International Journal Of
Industrial Electronics Control and Optimization
Print ISSN: 3517-2645
Online ISSN: 3568-2645
Homepage: https://ieco.usb.ac.ir/

Micro-Expression Recognition Using the Spatiotemporal Feature Extraction and Deep-Learning Methods

Vida Esmaeili[1] | Mahmood Mohassel Feghhi[2]

Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran[1,2]
Corresponding author's email: mohasselfeghhi@tabrizu.ac.ir

Article Info

Article type:
Research Article

Article history:
Received: 20-March-2024
Received in revised form: 14-July-2024
Accepted: 23-July-2024
Published online: 21-Dec-2024

Keywords:
Micro-expression recognition, Spatiotemporal hand-crafted, Spatiotemporal deep-learning.

ABSTRACT

The Micro-Expression (ME), which automatically reveals genuine human emotions, has gained significant attention. Recognizing the ME is crucial for many real-time applications. However, there are significant challenges to overcome. For instance, the number of ME frames is limited due to their short duration, and the subtle facial movements can be hard to detect due to their low intensity. These challenges need to be addressed to improve ME recognition. We propose a novel method for ME recognition in real-time. In this method, first, the apex frame is spotted using the rotated local binary pattern from six planes (RLBPS) and correlation coefficient (CC). Next, three hand-crafted methods such as the multi-color rotated local binary pattern from six planes (MRLBPS), the histograms of directed gradients from six planes (HDGS), and the histogram of image gradient direction from six planes (HIGDS) extract the features from the apex frame and its surrounding frames. Finally, the stacks of features as matrixes are fed into a three-dimensional convolutional neural network (3D-CNN), and the output is the maximum recognition rate by voting three results. The proposed method has shown promising results when compared to most state-of-the-art methods. According to the results, an average precision of 99% has been obtained using our proposed method. The combination of the RLBPS and the CC creates a strong method for spotting the apex frame. Also, feeding the stacks of spatiotemporal features into the 3D-ResNet increases the ME recognition rate in real-time.

I. Introduction

Over the past few years, the Micro-Expression (ME), as an automatic human emotion revelation, has attracted much attention. Recognizing it can provide insight into the emotional state (i.e., happy, scared, anxious, depressed, angry, and surprised). Also, analyzing it is valuable for various applications, such as social interaction, medicine, surveillance, law enforcement, teaching evaluations, psychology, interrogation, national security, etc. [1, 2].

The ME is the brief, spontaneous, subtle, involuntary, and fleeting facial expression that occurs in one-twenty-fifth to one-fifth of a second, which is challenging to detect by the naked eye and computer vision [3]. In other words, the ME is the facial muscle movements reacting to the emotional stimulus [4]. It may occur in high-risk situations when people attempt to conceal, manipulate, hide, or suppress their true feelings [5]. In fact, in these situations, the facial muscles work very cleverly to unveil the appearance of real emotions. Since this muscle action is uncontrollable, the ME can provide the essential clues to identify the criminal and detect lying.

Its association with thoughts and feelings was first declared by Ekman and Friesen in 1969 [6]. They found the MEs while examining videos of the psychiatric patients for lie detection. Indeed, one of the patients had decided to commit suicide. But she feigned to be glad and optimistic. While scrutinizing the patient interview video clip, Ekman surprisingly understood she was willing to hide her decision. He saw that the smile was replaced by a brief sadness. Later, the patient confessed to her lying.

In the following years, Ekman developed the ME Training Tool (METT) [7] to increase the human identification ability.

Training by the MEET could slightly increase the recognition rate, and it was time-consuming [1]. Hence, researchers have endeavored to increase the recognition rate with the artificial intelligence methods [8-10]. In addition, they have tried to reduce the total elapsed time of the process. The hand-crafted methods extract the hand-engineered features relying on prior knowledge and heuristics, and deep-learning methods learn the high-level features automatically from very large datasets [1, 8-21].

The ME research is not limited to the recognition task. The detection and spotting of the apex frame are considered as its study bottleneck. The ME detection aims to capture the frames including the ME, excluding the ordinary facial expressions. Most importantly, the detection of a frame (i.e., the apex) that displays the expressive emotional state and the most facial muscle motions throughout the video sequence could be surprisingly utilized for the recognition task [1].

Still, there are severe challenges, which need to be solved. For example, there are a low number of involuntary public datasets as the ME happens under high-stakes, and collecting it is an arduous task. A scarce number of ME frames due to its short duration makes it also impossible to discover with computer vision. In addition, indiscernible subtle motions are not simple to describe due to the low intensity of facial muscular movements.

To address these challenges, in this study, we propose a novel method for the ME recognition in real-time. In this method, first, the apex frame is spotted using the rotated local binary pattern from six planes (RLBPS) and correlation coefficient (CC). We named this new method the RLBPS-CC. Second, the apex frame and surround frames are converted to a video to add the missed frames using the temporal interpolation model. Third, the subtle movements in the video are magnified, and the faces are aligned. Fourth, three hand-crafted methods such as the multi-color rotated local binary pattern from six planes (MRLBPS), the histograms of directed gradients from six planes (HDGS), and the histogram of image gradient direction from six planes (HIGDS) extract the features. Finally, the stacks of features as matrixes are fed into a three-dimensional convolutional neural network (3D-CNN), and the output is the maximum recognition rate by voting three results. Fig. 1 shows a framework of our proposed method.

To summarize, our main contributions are as follows:

- We propose a novel method by combining the spatiotemporal hand-crafted and deep-learning methods for the ME recognition in real-time.
- We propose a new method called the RLBPS-CC for spotting the apex frame.
- Unlike [13], we use the CC rather than the sum of squared differences (SSD) for the apex spotting. Also, we utilize the temporal interpolation model for the addition of the missed frames, since a low number of ME frames is recorded due to its short duration. In addition, the faces are alignment to remove the differences in the face shapes during the video recording. Furthermore, we add the HDGS and MRLBPS for feature extraction, and the outputs of three feature descriptors (the MRLBPS, HDGS, and HIGDS) are fed to a 3D-CNN for the ME recognition. Moreover, we apply our proposed method to four datasets, and we use the large-scale and balanced dataset (i.e., the CAS(ME)3).
- We have employed the rotated local binary pattern (RLBP) [22] on the three color channels rather than the simple local binary pattern (LBP) [23] and uniform LBP [23], and we also use the RLBP and CC for the apex detection task. Since this method is not only invariant to rotation, but also has a high discriminative power. To our knowledge, the RLBP has not yet been used for color channels feature extraction in six planes.
- In our proposed method, the matrixes obtained from MRLBPS, HDGS, and HIGDS are fed to a 3D-CNN by applying transfer learning. Although this method has more planes than the [2], [9], and [10] works, it does not consider as many planes as the [11], [19], and [17] works. Accordingly, it gives powerful spatiotemporal features to the 3D-CNN.
- Our proposed method increases the ME recognition rate compared to the previous state-of-the-art methods.

