	37.33 ± 3.33	37.33 ± 3.33	63.96 ± 0.16	63.96 ± 0.16	78.04 ± 0.21	78.04 ± 0.21
Additive illumination (200)	5.40 ± 0.40	5.40 ± 0.40	11.86 ± 0.16	11.86 ± 0.16	28.01 ± 0.89	28.01 ± 0.89
Multiplicative illumination (1)	99.98	**100.00**	100.00	100.00	99.98	**100.00**

(Continues)

TABLE 5 (Continued)

Image degragations	VGGFace	HFR-VGGFace	VGGFace2	HFR-VGGFace2	ArcFace	HFR-ArcFace
Multiplicative illumination (1.5)	99.77 ± 0.03	**99.86 ± 0.03**	99.98	99.98	99.98	**100.00**
Multiplicative illumination (2)	97.57 ± 0.03	**98.14 ± 0.04**	99.48 ± 0.04	**99.73 ± 0.03**	99.98	**100.00**
Multiplicative illumination (2.5)	84.99 ± 0.29	**86.67 ± 0.07**	96.17 ± 0.17	**97.30 ± 0.06**	98.44 ± 0.04	**98.62 ± 0.04**
Multiplicative illumination (3)	60.83 ± 2.16	60.83 ± 2.16	82.97 ± 1.39	82.97 ± 1.39	91.11 ± 0.72	91.11 ± 0.72
Multiplicative illumination (4)	25.54 ± 2.28	25.54 ± 2.28	46.36 ± 0.66	46.36 ± 0.66	63.05 ± 1.05	63.05 ± 1.05
Multiplicative illumination (5)	9.82 ± 1.95	9.82 ± 1.95	20.89 ± 0.73	20.89 ± 0.73	39.74 ± 0.47	39.74 ± 0.47
Gaussian noise (var = 0.01)	99.95 ± 0.01	**99.98**	100.00	100.00	99.80 ± 0.02	**99.91 ± 0.01**
Gaussian noise (var = 0.05)	94.15 ± 0.15	**95.35 ± 0.05**	97.75 ± 0.07	**98.50 ± 0.05**	79.52 ± 1.12	**80.52 ± 0.42**
Gaussian noise (var = 0.1)	51.33 ± 1.27	**54.66 ± 1.06**	70.38 ± 0.58	**72.26 ± 0.48**	32.16 ± 0.88	**34.32 ± 0.62**
Gaussian noise (var = 0.15)	19.35 ± 3.07	19.35 ± 3.07	35.90 ± 2.68	35.90 ± 2.68	13.81 ± 0.58	13.81 ± 0.58
Gaussian noise (var = 0.2)	6.87 ± 1.76	6.87 ± 1.76	15.17 ± 2.82	15.17 ± 2.82	7.03 ± 0.35	7.03 ± 0.35
Gaussian noise (var = 0.25)	3.31 ± 0.27	3.31 ± 0.27	5.35 ± 1.59	5.35 ± 1.59	4.40 ± 0.10	4.40 ± 0.10
Contrast variation [0.1–0.9]	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Contrast variation [0.2–0.8]	99.98	**100.00**	100.00	100.00	99.98	**100.00**
Contrast variation [0.3–0.7]	99.61 ± 0.05	**99.73 ± 0.03**	99.98	**100.00**	99.98	**100.00**
Contrast variation [0.4–0.6]	93.78 ± 0.18	**94.92 ± 0.06**	97.60 ± 0.05	**98.16 ± 0.06**	99.68 ± 0.01	**99.73 ± 0.03**
Contrast variation [0.49–0.51]	55.86 ± 1.16	**62.05 ± 0.55**	77.66 ± 0.36	**82.56 ± 0.16**	91.00 ± 0.05	**95.92 ± 0.12**
Salt and pepper (0.1)	99.09 ± 0.06	**99.30 ± 0.04**	99.75 ± 0.03	**99.82 ± 0.02**	97.30 ± 0.30	**98.98 ± 0.04**
Salt and pepper (0.2)	75.05 ± 1.05	**79.36 ± 0.36**	90.54 ± 1.31	**92.72 ± 0.19**	62.33 ± 1.33	**70.65 ± 0.65**
Salt and pepper (0.3)	21.80 ± 1.80	**25.15 ± 0.65**	43.57 ± 0.57	**47.56 ± 0.56**	19.39 ± 1.39	**23.75 ± 0.75**
Salt and pepper (0.4)	3.08 ± 0.21	3.08 ± 0.21	6.96 ± 0.36	6.96 ± 0.36	5.42 ± 0.12	5.42 ± 0.12
Salt and pepper (0.5)	0.60 ± 0.06	0.60 ± 0.06	0.34 ± 0.04	0.34 ± 0.04	2.52 ± 0.07	2.52 ± 0.07

3.3.2 | Face identification and verification results using appearance and dynamics features

In order to investigate the face recognition performance on the UvA-NEMO database, an image database was formed by extracting 12 frames from each video at regular intervals regardless of the length of the video. Subjects with only one video were excluded, which resulted in 1215 videos of 370 subjects. Hence, a database containing 14,580 face images from 1215 videos was obtained. The training set consists of 9864 of these images and 4716 images were used for testing. This split was done considering the total number of videos for each subject and making sure that the train and test videos for each subject were distinct.

