IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 5, 2021, accepted February 19, 2021, date of publication February 24, 2021, date of current version March 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061744

Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features

NAJMEH SAMADIANI[1], GUANGYAN HUANG[1], (Member, IEEE), YU HU[2], (Member, IEEE), AND XIAOWEI LI[2], (Senior Member, IEEE)

[1]School of Information Technology, Deakin University, Burwood, VIC 3125, Australia
[2]State Key Laboratory of Computer Architecture, Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Guangyan Huang (guangyan.huang@deakin.edu.au)

This work was supported in part by the Australia Research Council (ARC) Discovery Project under Grant DP190100587.

ABSTRACT Facial expressions have been proven to be the most effective way for the brain to recognize human emotions in a variety of contexts. With the exponentially increasing research for emotion detection in recent years, facial expression recognition has become an attractive, hot research topic to identify various basic emotions. Happy emotion is one of such basic emotions with many applications, which is more likely recognized by facial expressions than other emotion measurement instruments (e.g., audio/speech, textual and physiological sensing). Nowadays, most methods have been developed for identifying multiple types of emotions, which aim to achieve the best overall precision for all emotions; it is hard for them to optimize the recognition accuracy for single emotion (e.g., happiness). Only a few methods are designed to recognize single happy emotion captured in the unconstrained videos; however, their limitations lie in that the processing of severe head pose variations has not been considered, and the accuracy is still not satisfied. In this paper, we propose a Happy Emotion Recognition model using the 3D hybrid deep and distance features (HappyER-DDF) method to improve the accuracy by utilizing and extracting two different types of deep visual features. First, we employ a hybrid 3D Inception-ResNet neural network and long-short term memory (LSTM) to extract dynamic spatial-temporal features among sequential frames. Second, we detect facial landmarks' features and calculate the distance between each facial landmark and a reference point on the face (e.g., nose peak) to capture their changes when a person starts to smile (or laugh). We implement the experiments using both feature-level and decision-level fusion techniques on three unconstrained video datasets. The results demonstrate that our HappyER-DDF method is arguably more accurate than several currently available facial expression models.

INDEX TERMS Facial landmarks, facial expression recognition, long short term memory, multi-layer neural networks, happy emotion recognition.

I. INTRODUCTION

Emotion recognition has become a hot, attractive research area, which has a wide range of applications. For example, it can be utilized for an emotional understanding of customers in the advertising industry. Lie detection can be eased by facial expression recognition and physiological states in the crime and court domain [1]. It can be useful in diagnosing some diseases like anxiety and Parkinson's in medical applications [2], [3]. In the web and connected world, we can also employ the emotion recognition systems for specifying the spectators' feelings and moods to recommend the music [4], videos [5], or even products in virtual recommender systems [6], [7].

There are several types of emotion recognition systems based on different cues for detecting human emotion states such as facial expression recognition (FER) [8], speech emotion recognition [9], physiological emotion recognition [10]; also, they can be combined into multimodal systems [11], [12] to detect human emotions. Among them, non-verbal cues like facial expressions play a more critical role in determining emotions than others [13]. The traditional FER systems were developed to recognize only the facial expressions on the lab-controlled datasets with high accuracy of over 97% when

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif.

the participants were asked to pose an emotion. By applying these methods on more complicated, real-world datasets captured from movies or TV series, the accuracy was reduced to a very low accuracy [14]; for example, the best was 40% [8] for the Acted Facial Expressions in the Wild (AFEW) dataset. Many factors, such as various backgrounds, severe head pose variation, illumination changes, and different kinds of occlusions and noises, introduce disparities between emotion recognition systems' accuracy.

Smile or laugh is known as the most common facial expression among human communications during daily life. Similar to FER systems, many applications can also be defined by utilizing a happy emotion recognition module. For instance, investigating a smile as a genuine or posed may lead to anticipate further traits and behaviors [15], [16]. Although some outstanding approaches with a relatively-high average accuracy are developed for recognizing all six basic emotions (happiness, surprise, disgust, anger, sadness and fear), the separated accuracy of single emotion detection is not high enough for real-world applications. In this paper, we aim to practice recognizing only the happy expression from the unconstrained videos for intensively improving the accuracy using focused optimal strategies for a single emotion. With further study in this field, we have found a few robust standalone methods for happy emotion recognition from videos. However, the videos captured in the wild comprise human images with significant variations in head poses; this challenges the existing techniques which were designed to process the ideal faces captured in the laboratories, and they cannot be extended to real-world scenarios. Different view angles for capturing human faces significantly challenge the process for extracting facial features, such as the facial landmarks and textural features, due to incomplete availability of the face. Hence, it is essential to develop methods capable of emotion detection in any circumstances from the unconstrained videos that are naturally captured in the wild.

In this paper, we propose a Happy Emotion Recognition using the 3D hybrid deep and distance features (HappyER-DDF) method, which utilizes visual information by extracting two different types of deep features. First, if considering incomplete human faces captured from arbitrary view angles, only textured features are not enough to distinguish the expressions. And facial landmarks are needed, which has been proven to play a critical role in expression recognition systems [17]. Hence, this paper considers the complementary information from both facial texture and landmarks. Second, borrowing the techniques of deep neural networks used in computer vision, we extract textured and spatial-temporal features from sequential frames in the videos using a hybrid deep neural network containing 3D Inception-ResNet and long short term memory (LSTM). Moreover, a convolutional neural network (CNN) is employed to process the facial landmarks. It captures the detailed facial change from the time series produced by computing the distance between the landmarks of mouth (or eyes) and a reference point on the face (e.g., nose peak). The landmarks-based facial change time series track the muscle movements among sequential frames and can effectively demonstrate the changes while an emotion occurs. After extracting these discriminative features, we fuse them at both feature and decision levels to evaluate the system performance. The feature-level fusion is conducted to form the final feature vector, and a fully connected layer classifies the videos into happy and non-happy classes. The decision-level fusion utilizes an SVM with radial basis function kernel as a classifier to recognize the emotions and a weighted sum based on the genetic algorithm and weighted mean for concatenating the decisions [18].

The FER systems most related to ours recognize all six basic emotions [17], [51]. Our proposed scheme aims to demonstrate the single (e.g., happy) emotion recognition using specific optimisation strategies and can intensively help improve accuracy; for example, selecting action units only relevant to a happy emotion. So, it is different from these studies on the following aspects: (1) we have considered both the 3D version of convolutional neural network and 3D facial landmarks only relevant to happy emotion, (2) the proposed deep neural network architecture is simpler but with an increased accuracy for happy emotion detection. Experimental results on three unconstrained video datasets, i.e., AM-FED+, AFEW and multi-party conversational (MELD), demonstrate the high accuracy of the proposed method. The contribution of the work is summarized as follows:

1) In contrast to the existing methods that mainly focus on using facial expression recognition to identify all six basic emotions, we propose a unique, lighter, fast standalone system for happy emotion detection in the videos.
2) We propose a novel HappyER-DDF method, which builds a deep recognition framework modeled by both spatial and temporal information. We use 3D visual information aligned with landmark trends, apply them to the hierarchical architecture long short-term memory recurrent neural network (LSTM) and thus improve the system performance to achieve more reliable results.
3) We design a recognition system that is able to accurately detect happy and non-happy expressions in a video captured in the wild (where human faces may be not always completely captured from random angles), including different apex frames.
4) We evaluate the trained system using two unconstrained datasets and a large multi-party conversational dataset, and the experimental results demonstrate the accuracy and effectiveness of our proposed HappyER-DDF method.

The rest of this paper is organized as follows: Section II reviews the existing facial expression recognition models. The HappyER-DDF method is detailed in Section III. Section IV reports the experimental results and Section VII concludes the paper.

II. RELATED WORK

A. TRADITIONAL FACIAL EXPRESSION RECOGNITION METHODS

Both images and videos can capture facial expressions. The static facial expression recognition methods focus on extracting features from the images and attempting to classify the input images to six basic emotions: happiness, surprise, disgust, anger, sadness and fear. Some texture processing techniques, for example, scale invariant feature transform (SIFT) [19], local binary pattern (LBP) [20], the pyramid of histograms of oriented gradients (PHOG) and local phase quantization (LPQ) [21] have been utilized to extract the facial textures as appearance-based features. Most of these types of FER systems have achieved high accuracy on both lab-controlled and unconstrained datasets.

By expanding the data from static images to sequential frames in the videos, facial textured features cannot provide enough information to recognize the emotions. As a result, geometric features are employed since we observe some meaningful relations between facial landmarks. Moreover, the temporal features are considered by applying new textured features to the videos. Local binary pattern from three orthogonal planes (LBP-TOP) [22], local Gabor binary patterns from three orthogonal planes (LGBP-TOP) [23] and histogram of oriented gradients from three orthogonal planes (HOG-TOP) [24] are examples of textured features from the videos. Since the audio is also recorded in the videos, many multimodal FER systems use the speech features as the complementary information, when extracting features from visual data is not feasible due to occlusion or noises [14], [25]. These systems obtain a high accuracy on the lab-controlled dataset, but a low accuracy on the dataset captured in the wild.