The rest of this article is organized as follows. Section II presents the related studies. Section III introduces the proposed method for the ME recognition in real-time. Section IV discusses the achieved experimental results. The conclusion and future directions are discussed in Section V.

II. Related works

In this section, initially, we review recent related works of literature in the field of ME. Then, we glimpse into the LBP on six planes (LBPS) which our proposed method derived from it.

A. Recent related works

Recently, it has been determined that temporal features can

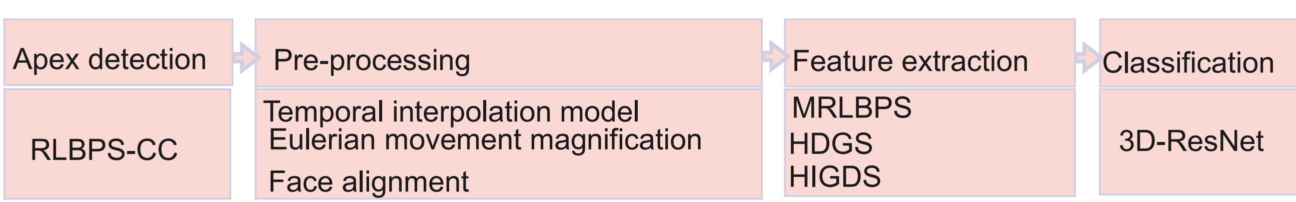

Fig. 1. A framework of our proposed method for the ME recognition in real-time.

reveal subtle movements [1]. In [9], four diagonal planes have been proposed for autonomous apex detection and ME recognition. In this work, these four temporal planes take the sequential images, which contain the magnified micro-motions. The pixels captured from these planes have been encoded using the LBP. The LBP histogram of the first plane is interconnected to the LBP histogram of the other planes for making a feature vector. The first feature vector, which obtains from three sequential images is compared with other feature vectors using the SSD for detecting the apex frame.

If the apex frame is the sixth frame, then, the fifth, sixth, and seventh frames are picked to feature extraction. The introduced feature extractor is a combination of the optical flow and the LBP from the four proposed planes [9]. Finally, the extracted features are classified using a linear support vector machine (LSVM) for the ME recognition.

In [10], the histogram of image gradient orientation (HIGO) method has been applied to the four diagonal planes, which consider sequential face images. Then, its histogram has been computed on each plane. Next, the obtained histograms have been concatenated to achieve the final histogram. Eventually, the final histogram has been fed to the LSVM for the ME recognition task.

In [11], the cubic-LBP feature descriptor has been presented for the first time. This method assumes fifteen planes in all possible directions. These planes can receive all temporal, spatial, and spatiotemporal features. Thus, the cubic-LBP is used for spotting the apex frame and other micro-movements [11, 17, 19]. Nonetheless, calculating the LBP code on the fifteen planes is time-consuming and expensive.

To reduce the processing time, in [19], one plane out of the 15 planes has been automatically selected using the partial differential equations (PDE) and the multivariate ridge regression as a simple linear classifier. This single plane specifies the main direction of the changes. Furthermore, in [17], the cubic-LBP has been made intelligent. In fact, a two-dimensional CNN (2D-CNN) intelligently learns to offer the particular planes for spotting the micro-movements. It identifies the direction/directions of the most movement.

To reduce the computational costs, the studies [13] and [8] choose the six planes with the best results according to several experiments. In [8], the apex frame is detected utilizing the LBP on the six planes and the SSD. Then, it is the input of a fast region-based 2D-CNN to recognize the ME. In addition to using these six planes for apex frame detection, these planes have been employed for feature extraction in [13]. At last, the features extracted from them have been classified by the LSVM.

Nevertheless, the spatiotemporal features can be learned by the spatiotemporal CNN (e.g., 3D-CNN) to lead to better performance. As far as the authors know, this work has not been done to recognize the ME.

A. Review of LBPS

Suppose a trail of images, and consider that they create a 3-dimensional (3D) volume (L, W, T). If $val_{\text{cen.pix.}}$ is corresponding to the gray value of the central pixel of the current frame, its coordinates are ($L_{\text{cen.pix.}}$, $W_{\text{cen.pix.}}$, $T_{\text{cen.pix.}}$). The coordinates of central pixel neighborhood in plane 1, plane 2, plane 3, plane 4, plane 5, and plane 6 are respectively given by:

$$\left(\mathrm{L}_{\text{cen.pix.}} + \cos\left(\frac{2\pi x}{8}\right), W_{\text{cen.pix.}}, T_{\text{cen.pix.}} + \sin\left(\frac{2\pi x}{8}\right)\right) \tag{1}$$

$$\left(L_{\text{cen.pix.}}, W_{\text{cen.pix.}} - \sin\left(\frac{2\pi x}{8}\right), T_{\text{cen.pix.}} + \cos\left(\frac{2\pi x}{8}\right)\right) \tag{2}$$

$$\left(L_{\text{cen.pix.}} + \cos\left(\frac{2\pi x}{8}\right), W_{\text{cen.pix.}} - \sin\left(\frac{2\pi x}{8}\right), T_{\text{cen.pix.}} + \cos\left(\frac{2\pi x}{8}\right)\right) \tag{3}$$

$$\left(L_{\text{cen.pix.}} + \cos\left(\frac{2\pi x}{8}\right), W_{\text{cen.pix.}} - \sin\left(\frac{2\pi x}{8}\right), T_{\text{cen.pix.}} - \cos\left(\frac{2\pi x}{8}\right)\right) \tag{4}$$

$$\left(L_{\text{cen.pix.}} - \cos\left(\frac{2\pi x}{8}\right), W_{\text{cen.pix.}} - \sin\left(\frac{2\pi x}{8}\right), T_{\text{cen.pix.}} - \sin\left(\frac{2\pi x}{8}\right)\right) \tag{5}$$

$$\left(L_{\text{cen.pix.}} + \cos\left(\frac{2\pi x}{8}\right), W_{\text{cen.pix.}} - \sin\left(\frac{2\pi x}{8}\right), T_{\text{cen.pix.}} + \sin\left(\frac{2\pi x}{8}\right)\right) \tag{6}$$

where x is 0,1, ..., 7. When we suppose these six planes on the 3D volume, they have been intersected each other in the central pixel.

Then, the gray value of these neighborhood pixels are taken and compared with the gray value of the central pixel for computing the LBP. If their value is more, 1 is put. Otherwise, 0 is put. At last, all LBP feature vectors obtained from the six planes and concatenated to create LBPS feature vector.

