Joint Local and Global Information Learning with Single Apex Frame Detection for Micro-expression Recognition

Yante Li, Xiaohua Huang *Member, IEEE,* and Guoying Zhao* *Senior Member, IEEE,*

***Abstract*—Micro-expressions (MEs) are rapid and subtle facial movements that are difficult to detect and recognize. Most recent works have attempted to recognize MEs with spatial and temporal information from video clips. According to psychological studies, the apex frame conveys the most emotional information expressed in facial expressions. However, it is not clear how the single apex frame contributes to micro-expression recognition. To alleviate that problem, this paper firstly proposes a new method to detect the apex frame by estimating pixel-level change rates in the frequency domain. With frequency information, it performs more effectively on apex frame spotting than the currently existing apex frame spotting methods based on the spatio-temporal change information. Secondly, with the apex frame, this paper proposes a joint feature learning architecture coupling local and global information to recognize MEs, because not all regions make the same contribution to ME recognition and some regions do not even contain any emotional information. More specifically, the proposed model involves the local information learned from the facial regions contributing major emotion information, and the global information learned from the whole face. Leveraging the local and global information enables our model to learn discriminative ME representations and suppress the negative influence of unrelated regions to MEs. The proposed method is extensively evaluated using CASME, CASME II, SAMM, SMIC, and composite databases. Experimental results demonstrate that our method with the detected apex frame achieves considerably promising ME recognition performance, compared with the state-of-the-art methods employing the whole ME sequence. Moreover, the results indicate that the apex frame can significantly contribute to micro-expression recognition.**

***Index Terms*—Micro-expression, 3D FFT, Facial expression recognition, Multi-instance learning, Deep learning**

I. INTRODUCTION

MICRO-EXPRESSIONS (MEs) are involuntary facial movements reacting to emotional stimulus [1]. MEs can reveal people's hidden feelings in high-stake situations and have many potential applications in different fields, such as clinical diagnosis, national security, and interrogations. Different from ordinary facial expressions that we see daily, MEs have short duration (1/25 to 1/3 second), low intensity, and occur with sparse facial action units [2]. All of the above characteristics make MEs difficult to detect and recognize.

Generally, two main tasks are included in ME analysis: spotting and recognition. The spotting task is aimed at identifying ME occurrence or finding the onset, offset and apex frames, while the recognition task classifies the MEs into specific emotion categories [3]–[5]. Most of the current research on ME recognition utilizes whole video clips [6]–[10]. Ekman declared that 'snapshot taken at a point when the expression is at its apex can easily convey the emotion message' [11]. This means that, the apex frame can contribute major information to facial expression recognition. Recently, Liong *et al.* discovered that the redundancy information in ME clips could decrease the performance of ME recognition [3]. In contrast, the onset, apex, and offset frames provide useful information to ME classification. Moreover, Liong *et al.* proposed a bi-weighted orientation optical flow feature extracted on the spotted apex frame for ME recognition [12]. However, so far there are few studies that analyze the contribution of the apex frame to ME recognition. On the other hand, as deep learning technology has achieved considerable performance in facial expression recognition [13], [14], some researchers have started to exploit deep neural networks for ME recognition [15], [16]. However, their proposed methods dramatically degrade the performance compared with hand-crafted methods [7]. This is explained by the fact that ME databases are very small and the changes in MEs are subtle. Motivated by the above-mentioned observations [3], [11], [15], [16], this paper provokes the three following discussions: (1) 'Does the single apex frame in ME contribute the important information for ME recognition?'; (2) 'How is the recognition result based on the apex frame compared with the methods employing the ME sequence?' and (3) 'Can deep learning achieve good performance of ME recognition with the apex frame?'.

