Received May 31, 2020, accepted June 27, 2020, date of publication July 3, 2020, date of current version July 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006958

Facial Micro-Expression Recognition Using Two-Dimensional Landmark Feature Maps

DONG YOON CHOI, (Graduate Student Member, IEEE), AND BYUNG CHEOL SONG, (Senior Member, IEEE)
Department of Electronic Engineering, Inha University, Incheon 22212, South Korea

Corresponding author: Byung Cheol Song (csong@inha.ac.kr)

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) [2020-0-01389, Artificial Intelligence Convergence Research Center(Inha University)] and Industrial Technology Innovation Program through the Ministry of Trade, Industry, and Energy (MI, Korea) [Development of Human-Friendly Human-Robot Interaction Technologies Using Human Internal Emotional States] under Grant 10073154.

ABSTRACT Emotion recognition based on facial expressions is very important for effective interaction of humans with artificial intelligence (AI) systems such as social robots. On the other hand, in real environment, it is much harder to recognize facial micro-expressions (FMEs) than facial general-expressions having rich emotions. In this paper, we propose a two-dimensional (2D) landmark feature map for effectively recognizing such FMEs. The proposed 2D landmark feature map (LFM) is obtained by transforming conventional coordinate-based landmark information into 2D image information. LFM is designed to have an advantageous property independent of the intensity of facial expression change. Also, we propose an LFM-based emotion recognition method that is an integrated framework of convolutional neural network (CNN) and long short-term memory (LSTM). Experimental results show that the proposed method achieves about 71% and 74% in the well-known micro-expression datasets, i.e., SMIC and CASME II, respectively, which outperforms the conventional methods. The performance of the proposed method was also verified through experiments on composite micro-expression dataset, which consists of SMIC, CAMSE II and SAMM, and cross-dataset validation using SMIC and CAMSE II. In addition, we prove that the proposed method is independent of facial expression intensity through an experiment on CK+ dataset. Finally, we demonstrate that the proposed method is valid even for the MAHNOB-HCI and MEVIEW datasets that are produced to monitor actual and wild emotional responses.

INDEX TERMS 2D landmark feature map, emotion recognition, facial micro-expression.

I. INTRODUCTION

Recently, emotion recognition technology through facial expression, action, and voice have been actively studied for advanced human-robot interaction (HRI). Especially, studies on facial expression-based emotion recognition (FER) are most active [1]–[6], [8]–[12], [33]–[35]. Conventional FER methods are focusing on facial macro-expressions as shown in Fig. 1 (a). In general, recognizing emotions corresponding to such macro-expressions is not a big deal. However, in actual situations (see Fig. 1 (b)), people may rarely express their emotions on the faces. In terms of simple facial expression metric (SFEM), i.e., one of the facial expression intensity metrics (see Section IV), the intensities of facial expression of Fig. 1 (a) and Fig. 1 (b) are 1.83 and 0.20, respectively, which indicate a big difference of about 9 times. In other words, even the same emotions may show significantly different quantitative values in terms of facial expression. Thus, we need to handle even such a facial micro-expression (FME) for ultimate emotion recognition. Unfortunately, there are not so many previous studies on recognition of FMEs. Also, the datasets for FER research purpose are usually collected from broadcast contents or movies where actors or actresses intentionally create their facial expressions [1], [2].

We have proposed an algorithm for recognizing FMEs in [16]. We presented a two-dimensional (2D) landmark feature map (LFM) robust to FME. LFM is defined by representing relative distances between facial landmarks and by properly normalizing them. We observed that LFMs of the same emotion tend to have a unique pattern, regardless of the intensity of facial expression, i.e., the intensity of emotion. Then, we proposed an LFM-based FME recognition algorithm which is based on CNN and LSTM. However,

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

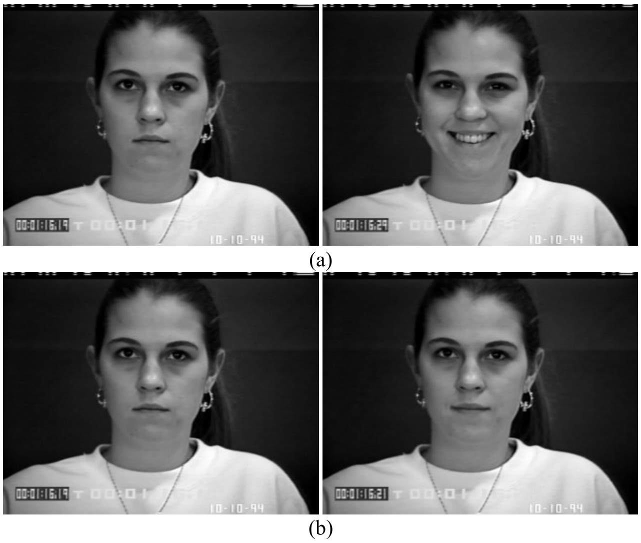


FIGURE 1. An example of change in facial expression for "happiness" (S094_004 sequence of CK + dataset) (a) before/after for facial macro-expression (SFEM: 1.83) (b) before/after for FME (SFEM: 0.20). Here, SFEM is the expression intensity metric. The larger the value, the stronger the facial expression.

LFM defined in [16] became burdensome to the network size based on feature maps and parameters. Also, the LFM-based network of [16] was not verified for actual FME datasets.

In order to reduce the network size without performance degradation, this paper presents a more compact LFM (CLFM) which is generated by appropriately sampling landmark points. In addition, we define a metric to evaluate the intensity of facial expression based on CLFM and propose a joint framework to select an appropriate method according to the facial expression metric. The joint framework keeps high emotion recognition accuracy even in a dataset with varying facial expression intensities

Experimental results show that the LFM/CLFM-based FER method provides high classification accuracy of about 71% and 74% for popular micro-expression datasets, SMIC [24] and CASME II [25]. This is superior performance over conventional FER techniques. We verified the superiority of the proposed method even through the experiments on composite micro-expression dataset and cross-dataset validation. For instance, the proposed method, which is trained using CK+ dataset [1] consisting of facial macro- expression images, shows about 80% classification performance for the FME test dataset. This is about 43% higher accuracy than a state-of-the-art (SOTA) scheme [11]. Finally, we observed that the proposed method works well even for actual (non-acting) and wild datasets, i.e., MAHNOB-HCI [19] and MEVIEW [57]. Note that the parameter size of the CLFM-based network amounts to only 8% of the LFM-based network [16]. This advantage will be of great help to the practical use of the proposed method.

The contribution points of this paper are organized as follows.

- This paper proposes an LFM that can effectively represent FMEs using landmark information and also presents its compressed version, i.e., CLFM.
- The proposed LFM/CLFM is robust to human appearance and facial expression intensity, and is suitable for FME recognition because it exhibits a unique pattern for each facial expression.
- When applying LFM/CLFM to conventional neural networks, LFM/CLFM provides outstanding FME recognition performance over SOTA.

The remainder of the paper is oargized as follows. Section II reviews existing micro-expression recognition methods. Section III details the proposed FER algorithm. Section IV describes the joint framework of the proposed method and a conventional emotion recognition technique, and Section V shows the experimental results. Finally, Section VI concludes.

II. RELATED WORK

Previous studies on FME can be summarized as follows: Pfister *et al.* proposed an FME recognition algorithm using temporal interpolation model and random forest [3]. Ngo and Wang used a so-called motion magnification to forcibly increase the intensity of FMEs [4], [5]. Then, they classified emotions using local binary pattern (LBP) and support vector machine (SVM). Zong *et al.* proposed a hierarchical spatio-temporal descriptor that controls the feature weight by searching the area where fine facial muscle movement exists [28]. Wang *et al.* proposed a new color space model to improve FME recognition performance [31], and recently proposed a deep learning-based method called transferring long-term convolutional neural network (TLCNN) [32].

Liu *et al.* proposed an FME recognition scheme using main directional mean optical flow (MDMO) [29]. They extracted the atomic feature representing the region of interest from the optical flow information, and applied sparse coding, and then classified the result using SVM. Peng *et al.* proposed a consolidated Eulerian frame that integrated independent motion magnification and frame interpolation into a single process [30]. Guo *et al.* proposed the extended local binary patterns on three orthogonal planes (ELBPTOP) as feature descriptors for recognizing FME [45]. In addition, Verma *et al.* proposed the dynamic representation of micro-expressions to preserve facial movement information in a single frame and also proposed a Lateral Accretive Hybrid Network (LEARNet) to capture micro-level features [15]. Esmaeili *et al.* proposed a feature extractor called Cubic-LBP for FME recognition [18]. Sun *et al.* proposed a knowledge distillation to transfer knowledge from action unit [58].