In the first part of the experiments, various distortions were applied to the test images obtained from UvA-NEMO smile video database and the face recognition performance of pre-trained deep neural networks under adverse conditions was investigated using these images.

In the second part of the experiments, the effect of combining the feature vectors obtained from pre-trained DCNNs and the statistical dynamic features obtained from smile videos was investigated. The facial landmarks could be detected for most of the tested distortions. In Figure 6, the images with the highest distortions for which the landmarks could be detected are shown for each of the six types of distortion categories. We can see from Figure 6 that, although the textures have deteriorated, the detection of landmarks enable us to make use of the dynamics features under severe distortions.

We also tested the face recognition performance of statistical facial dynamics under adverse conditions, which is given in Table 4. The table shows the face identification accuracy for each distortion with varying parameters. We can see that the face verification accuracy is 81.02% when there is no image degradation. The accuracy is especially stable for additive illumination and contrast variations.

The face identification results with both appearance-based and facial dynamics features for the UvA-NEMO smile database using 2-fold cross validation are shown in Table 5. Since subjects with at least two videos were used in the experiments, the k-value was chosen as 2 for k-fold cross validation. In Table 5, the first column indicates the applied distortions to the test images, where the distortion parameters are given in

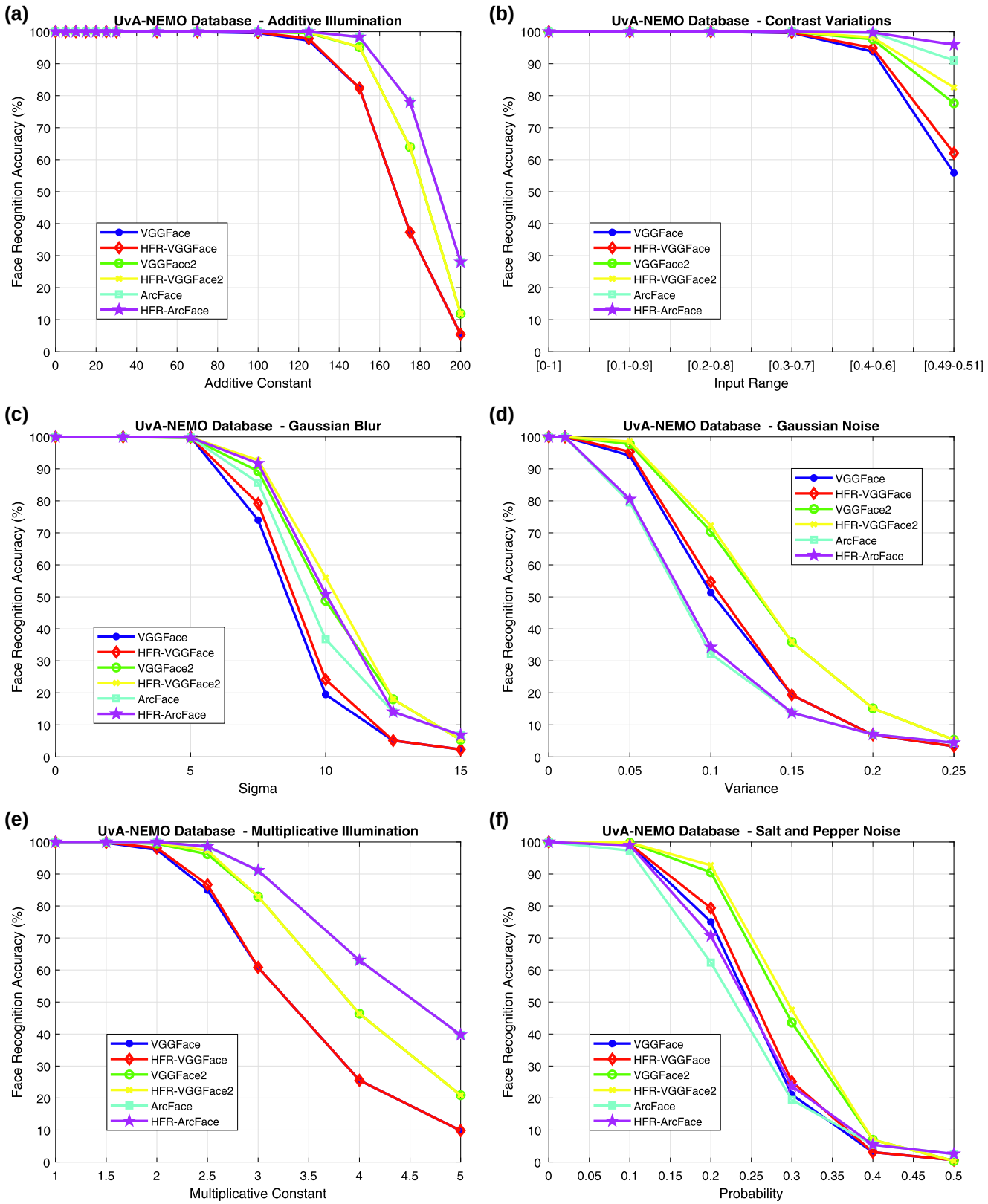

FIGURE 7 Plots for face recognition accuracies (%) on UvA-NEMO smile database for the three compared DCNNs (VGGFace, VGGFace2 and ArcFace) under various image degradations. HFR stands for 'Hybrid Face Recognition', which is the proposed method. Proposed method with ArcFace (HFR-ArcFace) gives the highest accuracies for (a) additive illumination, (b) contrast variations and (e) multiplicative illumination variations. Proposed method with VGGFace2 (HFR-VGGFace2) gives the highest accuracies for (c) Gaussian blur, (d) Gaussian noise and (f) salt-and-pepper noise degradations