III. Proposed method

This section presents the proposed method for ME recognition in real-time, which employs a combination of spatiotemporal hand-crafted and deep-learning methods. First, the apex frame is spotted using the RLBPS-CC. Then, the spatiotemporal features are extracted using spatiotemporal methods such as the MRLBPS, HDGS, and HIGDS from the apex and its surrounding frames.

Finally, the spatiotemporal features are fed to a spatiotemporal CNN (i.e., 3D-CNN). The flow diagram of the proposed method for the ME recognition is illustrated in Fig. 2.

A. Apex Frame Detection Using the RLBPS-CC

Initially, a video is recorded from the face of a person who wants to conceal her/his real emotions under special conditions. The video is converted into a trail of images, creating a 3D volume that includes the length (L) and width (W) of the images, as well as a time (T) dimension. Then, the spatiotemporal texture features are encoded using the RLBPS. In other words, the planes pictured in Fig. 3 take the pixels of the images in the spatiotemporal directions, and the RLBP operator codes them. Thus, the spatiotemporal directional changes are coded.

Clearly, the LBP weights are circularly shifting. Then, the maximum magnitude of the difference between neighbouring pixel's value ($\mathrm{val}_{\mathrm{neigh.pix.}}$) and the central pixel value ($\mathrm{val}_{\mathrm{cen.pix.}}$) in each plane is computed to find the dominant orientation (DO) as a reference. Next, the RLBP is calculated on each plane (see Equation 7). Finally, the RLBPS feature vector can be derived as follows:

$$\mathrm{DO} = \max\left|val_{neigh.pix.} - val_{cen.pix.}\right|$$
$$\mathrm{RLBP} = \sum \mathrm{sign}\left(\mathrm{val}_{\mathrm{neigh.pix.}} - \mathrm{val}_{\mathrm{cen.pix.}}\right) \times 2^{\mathrm{mod}(|\Lambda-\mathrm{DO}|,\Theta)}$$
$$\mathrm{RLBPS} = \sum v\{f(L,W,T) = \alpha\};$$
$$\alpha = 0,\dots,l_p - 1; p = 0,\dots.5;$$
$$v(\beta) = \begin{cases} 0, if\ \beta\ is\ false \\ 1, if\ \beta\ is\ true \end{cases} \tag{7}$$

where the Λ is the neighbouring pixel's number, the Θ is the whole number of the neighbours, the *f(.)* is the obtained RLBP code, and the different numbers of its labels are defined by the l_p.

We can find the difference between the feature vectors of the various frames using the CC. The frame with the most difference in feature vector is considered the apex. The CC is calculated by:

$$\mathrm{CC} = \frac{\sum \mathrm{RLBPS_N} \times \mathrm{RLBPS_O}}{\sqrt{\sum \mathrm{RLBPS_N^2} \times \sum \mathrm{RLBPS_O^2}}} \tag{8}$$

where the $\mathrm{RLBPS_N}$ is the feature vector of the neutral face frame, and the $\mathrm{RLBPS_O}$ represents the feature vector of other frames.

The process of apex frame detection using the RLBPS-CC is shown in Fig. 4. First, a video is recorded from a person's face to create a 3D volume. Six planes (see Fig. 3) are applied to the trail images of the 3D volume. The RLBP code is calculated on each plane. Here, the DO is 32 – 25 = 7, and rotation is not needed. Values above threshold and the high weights corresponding to the thresholded neighbors are shown in orange color. Then, the RLBP map and its histogram are obtained from each plane. The RLBP histograms of six planes are concatenated to obtain the RLBPS feature vector. Finally, all histograms achieved from the images are compared using the CC for spotting the apex frame.

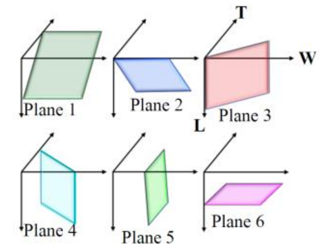

Fig. 3. The 6 planes.

B. Pre-processing

The apex frame, which was determined from section A, along with its surrounding frames are captured as a clip. The duration of ME short video clips should be increased and missed frames added using the temporal interpolation model [24] based on the Laplacian matrix. In this model, the video clip is a trail of images selected along a continuous curve, and the mapping between these frames has been learned to connect them. As a result, using this continuous-time function, it can generate by interpolation the frames that were not seen before at various time points (Fig. 5).

The next step is to import the video clip with more frames into a motion magnification method. We use the Eulerian movement

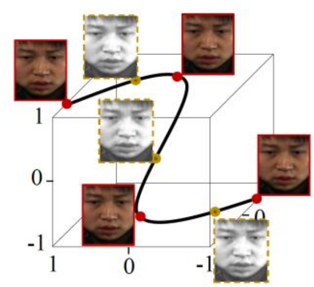

Fig. 5. The temporal interpolation model for generating the missed frames.

magnification [25]. In this method, the Laplacian pyramid structure is utilized for separating the different frequency bands and applying a bandpass filter on them. The resulting signal is then amplified by a η factor. Finally, the original signal is added back to the amplified one to reconstruct the video. Mathematically, the movement magnification is [25]:

$$I(L,W,T+1) = I(L,W,T) + \sum \eta \frac{\partial I(L,W,T)}{\partial T} \tag{9}$$

where motion intensity has been shown by *I*. The process of the Eulerian movement magnification is shown in Fig. 6.

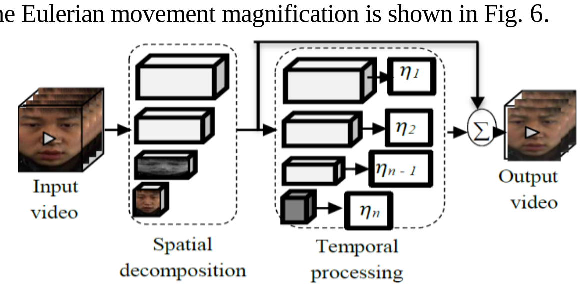

Fig. 6. The process of the Eulerian movement magnification.

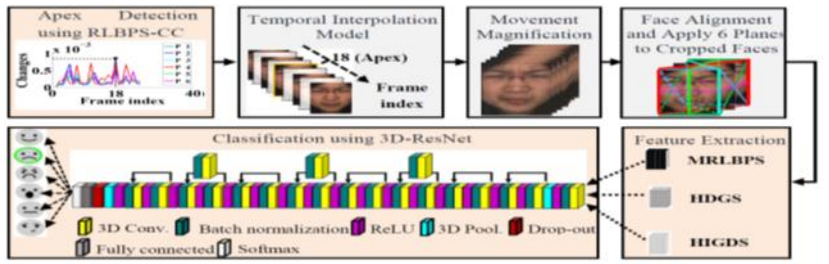

Fig. 2. The flow diagram of the proposed method for the ME recognition in real-time.