On the other hand, for a few years, a micro-expression grand challenge (MEGC) has been held to promote the competitive development of FME recognition techniques [37]. In MEGC, many FER techniques are evaluated using micro-expression-specific datasets such as SMIC [24], CASME II [25], SAMM [36], and their composite dataset. Here, we introduce several FME recognition algorithms published in MEGC 2019. Quang *et al.* employed capsule networks, which have been successful in general object recognition, for FME recognition [40]. Zhuo *et al.* proposed a two-stream

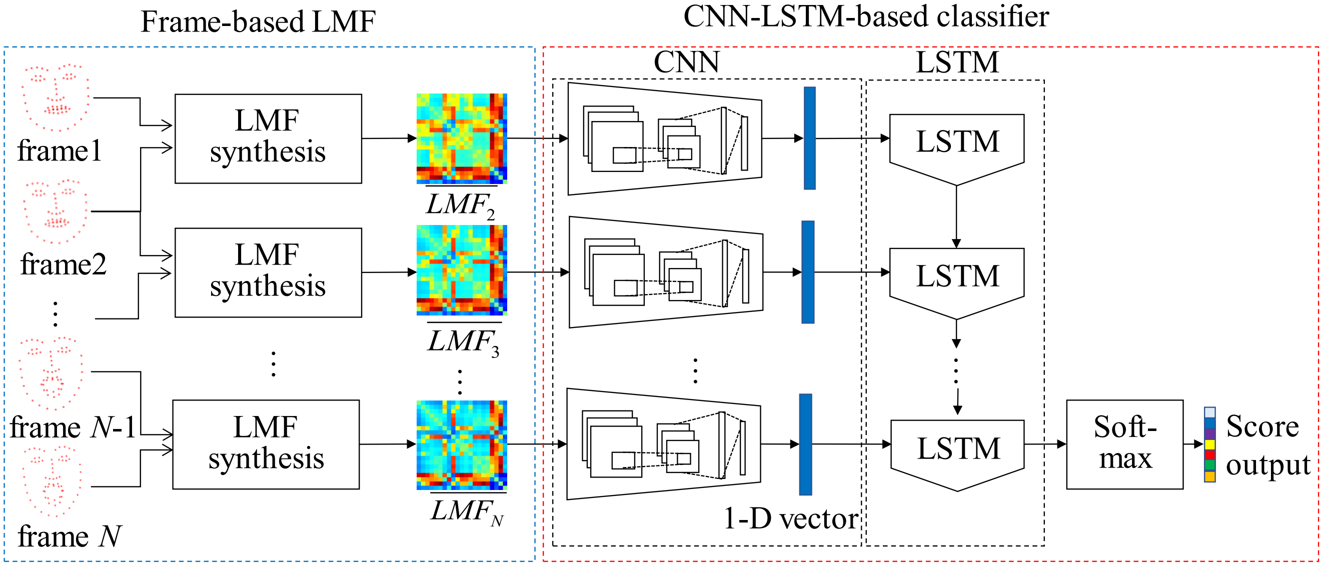


FIGURE 2. Block diagram of the proposed FER algorithm.

two-block variant of the Inception network to learn a robust feature representation from the horizontal and vertical components of TV-L1 optical flow information [42]. Liong *et al.* proposed a shallow 3D CNN which comprises of three parallel streams, each with a different number of feature maps to curb under-fitting. Liu *et al.* proposed a part- based deep neural network approach with two domain adaptation techniques–adversarial domain adaptation and motion magnification and reduction, which help to enrich the available training samples [44]. However, since the above-mentioned studies were verified only for the micro-expression datasets [6], [24], [25], [36], it has not been verified whether they recognize facial expressions of various intensity including FME.

III. LFM-BASED FACIAL EXPRESSION RECOGNITION

The proposed FER method consists of the LFM/CLFM generation step and the deep learning-based classifier (see Fig. 2). First, the point-wise distances between landmarks (LMs) of two adjacent video frames are computed and the computed distances are converted into a single 2D image, i.e., LFM. Here, a compact LFM (CLFM) can be generated. Note that LFM/CLFM is a sort of inter-frame distance. So, assuming a video sequence composed of N frames, N-1 LFMs/CLFMs can be generated for the video sequence. Second, an emotion corresponding to the video sequence is determined via classification using the LFMs/CLFMs. For this classification, we employ the CNN-LSTM-based network with the LFMs/CLFMs as input. Note that the CNN-LSTM network is a representative classifier for image sequences. In this paper, we adopt a structure combining VGG13 [13] and LSTM [7]. The CNN first transforms each LFM/CLFM into a one-dimensional (1D) feature vector. Then, if the consecutive 1D feature vectors pass through LSTM, the emotion class of each input video is determined.

A. GENERATION OF FRAME-BASED LFMs

First, take a look at the creation process of LFM. Figure 3 (a) is an example of 68 facial LMs. In this paper, we employed Dong *et al.*'s method [17] for LM extraction. Note that temporal change of LMs in a video sequence becomes an important clue to grasp the facial expression change. Also, LM has an advantage that it is rarely influenced by personal characteristics such as face shape, gender, age, and ambient illumination. Based on these advantages, Jung *et al.* [11] and Zhang *et al.* [12] proposed deep learning schemes that classify emotions using LMs. They used CNN receiving video data as the main network, and adopted recurrent NN (RNN) receiving the LM positions as the auxiliary network.

In order to apply LM information (instead of huge amount of pixel information) to CNN, we convert the LM information into a 2D LM feature map (LFM). In detail, LFM is defined as a time-varying pattern of distances between LMs of adjacent frames as follows.

$$LFM_n = \|p\,(i,n) - p\,(j,n)\|_2 - \|p\,(i,n-T) - p\,(j,n-T)\|_2 \tag{1}$$

where $p\,(i,n)$ denotes the coordinate of the i-th LM in the n-th frame. $LFM_n\,(i,j)$ shows how the distance between the i-th LM and the j-th LM changes between the n-th frame and the (n-T)-th frame. Assuming LM model consisting of 68 points, LFM_n becomes a 68 × 68 feature matrix. Since LFMs are 2D information, they can be learned by CNN. The experimental result in Section V-B shows that LFMs are visually distinguished according to emotions.

Next, we need to make LFM robust to the strength or intensity of emotion. Assume two facial expressions with the same emotion type but different emotion strengths. Without loss of generality, we could observe that even though the moving distances of the LMs are different each other,

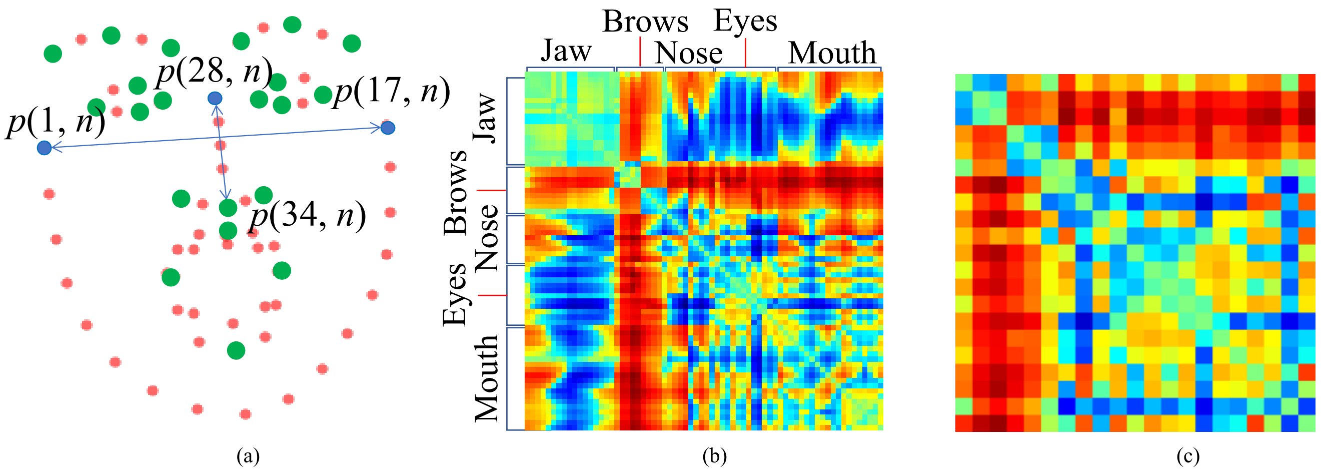


FIGURE 3. **(a) 68 LMs extracted from a face with "disgust" emotion (S097_004 sequence of CK+ dataset) (b) LFM generated with the 68 LMs (c) CLFM generated with 21 sampled LMs.**

their directions and patterns are similar. Therefore, by normalizing the LFM, i.e., dividing the LFM by the maximum value or the minimum value of the LFM as shown in Eq. (2), we make the same emotions have almost same LFMs irrespective of the emotion strengths. Here $\max(LFM_n)$ and $\min(LFM_n)$ indicate the farthest distance and the closest distance between LMs, respectively.

$$\overline{LFM_n}(i,j) = \begin{cases} 128\left\{\dfrac{LFM_n(i,j)}{\max(LFM_n)}\right\}^{0.5} + 127 & \textit{if } LFM_n(i,j) > 0 \\ -128\left\{\dfrac{LFM_n(i,j)}{\min(LFM_n)}\right\}^{0.5} + 127 & \textit{elseif } LFM_n(i,j) < 0 \\ 127 & \textit{else} \end{cases} \quad (2)$$

For instance, Fig. 3 (b) shows the result of transforming the LMs extracted from the disgusting expression of Fig. 3 (a) into an LFM according to Eq. (2). The red color means that the distance between the LMs is getting closer, and the blue color means that the distance is getting away. We could observe that the positions of jaw, eyes, nose, and mouth in the LFM are almost fixed due to the essential structure of human faces as in Fig. 3 (b). Thus, we can distinguish different emotions through activated patterns of LFM.

On the other hand, LFM outputs a unique pattern regardless of the strength of the facial expression thanks to the normalization effect. So, facial expressions of the same emotion type are characterized by similar LFMs. Using this characteristic, we can improve the FME recognition performance.

1) GENERATION OF CLFM VIA SAMPLING OF LANDMARKS

The number of facial LMs can normally vary depending on the LM extraction techniques. The most common method is to extract 68 LMs, but 49, 21, and 5 LMs are also available [20]–[22]. Since LFM aims to produce a unique pattern per emotion, it is enough to use only the LMs that can saliently represent the emotion change among all the LMs. That is, we do not have to entirely adopt 68 LMs as in [16]. For example, since facial expressions generally depend on changes in eyes, nose, mouth, eyebrows, etc., it is seldom affected by the LMs near jaw with little change. Also, since LMs are relatively densely distributed, a proper sampling of LMs does not adversely affect the character of facial expressions. Based on this philosophy, we sample only the minimum number of LMs that can represent the movement of each facial part. First, in the case of eyebrows, three LMs of both sides and in the middle of each eyebrow are chosen. Also, in order to sense behaviors such as eye closing or frowning, four LMs are selected at both sides, up and down of each eye. Then, four LMs are chosen to represent the mouth opening and the movement of mouth corner. Finally, three LMs at the bottom of the nose are selected, except for the vertical LM with little movement. As a result, 21 ($= 3 \times 2 + 4 \times 2 + 4 + 3$) LMs are sampled.