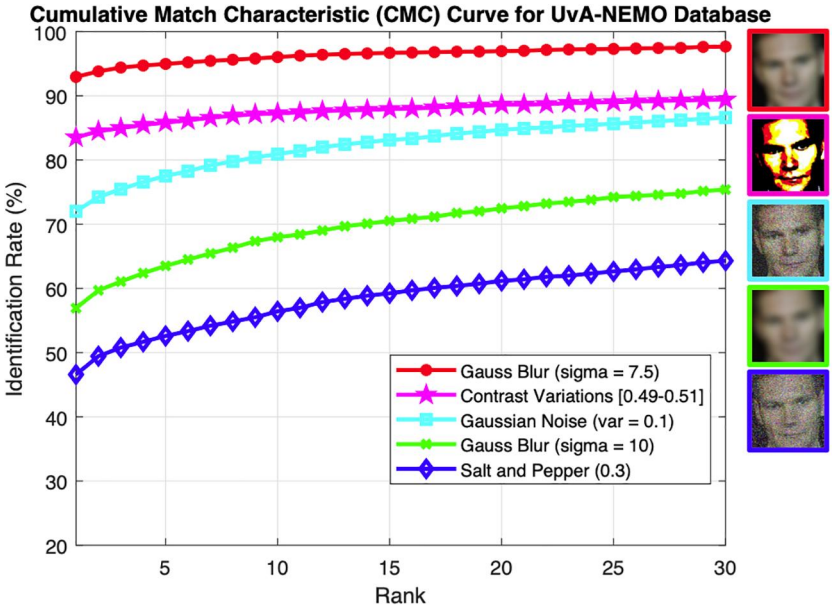

FIGURE 8 Cumulative match characteristic (CMC) curve for HFR-VGGFace2 using UvA-NEMO smile database under five different degradations. Example images are also shown

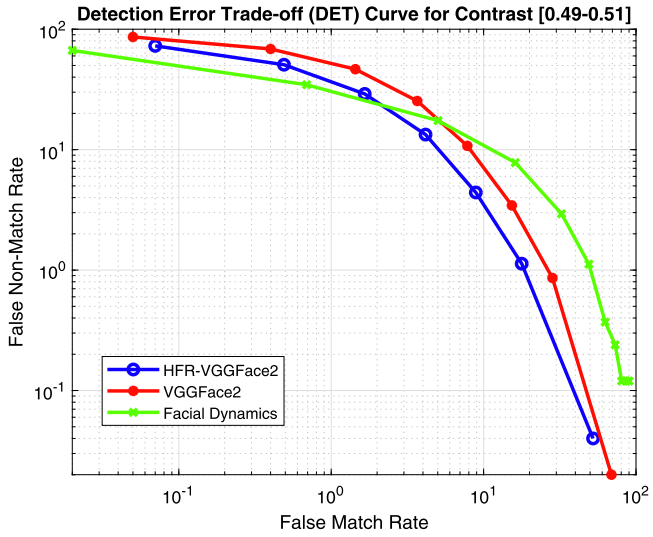

FIGURE 9 Detection error trade-off (DET) curve for contrast variations with range (0.49–0.51) on the UvA-NEMO smile database using HFR-VGGFace2, VGGFace2 and facial dynamics

parenthesis. The accuracies for the three DCNNs are given in columns 2, 4, and 6, which are labeled as VGGFace, VGGFace2 and ArcFace. We can see that the face recognition accuracies drop significantly, when the distortions become severe. In columns 3, 5, and 7, the accuracies for the proposed HFR method are given. The cases for which the facial dynamics features improve the accuracies are indicated in bold. For example, for Gaussian blur ($\sigma = 7.5$), the accuracy of VGGFace2 is 89.29%, but the proposed hybrid method (HFR-VGGface2) can achieve an accuracy of 92.72%. Another example is for contrast variation (0.49–0.51), the accuracy of VGGFace2 decreases to 77.66%, but the proposed method (HFR-VGGFace2) can sustain an accuracy of 82.56%. For the cases indicated in bold, the proposed hybrid method provides an increase in accuracy varying between (0.02%–4.90%).