To address the aforementioned questions, the first stage of this paper is to locate the apex frame in the ME sequence. Currently, most of the existing spontaneous ME apex frame spotting methods estimate the facial muscle change in the spatio-temporal domain to detect the apex frame in ME sequences, *e.g.*, optical flow-based methods [17]. Actually, for micro-expression, facial muscle change is not obvious along the temporal dimension. To a certain extent, these kinds of methods are prone to errors when spotting the apex frame. According to our empirical analysis [18], the apex frames in ME sequences are highly related to the high-frequency information. To this end, the frequency, even though rarely utilized by the current existing research, can provide rich and important information to apex frame spotting. This paper proposes a new method to locate the apex frame through frequency analysis. Different from commonly used ME spotting methods based

*Corresponding author.

Y. Li and G. Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Emails: {yante.li, guoying.zhao}@oulu.fi.

X. Huang is with the School of Computer Engineering, Nanjing Institute of Technology, Nanjing, China. Email: xiaohua.huang@njit.edu.cn.

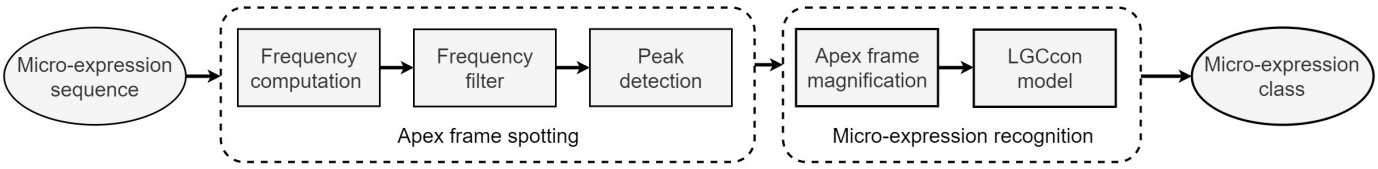


Fig. 1. Framework of the micro-expression recognition system.

on the change between frames, the proposed frequency-based method in this paper utilizes the frequency representation of facial muscle change in the frequency domain to locate the apex frame.

As previously discussed, most pioneers utilizing deep learning technology and video clips cannot obtain satisfactory results for micro-expression recognition [15], [16]. To gain an improvement, our previous work [18] used the deep VGG model [19] to classify the magnified ME apex frame, but still cannot achieve significant improvement, especially for the subjects with outliers. For example, our previous work [18] obtained an accuracy of 64.41% for the subjects without eyeglasses, 61.5% for the subjects with eyeglasses, and 33% for some subjects with big eyeglasses. It was found that eyeglasses have a seriously negative influence on the performance of ME recognition. The empirical experience and quantitative analysis in [20] also validated that observation. It is of importance to suppress the influence of outliers such as eyeglasses.

Some research has reduced the impact of regions without useful information by extracting features on the regions of interest (**ROIs**) [12]. Here, Liong *et al.* assumed all the ROIs have an equal contribution to micro-expression recognition. As far as we know, despite ROIs, the informative region could be attributed to different micro-expression. In other words, only specific AUs are triggered when facial expression occurs. For example, anger is mostly related to AU4 (Brow lower) or AU7 (Lids tight) [21]. Compared with the motionless regions, the local region related to AUs may contribute more information to ME recognition. Therefore, a novel learning framework termed LGCcon is proposed to join local and global information to emphasize the local informative region learning among global information for ME recognition. The architecture is designed based on the local informative region and the global face derived from the apex frame. Specifically, the local informative region contributing most ME information is automatically learned by following the concept of multi-instance learning (**MIL**) [22]. It aims to improve the discriminative representation ability and suppress the impact of outliers and motionless regions. Additionally, local and global information learning constraints are developed in our framework to improve the performance of local and global representations, respectively. Moreover, Centerloss [23] is used to increase the compactness of intra-class variations and separable inter-class differences.

The contribution of this paper is threefold:

(1) A new method, termed 3DF-N, is proposed to locate the apex frames in the frequency domain. Due to frequency information, which describes the rate of change clearly, the proposed method can effectively spot the apex frame of micro-expression videos.

(2) A local and global information joint learning module (LGCcon) is proposed to improve the ability and robustness of discriminative representation against the problem of outliers.