B. DEEP FACIAL EXPRESSION RECOGNITION METHODS

The FER systems applied to the real-world datasets result in low accuracy due to different backgrounds, illumination variations and other noises. The features extracted by traditional methods are correlated to the context and cannot ignore the light or background changing in the output feature vector. Consequently, some significant pre-processing techniques are applied to the videos for unifying them. On the other hand, a wide range of deep neural networks has an extensive ability to learn new patterns and extract detailed features without any pre-processing. Hence, they have been used in various research areas, such as image classification [26] and face recognition [27]. Among different deep architectures, convolutional neural networks (CNN), recurrent neural networks (RNN) and long short term memory (LSTM), as a special type of RNNs, are demonstrated to perform better in the sequential frames processing.

Many deep FER systems have been developed. A hybrid method using the CNN and DSIFT features and geometric relations has been provided to extract features from videos [28], where the spontaneous datasets are applied to verify the system performance for the real-world applications. Stacked CNN blocks are utilized in [29] to recognize the emotions. The input images are pre-processed to gray-scale form and passed through eight blocks containing convolutional, batch normalization, and dropout layers. Although the results are promising, it could not be applied to unconstrained datasets. Three different CNN frameworks are employed to extract the features in [30]. RNN has also been used with rectified linear hidden units (ReLu) to extract the temporal features. All three presented networks have shown an overfitting problem, whereas combining the features achieves higher accuracy. An attention network is proposed in [31], which extracts both local and global features. The method utilizes some convolutional filters and bilinear attention pooling for detecting the emotions in the images. A deep neural framework containing a spatial transformer network block (STN) is presented in [32] to track two main issues: 1) the input images to CNNs have different sizes, and 2) the CNNs are sensitive to the input image size. After scaling all inputs to a specific size, the model, including several VGG16 networks [33], detects the emotions in the images. Selecting a general scale value operating for all images is a challenge. The Inception layers have been established in [34] within the GoogLeNet network. These types of layers employ several convolution filters and concatenate them to extract facial features on various scales. Several variations of Inception architectures have been proposed and applied to facial expression recognition [35]. Another CNN has been provided combining inception and residual blocks [36], in which a wide and deep proposed framework removes the redundant filters and improves the training speed; but the problem is that using ordinary CNN could not extract the temporal features of consecutive frames. Hence, a 3D version of CNN is used to record the temporal information in addition to spatial features in various applications, such as 3D object detection [37] and video emotion recognition [18], [38].

As we know, the recurrent neural networks (RNNs) can map the temporal dynamic behavior using some internal states (memory) and hidden neurons. Although they have demonstrated a good capability to extract the features, they are not able to learn and memorize the long term sequences. Therefore, the LSTMs are utilized to track their vanishing gradients by adding some more remembering and forgetting units. Several FER systems have applied LSTM networks to a fine-tuned CNN for recognizing the emotions in the videos [18]. A hybrid framework is provided by applying the extracted VGG16 features to an LSTM layer [39]. Their system can capture both spatial and temporal features in various video sequences. Another hybrid network has been proposed, containing both local and global frameworks based on CNN-LSTM cascaded network, in which, the videos with large head pose variation and occlusions have been removed to get the better results [40]. However, this method can just achieve high accuracy on laboratory datasets.

C. HAPPY EMOTION RECOGNITION METHODS

Most happy emotion detection systems are applied to images and very few for videos. They have used different traditional

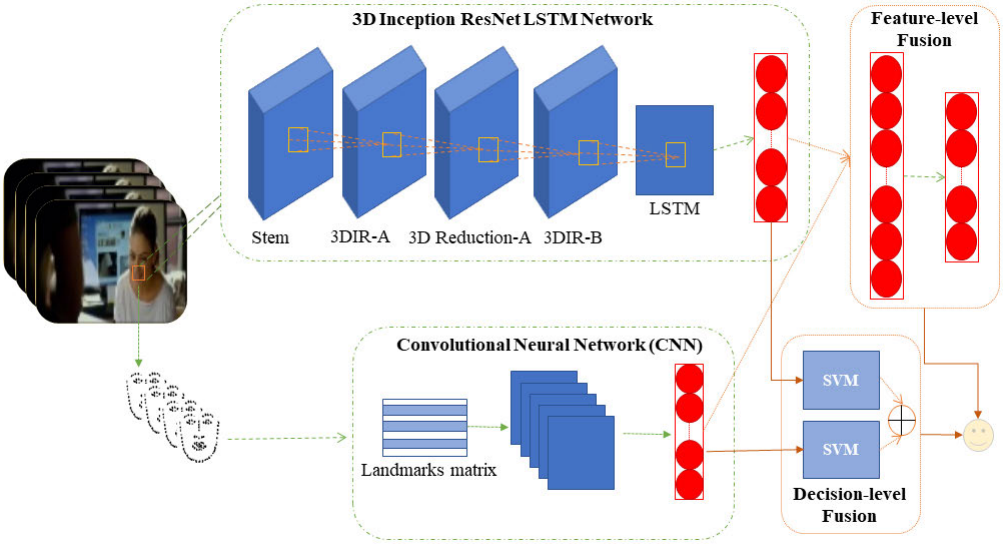


FIGURE 1. An overview of the proposed HappyER-DDF method. It comprises two different deep neural networks for extracting spatial-temporal features from videos and discriminative features from facial landmarks. The videos are classified as happy and non-happy emotions by concatenating the features in both feature and decision levels.

FER techniques and different deep neural networks to recognize the happiness in both lab-controlled and unconstrained situations.

An AdaBoost classifier is used to recognize smile faces from images using intensity differences between pixels in grayscale level [41]. A feature vector called Self-Similarity of Gradients (GSS) is extracted to find the similarities in a HOG feature map and then the AdaBoost algorithm with linear extreme learning machines (ELM) is used to recognize the smile and non-smile faces [42]. An image-based real-time smile detection method has been proposed that is invariant to head poses using conditional random regression forests [43]. Another smile detection system has been developed, which extracts a small dimension of features and uses the ELM method for classification [44]. A scale driven convolutional neural network (SD-CNN) has been utilized to extract deep features and then trained an extreme gradient boosting approach [45]. Their system could manage an imbalanced data of smile and non-smile faces. All the systems have evaluated their performance by testing the GENKI-4K smile image dataset [46] that includes 4,000 face images captured in unconstrained scenarios.

A happy emotion detection system from videos contains three components: different happy emotion detection, happy emotion intensity estimation and spontaneous versus posed (SVP) smile recognition [47]. By extracting Self-Similarity of Gradients(GSS) features and some spatial-temporal features from the face, a discriminative learning model (DML) has detected and recognized the happiness and its intensity. They have evaluated their system by the UvA-NEMO dataset [48], which contains 597 and 643 spontaneous and posed happy/smile videos captured in the laboratory with controlled daylight illuminations. Another happy emotion detection system has been proposed using a fuzzy approach [49], which employs geometrical features and evaluates the method on an unconstrained dataset, Affectiva-mit [50]. The dataset includes 242 facial videos captured from the viewers who recorded their faces when using the webcam. It is worth noting that the webcam videos' background is not as diverse as the videos captured from the movies. Also, there are no significant variations of head poses in the recorded videos. Among all of the existing methods, it is noticeable that developing a happy emotion detection system is necessary to employ in all of the real-world scenarios under large head pose variation and illumination changes.

III. METHODOLOGY

A. OVERVIEW

As mentioned above, many deep learning architectures have been applied to recognize facial expressions. Due to the proven excellent performance of ResNet neural networks in facial expression recognition, this paper adopts a 3D-Inception-ResNet neural network, which extracts spatial-temporal features from different sequential video frames. Fig. 1 illustrates the proposed HappyER-DDF method. The potential of using such a system for training the machine for the natural happiness expressing motivates us to focus on studying the single emotion recognition where specific optimisation strategies can be developed, for example selecting action units relevant to a single emotion. In the video processing, the temporal dynamics features play a critical role in expressing the emotions. So, we extract the temporal dynamics features by adding a long short term memory (LSTM) unit to the extracted features at the end of network architecture. Besides, the importance of action units cannot be ignored in the facial expression recognition

systems. We can consequently extract deep facial landmarks' features by employing a CNN. Finally, we fuse two models at both feature level and decision level. A simple feature fusion method combines these features, and a fully connected unit detects the happy and non-happy videos. A weighted sum based on the weighted mean and genetic algorithm determines the final class labels after a support vector machine (SVM) with radial basis function kernel classifies the features separately. The details of each module in the proposed method are explained as follows.

B. 3D INCEPTION-ResNet (3DIR) NEURAL NETWORK

The original Inception-ResNet neural network is introduced in [34]. This architecture aims to create a wider and deeper network whereas it costs computationally. As a result, it can significantly improve the efficiency of video facial expression recognition systems when we have a large amount of data located in the sequential frames. We use a 3D version of Inception-ResNet architecture, and our proposed network comprises less layers than both the original and the introduced system in [51].

We plot the different layers of the proposed network in Fig. 2. We have provided 24 frames per second, and the system will recognize the happy/non-happy emotions per second. Therefore, the video input size is $24 \times 199 \times 199 \times 3$ (24 frames, 199×199 frame size, and 3 color channels). At the first layer, the ''stem'' uses an initial set of convolutional and pooling operations and extracts the faces' general features, which is followed by a 3DIR-A for extracting the features related to the most vital face part. This Inception-ResNet block applies the factorized convolutional filters and preserves the input size as the same. According to [34], for speeding up the training process, the Reduction-A block is employed to change the width and height of the grid and reduce the 50×50 input size to 25×25. Another 3DIR unit, 3DIR-B, is utilized to extract more accurate and detailed features. There are two main difference between ours and the method in [51]. Firstly, due to a binary classification problem, there is no need to have deeper Inception-ResNet blocks and we only use two blocks. Secondly, we apply batch-normalization on the output of each 3DIR block to normalize the extracted features and directly use it as a shortcut in our network, since we reduce the input dimension to accelerate the training phase. Finally, Average Pooling, and Dropout, and a fully connected layer form a proper shape for feeding the data into the long short term memory (LSTM) unit. The size of filters and the layers outputs are shown in Fig. 2.