By sampling only 21 salient LMs among 68 LMs, we generate 21 × 21 LFM. Experimental results of Section V show that 21 × 21 LFM, i.e., CLFM has little difference in performance from 68 × 68 LFM. The indices of the sampled LMs among 68 LMs is [18], [20], [22], [23], [25], [27], [32], [34], [36], [37], [39], [40], [41], [43], [44], [46], [48], [49], [58] (see the green points in Fig. 3 (a)). Figure 3 (c) shows the CLFM derived from Fig. 3 (a). Note that the spatial resolution of Fig. 3 (c) is only 1/10 of Fig. 3 (b), but the main features are preserved well in spite of sampling. Section V-A will prove that CLFM can significantly decrease the parameter size as well as computation of the classifier network with negligible performance degradation.

B. CNN-LSTM-BASED CLASSIFIER

In order to classify the generated LFMs, we adopt a CNN-LSTM-based network which has been widely used for video classification. First, a network based on VGG13 [13] produces a 1D spatial feature vector from $\overline{LFM_n}$ of Eq. (2).

TABLE 1. The configuration of the CNN network based on LFM and CLFM. Here *C* is the number of classes.

LFM		CLFM	
Operation	Feature dimension	Operation	Feature dimension
Input(filter size)	224x224x1	Input(filter size)	21x21x1
Conv1(3x3x16)	224x224x16	Conv1(3x3x16)	21x21x16
Conv2(3x3x16)	224x224x16	Conv2(3x3x16)	21x21x16
MaxPool1(2x2)	112x112x16	MaxPool1(2x2)	10x10x16
Conv3(3x3x32)	112x112x32	Conv3(3x3x32)	10x10x32
Conv4(3x3x32)	112x112x32	Conv4(3x3x32)	10x10x32
MaxPool2(2x2)	56x56x32	MaxPool2(2x2)	5x5x32
Conv5(3x3x64)	56x56x64	Conv5(3x3x64)	5x5x64
Conv6(3x3x64)	56x56x64	Conv6(3x3x64)	5x5x64
Conv7(3x3x64)	56x56x64	Conv7(3x3x64)	5x5x64
MaxPool3(2x2)	28x28x64	Flatten	1x1600
Conv8(3x3x128)	28x28x128	FC1(1600x256)	1x256
Conv9(3x3x128)	28x28x128	FC2(256x*C*)	1x*C*
Conv10(3x3x128)	28x28x128	Softmax	1x*C*
MaxPool4(2x2)	14x14x128		
Conv11(3x3x128)	14x14x128		
Conv12(3x3x128)	14x14x128		
Conv13(3x3x128)	14x14x128		
MaxPool5(2x2)	7x7x128		
Flatten	1x6272		
FC1(6272x1024)	1x1024		
FC2(1024xC)	1x*C*		
Softmax	1x*C*		

TABLE 2. Comparison in terms of the parameter amount.

	CNN	LSTM	Total
LFM	7,351,223	1,838,855	9,190,078
CLFM	520,375	198,023	718,398

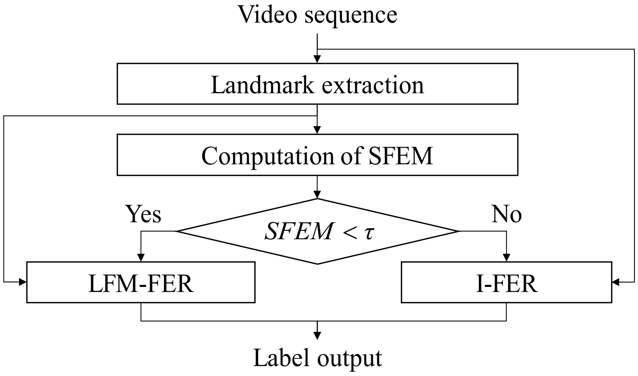


FIGURE 4. A generic framework that adaptively employs LFM-FER and I-FER according to SFEM. Here, SFEM is a metric to indicate expression intensity.

Then, the feature vectors are sequentially applied to LSTM [7] to extract spatio-temporal features, i.e., the hidden states of LSTM. Finally, an emotion class is determined by softmax.

In this paper, CNN network and LSTM network are learned separately. The former has a frame unit and the latter has a video sequence unit. The CNN is first trained using a pre-determined dataset, e.g., LFMs derived from target dataset. Next, the LSTM is trained using the feature vectors obtained from the CNN network. The details of each network architecture are as follows.

1) CNN NETWORK

As mentioned above, the CNN network is based on VGG13 model. As a CNN network that acts as an encoder to extract 1D feature vectors from LFMs, this paper chose VGG13, which has been widely used as a backbone or baseline in various computer vision tasks. The CNN network is trained with a target dataset, and the LFMs (or CLFMs) of all frames in the dataset are input with the corresponding labels. The detailed structure is shown in Table 1. The left and right columns of Table 1 are CNN networks for LFMs and CLFMs, respectively. Note that the 1D feature vectors from the fully connected 1 (FC1) layer are input to the subsequent LSTM network. In fact, the output of the last layer, i.e., FC2 layer can be regarded as probability values of emotion classes. Also, the goal of the CNN network is to analyze the spatial patterns of LFMs and generate the compressed features. Therefore, the intermediate features from FC1 layer are transferred to the LSTM network.

2) LSTM NETWORK

While CNN is used to extract the frame unit features, LSTM, that is a sort of RNN, is used to extract the sequence unit features. The LSTM receives the 1D feature vector from each cell, and then outputs the hidden state and cell state of 128 dimensions. If the hidden state of the last data of each video sequence passes through the FC layer and softmax, a final score vector is output as in Fig. 2.

3) ANALYSIS OF NETWORK ARCHITECTURE

In [16], a 68 × 68 LFM was resized to 224×224 size, and a feature vector of 1024 length was output from the CNN network. In this paper, if CLFMs are input to the CNN network, they are not resized and a single layer LSTM is adopted. Therefore, the parameter size of the entire network is drastically reduced. Table 2 compares the number of parameters of CLFM-based network with that of LFM-based network. Note that the parameter size of CLFM-based network amounts to only about 8% of LFM-based network. Section V shows that despite the reduced parameter size, the performance degradation is negligible.

IV. FACIAL EXPRESSION RECOGNITION USING IMAGE AND LFM

Since the LFM-based FER (LFM-FER) proposed in Section III was originally designed to recognize FMEs, its performance may be lower than that of traditional image-based FER (I-FER) to be designed for recognizing general macro-expressions (refer to Section V). Therefore, this section presents an adaptive framework that selects LFM-FER and I-FER according to the facial expression strength (see Fig. 4). In other words, if an input is determined as FME, LFM-FER is applied. Otherwise, the existing FER algorithm (I-FER) is applied.

A. A SIMPLE METRIC OF FACIAL EXPRESSION INTENSITY

In order to realize the framework of Fig. 4, a specific metric to measure the emotion intensities in the facial expressions is required. Several metrics have been proposed [8], [9].

However, since the existing metrics are computationally intensive, Eq. (3) is given as a simple facial expression metric (SFEM).

$$SFEM = \frac{\sum_{n=1}^{N}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|LFM_n(i,j)\right|}{\|p(1,n)-p(17,n)\|_2 \cdot \|p(28,n)-p(34,n)\|_2} \tag{3}$$

where N is the number of frames in the sequence and M is the number of LMs. The numerator in Eq. (3) is the sum of distances between LMs in the video sequence, and the denominator indicates the approximate face size around eyes and nose. Assuming 68 LMs as shown in Fig. 3 (a), LM 1 and LM 17 are located near both ears, and LM 28 and LM 34 are located at the top and bottom of the nose, respectively. Since these LMs are generally insensitive to emotion changes, the product of the two distances in the denominator can be assumed to be the relative size of a given face for normalization. So, SFEM can estimate the intensity of face expression irrespective of human identity and emotion class. In an example of Fig. 1, SFEM for a macro-expression was 1.83 and that of a micro-expression was 0.20. Therefore, we could empirically find that the SEFM is proportional to the expression intensity.

B. A GENERIC FRAMEWORK THAT ADAPTIVELY EMPLOYS LFM-FER AND I-FER ACCORDING TO SFEM

As in Fig. 4, if SFEM is first computed for each input video, it is compared with a pre-determined threshold τ. Through experiments on various videos, we set τ to 0.35 in this paper. Unlike straightforward ensemble methods which require inferencing two or more networks, this selective method chooses an appropriate processing type before inferencing. Thus, the proposed framework has an advantage of reducing overall computational cost since only one network is inferenced.

V. EXPERIMENTS

The following four kinds of experiments were performed to evaluate the performance of the proposed method. First, Section V-A verifies the FME recognition performance of the proposed method. To do this, we adopted SMIC [24], CASME II [25], SAMM [36], which are representative micro-expression datasets. In Section V-B, the characteristics of LFMs (or CLFMs) are analyzed. For this analysis, another micro-expression dataset is synthesized from CK+ dataset [1], and it is used together with CK+ itself. Performance verification of the proposed method for macro-expression datasets as well as micro-expression datasets given in Section V-C. Finally, Section V-D shows the results for the actual and wild emotional response.

