AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 54(1) (2022) 59-72
DOI: 10.22060/miscj.2022.21133.5272

Automatic Micro-Expression Recognition Using LBP-SIPl and FR-CNN

V. Esmaeili[1], M. Mohassel Feghhi[1*], S. O. Shahdi[2]

[1] Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran
[2] Department of Electrical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

ABSTRACT: Facial Expressions are one of the most effective ways for non-verbal communications, which can be expressed as the Micro-Expression (ME) in the high-stake situations. The MEs are involuntary, rapid, subtle, and can reveal real human intentions. However, their feature extraction is very challenging due to their low intensity and very short duration. Although Local Binary Pattern on Three Orthogonal Plane (LBP-TOP) feature extractor is useful for ME analysis, it does not consider essential information. To address this problem, we propose a new feature extractor called Local Binary Pattern from Six Intersection Planes (LBP-SIPl). This method extracts LBP code on Six Intersection Planes, and then combines them. Results show that the proposed method has superior performance in apex frame spotting automatically in comparison with the relevant methods on the CASME I and the CASME II databases. Afterwards, the apex frames are the input of the Fast Region-based Convolutional Neural Network (FR-CNN) to recognize the Facial Expressions. Simulation results show that the ME has been automatically recognized in 81.56% and 96.11% on the CASME I and the CASME II databases by using the proposed method, respectively.

Review History:

Received: Feb. 21, 2022
Revised: Jun. 12, 2022
Accepted: Aug. 07, 2022
Available Online: Oct. 01, 2022

Keywords:

Apex frame spotting
Local Binary Pattern
FR-CNN
Micro-Expression recognition

1- Introduction

The Micro-Expressions (MEs) have received increasing attention in situations where people are motivated to manipulate, conceal, or repress their true feelings, and there is no evidence to prove it [1, 2]. These types of expressions have a wide range of applications, such as fair law enforcement, lie detection, identifying potentially dangerous persons, and more reliable diagnosis in the clinical psychology and psychotherapy [3-5].

In fact, they show the emotion of fear, surprise, happiness, disgust, sadness, or contempt in the human face when people are trying to hide and neutralize them on purpose [6, 7]. Unlike ordinary Facial Expressions (FEs), MEs have a very brief, rapid, subtle, and involuntary reaction with short time (from 0.04 to 0.33 seconds) [3]. These characteristics make it challenging, not only by the unaided eye but also with tools in machine vision system [8].

Haggard and Isaacs [9] discovered micro-momentary Facial Expressions as repressed emotions, while scanning motion picture psychotherapy films for the first time in 1966. Three years later, Ekman et al. [10] used the phrase MEs instead of the micro-momentary, analyzing a video of a psychiatric patient interview who tried to hide commiting suicide from her psychiatrist. The patient pretended to be happy and optimistic throughout the recording, a fleeting look of anguish and despair that lasted for merely two frames (0.08s) was found, when the tape was examined curiously in the slow motion. Meanwhile, a brief anguish was quickly replaced by a smile when reviewing the video frame by frame. Afterwards, she confessed to lying to her doctor in another counselling session [11]. Thus, Ekman et al. spotted a relationship between the deception or lie and the ME.

Hence, there have been some efforts for ME detection and recognition. Among presented approaches, Local Binary Pattern-based (LBP) methods are more suitable for ME feature extraction [2, 12]. These methods are appearance-based methods that extract facial features from skin texture (i.e., lines around the mouth and the eyes, wrinkles on the chin) [2, 8]. In addition, deep-learning-based methods are powerful to learn the ME features [13-17]. However, their performance will be improved when fed raw data have the most relevant features.

The apex frame has a peak intensity of ME and maximum facial muscle movements throughout a video sequence. Thus, we can separate the FEs from the MEs only by computing the time of the apex frame and the neutral face frame throughout the ME image sequence. Additionally, the apex frame can be utilized to classify the FE. Hence, spotting the apex frame is important for ME recognition and spotting tasks. Since the MEs are subtle and short in intensity and time, the apex frame detecting is challenging.

*Corresponding author's email: mohasselfeghhi@tabrizu.ac.ir

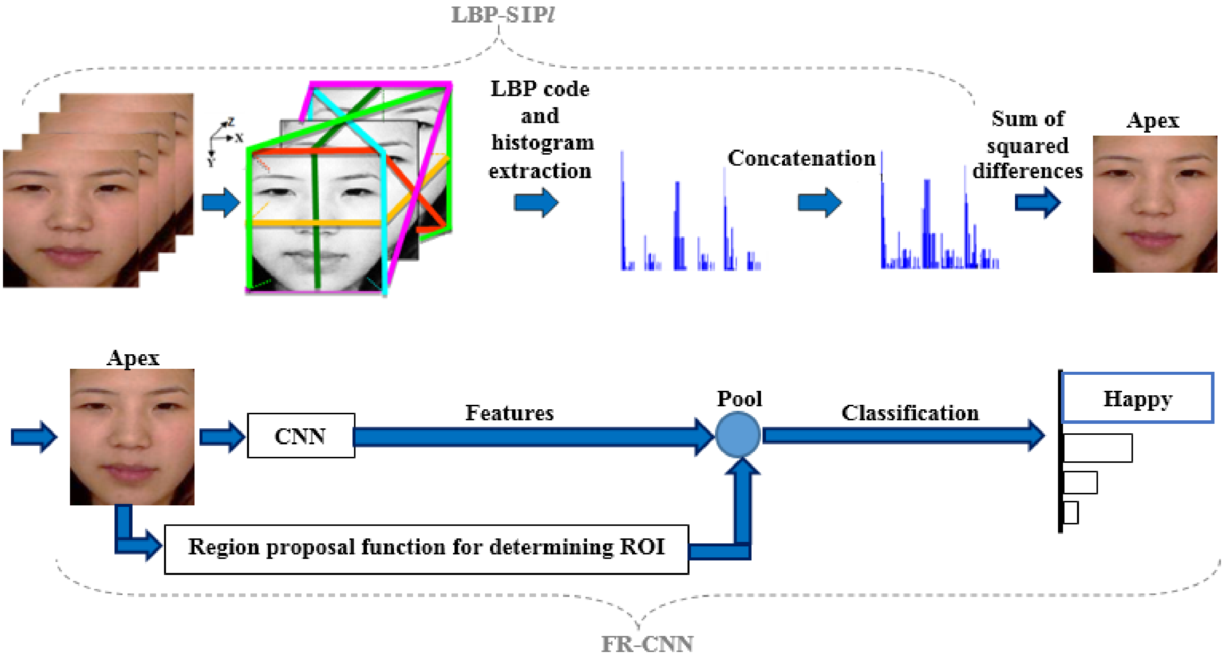
LBP-SIP*l*
LBP code and histogram extraction
Concatenation
Sum of squared differences
Apex
Apex
CNN
Features
Pool
Classification
Happy
Region proposal function for determining ROI
FR-CNN

Fig. 1. Framework for our method.