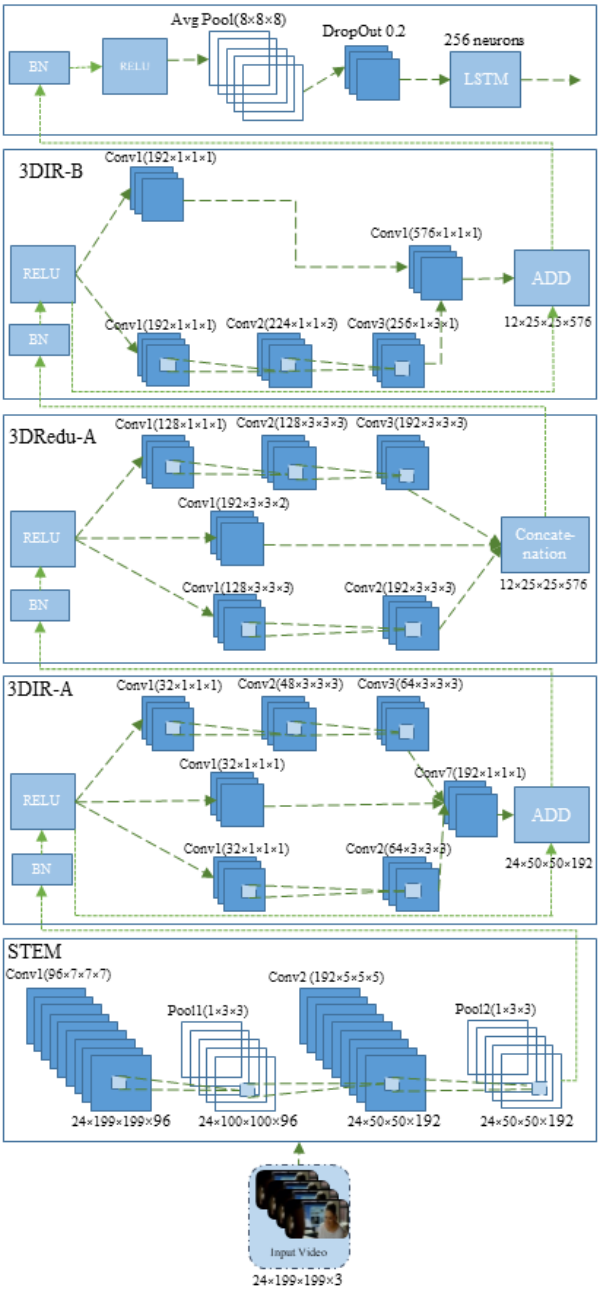


FIGURE 2. The 3D Inception-ResNet (3DIR)-LSTM architecture.(BN block shows a batch normalization node).The LSTM block containing 256 neurons is shown at the final stage of the structure, after extracting the spatial-temporal features from Inception-ResNet modules. The LSTM block output is fed to the classifier in the next phase.

C. LONG SHORT-TERM MEMORY (LSTM)

In processing still images, we only consider spatial information, but we also remark the temporal data for a video. Although 3D filters capture both spatial and temporal features, we need to consider dependencies between frames in a video. LSTM networks are a particular type of RNN that are able to learn long-term dependencies. It means that they can produce temporal dynamics features during the time and follow up on the sequential video frames' changes. According to [52], 3D CNNs learn spatiotemporal features, whereas RNN/LSTM networks learn long-term temporal information. Thus, we can explain that global temporal features are extracted by LSTMs while 3D CNNs extract local temporal features in addition to spatial features.

In the LSTM layer, per the memory cell, c, at the time-step, t, the input gate,(i), the forget gate,(f) and the output

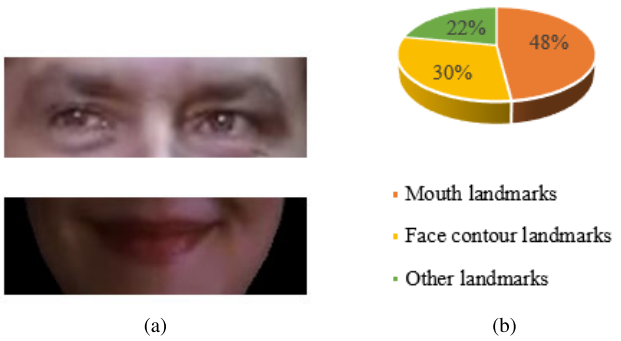


FIGURE 3. (a) AU6- cheek raiser, and AU12- lip corner puller from top (the images are from AFEW dataset), (b) Impact proportion of various facial landmarks when two smiling and neutral faces are compared [40].

gate,(o) are three gates responsible for overwriting, keeping and retrieving, respectively. We define the sigmoid function as $\sigma(x) = (1 + exp(-x))^{-1}$ and hyperbolic tangant as $\phi(x) = \frac{exp(x)-exp(-x)}{exp(x)+exp(-x)} = 2\sigma(2x) - 1$. The variables x, h, c, W and b are the input, output, cell state, parameter matrix and parameter vector, respectively. Then, the LSTM layer updates for time-step t can be defined as follows with inputs x_t, h_{t-1} and c_{t-1}:

$$\begin{aligned} f_t &= \sigma(W_f.[h_{t-1}, x_t] + b_f]) \\ i_t &= \sigma(W_f.[h_{t-1}, x_t] + b_i]) \\ o_t &= \sigma(W_o.[h_{t-1}, x_t] + b_o]) \\ g_t &= \phi(W_o.[h_{t-1}, x_t] + b_o]) \\ c_t &= f_t * (c_{t-1} + g_t * i_t) \\ h_t &= o_t * \phi(c_t) \end{aligned} \quad (1)$$

In our proposed method, we extract spatiotemporal features using 3D Inception-ResNet modules to record the face pixel changes among all the frames. The extracted spatiotemporal feature maps are then fed to an LSTM unit capturing the temporal dynamic features to discern the dependencies between frames in the videos. We tried several different unit numbers and finally have found 256 hidden neurons that are adequate for the unit. The LSTM layer is shown in Fig. 2 at the final stage of the 3D Inception-ResNet-LSTM structure. The LSTM block output is fed to the classifier in the next phase.

D. FACIAL CHANGE TIME SERIES

In addition to the visual data provided by the structure and texture of the face, tracking the facial muscle movement is complementary to facilitate the training process for achieving the goal. Fig. 3 illustrates the facial regions involving in expressing a happy emotion. The facial action coding system (FACS) [53] includes a set of action units related to each emotion. According to [53], there are two action units responsible for expressing happiness. They are cheek raiser (AU 6) and lip corner puller (AU12) related to two facial muscles, zygomatic major and the Orbicularis Oculi. Fig. 3a shows these AUs, and it is noticeable that the two most related facial elements to them are eyes and mouth. It means we can capture these movements by measuring the mouth and eyes changes. Inspired by [17], if we consider a reference point like nose peak on the face, and then calculate the distance between the mouth and reference point and the distance between the eyes and reference point, the changes can be recorded. Hence, we can optimize our system to utilize facial landmarks instead of finding AUs. As the system will determine the happy faces, we only take the eyes and mouth landmarks, a total of 24 points, into the account. However, according to the comparison of different facial components and their impact on neutral and happy expressions in [44], the face contour also plays a vital role (around 30% effective as shown in Fig. 3b) in distinguishing smile and neutral emotions. As a result, we add 16 points, representative of face contour, to the points set.

Fig. 4 and Fig. 5 illustrate changes of the mouth and eye landmarks during 28 consecutive frames, respectively, when the actor is expressing both happy and angry emotions. Fig. 4a shows the trends for mouth landmarks. It is noticeable that happy emotion mouth landmarks' changes are uniform and create a specific pattern, whereas the mouth points for angry expression (Fig. 4b) do not change constantly. Fig. 5 also confirms these trends for the eyes landmarks. According to Fig. 4b, although the mouth landmarks' changes for some landmarks of angry video are almost consistent, the entire mouth landmarks vary differently during 28 frames. The inconsistently can also be observed in Fig. 5b for eye landmarks from the starting point of the angry video. In contrast, the eye and mouth landmarks change homogeneously at all frames in the happy video, as shown in Fig. 4a and Fig. 5a. As a result, it is worth noting that discriminative features can be extracted from the landmarks patterns to distinguish between happiness and other emotions.

We apply a CNN to the resulted distance between the 40 landmarks and the reference point, i.e. the nose peak. We define the landmark points on the i^{th} frame as follows:

$$l^i = [x_1^i, y_1^i, z_1^i, \ldots, x_m^i, y_m^i, z_m^i, \ldots, x_{40}^i, y_{40}^i, z_{40}^i] \quad (2)$$

where (x_m^i, y_m^i, z_m^i) is the m^{th} landmark point.