A. EXPERIMENTAL RESULT ON MICRO-EXPRESSION DATASETS

The SMIC dataset consists of video sequences having micro-expressions. Each video sequence has a resolution of 640 × 480@100Hz. There are 164 video sequences for 16 subjects, and each sequence is labeled with three emotional classes such as ‘negative’, ‘positive’, and ‘surprise’. The average number of frames in each sequence is 33.7. Each video sequence of CASME II dataset has a resolution of 640×480@200Hz. There are 246 sequences for 26 subjects. CASME II consists of five emotional labels such as ‘disgust’, ‘happiness’, ‘others’, ‘repression’, and ‘surprise’. The average number of frames in each sequence is 67.2. Finally, the SAMM dataset consists of video sequences of 2040 × 1080@200Hz. There are 159 sequences for 29 subjects. SAMM consists of eight emotional labels such as ‘anger’, ‘contempt’, ‘disgust’, ‘fear’, ‘happiness’, ‘others’, ‘sadness’, and ‘surprise’. The average number of frames per video is 74.3.

For fair evaluation, we used a protocol called leave-one-subject-out cross validation (LOSO) in the same way as the conventional methods [26]–[30]. In LOSO, samples for one subject are used as test data, and samples for the remaining subjects are adopted as training data.

The frame interval T in Eq. (1) was set to 4 for SMIC and 8 for CASME II and SAMM. This is to match the general video frame rate of 25Hz. In addition, the length of the LFM sequence input to the CNN-LSTM classifier was fixed to 24 frames. Basically, 24 frames in the middle of each sequence are extracted. If the length of a particular sequence is shorter than 24 frames, repetitive padding is applied to both ends of the sequence to forcibly set the frame length to 24 frames.

1) EXPERIMENTS ON SINGLE DATASETS

This section evaluates the performance of the proposed method for SMIC and CASME II datasets, and then compares it with SOTA techniques. As the SOTA techniques for benchmarking, we chose Wang *et al.*’s [26], Li *et al.*’s [27], Zong *et al.*’s [28], Liu *et al.*’s [29], Peng *et al.*’s [30] Guo *et al.*’s [45], and Sun *et al.*’s [58]. LOSO validation was applied to all techniques. 16-fold validation was used for SMIC and 26-fold validation for CASME II as in conventional techniques. Quantitative evaluation was done in terms of classification accuracy and F1-score.

Table 3 compares the proposed method with the SOTA methods. The numerical values of each SOTA are taken as they are in the papers, and F1-scores of some methods are available. We can find that the proposed method provides the best accuracy and F1-score on average for both datasets. In case of SMIC, [26] provides the best performance among SOTAs, which is the same accuracy as the proposed method. However, note that the proposed LFM-based network has an improvement of about 8.6% over [26] for CASME II. For CASME II, [45] is the best among SOTAs. However, the proposed method shows about 0.04% higher accuracy than [45]. In case of SMIC, the proposed method provides higher accuracy of 2.28% than [45]. Although [45] uses three types of LBPs, which are much more complicated than the LFM of the proposed method,

TABLE 3. Performance comparison for SMIC and CASME II datasets. Red and blue mean the first and second places, respectively.

	SMIC		CASME II	
	Accuracy	F1-score	Accuracy	F1-score
CASME II baseline[25]	N/A	N/A	63.41%	N/A
Wang et al.'s [26]	71.34%	N/A	65.43%	N/A
Li et al.'s [27]	68.29%	N/A	67.21%	N/A
Zong et al.'s [28]	66.46%	0.6577	65.18%	0.6254
Liu et al.'s [29]	70.51%	0.7041	66.95%	0.6911
Peng et al.'s [30]	68.90%	N/A	70.85%	N/A
Guo et al's [45]	69.06%	N/A	**73.94%**	0.69
Sun et al's [58]	N/A	N/A	72.61%	0.67
LFM-based (68x68 LFM)	**71.34%**	**0.7134**	**73.98%**	**0.7165**
CLFM-based (21x21 LFM)	**71.34%**	**0.7134**	**71.54%**	**0.7026**

TABLE 4. Confusion matrices for SMIC dataset (a) LFM (b) CLFM.

	Neg.	Pos.	Sur.
Neg.	**75.7**	15.7	8.6
Pos.	29.4	**68.6**	2.0
Sur.	23.3	9.35	**67.4**

(a)

	Neg.	Pos.	Sur.
Neg.	**74.3**	17.1	8.6
Pos.	35.3	**60.8**	3.9
Sur.	14.0	7.0	**79.0**

(b)

it is encouraging that the proposed method performs better than [45].

On the other hand, in case of SMIC, there is no difference in performance between CLFM-based network and LFM-based network, whereas for CASME II, CLFM is about 2.4% less accurate than LFM. This is due to the difference in the number of emotion classes between the two datasets. SMIC has only three emotion classes: 'negative', 'positive', and 'surprise', but CASME II consists of five classes. So, in SMIC with a small number of classes, there is almost no performance degradation due to CLFM, which is slightly less discriminating than LFM. However, in CASME II with a large number of classes, the performance degradation due to the low discrimination power of CLFM becomes visible. In order to examine this, Table 4 and Table 5 compare the confusion matrices for the two datasets. Looking at Table 4 for SMIC, LFM and CLFM show opposite results in the two classes, but on average they show similar performance. However, in Table 5, CLFM is inferior to LFM in 3 out of 5 classes. So, in CASME II, CLFM is on average lower than LFM.

TABLE 5. Confusion matrices for CASME II dataset (a) LFM (b) CLFM.

	Dis.	Hap.	Oth.	Rep.	Sur.
Dis.	**74.6**	1.6	19.0	3.2	1.6
Hap.	12.5	**53.1**	12.5	15.6	6.3
Oth.	9.1	2.0	**84.8**	3.0	1.0
Rep.	3.7	3.7	22.2	**70.4**	0.0
Sur.	0.0	4.1	36.0	0.0	**60.0**

(a)

	Dis.	Hap.	Oth.	Rep.	Sur.
Dis.	**71.4**	4.8	22.2	1.6	0.0
Hap.	12.5	**43.8**	34.4	6.3	3.1
Oth.	7.1	2.0	**88.9**	2.0	0.0
Rep.	7.4	14.8	37.0	**40.7**	0.0
Sur.	4.0	0.0	20.0	4.0	**72.0**

(b)

Figure 3 (a) shows that CLFM has a higher LM density in eyes and eyebrows than LFM. In the case of 'surprise', the movements of eyes and eyebrows are dominant information. Since CLFM has a relatively large density for such a region, it is superior to LFM for 'surprise' emotion. In the case of 'positive', the information around the mouth is dominantly important. So, CLFM has a relatively low performance because its LM density around mouth is lower than LFM's. We experimentally chose the LMs of CLFM such that the average accuracy is maximized for various emotions. That is, the number of LMs in CLFM, i.e., 21 was an experimental minimum that maximizes the average accuracy in FME recognition. As a result, CLFM provides better FME performance than SOTAs as in Table 3, while CLFM provides significantly lower network complexity than LFM with negligible performance degradation as in Table 2 and Table 3.

2) EXPERIMENTS ON A COMPOSITE DATASET

This section shows the experimental results for a composite dataset combining SMIC, CASME II, and SAMM. This experiment followed the protocol of MEGC 2019 [37]. In order to unify the classes of three different datasets, only three classes of 'negative', 'positive', and 'surprise' were used. In CASME II, 'disgust' and 'repression' were merged into a single 'negative' class. In SAMM, 'anger', 'contempt', 'disgust', 'fear', and 'sadness' were all integrated into a single 'negative' class. As a result, the composite dataset consists of 442 sequences obtained from a total of 68 subjects. We adopted LOSO validation and analyzed performance in terms of unweighted F1-score (UF1) and unweighted average recall (UAR).

Figure 5 shows some examples of LFM and CLFM for three micro-expression datasets. We can observe that LFMs and CLFMs of the same class tend to have similar patterns even if the datasets are different each other. Therefore, we can find that the proposed LFM/CLFM can extract robust features regardless of gender or race.

Table 6 gives quantitative experimental results. EMR [44] shows the best performance in composite dataset, SMIC,

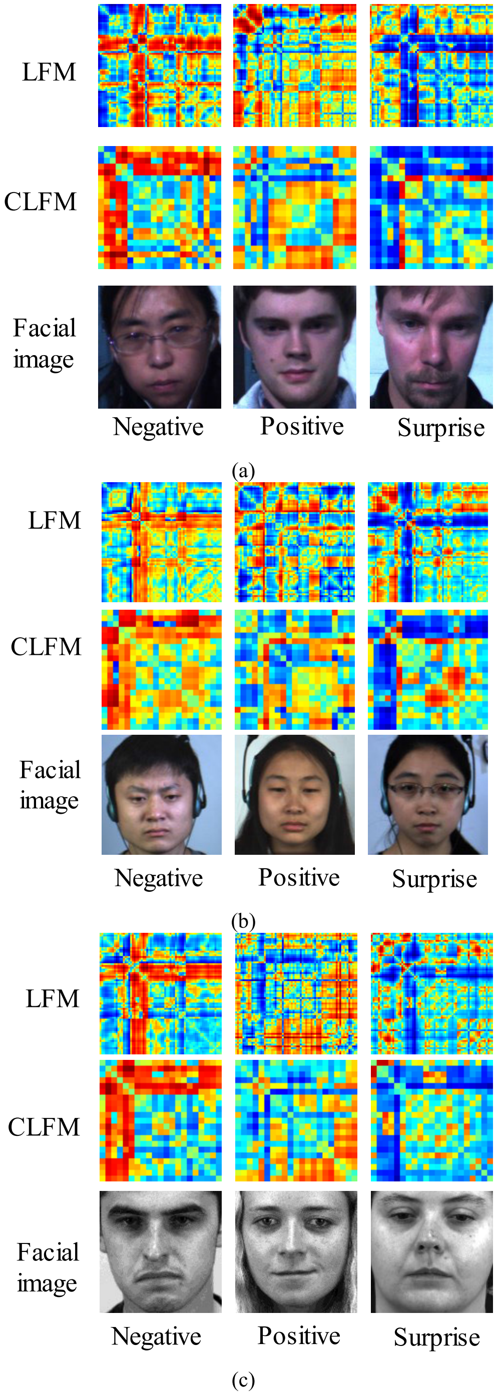


FIGURE 5. Examples of the proposed LFM and CLFM for micro-expression datasets. (a) SMIC, (b) CASME II, (c) SAMM dataset.

and SAMM datasets, but in CASME II it is inferior to the proposed method. ELBPTOP [45] gives the best performance in CASME II, but is inferior to the proposed method for the other datasets. Note that the proposed method shows good performance evenly in all datasets. For instance, LFM and CLFM provide only 0.02 and 0.04 less UF1 than EMR for the composite dataset. This is a very marginal degradation in performance. It is noteworthy that EMR additionally used CK+ dataset for training and adopted domain adaptation technique, whereas the proposed method employed a typical training technique without domain adaptation. Nevertheless, the good performance of LMF and CLMF demonstrates that they are effectively extracting key information about FME.