For this reason, we propose the Local Binary Pattern on Six Intersection Planes (LBP-SIP*l*) to spot the apex in this paper. In fact, the ME features are attained from six temporal and directional planes to determine the apex frame. Next, the apex frames are the input of the Fast Region-based Convolutional Neural Network (FR-CNN) to recognize the Facial Expressions from them. The general framework of the proposed method can be seen in Fig. 1.

The contributions of this paper are as follows:

- We propose the LBP-SIP*l* feature extractor to distinguish the apex frame. Although this method considers more planes than the LBP-TOP [18], the LBP-SIP [19], and the ST-CLQP [8], it does not have as many planes as the Cubic-LBP [20]. Accordingly, it takes major features in the low time.
- We suggest the FR-CNN for the ME recognition in this study, applying transfer learning. Since MEs occur in some local regions, the FR-CNN improves the recognition rate.
- We discard redundant and irrelevant data. For this purpose, we use the apex frames as raw data that are the input of FR-CNN. Most of the frames of a ME sequence produce minute appearance changes which cause them to seem as neutral. Hence, the apex frames are qualified for facial expression recognition.
- Unlike the work of Esmaeili et al. [21], spotting the apex is done on both CASME I and CASME II databases in this work. Additionally, we add an efficient method to classify ME emotions. In other words, in [21], only the task of the apex frame spotting on one database (CASME I) had been done. In this study, not only the task of spotting the apex frame has been done, but also the task of ME recognition has been performed on two databases (CASME I and CASME II).

The remainder of this paper is subdivided into the following sections: The next section discusses related works. Section III is the proposed method. Section IV presents the findings of the experiment and section V provides the conclusion

.

2- Related works

The LBP on Three Orthogonal Planes (LBP-TOP), introduced in [18], is used to extract the MEs in [22]. Since the LBP-TOP was succeeded in revealing subtle facial movements, Ruiz-Hernandez et al. [12] have employed re-parameterization of Gaussian on it to make more reliable and robust histograms. The robustness to lighting changes and spatial transformations are its two profits, although it has computational complexity, redundant information, and high-dimensional feature sets.

To reduce redundant information, Wang et al. [19] proposed the LBP with Six Intersection Points (LBP-SIP). Unlike the LBP-TOP, the LBP-SIP computes the surrounding points once. It provides not only a lightweight and more compact representation, but also low computation and low elapsed time. Nevertheless, there was not much difference between the LBP-SIP and the LBP-TOP in terms of accuracy [19].

Afterwards, Huang et al. [8] mentioned two problems of the LBP-TOP: 1) Extracting motion and appearance features from the sign of pixels without considering their orientation and magnitude, 2) Using an improper pattern for local structures. Therefore, they proposed the Spatio-Temporal Completed Local Quantization Patterns (STCLQP) [8], which encodes all three components (i.e., orientation, sign, and magnitude). The STCLQP, same as the other methods, extracts the components from only Three Orthogonal Planes without considering other planes.

Afterwards, a method with 12 extra planes compared to the previous planes was introduced to find the apex. It is named as Cubic-LBP [20]. However, the desired results can be achieved by fewer planes.

According to the importance of the apex frame, researchers have tried to find a way to reduce the mean distance of the predicted apex frame from the real apex. In other words, they have tackled to determine the apex frame number exactly. Recently, Liong et al. [1] have obtained the apex frame with a deviation of one frame from the ground-truth apex in a sample. Additionally, Ma et al. [23] have presented the Region Histogram of Oriented Optical Flow (RHOOF) to spot the apex frame automatically. They achieved improvements of 19.04% compared to related previous works in the CASME I database.

However, the Mean Absolute Error (MAE) can be decreased by considering the necessary information. Furthermore, the subtle changes and the maximum motions can appear in the image sequences, the temporal planes could be essential to detect them. In fact, these planes extract temporal features. To this end, the apex frame accurately has been determined using the Cubic-LBP method and the sum of squared difference algorithm in 38% of the CASME I database [20]. However, processing information of 15 planes is a time-consuming task.

In addition to the hand-crafted methods mentioned above, deep-learning methods are more widely used due to their processing speed and efficiency [14, 15, 17, 24, 25]. In recent years, a computationally light network called Shallow Triple Stream Three-dimensional CNN (STSTNet) [14] has been suggested to recognize the ME. It learns high-level features from the horizontal and vertical Optical Flow (OF) fields. The F score has been measured at 73% to recognize the ME, using this method on the CASME II.

Additionally, a Three-Stream CNN (TSCNN) has been introduced [13]. It has a local-spatial stream module and two streams named static-spatial and dynamic-temporal. It was observed that this method can achieve a higher F score (80%) than the STSTNet [14]. Against, three-stream combining 2D and 3D CNN [16] produces a lower F score (71%).

Sometimes, researchers combine the deep-learning and the hand-crafted methods. For example, the deep CNN with the OF has been presented in [15]. It gives 56.60% and 56.94% recognition rates on the CASME I and CASME II. Generally, the invisiblity of the ME in some data makes the weak performance of methods. Thus, it is needed to reduce the redundant and irrelevant data.

3- Proposed Method

In this section, first, the apex frame spotting using the proposed method is presented. Next, the suggested method for ME recognition is explained.