3D landmark coordinates are processed in our system because 3D landmarks supply more accurate results than 2D counterparts by proceeding geometry of rigid face features [54]. Hence, using 3D landmarks have the potential to achieve better accuracy, especially for the videos captured in the wild with the faces in various head positions. Due to the different landmark locations, the landmarks vector should be normalized by (3):

$$\bar{x}_m^i = \frac{x_m^i - x_c^i}{\sigma_x^i}, \quad \bar{y}_m^i = \frac{y_m^i - y_c^i}{\sigma_y^i}, \quad \bar{z}_m^i = \frac{z_m^i - z_c^i}{\sigma_z^i} \quad (3)$$

where (x_c^i, y_c^i, z_c^i) is landmark coordinate of the reference point (i.e., the nose peak) and $(\sigma_x^i, \sigma_y^i, \sigma_z^i)$ shows the standard deviation of (x, y, z) coordinates at i^{th} frame.

We produce 120 distance features per frame based on 40 landmark points. We combine all these features

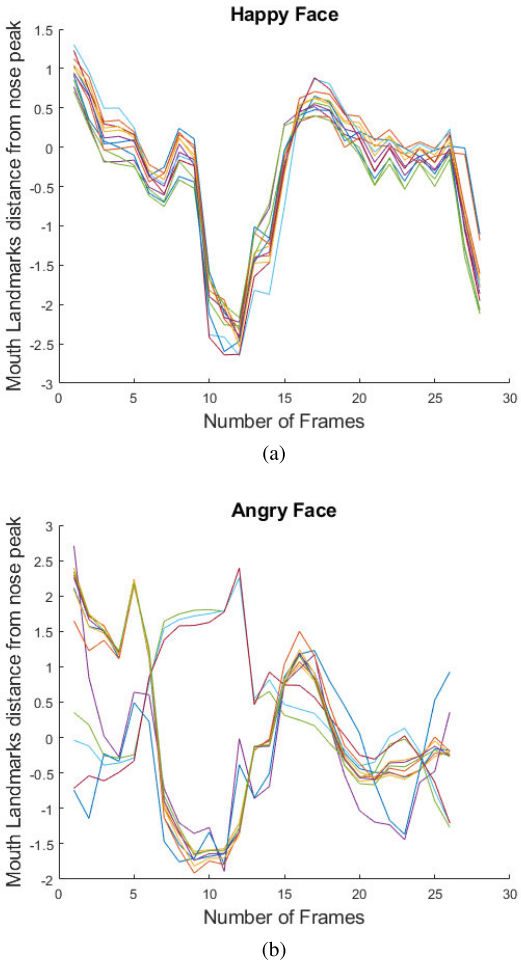


FIGURE 4. (a) and (b) illustrate the normalized distances between nose peak and different mouth landmarks among frames in two videos expressing happiness and anger, respectively. Each line of figures demonstrates the changes per mouth landmark in the video (12 mouth landmarks in total). It shows that the change during different frames in happy emotion follows a pattern separable from angry.

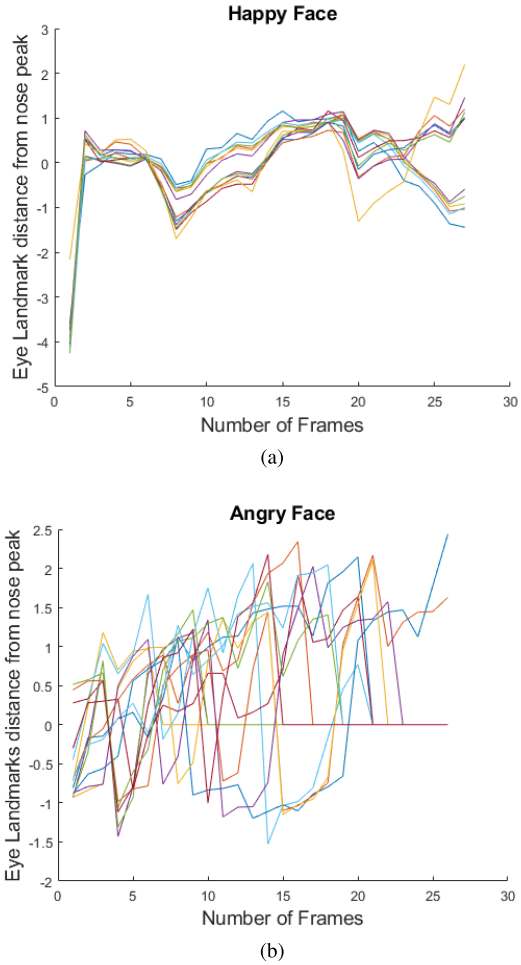


FIGURE 5. (a) and (b) illustrate the normalized distances between nose peak and different eye landmarks among frames in two videos expressing happiness and anger, respectively. Each line of figures demonstrates the changes per eye landmark in the video (12 eye landmarks in total). It shows that the change during different frames in happy emotion follows a pattern separable from angry.

of 24 frames in an image-like matrix as an input of a convolutional neural network. Fig. 6 illustrates the details of the proposed CNN. The input size is $24 \times 120 \times 1$ (24 frames, 120 distance features and one channel) that fed to four sequential convolutional filters with different sizes. As shown in Fig. 6, the pooling is applied after the first four layers and stacks the convolutional layers together to reduce the input to $2 \times 26 \times 12$. Since the pooling neurons intensify the spatial invariance [55], we only consider them at the end of the architecture. In this case, the network can find the attentional landmarks affective in expressing an emotion per video among all the landmarks input. The final fully connected layer forms our landmarks features vector with size ($f_1 \in \Re^{624}$).

E. MODEL FUSION

Since two proposed neural networks extract two various feature vectors, the model fusion plays a crucial role in the final result. Two main categorizations (feature-level and decision-level fusions [56]) are considered for fusion methods. In the feature-level fusion, also called early-fusion, different extracted feature vectors are combined to form a shared representation. The concept and type of features should be similar to ensure the feature-level fusion can be conducted. In the decision-level fusion, the decisions that are obtained by separately applying each feature vector to multiple classifiers are concatenated. Hence, it is also called the late-fusion technique. Due to each fusion method's independent impact on the data, we utilize both the feature-level and decision-level fusions. The fusion techniques are detailed as follows.

1) FEATURE-LEVEL FUSION

After extracting the desired feature vectors from both proposed deep neural networks by training on the videos, we use the vector ($f_1 \in \Re^{256}$) of LSTM output, and the vector ($f_1 \in \Re^{624}$) output of landmarks CNN and normalize these

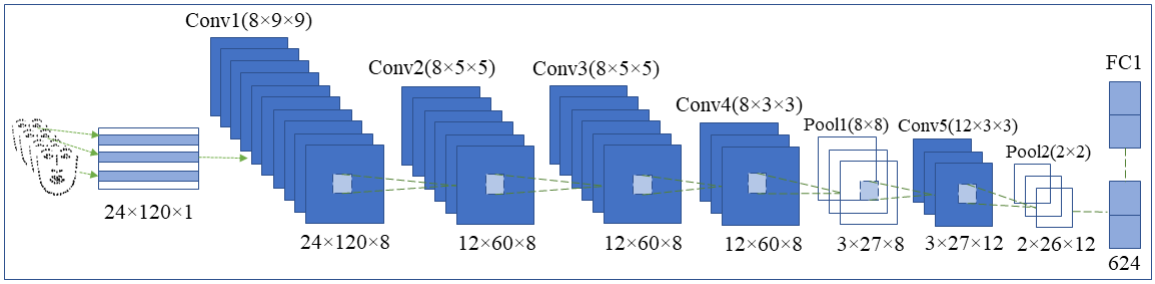


FIGURE 6. The proposed CNN architecture for extracting landmarks features. It comprises five convolution, two pooling layers and a 624 neurons fully connected layer.

features separately using $l2$ normalization. The normalized features are then concatenated into ($f \in \Re^{880}$) and form the final feature vector descriptive of the videos. The fused feature, f, is fed to a fully connected layer of 512 neurons and the final layer with a sigmoid function is used to recognize the happy and non-happy emotions.

2) DECISION-LEVEL FUSION

Due to the significant head pose variations, we may detect emotions from incomplete faces in the unconstrained videos. So, we cannot get the spatial-temporal features from the entire face in the side view faces, whereas an accurate estimation of the facial landmarks is accessible. As a result, it is worth weighing the impact of facial landmarks in the final decision. Inspired by the weighted sum fusion method introduced in [18], we define the weighted sum rule as follows:

$$d_{out} = \omega_1 d_{spatial-temporal} + \omega_2 d_{landmarks} \tag{4}$$

where $d_{out}, d_{spatial-temporal}$ and $d_{landmarks}$ are the final decision, the decision obtained by 3DIR-LSTM network, and facial landmarks classification, respectively. ω_1 and ω_2 are the weights assigned to our models for applying spatial-temporal and facial landmarks features and $\omega_1 + \omega_2 = 1$. We employ an SVM with radial basis function kernel to obtain a single decision per feature vector. For optimizing the weights, ω_1 and ω_2, the weighted mean method used in [57] and a genetic algorithm (GA) are applied to the validation sets. In weighted mean method [57], the weights are selected based on the validation sets that achieve to the highest performance. In GA, an evolutionary searching algorithm [58], a population for our parameters, ω_1 and ω_2, are randomly initialized. The random operators like selection, crossover and mutation are repeated to maximize the objective function (d_{out}) subject to $\omega_1 + \omega_2 = 1$. The values of $d_{spatial-temporal}$ and $d_{landmarks}$ are obtained by applying the validation sets. We limit creating the new generations by reaching to 200 repeats. This value is set based on the trial and error when we apply various validation sets of different datasets. After termination, we can obtain the optimized values of ω_1 and ω_2 for different datasets.