In Figure 7, the experimental results for UvA-NEMO database are plotted for ease of comparison. We can see that for additive illumination, contrast variations and multiplicative illumination variations, proposed method with ArcFace (HFR-ArcFace) gives the best results. On the other hand, for Gaussian blur, Gaussian noise and salt-and-pepper noise degradations, proposed method with VGGFace2 (HFR-VGGFace2) gives the best accuracies. Since the proposed hybrid facial recognition system is a closed-set identification system, the cumulative match characteristic curves (CMC) of HFR-VGGFace2 are given Figure 8 to evaluate the proposed method under the selected distortions. The CMC curves show rank-k identification rates versus k, where k varies between 1 and 30, and they are given for five different adverse conditions under which the proposed HFR system shows the highest improvements in the face recognition accuracy as compared to using DCNN methods only.

The DET curves obtained during the face verification experiments following a similar protocol described in Section 3.3.1 are given in Figure 9. In these curves, image distortions using contrast variations with the range (0.49–0.51) were used, for which the hybrid facial recognition system provided the most improvement. We can see that the DET curve of the proposed method (HFR-VGGface2) is below the others, which shows that HFR-VGGface2 always provides lower FMR as compared to using appearance-based features alone (VGGFace2).

3.4 | Results on FEEDTUM database

FEEDTUM is a relatively small database containing videos of 18 subjects. We extracted uniformly distributed 12 frames from each smile video (regardless of the length) to form an image database containing 648 images was from 54 videos. The

TABLE 6 Face recognition accuracies (%) on the FEEDTUM database for the three compared DCNNs (VGGFace, VGGFace2 and ArcFace) under various image degradations. HFR stands for 'Hybrid Face Recognition', which is the proposed method. Cases where the proposed method gives higher recognition rates are indicated with bold letters

Image degradation/Methods	VGGFace	HFR-VGGFace	VGGFace2	HFR-VGGFace2	ArcFace	HFR-ArcFace
No image degradation	100.00	100.00	100.00	100.00	100.00	100.00
Gaussian blur ($\sigma = 2.5$)	100.00	100.00	100.00	100.00	100.00	100.00
Gaussian blur ($\sigma = 5$)	100.00	100.00	100.00	100.00	100.00	100.00
Gaussian blur ($\sigma = 7.5$)	100.00	100.00	100.00	100.00	100.00	100.00
Gaussian blur ($\sigma = 10$)	88.42 ± 0.32	**89.81 ± 0.19**	100.00	100.00	83.33 ± 0.33	**84.26 ± 0.26**
Gaussian blur ($\sigma = 12.5$)	63.43 ± 2.23	63.43 ± 2.23	74.54 ± 0.64	74.54 ± 0.64	60.65 ± 1.35	60.65 ± 1.35
Gaussian blur ($\sigma = 15$)	47.22 ± 1.42	47.22 ± 1.42	53.24 ± 0.82	53.24 ± 0.82	42.59 ± 1.25	42.59 ± 1.25
Additive illumination (5)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (10)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (15)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (20)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (25)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (30)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (50)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (70)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (100)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (125)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (150)	100.00	100.00	100.00	100.00	100.00	100.00
Additive illumination (175)	95.37 ± 0.60	95.37 ± 0.60	100.00	100.00	100.00	100.00
Additive illumination (200)	13.90 ± 1.80	13.90 ± 1.80	38.43 ± 0.73	38.43 ± 0.73	51.39 ± 1.39	51.39 ± 1.39
Multiplicative illumination (1)	100.00	100.00	100.00	100.00	100.00	100.00
Multiplicative illumination (1.5)	100.00	100.00	100.00	100.00	100.00	100.00
Multiplicative illumination (2)	100.00	100.00	100.00	100.00	100.00	100.00
Multiplicative illumination (2.5)	100.00	100.00	100.00	100.00	100.00	100.00
Multiplicative illumination (3)	100.00	100.00	100.00	100.00	100.00	100.00
Multiplicative illumination (4)	85.19 ± 1.19	85.19 ± 1.19	96.76 ± 0.46	96.76 ± 0.46	99.54 ± 0.24	99.54 ± 0.24
Multiplicative illumination (5)	38.89 ± 2.39	38.89 ± 2.39	43.98 ± 1.44	43.98 ± 1.44	65.74 ± 1.24	65.74 ± 1.24
Gaussian noise (var = 0.01)	100.00	100.00	100.00	100.00	100.00	100.00
Gaussian noise (var = 0.05)	96.76 ± 0.32	96.76 ± 0.32	98.15 ± 0.18	98.15 ± 0.18	87.50 ± 0.32	85.65 ± 0.25
Gaussian noise (var = 0.1)	70.84 ± 0.82	**71.76 ± 0.46**	65.28 ± 0.38	**65.74 ± 0.32**	54.63 ± 0.47	48.61 ± 0.29
Gaussian noise (var = 0.15)	39.81 ± 1.21	39.81 ± 1.21	33.80 ± 1.60	33.80 ± 1.60	34.26 ± 1.42	34.26 ± 1.42
Gaussian noise (var = 0.2)	24.15 ± 2.52	24.15 ± 2.52	18.52 ± 0.90	18.52 ± 0.90	23.15 ± 2.60	23.15 ± 2.60
Gaussian noise (var = 0.25)	12.96 ± 1.42	12.96 ± 1.42	15.28 ± 0.64	15.28 ± 0.64	15.74 ± 0.76	15.74 ± 0.76
Contrast variation [0.1–0.9]	100.00	100.00	100.00	100.00	100.00	100.00
Contrast variation [0.2–0.8]	100.00	100.00	100.00	100.00	100.00	100.00
Contrast variation [0.3–0.7]	100.00	100.00	100.00	100.00	100.00	100.00
Contrast variation [0.4–0.6]	100.00	100.00	100.00	100.00	100.00	100.00
Contrast variation [0.49–0.51]	70.37 ± 0.42	**72.22 ± 0.32**	74.07 ± 0.47	**77.31 ± 0.46**	82.41 ± 0.56	**83.80 ± 0.30**
Salt and pepper (0.1)	100.00	100.00	100.00	100.00	98.61 ± 0.09	96.76 ± 0.23
Salt and pepper (0.2)	84.72 ± 0.44	**85.19 ± 0.21**	86.57 ± 0.47	**87.96 ± 0.34**	80.56 ± 0.74	72.22 ± 1.22
Salt and pepper (0.3)	45.83 ± 1.21	**45.91 ± 1.26**	41.67 ± 2.40	**43.06 ± 1.70**	41.67 ± 2.34	39.35 ± 1.24
Salt and pepper (0.4)	16.67 ± 2.37	16.67 ± 2.37	14.81 ± 1.10	14.81 ± 1.10	17.13 ± 0.73	17.13 ± 0.73
Salt and pepper (0.5)	9.26 ± 1.66	9.26 ± 1.66	7.87 ± 2.10	7.87 ± 2.10	9.72 ± 1.10	9.72 ± 1.10