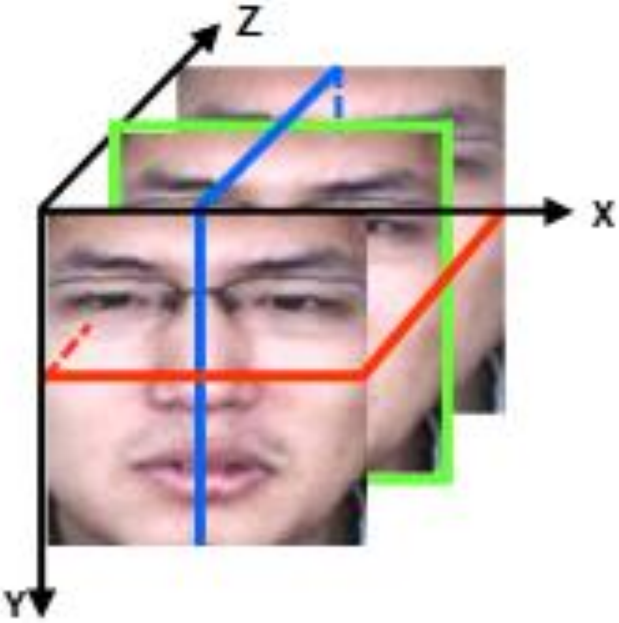

Fig. 2. The XZ, XY, and YZ planes.

3- 1- The Apex Frame Spotting

Initially, we glance into the base of the proposed method. Next, the proposed LBP-SIP*l* will be described completely.

3- 1- 1- A glimpse into the LBP-TOP

The LBP-TOP [18] has 3 orthogonal planes (XZ, XY, YZ), crossed each other in the middle pixel (see Fig. 2). In this method, the sequential images are presumed as the XY planes. Their vertical and horizontal pixels with corresponding locations are put into the YZ and XZ planes, respectively. Afterwards, the LBP code [26] is calculated on each plane. Finally, the achieved histograms from 3 planes are incorporated into a histogram. The mentioned approach generally extracts the appearance and the motion features from 3 planes, but it does not consider further information.

3- 1- 2- The Proposed Method

To overcome the constraints of the LBP-TOP, we suggest the LBP-SIP*l*, which has six planes. In this method, we exploit more meaningful and essential information by combining the histograms of the six temporal planes.

The six planes are containing two temporal planes of the LBP-TOP and four new planes. The four planes show the motions of the facial muscle simultaneously in both vertical and horizontal directions. All six planes intersect in the middle pixel.

Now, consider a sequence of images, and suppose that they make a 3-dimensional volume (*X, Y, Z*). If $g_{tcen,cen}$ is corresponding to the gray value of the central pixel of current frame, its coordinates are (x_{cen}, y_{cen}, z_{cen}). Similarly, the $g_{tcen-L,cen}$ and $g_{tcen+L,cen}$ are the gray value of the central pixel of the previous and the posterior frames, respectively.

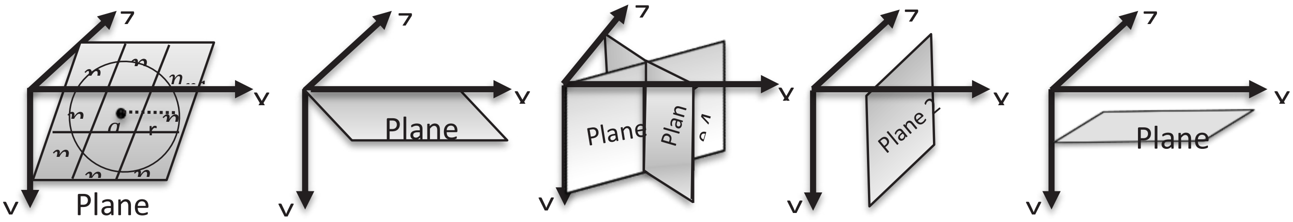

Fig. 3. six planes in the proposed LBP-SIPl.

The coordinates of the $g_{tcen,cen}$ are given by (x_{cen}+$rd_x cos(2\pi nu_p/Nu_P)$, y_{cen}-$rd_y sin(2\pi nu_p/Nu_P)$, z_{cen}), where nu_p ϵ $0, \ldots, Nu_P$-1 is the local neighbouring point around the central pixel. The Nu_P and rd are the number of local neighbouring points and the radius of the circle (distance from central pixel to the neighbour points on the circle), respectively. The average of nearest pixels are computed to estimate a new pixel value, whenever the values of the neighbors do not fall exactly on a circle. Therefore, the coordinates of the neighboring points around the central pixel in planes number 1, 2, …, 6, (Fig. 3) are given by:

$$(x_{cen}+rd_x cos(2\pi nu_p/Nu_P), y_{cen}, z_{cen}+r_z sin(2\pi nu_p/Nu_P)) \quad (1)$$

$$(x_{cen}, y_{cen}-rd_y sin(2\pi nu_p/Nu_P), z_{cen}+rd_z cos(2\pi nu_p/Nu_P)) \quad (2)$$

$$(x_{cen}+rd_x cos(2\pi nu_p/Nu_P), y_{cen}-rd_y sin(2\pi nu_p/Nu_P), z_c+rd_z cos(2\pi nu_p/Nu_P)) \quad (3)$$

$$(x_{cen}+rd_x cos(2\pi nu_p/Nu_P), y_{cen}-rd_y sin(2\pi nu_p/Nu_P), z_{cen}-rd_z cos(2\pi nu_p/Nu_P) \quad (4)$$

$$(x_{cen}-rd_x cos(2\pi nu_p/Nu_P), y_{cen}-rd_y sin(2\pi nu_p/Nu_P), z_{cen}-rd_z sin(2\pi nu_p/Nu_P)) \quad (5)$$

$$(x_{cen}+rd_x cos(2\pi nu_p/Nu_P), y_{cen}-rd_y sin(2\pi nu_p/Nu_P), z_{cen}+rd_z sin(2\pi nu_p/Nu_P)) \quad (6)$$

The LBP code is calculated for each plane (refer to [18] for more details). Subsequently, the histogram of the LBP-SIP*l* can be defined as:

$$ht_{k,j}=\Sigma_{x,y,z}\, I\{f_k(x,y,z)=k\}, \quad k=0,\ldots,n_j-1;\ j=0,1,\ldots,5 \quad (7)$$

where $f_k(x,y,z)$ is the LBP code of central pixel (x,y,z) in the j-th plane, n_j is the number of different labels processed by the LBP operator in the j-th plane and Finally, the histograms are normalized using:

$$I(\beta)=\begin{cases} 1, if & \beta \ \ is \ \ true \\ 0, \ if & \beta \ \ is \ \ false \end{cases} \quad (8)$$

$$Nm_{k,j}=ht_{k,j}/\sum_{m=0}^{n_j-1} ht_{m,j} \quad (9)$$

For apex frame spotting, the normalized histograms should be compared with the neutral frame histogram. Therefore, the frame which contains the most wrinkles and shadow changes is determined as the apex.