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the proposed method using three different video datasets captured in the wild and demonstrate its accuracy by comparing it with various state-of-the-art facial expression recognition methods. We first briefly introduce the datasets, then explain the networks settings and finally report the results when concatenating at both feature and decision levels. The results obtained by applying single features are also explained to prove the importance of each feature vector.

A. DATASETS

Among various benchmark facial expression datastets, we have selected three AM-FED+, AFEW, and MELD datasets to evaluate the generality of the proposed method on different levels of challenge. Some examples of these datasets have been shown in Fig. 7.

1) EXTENDED AFFECTIVA-MIT (AM-FED+)

The dataset is an extended version of naturalistic facial response videos collected in daily settings [59]. The participants from all over the world were asked to watch video advertisements. They permitted to use their webcam while watching videos to record their face. As a result, 1044 webcam videos have been streamed to the server. Among them, 545 videos were manually labeled by facial action units and smile. Since there were not any requirements for the environment that participants recorded their videos, the lighting varied in all videos. Most videos were recorded in frontal view, whereas several had some variations in the head pose. Almost all smile videos have three types (onset, apex and offset) of frames. Due to videos self-recording by users, some videos are very long (around 5-6 min) without any proper information in the first 5 minutes. We selected continuous frames (at least 24 frames) per video dependent on the video duration. It can be ensured that the chosen frames are the most important ones since they comprise onset, peak, and offset according to existing labels. Fig. 7(a) shows an example of AM-FED+ dataset. AM-FED+ is the least challenging dataset due to two reasons: 1) most faces are in frontal view, and contains videos including onset, apex, and offset frames, and 2) there is a slight difference in expressing the happiness between different videos (small intra-class distance).

2) ACTED FACIAL EXPRESSION IN THE WILD (AFEW)

This dataset contains videos captured from Hollywood movies and TV series and first introduced for the EmotiW

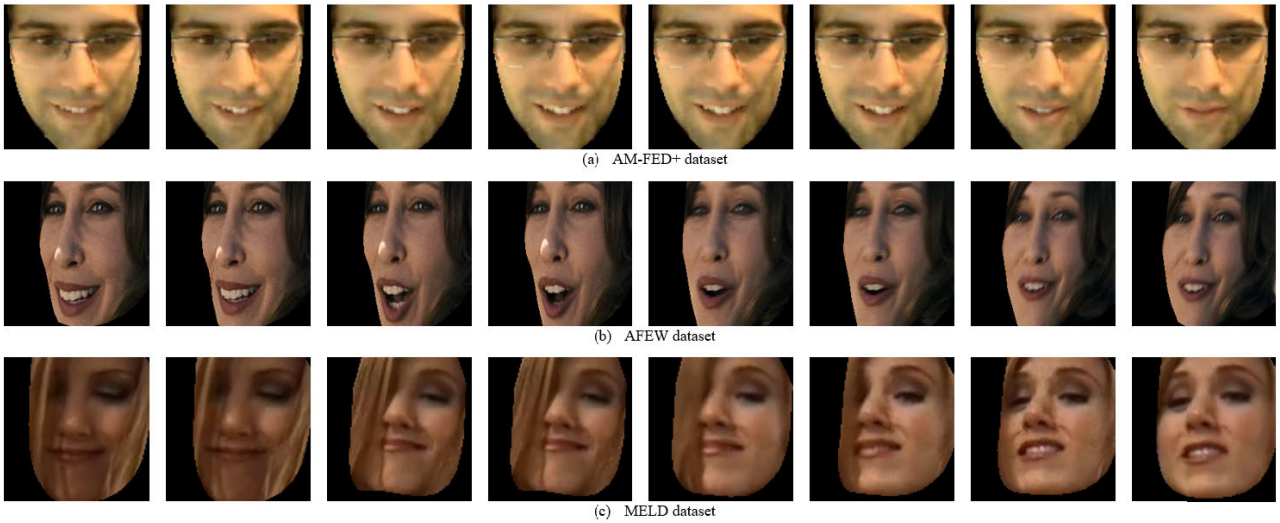


FIGURE 7. Some examples of onset, apex and offset frames from happy videos in three in-the-wild datasets. As it is seen, AM-FED+ videos have all three onset, apex, and offset frames. In contrast, AFEW and MELD videos do not include entire types of frames.

challenge [60]. There are 578, 383 and 307 videos for training, validation and testing, respectively. It is one of the most challenging datasets for emotion recognition as the 330 actors express seven emotions in different outdoor and indoor environments. The number of happy videos in the training and validation sets is 150 and 63, respectively. If we wanted to put happy videos in one category and all other emotions in another, we might face with imbalanced data. Hence, we created three different training and validation sets by randomly data selection from all the emotions (except happiness) to have a balanced dataset for our model. Although the video durations are various between 0.75 and 4 seconds, we have chosen 24 frames per second. For the videos less than 24 frames, we had to keep the frames to be sequential. Hence, we repeated first several frames and final several frames in the video to ensure at least 24 frames. Finally, the max voting is used for labeling videos into happy and non-happy groups since almost all the videos' frames contain the apex expressions. Fig. 7(b) shows an example of AFEW dataset. It is ranked the second challenging dataset since there is a big intra-class distance between happy videos. The emotions are expressed by many actors from various nationalities and races. Also, onset, apex, and offset frames may not constantly occur in the videos.

3) MULTIMODAL MULTI-PARTY DATASET IN EMOTION RECOGNITION IN CONVERSATIONS (MELD)

Multimodal Emotionlines Dataset (MELD) was created to provide multimodal multi-party conversational data by collecting three different visual, audio and text modalities [61]. MELD contains 13,708 utterances from 1433 dialogues of Friends TV series. Consequently, there are 9989, 1109 and 2610 videos in the training, validation and test sets, respectively. Among them, 1743, 163 and 402 videos in each set are labeled by happy emotion. Each utterance was annotated by an emotion and sentiment label. Neutral, positive and negative are three groups of sentiments, whereas the emotions categorization covers six basic emotions (joy, disgust, sadness, anger, fear, and surprise). Because videos (acoustic and visual data) in the MELD dataset were extracted based on the time-stamps of subtitle transcription, some videos did not include any faces and were ignored. Similar to the AFEW data preparation, we created two non-happy training groups by randomly selecting videos from other emotions. Fig. 7(c) shows an example of MELD dataset. Processing the unfocused faces in the scenes is the addition cause to AFEW reasons why the MELD dataset is the most challenging. Detecting the faces is more complicated, and aligning the faces leads to low-resolution input.

B. PRE-PROCESSING

Before training our happy emotion detection system, the faces in video frames are detected and resized to the network input size. Using OpenFace toolkit [62], a multi-task convolutional neural network (MTCNN) has been employed to detect the faces. Also, a piecewise affine warp was trained to estimate the landmarks error and accurately align the faces. The aligned faces were resized to 199×199, and considering 24 frames per second, we have the aligned video frames with size of $24 \times 199 \times 199 \times 3$ equal to the 3DIR-LSTM network input. Although convolutional experts constrained local model (CE-CLM) in the OpenFace facial detection module was trained at different head orientations, we noticed that the system could not sometimes detect any landmarks or fit the obtained landmarks on the face. It occurred when there was a significant variation in head poses. Fig. 8a shows some unsuccessful examples where landmarks are wrongly detected by OpenFace toolkit [62]. As we aim to develop a precise happy emotion detection system for the unconstrained videos, finding the true 3D facial landmarks is crucial in

(a) (b)

FIGURE 8. Two face examples from MELD dataset that we have applied OpenFace toolkit [62] and method in [63] for detecting 68 face landmarks. a) The OpenFace toolkit could not detect 68 face landmarks correctly, b) the correct landmarks were detected by the method in [63].

the proposed approach. We have utilized the method in [63], in which a proposed hierarchical, parallel and multi-scale block has been utilized instead of the Hour Glass bottleneck blocks (i.e., four different blocks of architectures in a Face Alignment Network [64]) to locate the 2D and 3D facial landmarks points. Fig. 8b shows the correct landmarks detected by the method in [63]. Among all 68 obtained facial landmarks, we chose 24 3D eyes and mouth landmarks adding to 16 3D face contour points. The image-like matrix with a size of 24×120 was created, equal to the proposed landmarks that are input of CNN network.

Since the AM-FED+ dataset has not been categorized to different training, validation, and test sets, we apply 10-fold cross-validation strategy similar to the methods in [31] and [65] to ensure the approach generalization. The data is divided into ten groups, selecting nine of them as training and one as testing sets. We consider a near-identical distribution for expressions in both groups to prevent the issues caused by imbalanced data and report the average results of 10 runs for the AM-FED+ dataset. For the AFEW and MELD datasets, we created three and two balanced training and validation sets and consequently describe the average results as well.