(3) We develop a new deep learning framework to recognize MEs based on apex frames, and to further study the contribution of apex frames to ME recognition. The experiments demonstrate that our method based on the apex frame can achieve comparable promising performance, when compared with ME sequences. The experimental results further indicate that deep learning can achieve good performance on ME recognition with the apex frame. Figure 1 shows our whole framework.

The remaining parts of this paper are organized as follows: Section II presents the related work, while Section III details two proposed methods for apex frame spotting and ME recognition, respectively. Section IV discusses the experiments and analyzes the results. Section V makes conclusions on the proposed methods and observations. A preliminary version of this work was presented in [18]. The work in this paper is substantially extended in four aspects: (1) The apex frame spotting method is further improved through reducing redundancy information, which is achieved by locating regions with large change rates. (2) To suppress the influence of outliers and motionless regions for ME recognition, a MIL-based method is proposed to automatically detect the most important information on the face. (3) To further gain discriminative representation ability, a local maximum and global context joint learning framework is designed to adaptively embed local and global information. (4) Intensive experiments are conducted on the CASME, CASME II, SAMM, SMIC, and composite databases to demonstrate the effectiveness and generalizability of LGCcon.

II. RELATED WORKS

This section briefly summarizes the existing study on micro-expression analysis. The techniques on micro-expression spotting and recognition are described to indicate the research focus. As the proposed local information learning method is following the concept of multi-instance learning, MIL is also presented briefly.

A. Micro-expression Spotting

For ME spotting, due to the subtle and rapid change characteristics of MEs, it is difficult to locate the onset, apex, and offset frames accurately. Patel *et al.* [24] used integrate optical flow vectors computed on small local spatial regions to spot the onset and offset frames from a long-term video.

Li *et al.* [25] proposed a training-free method based on the feature difference contrast and peak detection to spot MEs. On the other hand, Liong *et al.* [17] used the binary search strategy with a local binary pattern and optical flow on several interesting facial sub-regions to spot the apex frame in ME clips. And Ma *et al.* [26] further improved the performance of apex frame spotting by utilizing the histogram of oriented optical flow. However, these methods merely concerned the subtle spatial change between neighboring frames, but omitted the rapid change of frames along the temporal domain. In contrast, the proposed apex frame spotting method based on 3D Fast Fourier Transform (**FFT**) not only analyzes rapid changes of ME in the frequency domain, but also leverages the spatial and continuous temporal information.

B. Micro-expression Recognition

The ME recognition research can be traced to the work of Pfister *et al.* [27]. Pfister *et al.* proposed recognizing ME by using a local binary pattern from three orthogonal planes (LBP-TOP) [28] and classical classifiers. Following the work of [28], to increase the efficiency of LBP-TOP for ME recognition, Wang *et al.* [29] proposed a spatio-temporal descriptor with six intersection points (LBP-SIP), also suppressing the redundancy information of LBP-TOP. In order to improve the performance of ME recognition, certain spatio-temporal descriptors have been proposed, *e.g.*, the spatio-temporal completed local quantized pattern (STCLQP) [7] and a histogram of image gradient orientation (HIGO-TOP) [25]. Likewise, other feature types like main direction main optical (MDMO) [30] and tensor independent color space (TICS) [31] methods were proposed. All the aforementioned methods are mostly based on the whole video clip. However, there remains a query over which frame could significantly contribute to micro-expression recognition. Liong *et al.* [12] attempted to use apex frames for ME recognition, but unfortunately, this system based on apex frames cannot gain an improvement, and in contrast, it still behaved worse than the state-of-the-art methods [25], [30] throughout the whole video clip. Even so, using apex frame could obtain high efficiency to some extent in a real-world application.