To find the difference between two frames, we compute the sum of squared differences of two histograms as follows:

$$\sum_{ii=0}^{n}(h_1-h_{ii})^2 \quad (10)$$

where *h1* is the first frame (neutral face) histogram, and *hii* is the current frame histogram in an image sequence with *n* frames.

3- 2- The Suggested FR-CNN for ME Recognition

In this stage, the spotted apex frames, such as the training data, are organized according to their respective labels (such as sad, happy, fearful, angry, surprise, and disgust). Each emotion has a label. In fact, we create a folder for each emotion to contain the apex frames which show that emotion. To disregard background and etc., each face is cropped. Afterwards, the facial images are resized to the 224x224 size. Next, the whole file as an image data store is utilized for training.

The FR-CNN is used to train ME characteristics. It uses the Edge Boxes algorithm for generating efficient region. First, the input image is processed by this algorithm to generate Region of Interest (ROI) and is also processed by convolutional layers to extract features. Afterwards, a ROI pooling combines convoluted features with the region for the classification of emotions. The rectangular ROI is a cropped image in the apex frame, too. Thus, the region has a label of an emotion. Hence, using the FR-CNN results in rapid facial emotion recognition.

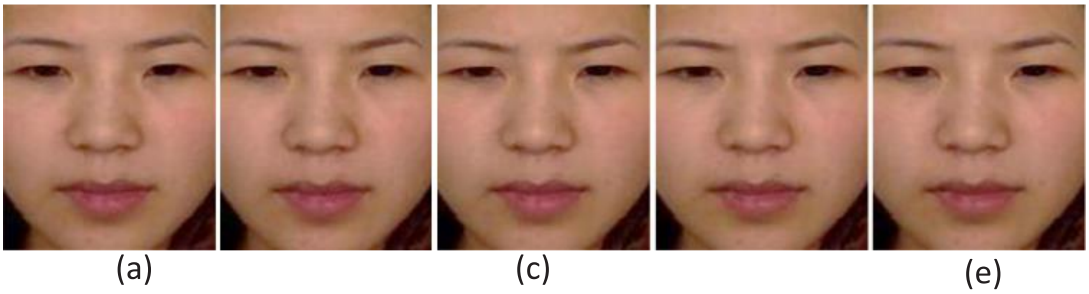

Fig. 4. A sample of the CASME I database. (a) Onset, (c) Apex, and (e) Offset.

By running through the vast numbers of experiments, we have picked the FR-CNN based on VGG-16. According to different experiments, it has the best results with a small number of data. In fact, it is a pre-trained CNN which exerts transfer learning technique. Hence, the last three layers of the model are exchanged with new layers, but the weights and layers are kept intact. This way, the FR-CNN would be able to generalize and converge by training with relatively small number of labeled data.

4- Experiments and Results

The details of the experimentation and analysis are given in this section. The implementation and performance of the proposed method and its comparison with other state-of-art methods are also provided. It is noteworthy that the implementation of the proposed method is performed with MATLAB 2020b, using an Intel Core i7 Duo processor onboard.

4- 1- Databases

We perform the experiments on two Chinese ME databases (i.e., CASME I [6] and CASME II [27]). An example of a part of the frame sequence from the CASME I database is shown in Fig. 4. This database is a collection of 195 videos of MEs at a frame rate of 60 fps. The videos were filmed using two Point Grey GRAS-03K2C and BenQ M31 cameras with the resolutions of 640×480 and 1280×720 pixels [6]. After that, they have been converted to the sequential images. Some of the advantages of the aforementioned database are as follows [6]:

- Including labels for Action Unit (AU), onset (the frame that shows ME is occurring), offset, and apex
- Showing neutral face after and before each clip
- Containing different emotions.

The CASME II [27] has labeling clips on the onset, apex, and offset frames. This database includes 247 samples of spontaneous MEs. Its images have 280 × 240 resolution. The camera recording rate is two hundred fps. It is noteworthy that the number of samples for each expression in the databases (CASME I and CASME II) is unbalanced and biased [28].

4- 2- The Implementation Process of the Proposed Method

The sequential color images of each sample from CASME I and CASME II databases are taken. The images are converted to the gray level. The pre-processing techniques are applied. The same size image sequences are received as a 3-dimensional array.

We consider circles with unit radius along Z, X, and Y axes. The LBP-SIP*l* histogram as a feature vector is achieved as a matrix with the size 6×2^{Nup}, in which six is the number of planes. According to the best result, we consider eight neighboring points (Nu_p=8) in this paper. The elapsed time for each the LBP-SIP*l* histogram is 0.8 seconds based on the simulation run time. Afterwards, the histograms are compared with the neutral frame histogram for apex frame spotting. The frame which contains the most wrinkles caused by muscular movements is the apex.

Now, we create a folder for each emotion to contain the cropped human face of each apex frame. In fact, each emotion has a folder that is labeled by the respective emotion. All images are resized to 224x224. Thus, the resolution is [224, 224, 3] for the entire RGB images. These images are the input of the network. It is noteworthy that we apply data augmentation for a better recognition rate.

The pre-trained VGG-16 network is transformed into the FR-CNN by adding two layers (i.e., bounding box regression and ROI pooling). Additionally, the last layers of the model are replaced for ME recognition task. The network architecture using the layers property is shown in Table 1. Next, the training options are defined (see Table 2). The solver is Stochastic Gradient Descent with Momentum (SGDM). This optimizing algorithm is to minimize the loss function with small steps (mini-batches). The momentum controls oscillations towards the optimum. According to our experiments, the SGDM obtains superior accuracy results.

Finally, we use a k-fold cross-validation process with k=10, where the sample is randomly partitioned into 10 equal size subsamples. One by one, a set is selected as the test set, and the other sets (9 sets) are combined into the corresponding training set. This is repeated for each of the ten sets. Then, the average is computed from the obtained results.