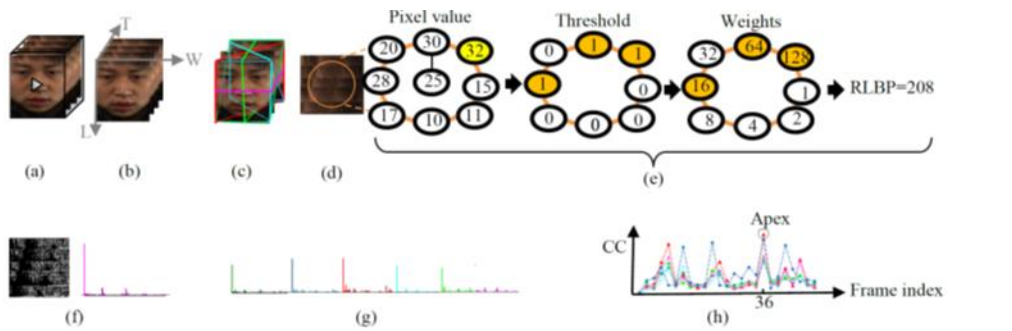

Fig. 4. Apex frame detection using the RLBPS-CC. (a) Video; (b) Creating a 3D volume using the trail of images; (c) Applying the six planes on the images of the 3D volume; (d) A sample of applying plane 6 to the trail of images; (e) Calculating the RLBP code on the plane 6; (f) The RLBP map (left) and its histogram (right) obtained from the plane 6; (g) Concatenating the RLBP histograms of six planes (obtaining RLBPS feature vector); (h) Comparing histograms achieved from all images of the trail (comparing RLBPS feature vectors) for spotting apex frame

Since the face shapes may be changed during the recording video, the face alignment should be performed. For the face alignment, the facial landmarks (with the *C* coordinates) of the front face (*F*) are detected by the discriminative response maps fitting (DRMF) [26]. The next frames (*R*) are reshaped using the local weighted mean (*Y*) for computing the transformation matrix (*M*) to register the face. Finally, the face regions are cropped according to the coordinates of the eyes in the first frame. Mathematically, the face alignment is as follows [2]:

$$M = Y\big(C(R), C(F)\big)$$

$$\mathrm{NR} = F \times M; \qquad (10)$$

where the NR is the normalized transformation frame. A sample of applying the DRMF on a face is illustrated in Fig. 7.

Fig. 7. A sample of applying the DRMF on a face.

B. Feature Extraction

In this step, the MRLBPS, HDGS, and HIGDS extract the features of the cropped face areas from the apex frame and its surrounding frames. The MRLBPS is applied to the color images to use their three channels (e.g., red, green, and blue). In this method, the RLBP (see Equation 7) separately codes the features of three channels on each of the six planes (Fig. 3). The results obtained from each channel are concatenated together. All achieved features from each plane are put into a matrix (Equation 11). This matrix represents the feature vector of the MRLBPS.

$$\mathrm{MRLBPS} = \begin{bmatrix} \mathrm{RedChannelRLBP1} & \mathrm{GreenChannelRLBP1} & \mathrm{BlueChannelRLBP1} \\ \mathrm{RedChannelRLBP2} & \mathrm{GreenChannelRLBP2} & \mathrm{BlueChannelRLBP2} \\ \mathrm{RedChannelRLBP3} & \mathrm{GreenChannelRLBP3} & \mathrm{BlueChannelRLBP3} \\ \mathrm{RedChannelRLBP4} & \mathrm{GreenChannelRLBP4} & \mathrm{BlueChannelRLBP4} \\ \mathrm{RedChannelRLBP5} & \mathrm{GreenChannelRLBP5} & \mathrm{BlueChannelRLBP5} \\ \mathrm{RedChannelRLBP6} & \mathrm{GreenChannelRLBP6} & \mathrm{BlueChannelRLBP6} \end{bmatrix} \qquad (11)$$

In the second feature descriptor (i.e., the HDGS), the Sobel operator with 90 and 0 direction masks as [-1 0 1] moves on the pixels in each plane to achieve the vertical and the horizontal derivatives (the D_L and D_W). Then, the gradient magnitude and its direction are calculated by

$$\text{Gradient magnitude} = \sqrt{{D_L}^2 + {D_W}^2}$$

$$\text{Gradient direction} = \tan^{-1}\frac{D_L}{D_W} \qquad (12)$$

To compute the feature vector, the pixels have the weighted voting for the direction channels based on the gradient calculation response. In the final feature vector, the histograms (the feature vectors) of the six planes are concatenated. Fig. 8 represents operating the HDGS.

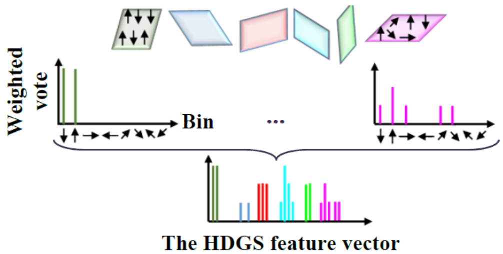

Fig. 8. The HDGS method.

The third feature extractor (i.e., the HIGDS) is applied to decrease the influence of the illumination and contrast. Although this method is similar to the HDGS, it utilizes the unweighted voting, since it gives up the gradient magnitude. The unweighted voting means a fixed amount of votes for each seen direction in the plane. Therefore, the gradient direction is independent of the brightness level of the pixel.

C. Classification for ME Recognition

The feature vectors obtained in the form of a 3D volume of matrices (224×224×224) are fed to a 3D-CNN (i.e., 3D-ResNet) by applying transfer learning. The core idea behind using the 3D-ResNet is the shortcut connections to resolve the vanishing gradient problem. In addition, we have assumed that the ResNet, which has already performed well on the ImageNet dataset, can be successfully applied to our achieved feature vectors as it

could be generalized to our ME recognition task.

We follow a transfer learning approach for the 3D-CNN architecture, since it is not only optimized, but also resolves the issues of computational expensiveness and an insufficient number of samples. To avoid over-fitting in the training process, the pre-trained model and its learned parameters such as weight and bias are recommended [20].

To build the 3D-ResNet, 2D filters were extended in the third dimension to have 3D filters. Also, the 3D-ResNet includes the 3D-pooling and 3D-convolution layers. The depth of this model is eighteen, and it has a total of 71 layers. The architecture of the 3D-ResNet has been illustrated in Fig. 2. Finally, the ME classification was performed by voting.

IV. Experimental results

In this section, the experiments and achieved results are discussed. First, the used datasets containing the CASME, CASME II, CAS(ME)^3, and SMIC-NIR are explained. Then, implementation detail is brought to provide more details about the hyper-parameter setting, data augmentation, and the hardware and software. Finally, the experimental results and discussion will be expressed.