In recent years, deep learning has achieved promising performance in many research fields [13], [14]. It has also been used in ME recognition [16], [32]. The work in [15] was the first to transfer deep convolutional neural network models from objects and facial expressions to small ME databases. However, its recognition rate on the CASME II database is 47.3%, which is worse than hand-crafted descriptors. Peng *et al.* [16] proposed a dual-template CNN model based on optical flows extracted from the ME sequences for ME recognition. The optical flow information over the whole video should first be extracted and then fed into CNN. Actually, the extraction of optical flow leads to heavy computation in real-world applications, which seriously degrades the efficiency of the dual-template CNN model. Li *et al.* [33] proposed a novel automatic ME analysis algorithm utilizing the Flownet 2.0 [32]. With the benefit of Flownet, Li *et al.* improved the performance of dual-template CNN [16], but it is still inferior to classical methods [25]. More recently, methods [34], [35] based on optical flow between the onset, apex, and offset frames considerably increased ME recognition performance compared with algorithms employing ME clips [15], [25], [33]. This indirectly indicates that not all frames in ME clips make a contribution to ME recognition. Thus, this paper puts the focus on the contribution of apex frames to ME recognition. The experimental results validate that the apex frame is the most contributing one for ME recognition.

C. Multi-instance Learning

Multi-instance learning (**MIL**) supplies a training framework for resolving the problem of inaccurate annotations. Different from supervised algorithms [13], [14] needing accurate annotations, MIL merely requires data in the form of bags with positive or negative labels [36]. If there is at least one positive instance in the bag, the bag is positive, otherwise, it is negative. With the obscure bags, MIL can annotate individual instances correctly [36].

MIL has achieved excellent performance in several computer vision tasks. For example, Oquab *et al.* [22], [37] proposed to regarding sub-regions in the complex image as instances in a bag by following MIL. The image is classified through combining sub-region scores through a max pooling layer. For action recognition, Gkioxari *et al.* [38] employed MIL to locate an action region and recognized the action with the contextual information.

For ME recognition, most MEs are related to one or two AUs. This means that the regions related to AUs contribute more to ME recognition. On the other hand, outliers, *e.g.*, eyeglasses, have a negative influence on ME recognition. Following the concept of MIL, the candidate facial local sub-regions are considered as the instances. The proposed LGCcon obtains the positive instance contributing the most ME information for inferring the ME class automatically via a maximum operation. It combines MIL and deep learning architecture to learn the local informative feature from the local facial region. It also largely suppresses the influence of some noisy regions, such as those occluded by eyeglasses.

III. MICRO-EXPRESSION RECOGNITION BASED ON THE APEX FRAME

In this section, the proposed methods for ME recognition are detailed. As we study the contribution of apex frame to ME recognition, the first stage is to locate the apex frame. Subsection A introduces the flow of spotting an apex frame in a ME sequence through frequency analysis. Then the LGCcon network is proposed to recognize ME based on the detected apex frame through joint local and global information learning. In Subsection B, the specific details of LGCcon are elaborated. Finally, we introduce the implementation of LGCcon in Subsection C.

A. Micro-expression apex frame spotting

As previously discussed in Section I, the subtle change of ME leads to a hard locating apex frame in the spatio-temporal domain. According to our empirical experience [18],

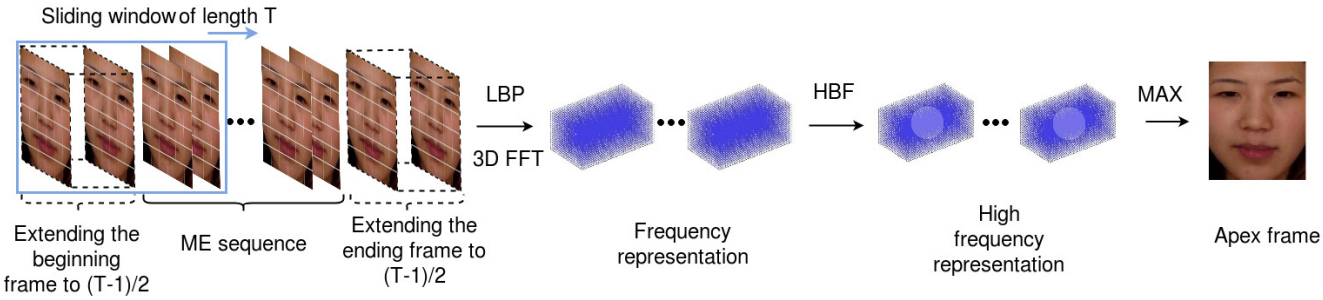


Fig. 2. The flowchart for the proposed 3DF-N for micro-expression spotting. HBF represents the high-band frequency filter.