Table 1. The proposed network architecture properties in details

Name	Property
'input'	224x224x3 images
'Conv1-1' + ' ReLU 1-1'	64 3x3x3 convolutions with stride 1 and padding 1
'Conv1-2' + 'ReLU 1-2'	64 3x3x64 convolutions with stride 1 and padding 1
'pool1'	2x2 max-pooling with stride 2 and padding 0
'Conv2-1' + 'ReLU 2-1'	128 3x3x64 convolutions with stride 1 and padding 1
'Conv2-2' + 'ReLU 2-2'	128 3x3x128 convolutions with stride 1 and padding 1
'pool2'	2x2 max-pooling with stride 2 and padding 0
'Conv3-1' + 'ReLU 3-1'	256 3x3x128 convolutions with stride 1 and padding 1
'Conv3-2' + 'ReLU 3-2'	256 3x3x256 convolutions with stride 1 and padding 1
'Conv3-3' + 'ReLU 3-3'	256 3x3x256 convolutions with stride 1 and padding 1
'pool3'	2x2 max-pooling with stride 2 and padding 0
'Conv4-1' + 'ReLU 4-1'	512 3x3x256 convolutions with stride 1 and padding 1
'Conv4-2' + 'ReLU 4-2'	512 3x3x512 convolutions with stride 1 and padding 1
'Conv4-3' + 'ReLU 4-3'	512 3x3x512 convolutions with stride 1 and padding 1
'pool4'	2x2 max-pooling with stride 2 and padding 0
'Conv5-1' + 'ReLU 5-1'	512 3x3x512 convolutions with stride 1 and padding 1
'Conv5-2' + 'ReLU 5-2'	512 3x3x512 convolutions with stride 1 and padding 1
'Conv5-3' + 'ReLU 5-3'	512 3x3x512 convolutions with stride 1 and padding 1
'pool5'	2x2 max-pooling with stride 2 and padding 0
'fc6' + 'ReLU 6' + 'drop6'	fully connected layer; 50% dropout
'fc7' + 'ReLU 7' + 'drop7'	fully connected layer; 50% dropout
'fc8'	New fully connected layer
'prob'	softmax
'output'	Classification Output

Table 2. Training hyper-parameters

Mini batch size	Initial learning rate	Max epoch	Positive overlapping range	Negative overlapping range	Weight learn rate factor in fc8	Bias learn rate factor in fc8	Validation Frequency
16	0.5e-5	30	[0.15 1]	[0.1 0.15]	10	10	30

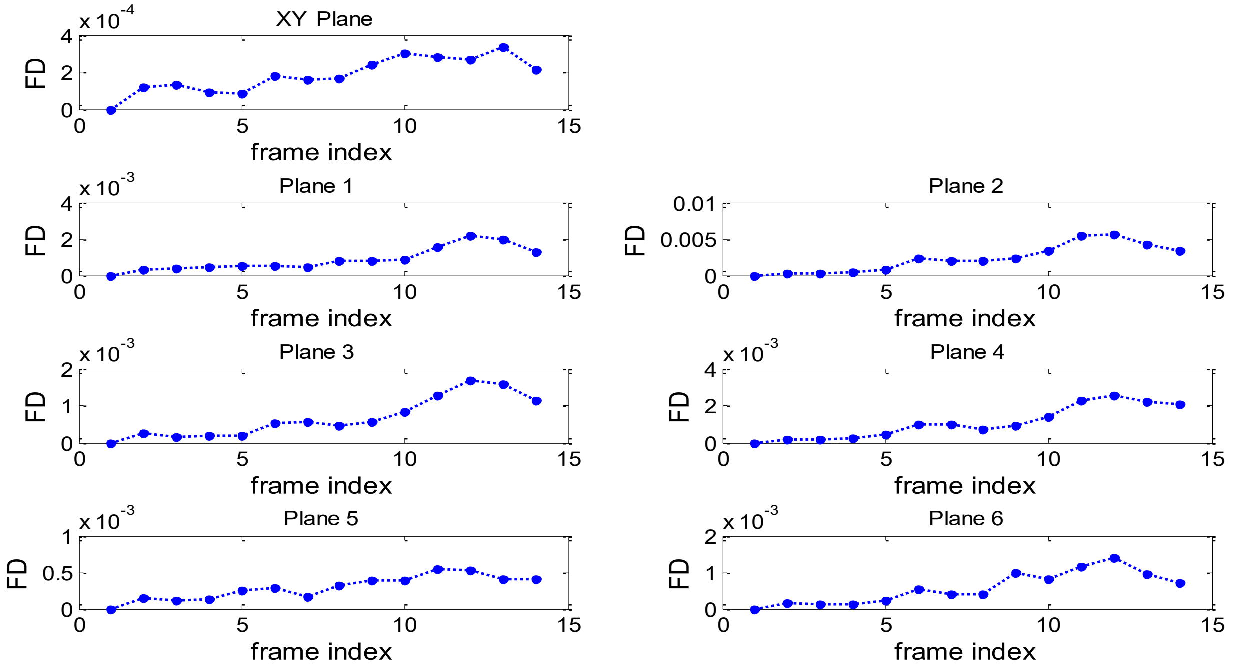

Fig. 5. The results of the apex frame finding in the XY plane and 6 planes of the proposed method.

4- 3- Results and Discussion

A sample of the apex frame spotting using each plane is shown in Fig. 5. In this sample (i.e., sub08-EP12-2-2 in CASME I database), the ground-truth apex is 12. Additionally, we put six planes in one figure (shown in Fig. 6) to find apex frame using the LBP-SIP*l*. Since each plane shows the changes and motions in located direction; thus, the frame with the most feature differences in every direction is the apex. As can be seen, the spotted apex frame is exactly 12. In Fig. 5 and Fig. 6, vertical axis shows the Feature Difference (FD) of each frame from first frame, and the horizontal axis is the number of frames.

The percentage of the apex frame finding using each six planes of the proposed method and their total are shown in Fig. 7. According to the results, the percentage of the apex frame spotting by the proposed method is 43%. It means that in 43% of all CASME I samples, spotted apex is the ground-truth apex.

Planes 3, 4, 5, and 6 display both of changes in the row and the column, while plane 2 shows one column of the pixels changing in the time. Since facial muscle movements are less in a vertical direction, plane 2 does not have a good performance. In addition, the temporal planes are more important than the XY plane, as the XY plane only shows the appearance changes and it does not contain the motion transition information.

In addition, we compute the SE and the MAE for comparing our method with other state-of-art methods. The MAE estimates the number frames that the spotted apex frame is off the ground-truth. It is defined for *SN* sample size as [18]:

$$M = (1/SN)\sum_{jj=1}^{SN}\left|e_{jj}\right| \tag{11}$$

where *e* is deviation of the ground-truth from the spotted apex frame. SE is the deviation from the mean of the distribution. It is defined as [23]:

$$S = \text{dev} \ / \ \surd \ SN \tag{12}$$

where *SN* and *dev* are the sample size and the sample deviation, respectively.