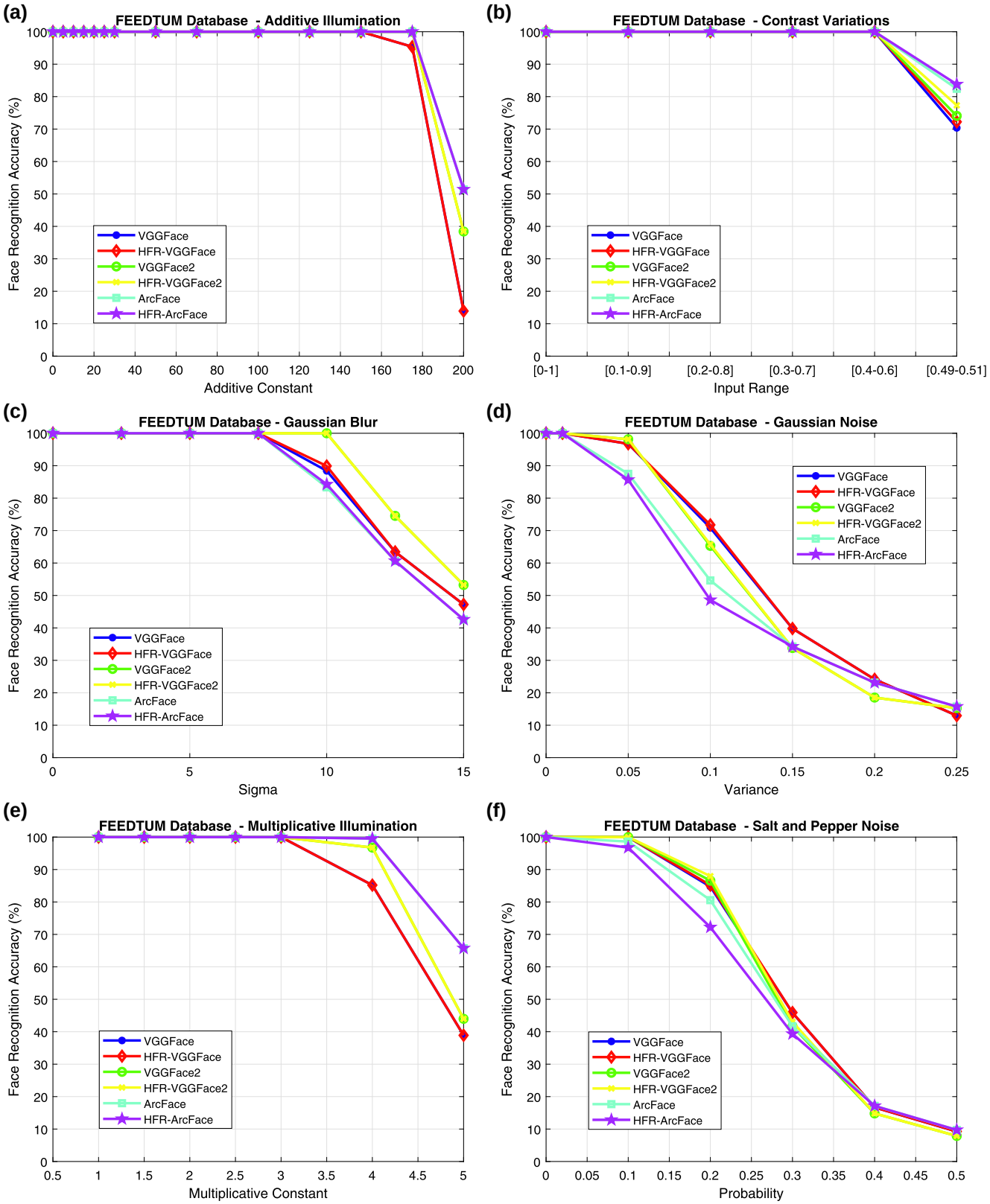

FIGURE 10 Plots for face recognition accuracies (%) on FEEDTUM smile database for the three compared DCNNs (VGGFace, VGGFace2 and ArcFace) under various image degradations. HFR stands for 'Hybrid Face Recognition', which is the proposed method