the frequency can clearly express the subtle but rapid pixel changes in ME sequences. Thus, this paper spots the apex frames in the frequency domain instead. The basic idea is to represent each ME frame with the frequency components at short intervals, and to locate the apex frame through comparing the frequencies afterwards. Figure 2 depicts the flowchart of the proposed ME apex frame spotting method.

According to [39], it is found that the frequency is sensitive to illumination variations. Prior to analyzing frequency, gray-scale invariant Local Binary Patterns (LBP) [40] is used to extract the texture map of the ME frame, which suppresses the influence of illumination change to frequency. Subsequently, the frequency of sequential video frames is obtained at a specified interval. For more details, the facial area is divided into equal-sized blocks (6×6 in the experiments). Afterwards, the video blocks are transformed into frequency domain through 3D Fast Fourier Transformation (**3D-FFT**) with a sliding time window. Given the sliding window of length T, for the i-th interval, the frequency values for the interval are computed on blocks by 3D-FFT. The frequency value of the j-th block in the i-th interval is obtained as follows:

$$F_{b_{ij}}(u,v,q) = \int_{\frac{-T}{2}}^{\frac{T}{2}} \int_{\frac{-L_b}{2}}^{\frac{L_b}{2}} \int_{\frac{-W_b}{2}}^{\frac{W_b}{2}} f_{b_{ij}}(x,y,z) \times e^{j2\pi(ux+vy+qz)} dx dy dz, \quad (1)$$

where (u, v, q) represents the position in the frequency domain; L_b and W_b represent the height and width of the j-th block b_{ij} in the i-th interval, respectively and $j = \{1, 2, \cdots, 36\}$.

Based on the observation [18], the apex frame with rapid pixel change is related to the higher frequency. On the other hand, MEs with subtle changes contain useless low-frequency information. Thus, a high-band frequency (HBF) filter is antecedently used to filter the higher frequency and reduce the influence of unchanging pixels in the frames. The HBF filter $H_{b_{ij}}$ is defined as follows:

$$H_{b_{ij}}(u,v,q) = \begin{cases} 1 & \text{if } \sqrt{u^2+v^2+q^2} \geq D_0 \\ 0 & \text{if } \sqrt{u^2+v^2+q^2} < D_0 \end{cases}, \quad (2)$$

where D_0 is the threshold.

The proposed 3DF-N obtains the high-frequency components of the j-th block in the i-th interval according to Equation 3,

$$G_{b_{ij}}(u,v,q) = F_{b_{ij}}(u,v,q) \times H_{b_{ij}}(u,v,q). \quad (3)$$

Due to sparse facial changes caused by MEs, the occurrence of apex frame leads to higher frequency in some specific blocks. To reduce redundancy information, 3DF-N uses the specific blocks with the N largest frequency values, and then sums up the high-frequency value $G_{b_{ij}}$ in the i-th video interval by the following formulation,

$$A_i = \sum_{j=1}^{N} \sum_{u=1}^{T} \sum_{v=1}^{L_b} \sum_{q=1}^{W_b} \mid G_{b_{ij}}(u,v,q) \mid, \quad (4)$$

where A_i represents the frequency amplitude of the i-th interval. A_i indicates the range of rapid facial movements at the i-th interval. In the same way, 3DF-N can obtain frequency information of all the video intervals. The interval with maximum amplitude indicates the frames with the most obvious facial movement, which is defined as follows:

$$A_{pi} = \max(A_i), \quad (5)$$

where A_{pi} represents the interval with the most rapid facial movements. The middle of the interval can be viewed as the apex frame.