The experimental results are illustrated in Table 3. As can be seen, the SE and MAE obtained from our method are smaller than the SE and the MAE obtained from the RHOOF [23], LBP-TOP, Cubic-LBP [20], and LBP methods on the CASME I and CASME II databases. They confirm the error decrease in our work compared to the aforementioned previous methods.

Furthermore, ME recognition accuracy using the FR-CNN is obtained. The obtained accuracy in each fold of 10 fold cross-validation is given in Table 4 and Table 5. The average ME recognition accuracy using our proposed method is compared to other state-of-art methods in Table 6. According to the results, the rate of ME recognition in our work is high compared to other methods.

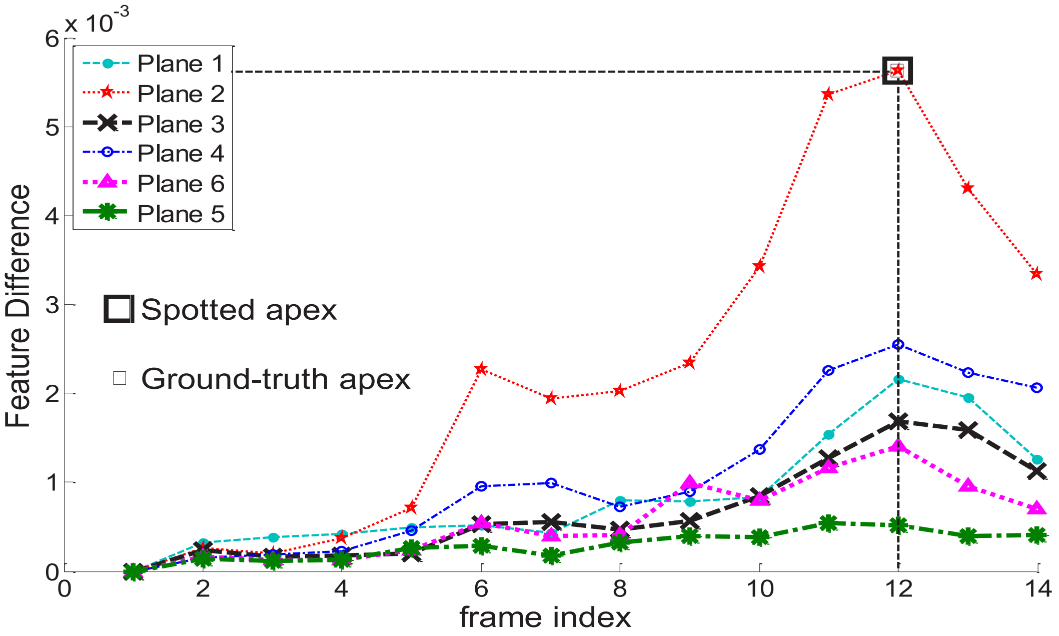

Fig. 6. Apex frame spotting using the LBP-SIPl.

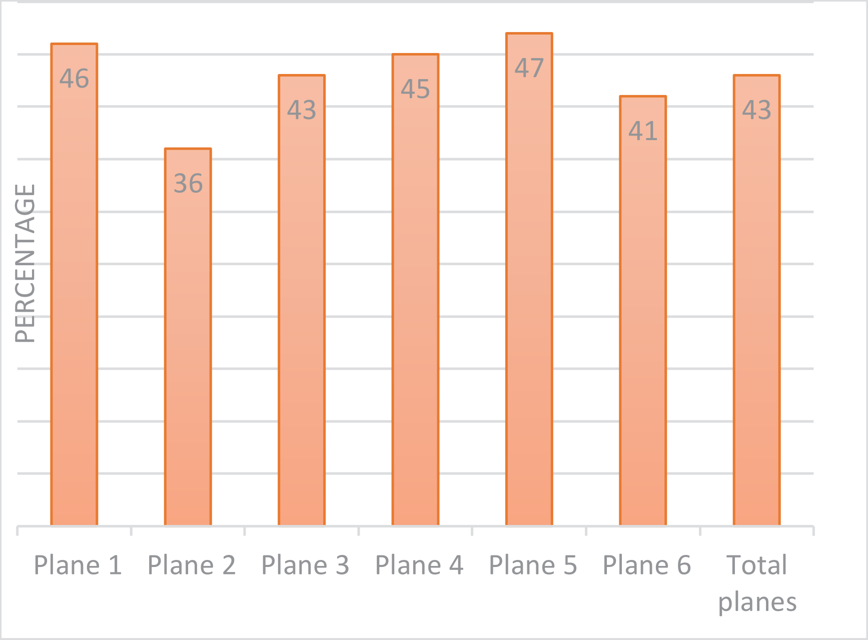

Fig. 7. The obtained percentage of apex frame spotting from the planes of the LBP-SIPl on the CASME I.

Table 3. The results of the apex frame spotting using the LBP-SIPl method on the CASME I and CASME II database.

Method	CASME II		CASME I	
	SE	MAE	SE	MAE
LBP (BS-RoIs)	0.79	13.55	0.58	5.20
RHOOF	0.73	10.97	0.35	3.60
LBP-TOP	0.67	8.38	0.23	2.54
Cubic-LBP	0.62	6.41	0.15	1.93
LBP-SIP*l*	**0.59**	**6.10**	**0.12**	**1.76**

Table 4. The results of the ME recognition accuracy using the FR-CNN method on CASME I.

Method	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Avg.
FR-CNN	0.80	0.79	0.82	0.80	0.83	0.81	0.82	0.81	0.79	0.83	0.81

Table 5. The results of the ME recognition accuracy using the FR-CNN method on CASME II.

Method	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Avg.
FR-CNN	0.94	0.96	0.96	0.98	0.94	0.95	0.98	0.97	0.95	0.97	0.96

Table 6. The ME recognition accuracy using our proposed method (FR-CNN) in comparison with other state-of-art methods.