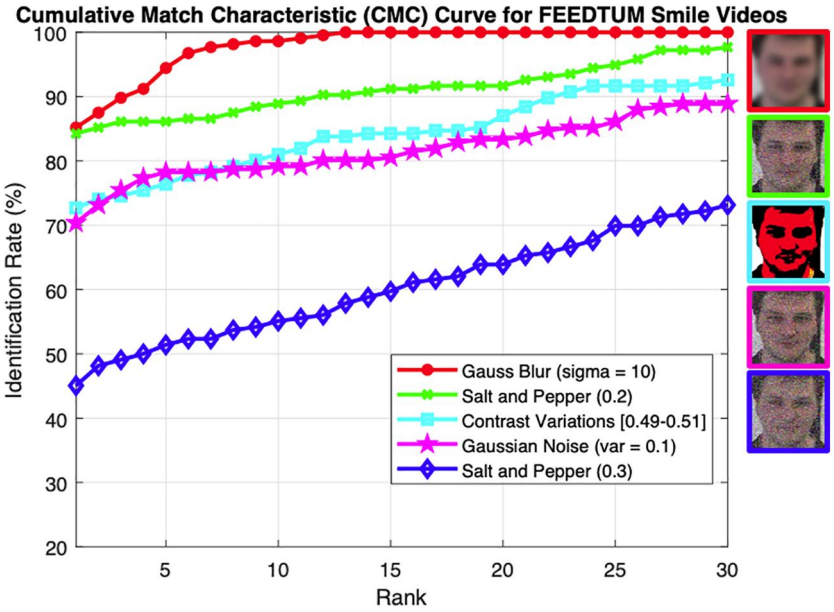

FIGURE 11 Cumulative match characteristic (CMC) curves for the proposed method HFR-VGGFace using FEEDTUM smile database under five different degradations. Example images are also shown

images extracted from one video of each subject were reserved for testing and the images obtained from the other two videos were used for training. As a result, the training database contains 432 images and the test database contains 216 images.

We would like to note that some smile videos in the FEEDTUM database do not contain the offset segment of the smile expression, hence they do not end with a neutral expression. Moreover, some subjects could not express the emotion well. Hence, it is difficult to extract the statistical dynamic features for such videos.

The face recognition results obtained for the smile videos in FEEDTUM database using twofold cross validation are shown in Table 6, where the first column indicates the applied degradations to the test images with the parameters indicated in parenthesis. In columns 3, 5 and 7, the accuracies for the proposed HFR method are given. The cases for which the facial dynamics features improve the accuracies are indicated in bold. We can see that the accuracies may decrease significantly using DCNN features alone, and dynamics-based features can compensate for the loss in some cases. For example, for contrast variations (0.49–0.51), the accuracy for VGGFace2 is 74.07%, whereas the accuracy of the proposed method HFR-VGGFace2 is 77.31%, with an increase of 3.24% in accuracy.

The experimental results are given as plots in Figure 10. We can see that in some of the cases, the proposed method using facial dynamics gives better results as compared to using DCNN features alone. Since FEEDTUM database contains videos, which does not contain the onset-apex-offset phases of the smile expression, the increases in face recognition accuracies are not as clear as in the UvA-NEMO database. Hence, the dynamical features may be as effective, when the offset phase of the smile is missing.

The CMC curves are also shown in Figure 11 for five different degradations under which the proposed face recognition system gives the highest improvement in face recognition accuracies as compared to using DCNN features alone.

4 | CONCLUSIONS AND FUTURE WORK

In this work, we presented a HFR method, which uses appearance-based features extracted using deep DCNNs and statistical facial dynamics features extracted from smile expression. First, we evaluated the performance of three different state-of-the-art pre-trained deep DCNNs under six different image distortion types (different illumination variations, blur and noise) with varying parameters, which showed that their accuracies drop significantly under adverse conditions. The used databases (UvA-NEMO and FEEDTUM) are easy however they contain frontal, high resolution videos therefore the decrease in accuracy was significant under the tested adverse conditions.

The statistical facial dynamics features were extracted using the positions of 68 facial landmarks detected on each frame of the video containing a smile expression consisting of onset-apex-offset temporal phases. The utilization of facial dynamics features was shown to compensate for the performance loss and increased the accuracy significantly under severe image distortions. Hence, the temporal dynamics of smile expression contain identity-related information. We believe the proposed system can be useful when face recognition is performed using videos obtained from systems, which may contain blurry and noisy images and a wide range of illumination variations.