C. NETWORK SETTINGS

In the proposed approach, two 3DIR-LSTM and CNN networks need to be trained to extract the proper features from faces and landmarks. There are around 409,068,581 trainable parameters in total for training of the proposed algorithm. Network initialization plays a fundamental role in network training, prohibiting over-fitting and early convergence. The weight matrixes were initialized with Xavier uniform initializer, and the training process starting with the learning rate, 0.005, was decayed to 0.0003 after the first 20 epochs, 0.0002 in the next 20 epochs and 0.00001 in the remaining epochs. All layers except pooling and final layer were fired by Rectifier Linear Unit (ReLu) activation function and a linear function was used when concatenated the output of inside layers. Sigmoid function determined the final output. We optimized the training process with Adam optimizer and estimated the loss employing the binary cross-entropy. It was repeated till the model was convergence to the minimum loss error. At this stage, we froze it as the spatial-temporal features and apply to both the feature and decision level fusions in the next steps. The training phase was conducted on a computer (NVIDIA GPU, GeForce GTX 1080 Ti), where the batch size was set to 16. Table 1 shows the time complexity of the proposed method. We report the running times when feature-level fusion concatenates the features and the decision-level method is applied by both weighted mean and GA. Dependent on the size of each dataset, the proposed approach at feature-level fusion was executed in about 3, 2, and 30 hours and a half for AM-FED+, AFEW, and MELD datasets, respectively. Due to the complexity of the Genetic algorithm, the decision-level fusion models were the slowest and took around 4 hours, 2 hours and a half, and 36 hours and a half for AM-FED+, AFEW, and MELD datasets, respectively. The running times for decision-level models by weighted mean were approximately close to feature-level fusion with 3, 2, and 30 hours and a half for AM-FED+, AFEW, and MELD datasets, respectively.

TABLE 1. Time complexity of the proposed method.

Dataset	Feature-level model running time	Decision-level fusion model with GA running time	Decision-level fusion model with weighted mean running time
AM-FED+	$\sim$ 3 hours	$\sim$ 4 hours	$\sim$ 3 hours
AFEW	$\sim$ 2 hours	$\sim$ 2 hours and a half	$\sim$ 2 hours
MELD	$\sim$ 30 hours and a half	$\sim$ 36 hours and a half	$\sim$ 30 hours and a half

Meanwhile, we have an image-like matrix with 24×120 formed by facial landmarks. The numbers of batch sizes, epochs, activation and loss functions and optimizer were the same as the 3DIR-LSTM network except for the learning rate that is set to 0.001.

D. COUNTERPARTS METHODS FOR HAPPY EMOTION DETECTION

There are a few happy emotion detection systems/approaches from videos as discussed in Section II. To verify our proposed method, in contrast, we implemented some popular feature extraction and classification methods in happy emotion detection from images and extended them to process the videos. Also, we modified some general facial expression recognition systems to recognize only happy faces. To have a fair comparison, we have adopted the same settings as in their original papers. The counterpart methods are introduced as follows.

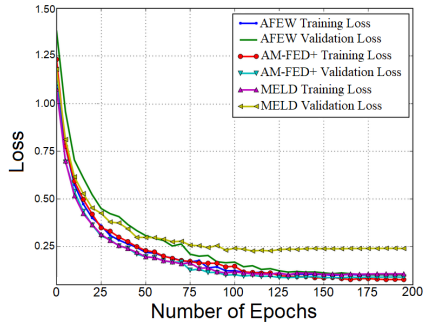
- According to the baseline report of the AM-FED+ dataset, the OpenFace toolkit is utilized to recognize the smile [62].
- A method for detecting smile from AM-FED videos recognizes the smile by employing the facial landmarks and defining rules through a fuzzy system approach [49].
- A system is proposed for recognizing emotions in wild videos [12]. LBP-TOP and SIFT features are extracted and the sparse representation is used to classify seven emotions. We have modified their classification method to enable it to detect happy and non-happy emotions.
- Another facial expression recognition method has been proposed by extracting 3DIR-LSTM features and conducting the element-wise operation with facial landmarks [51]. In this work, ten apex frames have been manually selected for all the datasets. We have modified the system to classify the happy emotions. As manually selecting takes much time, we chose the ten frames in the middle of each video.
- The decision level fusion has been employed to classify the emotions among two feature vectors extracted by CNN-BRNN and trajectory matrix from facial landmarks [17]. Forty frames are used as input videos, and if a video was shorter, they have repeated the first frame at the beginning of the sequence. We have modified their method to classify only two classes (happy and non-happy expressions).

Also, other two techniques were considered for evaluation. Extreme learning machine (ELM) is one adopted classifier containing a hidden layer feed-forward neural network that has been utilized in smile detection from images [44]. We set the number of hidden layer neurons to two times of the input dimension and extract LBP-TOP and HOG-TOP from the sequential frames in the videos. The ELM then classifies the videos into happy or non-happy groups. We have also adopted a convolutional neural network similar to [66]. We use it in two ways: firstly, passing the extracted HOG-TOP features to the CNN; secondly, converting it to a 3D-CNN network to cover the video frames. Finally, we study the performance of our proposed method by comparing with all counterpart methods using the three in-the-wild datasets.

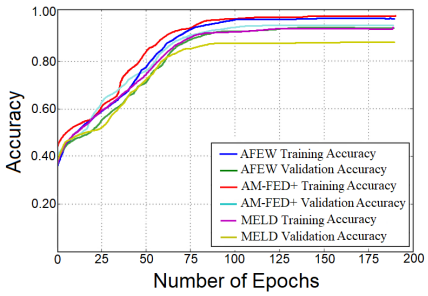
E. HAPPY EMOTION DETECTION RESULTS

1) OPTIMAL MODELS USING THE PROPOSED HAPPYER-DDF METHOD/STRATEGY

As mentioned earlier, we evaluated the proposed method at both feature and decision level fusions. We used the trained models for testing AM-FED+, AFEW and MELD datasets. Fig.9a illustrates the loss values for the training and validation process of three datasets when it was repeated for 195 epochs. It is clear that the training process of three datasets has been converged by reaching the minimum possible loss. Table 2 shows that the average accuracy of evaluating the trained models on the AM-FED+, AFEW and MELD datasets when we obtained the results using a fully connected layer (feature-level fusion) and weighted sum rule with radial basis function kernel (decision-level fusion). It is noticeable that we have reported the results with both extracted features and single feature vectors alone. In this case, we separately fed each feature vector to the final fully connected layer and SVM classifier. Although the videos in AM-FED+ have been labeled by action units and smile, we only considered the smile classification in the evaluation phase. According to Table 2, the highest accuracy for AM-FED+, AFEW and MELD datasets are 95.97%, 94.89%, and 91.14%, respectively, when the extracted features have been fused at feature level. We can see that concatenating features provide complementary information for the system to resist possible noises. The accuracy changes during 195 epochs are shown in Fig. 9b for the training round which leads getting the best



(a)



(b)

FIGURE 9. (a) and (b) show the calculated losses and accuracies for both training and validation processes. The accuracy plot is for the best results obtained by the proposed method during different runs.

TABLE 2. Average accuracy of the proposed trained models.

	AM-FED+	AFEW	MELD
3DIR- LSTM features + FC	90.24%	84.54%	82.18%
Facial distance time series features + FC	92.68%	87.65%	86.21%
3DIR- LSTM features + SVM	93.15%	91.4%	81.01%
Facial distance time series features + SVM	93.86%	92.76%	87.44%
3DIR- LSTM and Facial distance time series features + FC	**95.97%**	**94.89%**	**91.14%**
3DIR- LSTM and Facial distance time series features + Weighted sum rule based on GA	94.52%	93.29%	87.73%
3DIR- LSTM and Facial distance time series features + Weighted mean	93.03%	93.74%	85.36%

accuracy for AM-FED+, AFEW, and MELD. The obtained accuracy for AM-FED+ is higher than the accuracy of two other datasets since most faces are focused and in frontal view with a slight variation between different videos. Moreover, it is apparent that the results on the AFEW dataset are higher than MELD dataset because the primary goal of creating the MELD dataset was emotion recognition based on the text, the scenes were not chosen only dependent on the faces, and the aligned faces are in low-resolution. It is a great instance of the in-the-wild dataset where the faces were not on the scenes' concentration and were far from the camera; hence, detecting the faces and aligning them resulted in a low-resolution face as Fig. 10 shows. Getting acceptable and significant-good results on a large unseen and unexpected dataset (MELD) demonstrates that the proposed system is robust and well generalized. Although classifying the features separately by SVM achieved more accuracy, with 93.86%, 92.76% and 87.44% for AM-FED+, AFEW and MELD datasets, the concatenated features resulted in better accuracy when we used the fully connected layers for fusion and the classifier. It means feature-level fusion performs superior to decision-level based on both GA and weighted mean in recognizing happy expressions, with 95.97%, 94.89%, and 91.14% in contrast to 94.52%/93.03%, 93.29%/93.74%, and 87.73%/85.36% for AM-FED+, AFEW and MELD datasets.

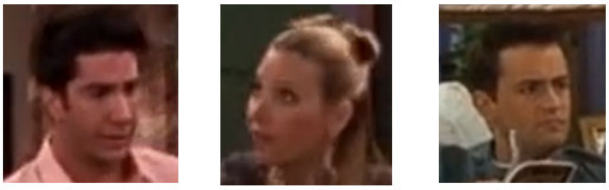


FIGURE 10. Some low-resolution examples from MELD dataset that challenge the proposed method performance.