Method	CASME II	CASME I
Deep CNN with OF [15]	56.94%	56.60%
3D and 2D CNN [16]	-	66.67%
Feature refinement [30]	68.38%	-
TSCNN [13]	80.97%	73.88%
STSTNet [14]	94.32%	-
Ours	**96.11%**	**81.56%**

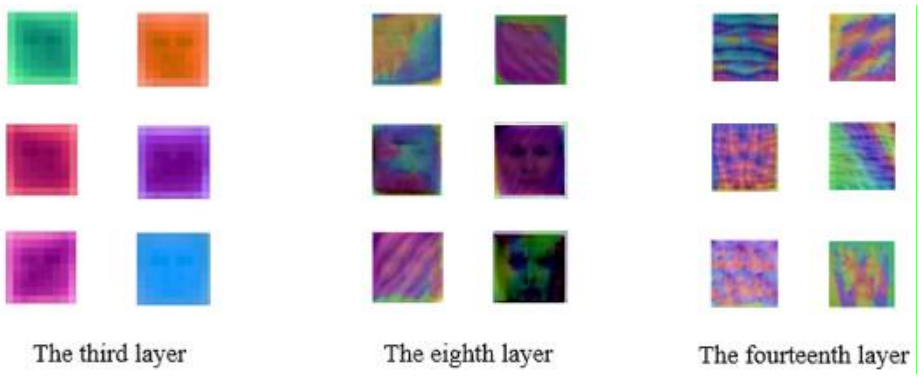

Fig. 8. The FR-CNN's learned features of different layers by Deep Dream.

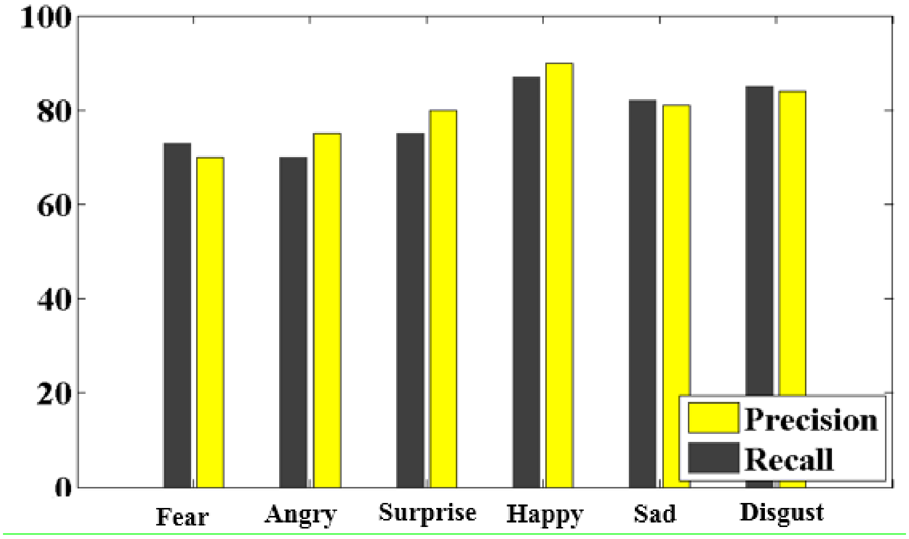

Fig. 9. Evaluating the ME recognition precision and recall of the FR-CNN trained on the CASME I.

Fig. 8. shows the learned features of different layers of FR-CNN by Deep Dream. Evaluating the ME recognition precision and recall of the FR-CNN trained on the CASME I and the CASME II are illustrated in Fig. 9 and Fig. 10. The precision and recall have been calculated from [29]. Some expressions might be seen only in one face region due to the ME local movements. However, the happiness expression occurs in the multi-regions. Muscle movements in the face with the lip corners pulled up, wrinkles around the nose, rising cheeks, and crow's feeds at the outer edges of the eyes correspond to 'happiness'. When comparing two happy-normal people, it is observed that neither the lips nor the eyes are similar. Additionally, the smile is a key and salient feature of happiness. Hence, the happiness expression usually provides the highest precision. Moreover, the F score result of the FR-CNN is benchmarked against the results obtained from other methods. Table 7 and Table 8 show the F measurement results. The results confirm the efficiency and performance of our proposed method.

5- Conclusion

The MEs help us understand the main human intention since they are uncontrollable and spontaneous. These characteristics can be applied in a wide range of applications. However, it could only be fruitful when the feature descriptor is able to extract all low-intensity changes and short duration. For this reason, in this paper we proposed two methods. The first one is the LBP-SIP*l*, comprising six planes, which could determine the apex frame. The other is FR-CNN for ME recognition. According to our numerical experiments, the proposed methods have superior performance in apex frame spotting and ME recognition compared to the related existing methods. For future research, the LBP-SIP*l* and the FR-CNN .could be applied for the detection of other subtle variations

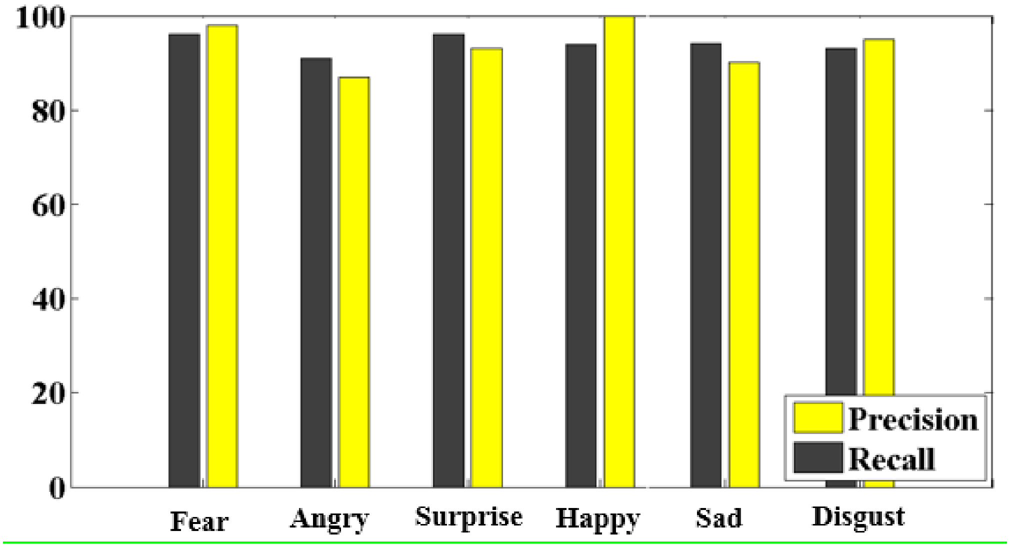

Fig. 10. Evaluating the ME recognition precision and recall of the FR-CNN trained on the CASME II.