As for future work, the usefulness of statistical dynamic features for face identification obtained from facial expressions of other emotions (anger, surprise, disgust, sadness etc.) can be investigated. To this effect, collection of a new database which contains many repetitions of the facial expression for an emotion could be required for each subject. A more challenging problem could be to extract facial dynamics features when the person is talking and showing an emotional facial expression simultaneously.

ACKNOWLEDGEMENTS

This work was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under project EEAG-116E088.

ORCID

Murat Taskiran https://orcid.org/0000-0002-6436-6963

REFERENCES

1. Dibeklioğlu, H., et al.: Combining facial dynamics with appearance for age estimation. IEEE Trans. Image. Process. 24(6), 1928–1943 (2015)
2. Dantcheva, A., Brémond, F.: Gender estimation based on smile-dynamics. IEEE Trans. Inf. Forensics Secur. 12(3), 719–729 (2016)
3. Erdem, C.E., Turan, C., Aydin, Z.: Baum-2: a multilingual audio-visual affective face database. Multimed. Tool. Appl. 74(18), 7429–7459 (2015)
4. Zhalehpour, S., Akhtar, Z., Erdem, C.E.: Multimodal emotion recognition based on peak frame selection from video. Signal Image Video Process. 10(5), 827–834 (2016)
5. Martinez, B., et al.: Automatic analysis of facial actions: a survey. IEEE Trans. Affect. Comput. 10(3), 325–347 (2017)
6. Zhalehpour, S., et al.: Baum-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. 8(3), 300–313 (2017)
7. Ulukaya, S., Erdem, C.E.: Gaussian mixture model based estimation of the neutral face shape for emotion recognition. Digit. Sig. Process. 32, 11–23 (2014)
8. Barr, J.R., et al.: Face recognition from video: a review. Int. J. Pattern Recogn. Artif. Intell. 26(05), 1266002 (2012)
9. Hassaballah, M., Aly, S.: Face recognition: challenges, achievements and future directions. IET Comput. Vis. 9(4), 614–626 (2015)
10. Wang, X., Tang, X.: Dual-space linear discriminant analysis for face recognition. IN: IEEE CVPR, vol. 2 (2004)
11. Zhao, H., Yuen, P.C., Kwok, J.T.: A novel incremental principal component analysis and its application for face recognition. IEEE Trans. Syst. Man Cy. B. 36(4), 873–886 (2006)
12. Déniz, O., Castrillon, M., Hernández, M.: Face recognition using independent component analysis and support vector machines. Pattern Recogn. Lett. 24(13), 2153–2157 (2003)
13. Taigman, Y., et al.: Closing the gap to human-level performance in face verification. In: IEEE CVPR, pp. 1701–1708 (2014)
14. Lu, J., Wang, G., Zhou, J.: Simultaneous feature and dictionary learning for image set based face recognition. IEEE Trans. Image. Process. 26(8), 4042–4054 (2017)
15. Rao, Y., Lu, J., Zhou, J.: Attention-aware deep reinforcement learning for video face recognition. In: IEEE ICCV, pp. 3931–3940 (2017)
16. Rao, Y., et al.: Learning discriminative aggregation network for video-based face recognition. In: IEEE ICCV (2017)
17. Hörmann, S., et al.: Outlier-robust neural aggregation network for video face identification. In: IEEE ICIP, pp. 1675–1679 (2019)
18. Gong, S., et al.: Video face recognition: component-wise feature aggregation network (c-fan). In: 2019 International Conference on Biometrics (ICB), pp. 1–8 (2019)
19. Liu, X., et al.: Dependency-aware attention control for image set-based face recognition. IEEE Trans. Inf. Forensics Secur. 15, 1501–1512 (2020)
20. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE Trans. PAMI. 40(4), 1002–1014 (2018)
21. Hadid, A., Dugelay, J., Pietikainen, M.: On the use of dynamic features in face biometrics: recent advances and challenges. Signal Image Video Process. 5, 495–506 (2011)
22. Rashedi, E., et al.: Stream loss: convnet learning for face verification using unlabeled videos in the wild. Neurocomputing. 329, 311–319 (2019)
23. Masi, I., et al.: Deep face recognition: a survey. In: SIBGRAPI—Conference on Graphics, Patterns and Images (2018)
24. Ranjan, R., et al.: Deep learning for understanding faces: machines may be just as good, or better, than humans. IEEE Signal Process. Mag. 35(1), 66–83 (2018)
25. Wang, M., Deng, W.: Deep face recognition: a survey. arXiv:180406655 (2019)
26. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. 189, 102805 (2019)
27. You, M., et al.: Systematic evaluation of deep face recognition methods. Neurocomputing. 388, 144–156 (2020)
28. Grm, K., et al.: Strengths and weaknesses of deep learning models for face recognition against image degradations. IET Biom. 7(1), 81–89 (2017)
29. Pala, G., Erdem, C.E.: Performance comparison of deep learning based face identification methods for video under adverse conditions. In: The 15th Int. Conf. On Signal-Image Technology and Internet Based Systems (SITIS) (2019)
30. Mohapatra, J., et al.: Towards verifying robustness of neural networks against semantic perturbations. In: IEEE CVPR (2020)
31. Pilz, K.S., Thornton, I.M., Bulthof, H.H.: A search advantage for faces learned in motion. Exp. Brain. Res. 171(4), 436–447 (2006)
32. Xiao, N.G., et al.: On the facilitative effects of face motion on face recognition and its development. Front. Psychol. 5, 633 (2014)
33. Schmidt, K.L., Cohn, J.F.: Dynamics of facial expression: normative characteristics and individual differences. In: IEEE ICME, pp. 728–731 (2001)
34. Cohn, J.F., et al.: Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification. In: IEEE ICMI, pp. 491–496 (2002)
35. Tsai, P., Hintz, T., Jan, T.: Facial behavior as behavior biometric? an empirical study. In: IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 3917–3922 (2007)
36. Tulyakov, S., et al.: Facial expression biometrics using tracker displacement features. In: IEEE CVPR (2007)
37. Tsai, P., et al.: An evaluation of bi-modal facial appearance+facial expression face biometrics. In: ICPR (2008)
38. Ning, Y., Sim, T.: You're on identity camera. In: ICPR (2008)
39. Zafeiriou, S., Pantic, M.: Facial behaviometrics: the case of facial deformation in spontaneous smile/laughter. IEEE CVPR Workshops, pp. 13–19 (2011)
40. Tubbs, D.J., Rahman, K.A.: Facial expression analysis as a means for additional biometric security in recognition systems. Int. Conf. MCSS. 113–123 (2015)
41. Shreve, M., et al.: A study on the discriminability of facs from spontaneous facial expressions. IEEE ICIP, pp. 1674–1678 (2016)
42. Gavrilescu, M.: Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks. IET Biom. 5(3), 236–242 (2016)
43. Haque, M.A., Nasrollahi, K., Moeslund, T.B.: Pain expression as a biometric: why patients' self-reported pain doesn't match with the objectively measured pain?. In: IEEE Int. Conf. on Identity, Security and Behavior Analysis (ISBA) (2017)
44. Haamer, R.E., et al.: Changes in facial expression as biometric: a database and benchmarks of identification. In: IEEE Int. Conf. Automatic Face and Gesture Recognition (FG), pp. 621–628 (2018)
45. Taskiran, M., et al.: Face recognition using dynamic features extracted from smile videos. In: IEEE INISTA (2019)
46. Kim, S.T., Kim, D.H., Ro, Y.M.: Facial dynamic modelling using long short-term memory network: analysis and application to face authentication. In: IEEE Int. Conf. Biometrics Theory Applied Systems (2016)
47. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. 57(2), 137–154 (2004)
48. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE CVPR, pp. 886–893 (2005)
49. Liu, W., et al.: Ssd: single shot multibox detector. In: ECCV, pp. 21–37 (2016)
50. King, D.E.: Max-margin object detection. arXiv:150200046 (2015)
51. Xiong, X., De la Torre, F.: Global supervised descent method. In: IEEE CVPR, pp. 2664–2673 (2015)