Used datasets

We employ four spontaneous and publicly available datasets. The first dataset, the CASME [27], has been tagged with the offset, onset, and apex frames. This dataset includes 195 samples of MEs at 60 fps with 1280 × 720 and 640 × 480 resolutions from 22 males and 13 females. The CASME II [28] contains 247 samples recording with a frame rate of two hundred fps from only Chinese participants. It consists of five classes of MEs. The resolution of the data is 280 × 240. The third is a large-scale dataset: the CAS(ME)^3 [29]. It offers 1,109 ME frames, which have been manually labeled. The CASME, CASME II, and CAS(ME)^3 have RGB images.

The fourth dataset is the SMIC-NIR [2], which is very close to the real situation. It contains 71 samples elicited from eight Caucasians and eight Asians participants (10 males and 6 females). In this dataset, the infrared data with a 640 × 480 resolution have been produced using a near-infrared camera. We employ this dataset for extracting the features using the HIGDS. A sample of the used datasets is shown in Fig. 9. Table 1 illustrates the summary of these datasets.

Fig. 9. A sample of the used datasets.

TABLE 1
COMPARING THE SPECIFICATIONS OF USED DATASETS

Dataset	The number of Ethnicities	Emotion Classes	Frame rate (fps)	The ME samples
SMIC-NIR	2	3	25	71
CASME II	1	5	**200**	247
CASME	1	**7**	60	195
CAS(ME)^3	1	**7**	30	**1,109**

A. Implementation detail

Our method is implemented by MATLAB 2020 using a 3.5GHZ Intel Core i7 Duo processor. The experiments for training the model use Nvidia GTX950 GPU accelerator 8 GB memory. The hyper-parameter setting in the training step are listed in Table 2. To reduce the class imbalance and data augmentation, we have used rotation, flipping, and mirroring.

TABLE 2
TRAINING HYPER-PARAMETERS

Hyper-parameter	Value
Input size	224×224×224
Learning rate	0.0003
L2 regularization	0.0005
Max epoch	30
Bias/weight learning rate factor	10
Mini-batch size	8

B. Experiments, results and discussion

Initially, a video clip is recorded from the face of a person who wants to conceal her/his real emotions under special conditions. The video is converted into sequential images, which create a 3D volume. Then, the spatiotemporal texture features are encoded using the RLBPS. Next, the obtained feature vectors are compared using the CC, and the apex frame with a maximum changes is spotted using the RLBPS-CC. Fig. 10 shows a sample of the apex frame spotting using our RLBPS-CC method. In this sample (i.e., CASME\sub07\EP08_2), the frame index of the ground-truth apex is 11. Fig. 11 and Fig. 12 illustrate the errors of the apex frame spotting using our RLBPS-CC method and other previous methods.

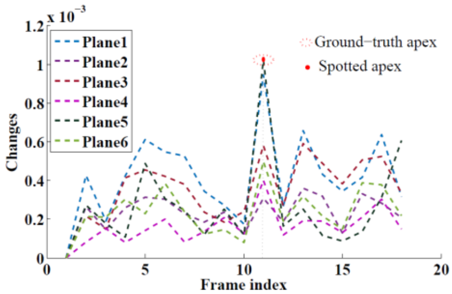

Fig. 10. The apex frame spotting using our RLBPS-CC method.

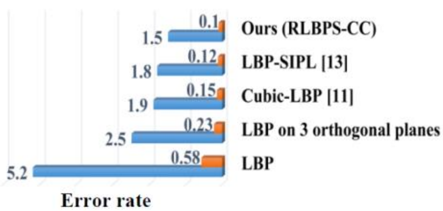

Fig. 11. The standard error (shown in orange color) and the mean absolute error (illustrated in blue color) for spotting the apex on the CASME.

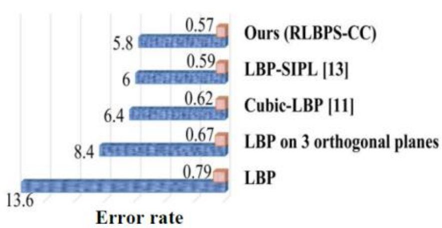

Fig. 12. The standard error (depicted in orange color) and the mean absolute error (pictured in blue color) for spotting the apex on the CASME II.

The apex frame along with its surrounding frames are captured as a clip video. The missed frames are added using the temporal interpolation model, and the sequential frames as a clip are imported into the Eulerian movement magnification method. The best motion magnification factor (η) based on several experiments is 8 (Fig. 13). According to the results, magnifying movements with a suitable factor improves the ME recognition performance. As seen in Fig. 13, excessive amplification reduces accuracy due to strengthening unwanted movements and artifacts, since these cases will hide the micro-movements in the MEs.

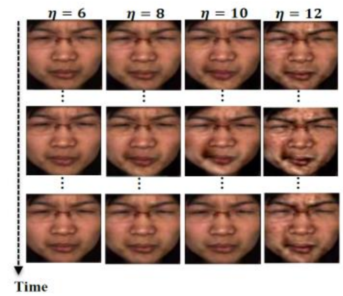

Fig. 13. Movement magnification with several η factors.

Since the face shapes may be changed, the face alignment is performed. For this task, the landmarks are determined using the DRMF. After the face alignment, the face regions are cropped according to the coordinates of the eyes in the first frame. Then, the spatiotemporal features of the cropped face areas are extracted using spatiotemporal methods such as the MRLBPS, HDGS, and HIGDS from the apex and its surrounding frames. The MRLBPS is applied to the color images of the CASME, CASME II, and CAS(ME)^3 to use their three channels (e.g., red, green, and blue). In this method, the RLBP separately codes the features of three channels on each of the six planes. The results obtained from each channel are concatenated together. All achieved features from each plane are put into a matrix, which represents the feature vector of the MRLBPS.

In the HDGS, the Sobel operator with 90 and 0 direction masks as [-1 0 1] moves on the pixels in each plane to achieve the vertical and the horizontal derivatives. To compute the feature vector, each pixel has a weighted vote for the direction channel, according to the gradient calculation response. In the final feature vector, the feature vectors of the six planes are concatenated.

The HIGDS ignores the gradient magnitude, and it utilizes the simple vote. This method is applied to the SMIC-NIR images.

The spatiotemporal feature vectors as a 3D volume of matrices (224×224×224) are fed to the pre-trained 3D-ResNet by applying transfer learning for the ME classification task. The fully-connected layer has been changed to adapt our work. Training the 3D-ResNet requires several hyper-parameters to be used in the back-propagation learning process. The optimization solver is an important parameter. We utilize the stochastic gradient descent (SGD) solver in the training process. It minimizes the loss function by moving in the direction of the negative gradient at each iteration. The other hyper-parameters are the learning rate, L2 regularization, mini-batch size, and drop-out. We set the L2 regularization=0.0005 and the drop-out=0. Stride is 1. Different values of the learning rate and the mini-batch size are implemented, and the best one is chosen. The obtained results are pictured in Fig. 14.