Table 7. The F score of the FR-CNN and other methods on the CASME II.

Method	CASME II
3-Stream combining 3D and 2D CNN [17]	0.6142
STSTNet [14]	0.7353
TSCNN [13]	0.8070
Feature refinement [30]	0.8915
Ours	**0.9566**

Table 8. The F score of the FR-CNN and other methods on the CASME.

Method	CASME
LBP-TOP	0.32
STCLQP	0.56
TSCNN	0.72
Ours	**0.80**

References

[1] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Less is more: Micro-Expression recognition from video using apex frame, Signal Processing: Image Communication, 62 (2018) 82-92.

[2] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikäinen, Discriminative spatiotemporal Local Binary Pattern with revisited integral projection for spontaneous facial Micro-Expression recognition, IEEE Transactions on Affective Computing, 10(1) (2017) 32-47.

[3] X. Li et al., Towards reading hidden emotions: A comparative study of spontaneous Micro-Expression spotting and recognition methods, IEEE transactions on affective computing, 9(4) (2017) 563-577.

[4] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, Autonomous Apex Detection and Micro-Expression Recognition using Proposed Diagonal Planes, International Journal of Nonlinear Analysis and Applications, 11 (2020) 483-497.

[5] V. Esmaeili, M. M. Feghhi, 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) 1-5: IEEE.

[6] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, 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) 1-7: IEEE.

[7] W. Merghani, A. K. Davison, M. H. Yap, A review on facial Micro-Expressions analysis: datasets, features and metrics, arXiv preprint arXiv:1805.02397, (2018).

[8] X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikäinen, Spontaneous facial Micro-Expression analysis using spatiotemporal completed local quantized patterns, Neurocomputing, 175 (2016) 564-578.

[9] E. A. Haggard, K. S. Isaacs, Micromomentary Facial Expressions as indicators of ego mechanisms in psychotherapy, in Methods of research in psychotherapy: Springer, (1966) 154-165.

[10] P. Ekman, W. V. Friesen, Nonverbal leakage and clues to deception, Psychiatry, 32(1) (1969) 88-106.

[11] F. Xu, J. Zhang, J. Z. Wang, Microexpression identification and categorization using a facial dynamics map, IEEE Transactions on Affective Computing, 8(2) (2017) 254-267.

[12] J. A. Ruiz-Hernandez, M. Pietikäinen, Encoding Local Binary Patterns using the re-parametrization of the second order gaussian jet, in 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), (2013) 1-6: IEEE.

[13] B. Song et al., Recognizing spontaneous Micro-Expression using a three-stream Convolutional Neural Network, IEEE Access, 7 (2019) 184537-184551.

[14] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, 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) 1-5: IEEE.

[15] Q. Li, J. Yu, T. Kurihara, H. Zhang, S. Zhan, Deep Convolutional Neural Network with Optical Flow for facial Micro-Expression recognition, Journal of Circuits, Systems and Computers, 29(01) (2020) 2050006,.

[16] L. Wang, J. Jia, N. Mao, Micro-Expression Recognition Based on 2D-3D CNN, in 2020 39th Chinese Control Conference (CCC), (2020) 3152-3157: IEEE.

[17] C. Wu, F. Guo, TSNN: Three-Stream Combining 2D and 3D Convolutional Neural Network for Micro-Expression Recognition, IEEJ Transactions on Electrical and Electronic Engineering, 16(1) (2021) 98-107.

[18] G. Zhao, M. Pietikainen, Dynamic texture recognition using Local Binary Patterns with an application to Facial Expressions, IEEE transactions on pattern analysis and machine intelligence, 29(6) (2007) 915-928.

[19] Y. Wang, J. See, R. C.-W. Phan, Y.-H. Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for Micro-Expression recognition, in Asian conference on computer vision, (2014) pp. 525-537: Springer.

[20] V. Esmaeili, S. O. Shahdi, Automatic Micro-Expression apex spotting using Cubic-LBP, Multimedia Tools and Applications, 79(27) (2020) 20221-20239.

[21] V. Esmaeili, M. M. Feghhi, S. O. Shahdi, Automatic Micro-Expression Apex Frame Spotting using Local Binary Pattern from Six Intersection Planes, in Proc. of the 2020 International Conference on Machine Vision and Image Processing (MVIP), University of Tahran, Iran, Feb. 2020, available on: https://arxiv.org/pdf/2104.02149, (2021).

[22] T. Pfister, X. Li, G. Zhao, M. Pietikäinen, Recognising spontaneous facial Micro-Expressions, in 2011 international conference on computer vision, (2011) 1449-1456: IEEE.

[23] H. Ma, G. An, S. Wu, F. Yang, A Region Histogram of Oriented Optical Flow (RHOOF) feature for apex frame spotting in Micro-Expression, in 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), (2017) 281-286: IEEE.

[24] R. Girshick, Fast r-cnn, in Proceedings of the IEEE international conference on computer vision, (2015) 1440-1448.

[25] S. O. Shahdi, S. A. R. Abu-Bakar, Neural network-based approach for face recognition across varying pose, International Journal of Pattern Recognition and Artificial Intelligence, 29(08) (2015) 1556015.

[26] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern recognition, 29(1) (1996) 51-59.

[27] W.-J. Yan et al., CASME II: An improved spontaneous Micro-Expression database and the baseline evaluation,

PloS one, 9(1) (2014) e86041.

[28] V. Esmaeili, M. Mohassel Feghhi, S. O. Shahdi, A comprehensive survey on facial Micro-Expression: approaches and databases, Multimedia Tools and Applications, (2022) 1-46.

[29] C. Nicholson, Evaluation metrics for machine learning—accuracy, precision, recall, and F1 defined, ed: Pathmind. http://pathmind.com/wiki/accuracy-precision-recall-f1, (2019).

[30] L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, Feature refinement: An expression-specific feature learning and fusion method for Micro-Expression recognition, Pattern Recognition, 122 (2022) 108275.

HOW TO CITE THIS ARTICLE

V. Esmaeili, M. Mohassel Feghhi, S.O.Shahdi, Automatic Micro-Expression Recognition Using LBP-SIPl and FR-CNN , AUT J. Model. Simul., 54(1) (2022) 59-72.

DOI: 10.22060/miscj.2022.21133.5272