3) EXPERIMENTS ON CROSS-DATASET VALIDATION

This section describes the experimental results for cross-dataset validation using SMIC and CASME II. The experiment was performed as follows. After setting one dataset as the source dataset, a given model is trained and the performance of the trained model is tested with another dataset. We followed the protocol of [46], which preceded the cross-dataset validation of FER algorithms for FME recognition. In [46], all combinations of four datasets, i.e., SMIC (SMIC-HS), SMIC-VIS, SMIC-NIR, and CASME II, were verified. This paper performed experiments on cross-dataset validation between SMIC (SMIC-HS) and CASME II datasets, i.e., 'SMIC (source) → CASME II (target)' and 'CASME II (source) → SMIC (target)'. For efficiency of this experiment, the classes of the two datasets were adjusted into three classes: 'positive', 'negative', and 'surprise'. 164 data from SMIC and 130 data from CASME II were used in this experiment. The performance of each algorithm was evaluated in terms of the mean F1-score and accuracy.

The domain adaptation (DA) with respect to the SVM-based technique [47] was applied to all benchmarking methods [46], [48]–[56]. During the DA process, the label information of the target dataset was not used, but the feature information of the input was used. Meanwhile, the LFM- and CLFM-based techniques did not use the target dataset information in the learning process. That is, the model was trained using only the source dataset, and the performance of the trained model was evaluated for the target dataset.

Table 7 shows the experimental results. In the 'CASME II → SMIC' experiment, the LFM-based method provides the best performance with mean F1-score of 0.5876 and accuracy of 58.54, and the CLFM-based method has an accuracy of 55.49, which is the second highest accuracy. Next, in case of 'SMIC → CASME II' experiment, the LFM-based method shows mean F1-score of 0.6066 and accuracy of 63.57, which is the second best performance. The CLFM-based method shows the third highest mean F1-score and the fourth highest accuracy. Given that DA was not applied to the proposed method, such a performance improvement is very encouraging. This experimental result indirectly proves that the LFM and CLFM features not only have very robust characteristics in the dataset acquisition environment, but also extract unique features according to emotions.

B. ANALYSIS OF LANDMARK-BASED FEATURE MAP

This section demonstrates experimentally that LFM/CLFM has robust properties to macro-expressions as well as micro-expressions. For this experiment, we employ two kinds of datasets. The first dataset is a famous CK+

TABLE 6. Experimental results on SMIC, CASME II, SAMM, and their composite datasets. Red and blue mean the first and second places, respectively.

Method	Composite		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBPTOP[38]	0.59	0.58	0.20	0.53	0.70	0.74	0.40	0.41
Bi-WOOF[39]	0.63	0.62	0.57	0.58	0.75	0.80	0.52	0.51
CapsuleNet[40]	0.65	0.65	0.58	0.59	0.71	0.70	0.62	0.60
OFF-ApexNet[41]	0.72	0.71	0.68	0.67	0.88	0.87	0.54	0.54
Dual-Inception Network[42]	0.73	0.73	0.66	0.67	0.86	0.86	0.59	0.57
STSTNet[43]	0.74	0.76	0.68	0.70	0.84	0.87	0.66	0.68
EMR with Adv. Training[44]	0.79	0.78	0.75	0.75	0.83	0.82	0.78	0.72
ELBPTOP[45]	0.71	0.69	0.65	066	0.89	0.88	0.49	0.49
LFM-based (68x68 LFM)	**0.77**	**0.75**	**0.72**	**0.71**	**0.87**	**0.84**	**0.67**	**0.60**
CLFM-based (21x21 LFM)	**0.75**	**0.72**	**0.71**	**0.71**	**0.72**	**0.77**	**0.65**	**0.51**

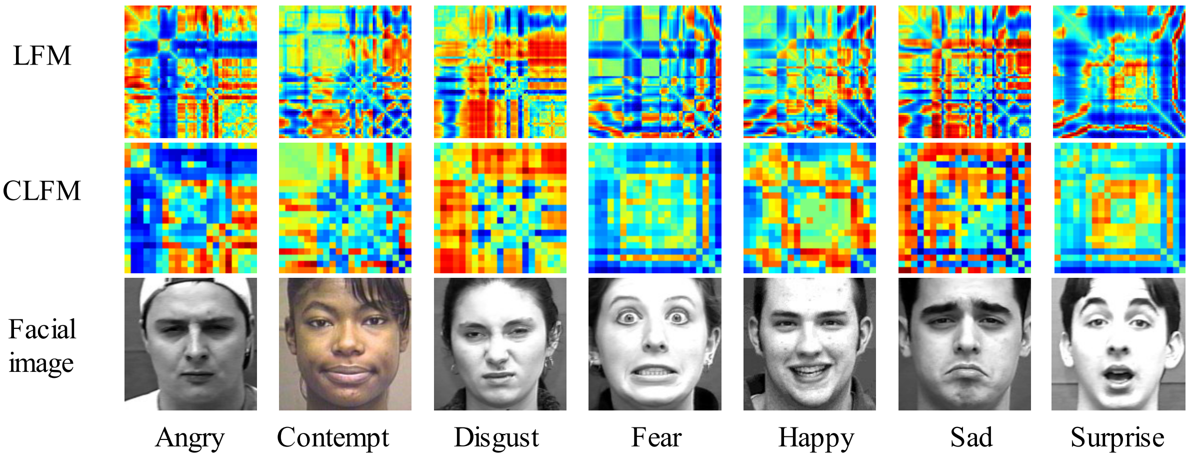


FIGURE 6. LFMs and CLFMs for seven emotions. (1st row) LFM [16] (2nd row) CLFM.

TABLE 7. Results of cross dataset experiment for SMIC and CASME II datasets (mean F1-score / accuracy). Here, domain adaptation (DA) technique was applied to all the techniques except SVM and the proposed method. Red and Blue mean the first and second places, respectively.

Method	CASME II → SMIC	SMIC → CASME II
SVM (w/o DA) [47]	0.3697 / 45.12%	0.3245 / 48.46%
IW-SVM [48]	0.3541 / 41.46%	0.5429 / 62.31%
TCA [49]	0.4637 / 46.34%	0.4870 / 53.08%
GFK [50]	0.4126 / 46.95%	0.4776 / 50.77%
SA [51]	0.4302 / 47.56%	0.5447 / 62.31%
STM [52, 53]	0.3640 / 43.90%	0.6115 / 63.85%
TKL [54]	0.4582 / 46.95%	0.4657 / 45.38%
TSRG [55]	0.5042 / 51.83%	0.5171 / 60.77%
DRFS-T [56]	0.4524 / 46.95%	0.5460 / 60.00%
DRLS [56]	0.4924 / 53.05%	0.5267 / 59.23%
RSTR [46]	0.5297 / 54.27%	0.5622 / 60.77%
LFM-based (68x68 LFM, w/o DA)	**0.5876 / 58.54%**	**0.6066 / 63.57%**
CLFM-based (21x21 LFM, w/o DA)	**0.4978 / 55.49%**	**0.5779 / 62.79%**

dataset [1] which is usually employed for evaluating facial general-expression (FGE), i.e., macro-expression recognition performance. CK+ consists of 327 sequences which are labeled with seven emotion classes. Generally, FME/FGE do not exist simultaneously for the same person. Therefore, to generate FME as well as FGE for the same person, we artificially synthesized FME sequences from FGE sequences of CK+ dataset. This becomes the second dataset. Although the synthesized FME may differ from the real one, we assumed that they are similar each other. The FME synthesis process is as follows. The first frame in which a facial expression starts in each sequence of the CK+ is detected, and three consecutive frames from that frame are selected. Since the starting points of the facial expressions usually vary from sequence to sequence, the starting point of each sequence is found manually in this paper. Since the movements of LMs is small at the beginning of emotional expression, we consider these selected frames as FME sequences without loss of generality. Next, each FME sequence is four times interpolated through Simony's method [14]. In other words, three-frame sequence having a specific FME is converted into a sequence of twelve frames having the same FME. The synthesized FME dataset is available at https://github.com/pride0723/LFM2D. The mean SFEM of the FME dataset is 0.32, which amounts to 13.2% of the FGE dataset.