Finally, the best ME recognition rate is reported by voting. It is worth mentioning that we use the 10-fold cross-validation strategy, and we report the average of folds. Table 3 details the results for the ME recognition using our proposed method, which shows improvements over other methods.

Since the ME is more effectively revealed in a sequence of frames than in a single frame, it is best to use multiple frames. Our feature extraction methods can reveal changes in the previous and next frames using the six temporal planes. The six introduced planes code the changes over time in different directions. These planes have pixels, which carry valuable and impactful information. Moreover, the spatiotemporal features can be quickly classified with a high-accuracy rate using a 3D-CNN (i.e., the 3D-ResNet) compared to the LSVM.

The recognition rate of the ME using the feature extraction

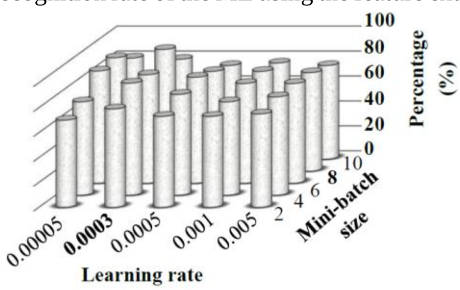

Fig. 14. The obtained results from different values of the learning rate and the mini-batch size.

methods on the different planes (e.g., the 3 orthogonal planes, the 4 diagonal planes, and the 6 introduced planes) has been summarized in Fig. 15. As we have seen, the six planes (shown in Fig. 3) produce the best result. In addition, Fig. 16 shows the elapsed time for training of different 3D-CNN models.

TABLE 3

THE RESULTS FOR ME RECOGNITION ON THE COMPOSITE DATASETS. BOLD TEXT SHOWS THE BEST RESULT. UNDERLINE TEXT INDICATES SECOND-BEST PERFORMANCE.

Method	Accuracy (%)
HIGD on three orthogonal planes + LSVM [2]	55.9
LBP on three orthogonal planes + LSVM [2]	55.9
Dual-stream combining optical flow and dynamic image CNNs [30]	73.1
Combination of optical flow and LBP on diagonal planes [9]	79.6
HIGO on diagonal planes + LSVM [10]	82.6
HIGDS + LSVM [13]	83.7
Multi-color LBP on six intersection planes + LSVM [13]	86.2
µ-Pre-training of Deep Bidirectional Transformers (Micron-BERT) [31]	89.1
DenseNet + Efficient Channel Attention (ECA) [32]	<u>95.76</u>
Ours	**97.6**

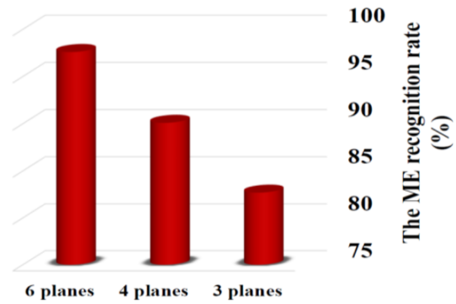

Fig. 15. The ME recognition rate using the feature extraction methods on the different planes.

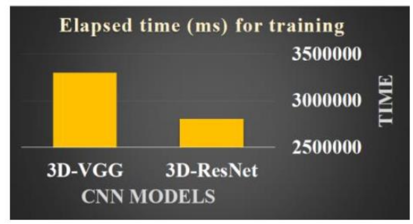

Fig. 16. The elapsed time (ms) for training of two different 3D-CNN models.

A. Experimental results using the evaluation metrics

In another experiment, the proposed method's usefulness is measured using the precision, F1-score, UF1, specificity, and sensitivity metrics [33]. These evaluation metrics can computed using the mathematical formulas in Table 4. Table 5 represents the achieved results of cross-validation using the proposed method. According to the results, an average precision of 99% has been obtained using our proposed method. Also, the average of other metrics is 97%. In addition, Table 6, Table 7, Table 8, and Table 9 compare the results of ours and other state-of-the-art methods on different datasets. The results show superior performance of our proposed method in comparison with others.

The possible reasons for increasing the ME recognition rate could be that (1) Unlike LBP, the RLBP circularly moves the weights to obtain the dominant direction; (2) Computing the RLBP on three channels of RGB images can provide the meaningful features; (3) Feature extraction on six small and temporal planes using the MRLBPS, HDGS, and HIGDS can reveal micro variants and changes; (4) Feeding the extracted temporal features to a 3D-ResNet can increase the ME recognition rate. Unlike 3D-VGG, the 3D-ResNet have advantages include requiring the feature learning only once, solving vanish/detonate gradients problem, reducing the training process time.

TABLE 4

THE EVALUATION METRICS

Name	Formula
Sensitivity	(true positives)/(false negatives+true positives)
Specificity	(true negatives)/(true negatives+ false positives)
Precision	(true positives)/(false positives+true positives)
F1-score	2×((Precision×Sensitivity)/(Precision+Sensitivity))
UF1	2×(∑ ((Precision×Sensitivity)/(Precision+Sensitivity)) /the number of classes)

TABLE 5
THE ACHIEVED RESULTS OF CROSS-VALIDATION USING THE PROPOSED METHOD ON THE COMPOSITE DATASETS

Fold	UF1	F1-score	precision	specificity	sensitivity
1	97%	98%	98%	98%	98%
2	96%	97%	99%	97%	97%
3	98%	98%	98%	98%	98%
4	96%	97%	99%	97%	95%
5	97%	96%	100%	98%	96%
6	97%	97%	99%	97%	97%
7	96%	97%	98%	96%	97%
8	98%	98%	99%	95%	98%
9	97%	98%	100%	99%	98%
10	98%	97%	99%	98%	97%
Average	97%	97%	99%	97%	97%

It is noteworthy that calculating the RLBP is time-consuming compared to the LBP. Also, feature extraction on six planes has computation complexity. Thus, the computational complexity of the proposed method is more than implementation a typical method or feeding raw images to a 3D-CNN.

TABLE 6
THE RESULTS FOR ME RECOGNITION IN TERM OF UF1 ON THE SMIC-NIR AND CASME II DATASETS. BOLD TEXT SHOWS THE BEST RESULT. UNDERLINE TEXT INDICATES SECOND-BEST PERFORMANCE.

Method	SMIC-NIR	CASME II
	UF1 (%)	UF1 (%)
Bi-Weighted Oriented Optical Flow (Bi-WOOF) [34]	57.27	78.05
Optical Flow Features from Apex frame Network (OFF-ApexNet) [35]	68.17	87.64
Shallow Triple Stream Three-dimensional CNN (STSTNet) [36]	68.01	83.82
Lightweight Apex-based Enhanced Network (LAENet) [37]	66.2	91.01
Hierarchical Transformer Network (HTNet) [38]	89	95.32
Ours	**96.8**	**97.5**

TABLE 7
THE RESULTS FOR ME RECOGNITION IN TERMS OF F1-SCORE AND ACCURACY ON THE SMIC-NIR AND CASME II DATASETS. BOLD TEXT SHOWS THE BEST RESULT. UNDERLINE TEXT INDICATES SECOND-BEST PERFORMANCE.