52. Asthana, A., et al.: Incremental face alignment in the wild. In: IEEE CVPR, pp. 1859–1866 (2014)
53. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: IEEE CVPR, pp. 1867–1874 (2014)
54. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE CVPR, pp. 532–539 (2013)
55. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE CVPR, pp. 2879–2886 (2012)
56. Sandıkcı, E.N., Erdem, Ç.E., Ulukaya, S.: A comparison of facial landmark detection methods. In: IEEE SIU (2018)
57. Dibeklioğlu, H., Salah, A.A., Gevers, T.: Are you really smiling at me? spontaneous versus posed enjoyment smiles. ECCV, pp. 525–538 (2012)
58. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC (2015)
59. Cao, Q., et al.: A dataset for recognising faces across pose and age. In: IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018)
60. Deng, J., et al.: Additive angular margin loss for deep face recognition. In: IEEE CVPR, pp. 4690–4699 (2019)
61. Huang, G.B., Learned.Miller, E: Labeled faces in the wild: updates and new reporting procedures. Univ. of Massachusetts, Amherst (2014). UM-CS-2014-003
62. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: IEEE CVPR, pp. 529–534 (2011)
63. He, K., et al.: Deep residual learning for image recognition. In: IEEE CVPR, pp. 770–778 (2016)
64. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE CVPR, pp. 7132–7141 (2018)
65. Guo, Y., et al.: Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: ECCV, pp. 87–102 (2016)
66. Dibeklioğlu, H., Salah, A.A., Gevers, T.: Recognition of genuine smiles. IEEE Trans. Multimed. 17(3), 279–294 (2015)
67. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. 63(1), 3–42 (2006)
68. Wallhoff, F., et al.: Efficient recognition of authentic dynamic facial expressions on the feedtum database. In: IEEE Int. Conf. on Multimedia and Expo, pp. 493–496 (2006)

How to cite this article: Taskiran M, Kahraman N, Eroglu Erdem C. Hybrid face recognition under adverse conditions using appearance-based and dynamic features of smile expression. *IET Biom.* 2021;10:99–115. https://doi.org/10.1049/bme2.12006