1) VISUAL ANALYSIS OF LFM PATTERN ACCORDING TO EMOTION CLASS

In this section, we visually investigate LFM patterns according to facial expressions. First, Fig. 6 shows LFMs

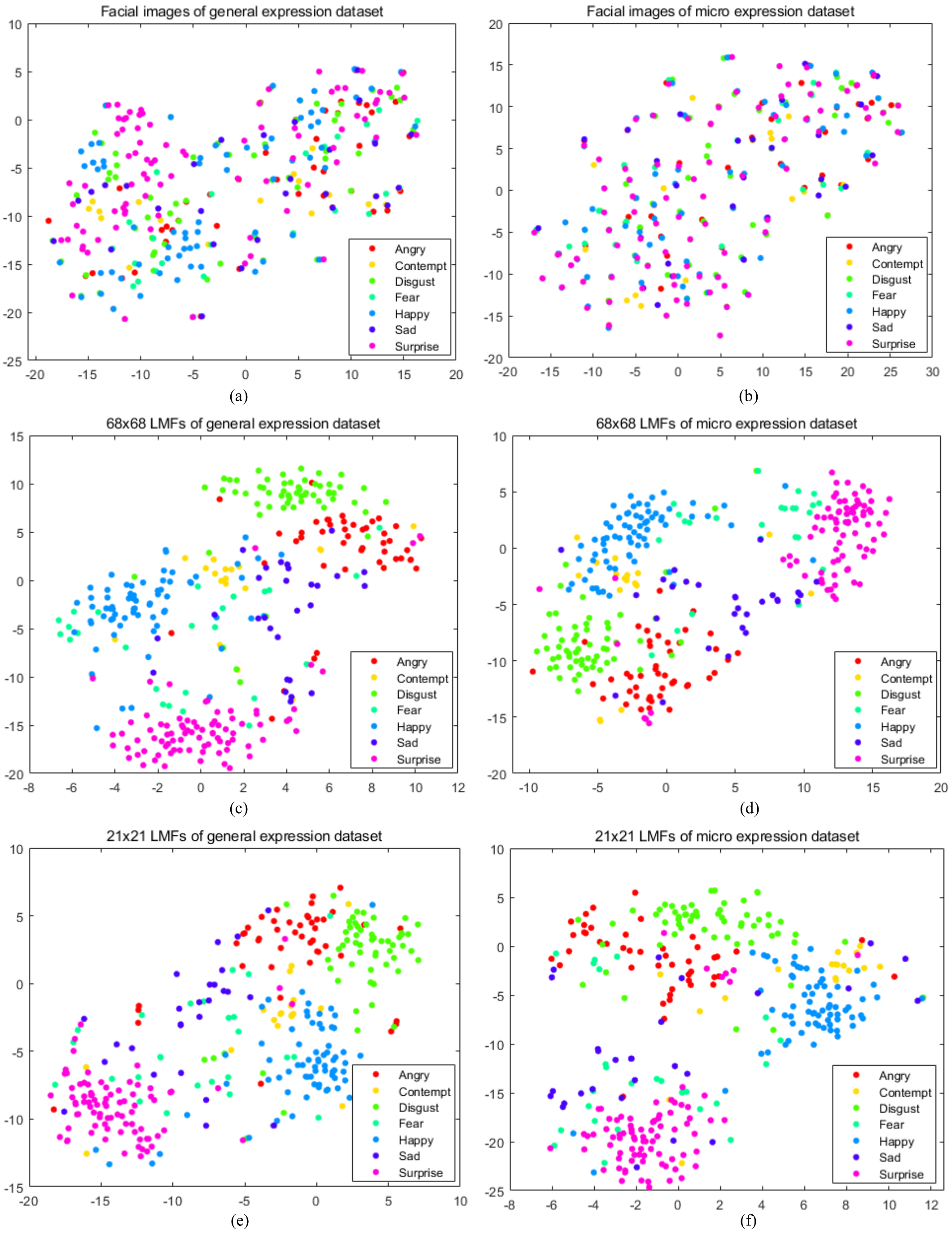


FIGURE 7. T-SNE comparison for FGE and FME datasets (a) Facial images for FGE dataset, (b) facial images for FME dataset, (c) LFMs for FGE dataset, (d) LFMs for FME dataset, (e) CLFMs for FGE dataset, (f) CLFMs for FME dataset.

generated according to seven emotions of FGE dataset, i.e., the first dataset. We can observe that the CLFM (the second row) as well as the LFM (the first row) shows a unique pattern depending on the emotion class.

Second, we adopted t-SNE [23] to examine the similarity between emotion classes. Figure 7 are the t-SNEs of facial images, LFMs, and CLFMs for both FGE and FME datasets. Figure 7 (a) shows the t-SNE distribution of facial images of the FGE dataset. Here, facial images were cropped as shown in the 3rd row of Fig. 6 and were resized to 128 × 128. We can find that the distribution of ‘surprise’ class has some concentration, but is not properly separated. Figure 7 (b) shows an example for FME dataset. We cannot observe any concentration according to emotion class. The distribution of some data may be crowded because different emotions for the same person have very high correlation. On the other hand, Fig. 7 (c) to (f) show t-SNEs of LFM and CLFM. We can see that the data of the same emotion class is properly clustered

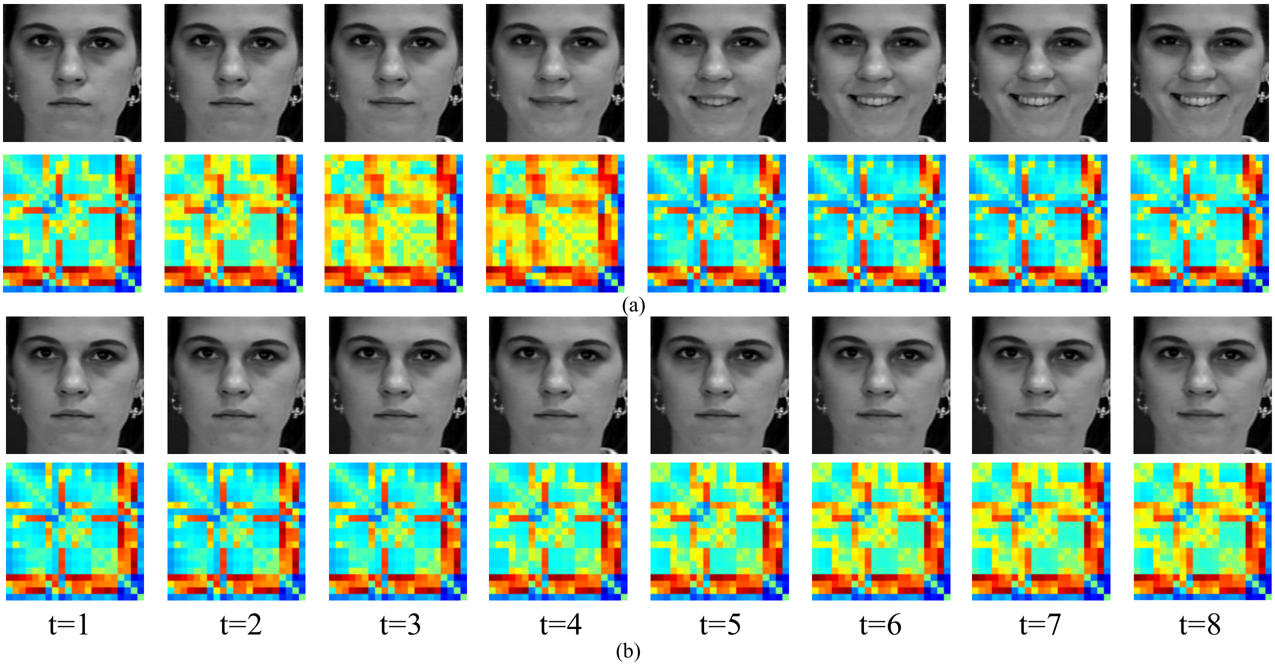


FIGURE 8. The facial images and CLFMs for "happiness" (S097_004 sequence in CK+ dataset) (a) CLFMs for FGE, (b) CLFMs for FME.

in case of FME as well as FGE. Since even different people have statistically similar movements of LMs, LFM shows a unique pattern depending on the emotion. As a result, unlike the facial images, LFMs have an advantage of eliminating redundant personal characteristics such as age and gender. In other words, the data distributions of Fig. 7 (c)-(f) are much less difficult in classification compared to those of Fig. 7 (a) and (b). Thus, the proposed method provides better performance.

2) LFM COMPARISON ACCORDING TO FACIAL EXPRESSION STRENGTH

In this section, we examine whether different LFMs of the same emotion show similar patterns regardless of the emotion intensity. For this experiment, we extracted CLFMs as shown in Fig. 8 (a) by selecting a video clip of 'happiness' (see Fig. 1 (a)) from CK+ dataset. This corresponds to FGE. On the other hand, CLFMs of the corresponding synthesized FME are shown in Fig. 8 (b). Similar to Fig. 8 (a), we can observe the emotion-specific pattern even in Fig. 8 (b). Therefore, LFM/CLFM works well not only for FGEs but also for FMEs because it can generate similar patterns for the same emotion regardless of the emotion intensity.

C. PERFORMANCE EVALUATION OF MICRO-EXPRESSION AND MACRO-EXPRESSION

In this section, the performance of the proposed method is verified for the FGE and FME datasets. The classifier network was trained only with FGE dataset. Using the same trained model, the test for FGE was performed with the original CK+ dataset, but the test for FME was performed with the synthesized FME dataset. Additionally, we made 'mixed expression' dataset which were randomly sampled in 50:50 from FGE and FME datasets, and tested the proposed method even for the mixed expression dataset. For all the datasets, we employed the most popular 10-fold validation protocol. That is to say, nine subsets were used for training the networks, and the remaining subset was used for validation.