Method	SMIC-NIR		CASME II	
	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)
Key Facial Components (KFC) [39]	78	77	73.75	72.76
Attention-based magnification-adaptive networks [40]	77	79.87	71	75.4
Shallow Triple Stream Three-dimensional CNN (STSTNet) [36]	68.01	70.13	83.82	86.86
Multi-scale 3D-ResNet50 [41]	75.1	74.6	88.9	91.35
Local Binary Pattern on Six Intersection Planes (LBP-SIP*l*) and Fast Region-based Convolutional Neural Network (FR-CNN) [8]	92	91.2	95.66	96.11
Ours	**97**	**97.4**	**97.5**	**97.8**

TABLE 8
THE RESULTS FOR ME RECOGNITION IN TERMS OF F1-SCORE AND ACCURACY ON THE CASME DATASETS. BOLD TEXT SHOWS THE BEST RESULT. UNDERLINE TEXT INDICATES SECOND-BEST PERFORMANCE.

Method	CASME	
	F1-score (%)	Accuracy (%)
Three-stream CNN (TSCNN) [42]	72.7	73.88
Knowledge distillation [43]	77	81.8
Dual-stream SpatioTemporal Attention Network (DSTAN) [44]	75	78
Local Binary Pattern on Six Intersection Planes (LBP-SIP*l*) and Fast Region-based Convolutional Neural Network (FR-CNN) [8]	80	81.56
Intelligent cubic-LBP [17]	93	93
Ours	**97.1**	**97.5**

TABLE 9
THE RESULTS FOR ME RECOGNITION IN TERMS OF F1-SCORE AND ACCURACY ON THE $CAS(ME)^3$ DATASETS. BOLD TEXT SHOWS THE BEST RESULT. UNDERLINE TEXT INDICATES SECOND-BEST PERFORMANCE.

Method	$CAS(ME)^3$
	UF1 (%)
Shallow Triple Stream Three-dimensional CNN (STSTNet) [36]	37.95
μ-Pre-training of Deep Bidirectional Transformers (Micron-BERT) [31]	56.04
Three-stream temporal-shift attention network based on self-knowledge distillation (SKD-TSTSAN) [45]	86.48
Ours	**97**

V. Conclusions

In this paper, a novel method called the RLBPS-CC has been proposed for the apex frame spotting. Also, a new method has been suggested for the ME recognition in real-time. In this method, three hand-crafted methods (i.e., the MRLBPS, HDGS, and the HIGDS) extract the features from the detected apex frame and its surrounding frames. Then, the stacks of the spatiotemporal features are fed into the 3D-ResNet, and the output is the maximum recognition rate by voting three results. The inference processing time took approximately 0.9 seconds. Based on the obtained results, the proposed methods show promising outcomes compared to most state-of-the-art techniques. In the future, this method could be used in other applications where subtle variations in features are required to be spotted.

ACKNOWLEDGMENT

This research was supported by Research grant of the University of Tabriz (number و/۲۷/۹۸۰).

REFERENCES

[1] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "A comprehensive survey on facial micro-expression: approaches and databases," Multimedia Tools and Applications, Vol. 81, No. 28, pp. 40089-40134, 2022.

[2] X. Li et al., "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," IEEE transactions on affective computing, Vol. 9, No. 4, pp. 563-577, 2017.

[3] G. Zhou et al., "Micro-expression action unit recognition based on dynamic image and spatial pyramid," The Journal of Supercomputing, pp. 1-24, 2023.

[4] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," IEEE Transactions on Affective Computing, 2022.

[5] G. Zhao, X. Li, Y. Li, and M. Pietikäinen, "Facial Micro-Expressions: An Overview," Proceedings of the IEEE, 2023.

[6] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," Psychiatry, Vol. 32, No. 1, pp. 88-106, 1969.

[7] P. Ekman, Micro expressions training tool Emotionsrevealed. com, 2003.

[8] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "Automatic Micro-Expression Recognition Using LBP-SIPl and FR-CNN," AUT Journal of Modeling and Simulation, Vol. 54, No. 1, pp. 59-72, 2022.

[9] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "Autonomous apex detection and Micro-expression recognition using proposed diagonal Planes," International Journal of Nonlinear Analysis and Applications, Vol. 11, pp. 483-497, 2020.

[10] V. Esmaeili, M. M. Feghhi, and S. O. Shahdi, "Micro-expression recognition using histogram of image gradient orientation on diagonal planes," in 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2021: IEEE, pp. 1-5.

[11] V. Esmaeili and S. O. Shahdi, "Automatic micro-expression apex spotting using Cubic-LBP," Multimedia Tools and Applications, Vol. 79, pp. 20221-20239, 2020.

[12] V. Esmaeili, M. M. Feghhi, and S. O. Shahdi, "Automatic micro-expression apex frame spotting using local binary pattern from six intersection planes," arXiv preprint arXiv:2104.02149, 2021.

[13] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "Micro-Expression Recognition based on the Multi-Color ULBP and Histogram of Gradient Direction from Six Intersection Planes," Journal of Iranian Association of Electrical and Electronics Engineers, Vol. 19, No. 3, pp. 123-130, 2022.

[14] V. Esmaeili, M. M. Feghhi, and S. O. Shahdi, "Early COVID-19 Diagnosis from Lung Ultrasound Images Combining RIULBP-TP and 3D-DenseNet," in 2022 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 2022: IEEE, pp. 1-5.

[15] V. Esmaeili and M. Mohassel Feghhi, "Real-time Authentication for Electronic Service Applicants using a Method Based on Two-Stream 3D Deep Learning," Soft Computing Journal, 2023.

[16] V. Esmaeili and M. Mohassel Feghhi, "Diagnosis of Covid-19 Disease by Combining Hand-crafted and Deep-learning Methods on Ultrasound Data," Journal of Machine Vision and Image Processing, Vol. 9, No. 4, pp. 31-41, 2022.

[17] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "Spotting micro - movements in image sequence by introducing intelligent cubic - LBP," IET Image Processing, Vol. 16, No. 14, pp. 3814-3830, 2022.

[18] V. Esmaeili and M. M. Feghhi, "COVID-19 Diagnosis: ULBPFP-Net Approach on Lung Ultrasound Data," Iranian Journal of Electrical & Electronic Engineering, Vol. 19, No. 3, 2023.