As Table 2 shows, we also obtained the optimized weights of decision-level model using two methods of the weighted mean and GA. The recognition accuracy of 94.52% and 87.73% for AM-FED+ and MELD using GA-based decision-level is higher than the accuracy rate obtained by weighted mean decision-level with 93.03% and 85.36%. For the AFEW dataset, we achieved a similar result to work [57] and the accuracy obtained by the weighted mean decision-level model (93.74%) is slightly higher than the GA-based model accuracy (93.29%). Regarding the runtimes shown in Table 1, although GA has achieved better accuracies (around 2%), it has sacrificed the time. It seems that optimizing the weights using GA help achieve higher accuracy by sacrificing a reasonable efficiency. As in any real-world application, we use the pre-trained models; it is worth to employ a slower method with higher accuracy than a fast lower rate one.

2) COMPARISON WITH OTHERS' METHODS

For demonstrating the effectiveness of the proposed method, we compare the obtained results with the existing happy emotion detection approaches in Table 3. Table 3 shows

TABLE 3. The comparison with the state-of-the-art methods.

Methods	Accuracy% on AM-FED+	Accuracy% on AFEW	Accuracy% on MELD
Baseline method [62]	87.9	-	-
Facial Landmarks + fuzzy system [49]	88.9	-	-
LBP-TOP + ELM	88.2	70.22	68.34
HOG-TOP + ELM	89.94	73.62	70.57
3D-CNN	89.35	64.81	65.28
HOG-TOP + CNN	91.76	74.39	75.19
LBP-TOP, SIFT + Sparse representation [12]	90.33	91.84	84.15
CNN-BRNN + trajectory matrix [17]	92.61	79.47	76.28
3DIR-LSTM + element-wised landmarks [51]	92.87	89.93	85.36
3DIR- LSTM and Facial distance time series features + FC (our best proposed method)	**95.97%**	**94.89%**	**91.14%**

that the best proposed approach (both feature and decision level fusion) applied to AM-FED+ has remarkably achieved an 95.97% using feature level fusion (note that our another proposed method achieves 94.52% using decision level fusion as shown in Table 2), higher than the baseline (achieving 87.9%) and all other counterpart methods. Comparing the methods applied to AM-FED+, the method [17] ranked the second only achieves 92.87% using a recurrent Inception-ResNet neural network and element-wised landmarks. Considering concatenated facial landmarks features and spatial-temporal in videos caused it to be better among others by improving the performance robustly and significantly. Similar to work [17], the method [51] applied both landmarks and textural features using CNN-BRNN structure to achieve a close accuracy of 92.61%, placed the third rank. Combining HOG-TOP features (traditional techniques) with CNN (deep learning-based) performed well and ranked forth with accuracy of 91.76%. Utilizing both local binary pattern (LBP-TOP) and scale invariant features (SIFT) as a single traditional method achieves a relatively high accuracy of 90.33% after all deep learning methods. After the baseline method, LBP-TOP demonstrates the worst accuracy among all, with 88.2%. It is worth noting that the result (88.9%) for the method [49] is obtained by the small AM-FED (including 242 videos) where the facial landmarks are employed only in the fixed number of changed frames, not processing all the sequences. Hence, it cannot be utilized for real-world applications as handling a large number of frames is a challenge.

Table 3 also shows the performance for AFEW and MELD as two very challenging datasets. They contained all the real-world conditions and were captured from TV series and movies where the actors/actresses express genuine emotions. According to Table 3, our proposed method is better than all counterpart methods, including traditional and deep learning methods, with the accuracies of 94.89% for AFEW and 93.29% for MELD. In [12], the traditional LBP-TOP and

SIFT feature extraction methods are proposed, and the data are classified using a modified sparse representation. It is noticeable that although their system had a significant performance on the AFEW test set (ranked the second with 91.84% accuracy), it was not generalized on unseen, unexpected data, and the performance on the MELD dataset has reduced much to around 84.15% accuracy. However, work [51] demonstrated a very high accuracy on chosen datasets on their study; its performance did not show superiority and was placed at the third, with accuracies of 89.93% and 85.39%. This problem can be referred to as the manual frame selection. By applying the work [17] settings and unifying the data accordingly, we again got the better result for AFEW test data, with 79.47% contrary to 76.28% for MELD. As mentioned above, the RNN cannot memorize the long sequences, and it is clear that the LSTM layer has a better performance on the videos. 3D-CNN showed the worst performance, with 64.81% and 65.28% accuracy, because the videos were captured in the unconstrained scenarios, and it cannot track all the existing challenges, such as significant variation in head poses. Overall, the proposed system performs well, especially in harsh situations and in-the-wild datasets, and outperforms at least seven state-of-the-art approaches. Furthermore, according to the results reported in Table 3, it can be claimed that the type of features along with the used classifier plays a key role in the recognition accuracy rate of happy emotion. The same classifier can achieve close accuracy values by different feature extraction methods. For instance, the ELM classifier could lead to higher accuracy when the features were changed from LBP-TOP to HOG-TOP. Also, using an additional feature to CNNs may achieve higher accuracy than applying a single CNN network.

Regarding the network size comparison, our method can work better for larger datasets than a counterpart method, even reducing the number of layers. We compared our proposed network with a similar method introduced in [51]; they have trained and tested their network with MMI, CK+, FERA, and DISA datasets that contain 11,500, 3270, 7000, and 89,000 images, respectively. We trained and tested our proposed method using AM-FED+, AFEW, and MELD datasets that contain 13,080, 7200, and 110,784 images, respectively. (Note that we calculated the number of frames considering at least 1 second (24 frames) per video, whereas some videos are longer (2-4 seconds)). According to Table 3, the proposed method achieves higher accuracy than the counterpart method in [51] (95.97%, 94.89%, 91.14% in contrast to 92.87%, 89.93%, and 85.36% for AM-FED+, AFEW, and MELD, respectively). As a result, the fewer layers in the proposed approach work more accurately even with larger-scale data compared to the method [51].

We face two different types of occlusion in the AFEW and MELD datasets. In the first type, half or top of the face is covered by hair. In this case, the system decides on the same as when we address a significant variation in the head pose. Thus, the proposed CNN of landmarks plays an essential role in the final decision. In the second type of face occlusion, in which the happy mouth is covered by something such as the actors' hand, the predicted label is wrong in most cases because the mouth landmarks are estimated similar to normal lips that express neutral emotion. Hence, we can claim that mouth is the most vital face part that misleads the system when it is under covered.

In summary, the obtained results demonstrate that our proposed method has considerably achieved the highest accuracy of 95.97%, 94.89%, and 91.14% among state-of-the-art FER systems on three AM-FED+, AFEW, and MELD datasets, respectively. Generally speaking, it has been proven that combining both textural and landmarks' features leads to higher accuracy. As we reported the results in both feature and decision levels, the most significant accuracies were obtained when feature-level fusion was applied. We tested our proposed method with different datasets in the wild that videos contain various challenging conditions such as large head pose variations and illumination changes. Moreover, race diversity existed among the actors who played in the videos expressing spontaneous facial expressions. The results demonstrate the well-generality of the proposed approach even on these datasets. It could achieve high accuracy by addressing the head pose problem and was lighting-invariant in most cases. Recognizing the spontaneous happy/smile expressions can also be considered as a significant achievement since they are natural rather than posed emotions expressed in the lab-controlled datasets. Therefore, the proposed approach could obtain similar results on similar real-world datasets in challenging situations that faces are captured at random angles.

V. CONCLUSION

This paper proposes a novel HappyER-DDF method that adopts a hybrid deep neural network to recognize the happy emotions from unconstrained videos. Due to the excellent result of ResNet frameworks on facial expression recognition, in our HappyER-DDF method, we have used a 3D version of Inception-ResNet architecture to extract the spatial-temporal features. An LSTM layer has been applied to the extracted features to consider the temporal dynamic features in the sequential frames. Since geometric features formed by facial landmarks are effective in expression recognition, we have employed a CNN to extract deep features from facial distance time series. By concatenating these feature vectors at both feature and decision level fusions, our methods have classified the happy and non-happy groups. The experiments on three unconstrained video datasets have demonstrated that the proposed HappyER-DDF approach detects happy emotions with an accuracy of 95.97% for AM-FED+ dataset, 94.89% for AFEW dataset and 91.14% for MELD dataset, which has a better accuracy than several counterpart methods such as methods in [12]–[17], and [47]. Due to overall low accuracy resulted in the emotion recognition systems in the wild, creating such a successful standalone happy emotion recognition approach can inspire us to develop the single emotion recognition strategy further for demonstrating in

other five emotions in the near future. It is worth noting that there is still room to improve the proposed method's performance by adding some attentional blocks to the mentioned framework to discover more important face parts in emotion recognition. Also, testing and developing the proposed model on other large-scale challenging unconstrained datasets can be considered in future work.

REFERENCES

[1] A. Curci, T. Lanciano, F. Battista, S. Guaragno, and R. M. Ribatti, "Accuracy, confidence, and experiential criteria for lie detection through a videotaped interview," *Frontiers Psychiatry*, vol. 9, p. 748, Jan. 2019.

[2] G. Mattavelli, E. Barvas, C. Longo, F. Zappini, D. Ottaviani, M. C. Malaguti, M. Pellegrini, and C. Papagno, "Facial expressions recognition and discrimination in Parkinson's disease," *J. Neuropsychol.*, 2020.