TABLE 8. Performance comparison of FER algorithms according to facial expression strength in CK+ dataset. Here, the parenthesis of DTAGN indicates the value published in [11].

	Test dataset		
	FGE dataset	FME dataset	Mixed dataset
DTAGN [11]	**93.88%** (97.25%)	43.34%	70.95%
LFM-based (68x68 LFM)	92.66%	77.98%	87.46%
CLFM-based (21x21 LFM)	88.78%	**78.60%**	85.63%
LFM and DTGAN [11] (adaptive method)	N/A	N/A	88.69%
CLFM and DTGAN [11] (adaptive method)	N/A-	N/A	**89.30%**

The proposed method was compared with DTAGN [11]. The code released by the authors was used and the augmentation was applied to the training data of DTAGN. To our knowledge, [11] is the only FER method whose source code is available. Table 8 shows emotion classification result for both FGE and FME. For FGE dataset, LFM and CLFM provide the accuracies of about 93% and 89%, respectively. The performance of LFM-based network is close to that of DTAGN. CLFM-based network shows about 4% lower accuracy than LFM-based one. This performance degradation is caused by LM sampling as mentioned in Section V-A. However, note that as shown in Table 2, the number of

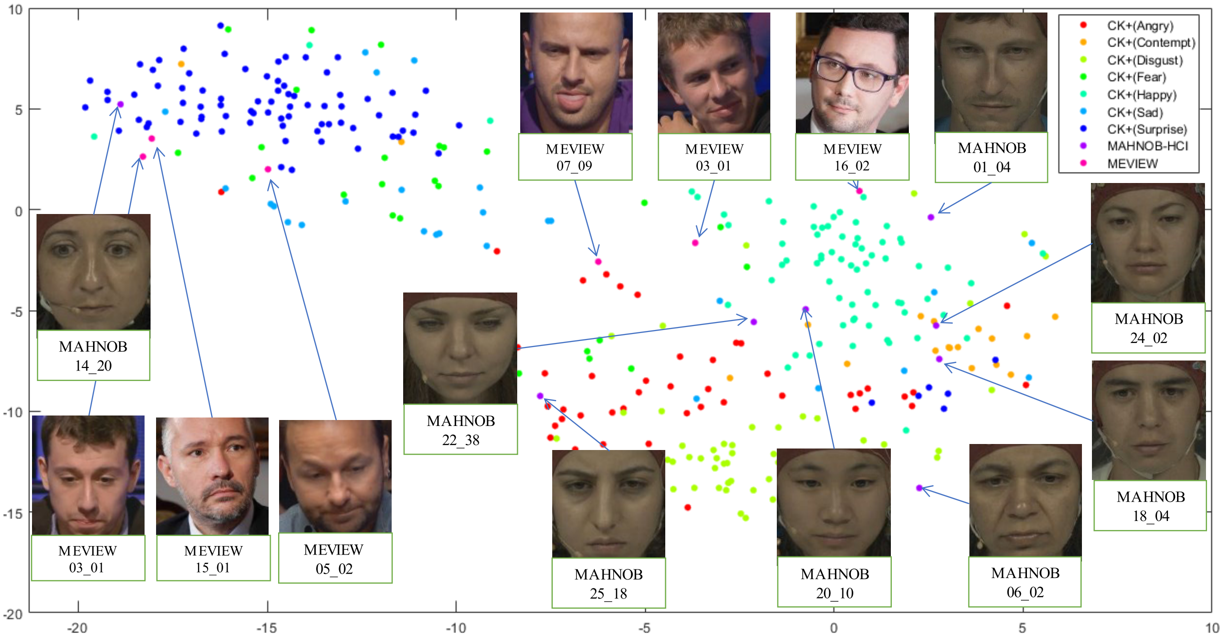


FIGURE 9. T-SNE map of 21 × 21 2D LMFs extracted from the CK+ FME, MAHNOB-HCI and MEVIEW dataset.

TABLE 9. Confusion matrix for FGE in case of the 68 × 68 LMF (a) LFM and (b) CLFM.

	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
An.	**86.7**	0.0	4.4	2.2	0.0	6.7	0.0
Co.	5.6	**77.8**	0.0	0.0	11.1	0.0	5.6
Di.	1.7	0.0	**96.6**	1.7	1.7	0.0	0.0
Fe.	0.0	0.0	4.0	**84.0**	8.0	0.0	4.0
Ha.	0.0	0.0	2.9	0.0	**97.1**	0.0	0.0
Sa.	10.7	3.6	0.0	0.0	0.0	**82.1**	3.6
Su.	0.0	1.2	0.0	0.0	0.0	0.0	**98.8**

(a)

	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
An.	**84.4**	0.0	11.1	2.2	0.0	2.2	0.0
Co.	5.6	**55.6**	5.6	0.0	22.2	5.6	5.6
Di.	3.4	0.0	**94.9**	0.0	1.7	0.0	0.0
Fe.	0.0	0.0	0.0	**60.0**	0.0	16.0	24.0
Ha.	0.0	0.0	0.0	0.0	**98.6**	0.0	1.4
Sa.	17.9	0.0	0.0	14.3	0.0	**64.3**	3.6
Su.	0.0	1.2	0.0	0.0	0.0	0.0	**98.8**

(b)

parameters of CLFM is only 1/12 of LFM. In case of FME, LFM and CLFM provide the accuracies of about 78% and 78.6%, respectively. Here, their performances become similar each other. Note that CLFM is superior to DTAGN by about 35%. Also, CLFM shows about 15% higher performance than DTAGN even for mixed dataset. Finally, the joint framework of CLFM and DTAGN as shown in Section IV-B shows a performance improvement of about 3.7% over CLFM alone for the mixed dataset. This is because DTAGN maintains reasonable performance in the general macro-expression section of the mixed expression sequence.

Table 9 shows the confusion matrices for the FGE dataset. The recognition performance of CLFM (Table 9 (b)) is lowered in 'contempt', 'fear' and 'sad' compared to LFM (Table IX (a)). Seeing facial images in the 3rd row of Fig. 6, we can find that the movements around the mouth are dominant information for those emotions. As mentioned in Section V-A, the LM density of CLFM decreases around the mouth and thus the recognition performance of the related emotions deteriorates. Similarly, Table 10 shows the confusion matrices for the FME dataset. In case of CLFM, the accuracies of 'contempt', 'fear' and 'sad' emotions somewhat decrease. However, the performance of CLFM improves for 'disgust' emotion where eyes and eyebrows become dominant information. As a result, the FME accuracy of CLFM is better than that of LFM by 0.62% on average as shown in Table 8.

D. RESULTS FOR REAL EMOTIONAL DATASET

The CK+ dataset consists of acted emotional images. So, this section shows the qualitative experimental result for actual and wild datasets, i.e., MAHNOB-HCI [19] and MEVIEW [57]. Each data of MAHNOB-HCI was produced by sensing the faces and bio-signals generated while the subjects are watching the emotional stimulus contents, and it was annotated in the continuous domain of arousal and valance. On the other hand, unlike the datasets such as SMIC and CASME II, MEVIEW dataset was produced by focusing on FME in wild conditions. Each content of MEVIEW dataset was mostly captured from poker games and TV interviews.

Figure 9 shows the t-SNE map which displays CLFMs extracted from the CK+ FME, MAHNOB-HCI and MEVIEW dataset. We can observe that images having the same emotion are mapped to similar positions (compare the displayed images). In addition, Fig. 10 shows the examples of CLFMs obtained from video clips from three datasets.

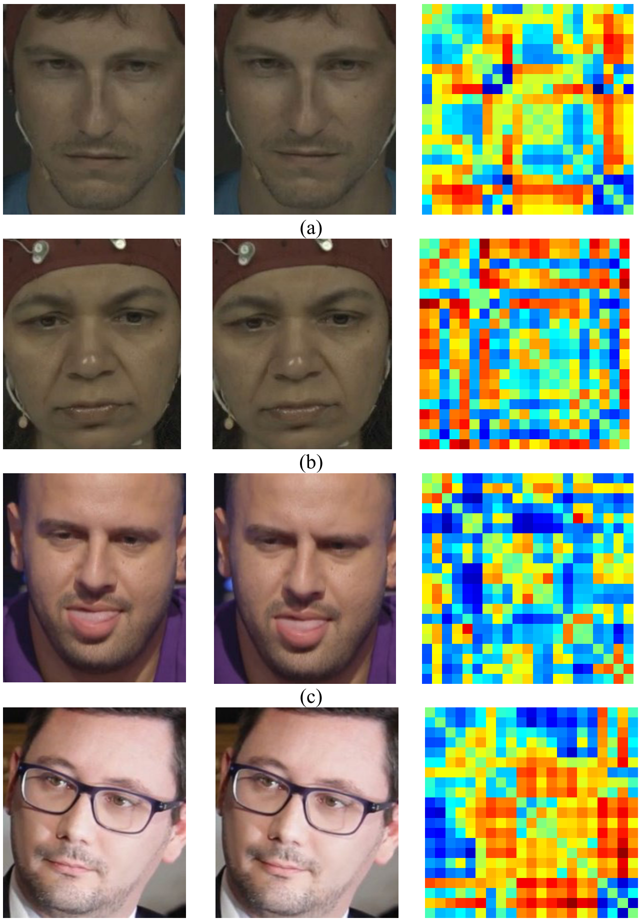
TABLE 10. Confusion matrices for FME dataset (a) LFM and (b) CLFM.