[19] V. Esmaeili, M. Mohassel Feghhi, and S. Shahdi, "Applying Partial Differential Equations on Cubic Uniform Local Binary Pattern to Reveal Micro-Changes," Journal of Electrical and Computer Engineering Innovations (JECEI), pp. 259-270, 2023.

[20] A. Ebrahimi, S. Luo, and R. Chiong, "Introducing transfer learning to 3D ResNet-18 for Alzheimer’s disease detection on MRI images," in 2020 35th international conference on image and vision computing New Zealand (IVCNZ), 2020: IEEE, pp. 1-6.

[21] Z. Raisi and J. Zelek, "Investigation of Deep Learning Optimization Algorithms in Scene Text Detection," International Journal of Industrial Electronics Control and Optimization, Vol. 6, No. 3, pp. 171-182, 2023.
and Optimization, Vol. 6, No. 3, pp. 171-182, 2023.

[22] R. Mehta and K. Egiazarian, "Rotated local binary pattern (RLBP): rotation invariant texture descriptor," in 2nd International Conference on Pattern Recognition Applications and Methods, ICPRAM 2013, Barcelona, Spain, 15.-18.2. 2013, 2013, pp. 497-502.

[23] T. Ojala, M. Pietikäinen, and T. Mäenpää, "A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification," in Advances in Pattern Recognition—ICAPR 2001: Second International Conference Rio de Janeiro, Brazil, March 11–14, 2001 Proceedings 2, 2001: Springer, pp. 399-408.

[24] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in 2011 international conference on computer vision, 2011: IEEE, pp. 1449-1456.

[25] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," ACM transactions on graphics (TOG), Vol. 31, No. 4, pp. 1-8, 2012.

[26] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3444-3451.

[27] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), 2013: IEEE, pp. 1-7.

[28] W.-J. Yan et al., "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," PloS one, Vol. 9, No. 1, p. e86041, 2014.

[29] J. Li et al., "CAS (ME) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 3, pp. 2782-2800, 2022.

[30] J. Tang, L. Li, M. Tang, and J. Xie, "A novel micro-expression recognition algorithm using dual-stream combining optical flow and dynamic image convolutional neural networks," Signal, Image and Video Processing,

Vol. 17, No. 3, pp. 769-776, 2023.
[31] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, and K. Luu, "Micron-bert: Bert-based facial micro-expression recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1482-1492.
[32] H. Yang, S. Sun, and J. Chen, "Deep Learning-Based Micro-Expression Recognition Algorithm Research," International Journal of Computer Science and Information Technology, Vol. 2, No. 1, pp. 59-70, 2024.
[33] C. Nicholson, "Evaluation metrics for machine learning—accuracy, precision, recall, and F1 defined,"Pathmind.http://pathmind.com/wiki/accuracy-precision-recall-f1, 2019.
[34] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," Signal Processing: Image Communication, Vol. 62, pp. 82-92, 2018.
[35] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "OFF-ApexNet on micro-expression recognition system," Signal Processing: I
mage Communication, Vol. 74, pp. 129-139, 2019
[36] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), 2019: IEEE, pp. 1-5.
[37] Y. Gan, S.-E. Lien, Y.-C. Chiang, and S.-T. Liong, "LAENet for micro-expression recognition," The Visual Computer, Vol. 40, No. 2, pp. 585-599, 2024.
[38] Z. Wang, K. Zhang, W. Luo, and R. Sankaranarayana, "HTNet for micro-expression recognition," arXiv preprint arXiv:2307.14637, 2023.
[39] Y. Su, J. Zhang, J. Liu, and G. Zhai, "Key facial components guided micro-expression recognition based on first & second-order motion," in 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021: IEEE, pp. 1-6.
[40] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, and J. Liu, "A novel micro-expression recognition approach using attention-based magnification-adaptive networks," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: IEEE, pp. 2420-2424.
[41] H. Jin, N. He, Z. Li, and P. Yang, "Micro-expression recognition based on multi-scale 3D residual convolutional neural network," Mathematical Biosciences and Engineering, Vol. 21, No. 4, pp. 5007-5031, 2024.
[42] B. Song et al., "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," Ieee Access, vol. 7, pp. 184537-184551, 2019.
[43] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," IEEE Transactions on Affective Computing, Vol. 13, No. 2, pp. 1037-1043, 2020.
[44] Y. Wang et al., "Micro expression recognition via dual-stream spatiotemporal attention network," Journal of Healthcare Engineering, Vol. 2021, No. 1, 2021.
[45] G. Zhu et al., "SKD-TSTSAN: Three-Stream Temporal-Shift Attention Network Based on Self-Knowledge Distillation for Micro-Expression Recognition," arXiv preprint arXiv:2406.17538, 2024.

Vida Esmaeili received her B.S. degree in electrical engineering from the Azad University of Abhar, Iran, in 2015 and the M.S. degree in electrical engineering from the Azad University of Qazvin, Iran, in 2018. Since 2019, she has been working toward the Ph.D. degree in electrical engineering at the Tabriz University, Iran. Her research interests include the area of image processing, machine learning, data science, and pattern recognition. She has published more than 18 technical papers both at the national and international levels. Moreover, she served as a reviewer for several international journals, such as IEEE transactions on affective computing, Multimedia Systems, Scientific Reports, Visual Computer, Machine Vision and Applications, Signal, Image and Video Processing, etc.

Mahmood Mohassel Feghhi received the B.S. and M.S. degrees (Hons.) in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 2006 and 2009, respectively, and the Ph.D. degree in electrical engineering from the College of Engineering, University of Tehran, Tehran, in 2015. From 2007 to 2016, he was a Senior Design Engineer in communication systems design with several communications industries and Inc. Since 2016, he has been with the Faculty of Electrical and Computer Engineering, University of Tabriz, Iran, where he is currently an Associate Professor. He has published more than 60 technical papers in international journals and conferences in the fields of information theory, wireless communications, signal and image processing, machine learning, data science, 5G/6G cellular networks, the Internet of Things (IoT), scheduling, and optimization . He is the Director-in-Charge of the Journal of Advanced Signal Processing (JASP), and the Executive Manager of the Tabriz Journal of Electrical Engineering (TJEE) . His current research interests include information theory, wireless communication networks, signal processing, machine learning, and optimization.

Dr. Mohassel Feghhi was the Chair of Scientific Committee at the 4th West Asian Symposium on Optical and Millimeter-wave Wireless Communications (WASOWC 2022), and the Scientific Committee Chair of the Communications section at the 28th Iranian Conference on Electrical Eng. (ICEE 2020) .Moreover, he served as a TPC member of several international conferences, and also serves as a reviewer for several international journals, such as IEEE Transactions on Green Communications and Networking, IEEE Transactions on Vehicular Technology (IEEE-TVT), IEEE Systems Journal, IET Communications, Ad Hoc Networks, Journal of the Franklin Institute, etc.