[3] N. Fayyazifar and N. Samadiani, "Parkinson's disease detection using ensemble techniques and genetic algorithm," in *Proc. Artif. Intell. Signal Process. Conf. (AISP)*, Oct. 2017, pp. 162–165.

[4] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, May 2018.

[5] I. Y. Choi, M. G. Oh, J. K. Kim, and Y. U. Ryu, "Collaborative filtering with facial expressions for online video recommendation," *Int. J. Inf. Manage.*, vol. 36, no. 3, pp. 397–402, Jun. 2016.

[6] S. Jaiswal, S. Virmani, V. Sethi, K. De, and P. P. Roy, "An intelligent recommendation system using gaze and emotion detection," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 14231–14250, Jun. 2019.

[7] X. Su, M. Gao, J. Ren, Y. Li, and M. Rätsch, "Personalized clothing recommendation based on user emotional analysis," *Discrete Dyn. Nature Soc.*, vol. 2020, pp. 1–8, Mar. 2020.

[8] N. Samadiani, G. Huang, W. Luo, Y. Shu, R. Wang, and T. Kocaturk, "A novel video emotion recognition system in the wild using a random forest classifier," in *Proc. Int. Conf. Data Sci. (ICDS)*. Singapore: Springer, 2019, pp. 275–284.

[9] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao, and B. W. Schuller, "Connecting subspace learning and extreme learning machine in speech emotion recognition," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 795–808, Mar. 2019.

[10] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, May 2019.

[11] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *Int. J. Social Robot.*, vol. 11, no. 4, pp. 555–573, 2019.

[12] N. Samadiani, G. Huang, W. Luo, C.-H. Chi, Y. Shu, R. Wang, and T. Kocaturk, "A multiple feature fusion framework for video emotion recognition in the wild," *Concurrency Comput. Pract. Exp.*, p. e5764, Apr. 2020.

[13] A. Mehrabian, "Communication without words," in *Communication Theory*. New York, NY, USA: Taylor & Francis, 2008, pp. 193–200.

[14] N. Samadiani, G. Huang, B. Cai, W. Luo, C. H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, Apr. 2019.

[15] M. J. Bernstein, D. F. Sacco, C. M. Brown, S. G. Young, and H. M. Claypool, "A preference for genuine smiles following social exclusion," *J. Exp. Social Psychol.*, vol. 46, no. 1, pp. 196–199, Jan. 2010.

[16] H. Ugail and A. Al-Dahoud, "A genuine smile is indeed in the eyes—The computer aided non-invasive analysis of the exact weight distribution of human smiles across the face," *Adv. Eng. Informat.*, vol. 42, Oct. 2019, Art. no. 100967.

[17] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, Oct. 2018.

[18] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 445–450.

[19] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *J. Electron. Imag.*, vol. 25, no. 6, Jun. 2016, Art. no. 061407.

[20] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. Face Gesture*, Mar. 2011, pp. 921–926.

[21] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. Face Gesture*, Mar. 2011, pp. 878–883.

[22] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[23] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 356–361.

[24] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 38–50, Jan. 2018.

[25] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 497–502.

[26] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[27] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.

[28] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 552–558.

[29] A. T. Kabakus, "PyFER: A facial expression recognizer based on convolutional neural networks," *IEEE Access*, vol. 8, pp. 142243–142249, 2020.

[30] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 467–474.

[31] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local-global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, 2020.

[32] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for In-the-Wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[36] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: Towards robust emotion recognition in the wild," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 472–478.

[37] M. M. Rahman, Y. Tan, J. Xue, L. Shao, and K. Lu, "3D object detection: Learning 3D bounding boxes from scaled down 2D bounding boxes in RGB-D images," *Inf. Sci.*, vol. 476, pp. 147–158, Feb. 2019.

[38] M. K. Lee, D. Y. Choi, D. H. Kim, and B. C. Song, "Visual scene-aware hybrid neural network architecture for video-based facial expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.

[39] S. Li, W. Zheng, Y. Zong, C. Lu, C. Tang, X. Jiang, J. Liu, and W. Xia, "Bimodality fusion for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 589–594.

[40] M. Hu, H. Wang, X. Wang, J. Yang, and R. Wang, "Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 176–185, Feb. 2019.

[41] C. Shan, "Smile detection by boosting pixel differences," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 431–436, Jan. 2012.

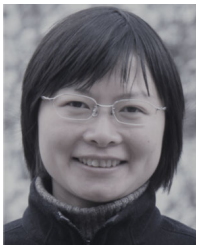
[42] Y. Gao, H. Liu, P. Wu, and C. Wang, "A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios," *Neurocomputing*, vol. 174, pp. 1077–1086, Jan. 2016.

[43] L. Liu, W. Gui, L. Zhang, and J. Chen, "Real-time pose invariant spontaneous smile detection using conditional random regression forests," *Optik*, vol. 182, pp. 647–657, Apr. 2019.

[44] D. Cui, G.-B. Huang, and T. Liu, "ELM based smile detection using distance vector," *Pattern Recognit.*, vol. 79, pp. 356–369, Jul. 2018.
[45] T. Vo, T. Nguyen, and C. T. Le, "A hybrid framework for smile detection in class imbalance scenarios," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8583–8592, Dec. 2019.
[46] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2106–2111, Nov. 2009.
[47] P. Wu, H. Liu, C. Xu, Y. Gao, Z. Li, and X. Zhang, "How do you smile? Towards a comprehensive smile analysis system," *Neurocomputing*, vol. 235, pp. 245–254, Apr. 2017.
[48] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 525–538.
[49] C. Vinola and K. Vimala Devi, "Smile intensity recognition in real time videos: Fuzzy system approach," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15033–15052, Jun. 2019.
[50] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 881–888.
[51] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 30–40.
[52] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3120–3128.
[53] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford Univ. Press, 1997.
[54] I. P. Okuwobi, Q. Chen, S. Niu, and L. Bekalo, "Three-dimensional (3D) facial recognition and prediction," *Signal, Image Video Process.*, vol. 10, no. 6, pp. 1151–1158, Sep. 2016.
[55] P. Barros, G. I. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: A deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, Aug. 2017.
[56] F. Castanedo, "A review of data fusion techniques," *Sci. World J.*, vol. 2013, Oct. 2013, Art. no. 704504.
[57] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 569–576.
[58] N. Samadiani and S. Moameri, "Diagnosis of coronary artery disease using cuckoo search and genetic algorithm in single photon emision computed tomography images," in *Proc. 7th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2017, pp. 314–318.
[59] D. McDuff, M. Amr, and R. E. Kaliouby, "AM-FED+: An extended dataset of naturalistic facial expressions collected in everyday settings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 7–17, Jan. 2019.
[60] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia-Mag.*, vol. 19, no. 3, pp. 34–41, Jul. 2012.
[61] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*. [Online]. Available: http://arxiv.org/abs/1810.02508
[62] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.
[63] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial Landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
[64] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 483–499.
[65] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2852–2861.
[66] J. Chen, Q. Ou, Z. Chi, and H. Fu, "Smile detection in the wild with deep convolutional neural networks," *Mach. Vis. Appl.*, vol. 28, nos. 1–2, pp. 173–183, Feb. 2017.

NAJMEH SAMADIANI received the bachelor's degree in computer engineering in 2012 and the master's degree in artificial intelligence in 2014. She was a Lecturer with the Kosar University of Bojnord, Iran, from 2015 to 2018. She is currently pursuing the Ph.D. degree with Deakin University, Australia. Her research interests include image/video processing, human emotion modeling, expert systems, and pattern recognition.

GUANGYAN HUANG (Member, IEEE) received the Ph.D. degree in computer science from Victoria University (VU), in 2012. From 2012 to 2013, she was a Research Fellow with VU, where she was a Senior Lecturer in 2014. She was with the Chinese Academy of Sciences, from 2003 to 2009, and visited the Platforms and Devices Centre, Microsoft Research Asia, in the last half of 2006. She is currently an Associate Professor with the School of Information Technology, Deakin University, Australia. She has over 100 publications mainly in data mining, IoT/sensor networks, text analytics, image/video processing, emotion modeling, and multimodal data fusion. She was a recipient of the ARC Discovery Early Career Researcher Awards (DECRA) Fellowship.



YU HU (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1997, 1999, and 2003, respectively. She is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her current research interests include autonomous driving, deep learning, and neural architecture search. She is also a member of the Association for Computing Machinery (ACM), the Institute of Electronics, Information and Communication Engineers (IEICE), and the China Computer Federation (CCF).

XIAOWEI LI (Senior Member, IEEE) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, in 1991. He was an Associate Professor with Peking University, in 1993. He was a Visiting Research Fellow with the University of Hong Kong, from 1997 to 1998, and the Nara Institute of Science and Technology, Japan, in 1999. He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also the Deputy Director of the State Key Laboratory of Computer Architecture. He has over 340 publications mainly in VLSI testing, the IoT/sensor networks, hardware security, and fault-tolerant computing. He also serves as the Vice Chair of the IEEE CS Test Technology Technical Council for the term 2018–2021 and the Chairman of the Supervisory Board of China Computer Federation for the term 2020–2024. He also serves as an Associate Editor for IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on Circuits and Systems—II: Express Briefs, and ACM TODAES.