	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
An.	**77.8**	0	11.1	4.4	0	4.4	2.2
Co.	0.0	**72.2**	0	0	11.1	5.6	1.1
Di.	8.5	0	**79.7**	0	10.1	1.7	0
Fe.	12.0	0	8.0	**48.0**	12.0	8.0	12.0
Ha.	1.4	0	4.3	5.8	**81.2**	1.5	5.8
Sa.	10.7	3.6	3.6	10.7	0	**71.4**	0
Su.	1.2	3.61	2.4	1.2	0	4.8	**86.7**

(a)

	An.	Co.	Di.	Fe.	Ha.	Sa.	Su.
An.	**75.6**	0	17.8	4.4	0	2.2	0
Co.	11.1	**50.0**	5.6	0	22.2	0	11.1
Di.	1.7	1.7	**89.8**	0	3.4	1.7	1.7
Fe.	4.0	0	0	**36.0**	16.0	8.0	36.0
Ha.	2.9	4.4	1.5	0	**87.0**	0	4.4
Sa.	10.7	0	3.6	3.6	3.6	**67.7**	10.7
Su.	4.8	2.4	0	1.2	0	3.6	**88.0**

(b)



(a)

(b)

(c)

(d)

FIGURE 10. Facial images and their CLFMs for some video clips. (a) Index: MAHNOB_HCI 01_04, (b) index: MAHNOB_HCI 06_02. (c) MEVIEW 07_09, (d) MEVIEW 16_02, Here, left, center, and right columns indicate the 1st frame, the last frame of the sequence, and CLFM, respectively.

Figure 10 (a) corresponds to 'happy' emotion in discrete domain view. The CLFM obtained here is very similar to the 'happy' emotional pattern in Fig. 5. Also, the CLFM of Fig. 10 (b) corresponding to 'sad' emotion is similar to the pattern corresponding to 'sad' of Fig. 6. On the other hand, Fig. 10 (c) for the MEVIEW dataset is similar to the pattern of 'angry' in Fig. 6, and Fig. 10 (d) is similar to the pattern of 'happy'. Therefore, we can qualitatively find that the proposed method is effective not only on the facial expression dataset in the artificial environment such as CK+ dataset, but also on the micro-expressions existing in the natural emotional response dataset, i.e., MAHNOB-HCI and MEVIEW datasets. This shows that the proposed LFM achieves effective emotion recognition by extracting robust features irrespective of the characteristics of the facial datasets.

VI. CONCLUDING REMARKS

The proposed LFM generates a unique pattern according to emotion class regardless of the strength of facial emotion, thus enabling effective emotion recognition. For micro-expression datasets such as SMIC and CASME II, the proposed method showed superiority in accuracy and F1-score over SOTAs. Even in the experiments on composite dataset and cross dataset validation, we proved that the proposed method outperforms the conventional techniques. Also, even if the proposed method is learned with the FGE training set, it can cope with the test set such as FME. We also proposed a compact LFM (CLFM) that consists only of LM points that have a major effect on emotion. Compared with our previous work [16], CLFM reduces the classifier cost to only 8% and concurrently minimizes the performance degradation. In the case of FME, the proposed method shows a performance gain of more than 35% over the conventional method [11]. In addition, the adaptive scheme that can selectively choose between the proposed method and [11] shows an accuracy improvement of about 18% over [11] for the mixed expression dataset.

REFERENCES

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[2] 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.

[3] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1449–1456.

[4] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan, and J. See, "Eulerian emotion magnification for subtle expression recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1243–1247.

[5] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21665–21690, Oct. 2017.

[6] 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 *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] R. Zhao, Q. Gan, S. Wang, and Q. Ji, ''Facial expression intensity estimation using ordinal information,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3466–3474.

[9] K. Keung Lee and Y. Xu, ''Real-time estimation of facial expression intensity,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Sep. 2003, pp. 2567–2572.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, ''Active appearance models,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[11] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, ''Joint fine-tuning in deep neural networks for facial expression recognition,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.

[12] K. Zhang, Y. Huang, Y. Du, and L. Wang, ''Facial expression recognition based on deep evolutional spatial-temporal networks,'' *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[13] 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

[14] S. Niklaus, L. Mai, and F. Liu, ''Video frame interpolation via adaptive separable convolution,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 261–270.

[15] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, ''LEARNet: Dynamic imaging network for micro expression recognition,'' *IEEE Trans. Image Process.*, vol. 29, pp. 1618–1627, 2020.

[16] D. Y. Choi, D. H. Kim, and B. C. Song, ''Recognizing fine facial micro-expressions using two-dimensional landmark feature,'' in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1962–1966.

[17] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, ''Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 360–368.

[18] V. Esmaeili and S. O. Shahdi, ''Automatic micro-expression apex spotting using cubic-LBP,'' *Multimedia Tools Appl.*, pp. 1–19, Apr. 2020.

[19] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, ''A multimodal database for affect recognition and implicit tagging,'' *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[20] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, ''Incremental face alignment in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1859–1866.

[21] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, ''Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

[22] W.-T. Chu and Y.-H. Liu, ''Thermal facial landmark detection by deep multi-task learning,'' in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 94–108.

[23] L. Maaten and G. Hinton, ''Visualizing data using t-SNE,'' *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[24] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, ''A spontaneous micro-expression database: Inducement, collection and baseline,'' in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[25] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, ''CASME II: An improved spontaneous micro-expression database and the baseline evaluation,'' *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041.

[26] S. J. Wang, W. J. Yan, G. Zhao, X. Fu, and C. G. Zhou, ''Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features,'' in *Proc Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 325–338.

[27] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen, ''Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods,'' *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, Oct. 2018.

[28] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, ''Learning from hierarchical spatiotemporal descriptors for micro-expression recognition,'' *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3160–3172, Nov. 2018.

[29] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, ''Sparse MDMO: Learning a discriminative feature for spontaneous micro-expression recognition,'' *IEEE Trans. Affect. Comput.*, early access, Jul. 9, 2018, doi: 10.1109/TAFFC.2018.2854166.

[30] W. Peng, X. Hong, Y. Xu, and G. Zhao, ''A boost in revealing subtle facial expressions: A consolidated Eulerian framework,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[31] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, ''Micro-expression recognition using color spaces,'' *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, Dec. 2015.

[32] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, ''Micro-expression recognition with small sample size by transferring long-term convolutional neural network,'' *Neurocomputing*, vol. 312, pp. 251–262, Oct. 2018.

[33] S. Xie and H. Hu, ''Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks,'' *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211–220, Jan. 2019.

[34] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, ''A deep neural network-driven feature learning method for multi-view facial expression recognition,'' *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[35] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, ''Learning personalized models for facial expression analysis and gesture recognition,'' *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775–788, Apr. 2016.

[36] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, ''SAMM: A spontaneous micro-facial movement dataset,'' *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.

[37] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, ''MEGC 2019—The second facial micro-expressions grand challenge,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[38] 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.

[39] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, ''Less is more: Micro-expression recognition from video using apex frame,'' *Signal Process., Image Commun.*, vol. 62, pp. 82–92, Mar. 2018.

[40] N. V. Quang, J. Chun, and T. Tokuyama, ''CapsuleNet for micro-expression recognition,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.

[41] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, ''OFF-ApexNet on micro-expression recognition system,'' *Signal Process., Image Commun.*, vol. 74, pp. 129–139, May 2019.

[42] L. Zhou, Q. Mao, and L. Xue, ''Dual-inception network for cross-database micro-expression recognition,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[43] 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 *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[44] Y. Liu, H. Du, L. Zheng, and T. Gedeon, ''A neural micro-expression recognizer,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–4.

[45] C. Guo, J. Liang, G. Zhan, Z. Liu, M. Pietikainen, and L. Liu, ''Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition,'' *IEEE Access*, vol. 7, pp. 174517–174530, 2019.

[46] T. Zhang, Y. Zong, W. Zheng, C. L. P. Chen, X. Hong, C. Tang, Z. Cui, and G. Zhao, ''Cross-database micro-expression recognition: A benchmark,'' *IEEE Trans. Knowl. Data Eng.*, early access, Apr. 6, 2020, doi: 10.1109/TKDE.2020.2985365.

[47] C.-C. Chang and C.-J. Lin, ''LIBSVM: A library for support vector machines,'' *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[48] A. Hassan, R. Damper, and M. Niranjan, ''On acoustic emotion recognition: Compensating for covariate shift,'' *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1458–1468, Jul. 2013.

[49] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, ''Domain adaptation via transfer component analysis,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[50] B. Gong, Y. Shi, F. Sha, and K. Grauman, ''Geodesic flow kernel for unsupervised domain adaptation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[51] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, ''Unsupervised visual domain adaptation using subspace alignment,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.

[52] W.-S. Chu, F. De la Torre, and J. F. Cohn, ''Selective transfer machine for personalized facial action unit detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3515–3522.

[53] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 529–545, Mar. 2017.

[54] M. Long, J. Wang, J. Sun, and P. S. Yu, "Domain invariant transfer kernel learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1519–1532, Jun. 2015.

[55] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning a target sample re-generator for cross-database micro-expression recognition," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 872–880.

[56] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2484–2498, May 2018.

[57] P. Husák, J. Cech, and J. Matas, "Spotting facial micro-expressions 'in the wild,"' in *Proc. 22nd Comput. Vis. Winter Workshop (Retz)*, 2017, pp. 1–9.

[58] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Trans. Affect. Comput.*, early access, Apr. 13, 2020, doi: 10.1109/TAFFC.2020.2986962.

DONG YOON CHOI (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Inha University, Incheon, South Korea, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree in electronic engineering. His research interests include image processing, computer vision, semi-supervised learning, and multimodal deep-learning.

BYUNG CHEOL SONG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1994, 1996, and 2001, respectively. From 2001 to 2008, he was a Senior Engineer with Digital Media Research and Development Center, Samsung Electronics Company Ltd., Suwon, South Korea. In March 2008, he joined the Department of Electronic Engineering, Inha University, Incheon, South Korea, where he is currently is a Professor. His research interests include image processing and computer vision.