B. Micro-expression recognition based on joint local and global information learning

For the majority of MEs, not all facial regions contribute to ME recognition. In order to emphasize emotion learning from informative regions and reduce the influence of outliers, the proposed LGCcon discovers that local facial regions contribute ME information and learns the local and global facial information jointly to increase the discrimination and robustness of features against the problem of outliers. Besides, multi-constraints on local and global information learning are developed to raise the discrimination of local and global representations, respectively. Furthermore, Centerloss [23] is employed to enhance inter-class dispersion and intra-class compactness for ME recognition.

Figure 3 illustrates the framework of LGCcon. The backbone of LGCcon is based on VGG-16 CNN architecture [19]. LGCcon consists of the Global Information path (**GI**) and Local Information path (**LI**), which extract global and local features, respectively. Specifically, the GI aims to extract contextual features from the whole facial image. Meanwhile, the LI aims to extract features from the local region contributing the most ME information. Below, the details of the LI and GI in LGCcon are presented.

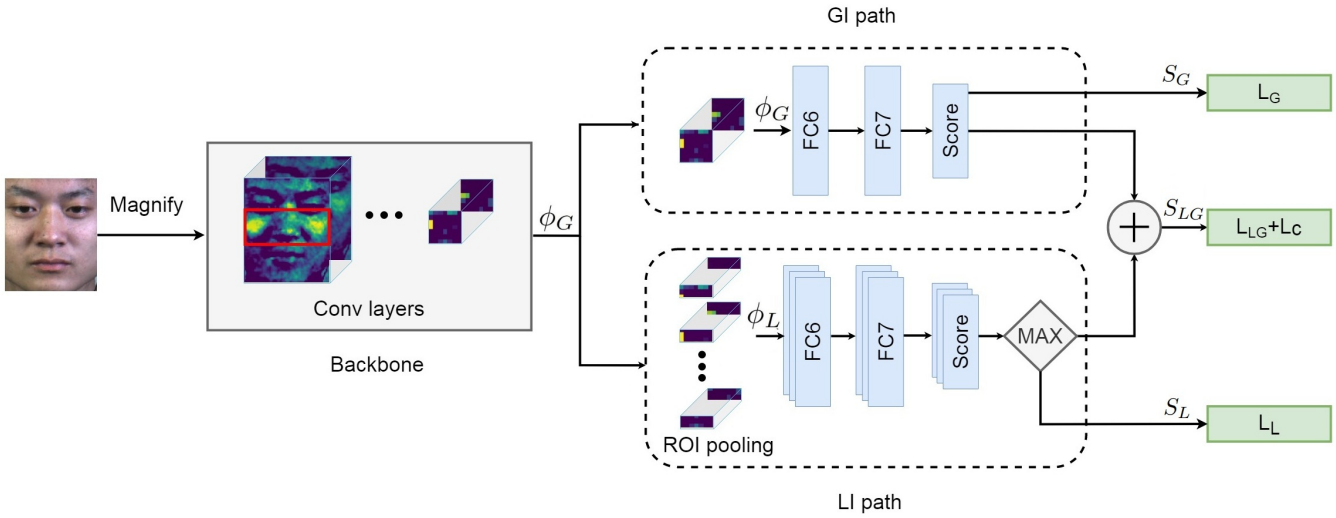


Fig. 3. The proposed ME recognition framework LGCcon. GI path and LI paths extract global features of the whole face and local features of sub-regions, respectively.

As Figure 3 shows, given a face image I, it passes through 16 convolutional layers and three fully connected (**FC**) layers. The feature of the last FC layer is represented as ϕ_G. For the GI, the score of ME based on the whole face is defined as S_G:

$$S_G(\theta; I) = w_G^\theta \cdot \phi_G(I), \tag{6}$$

where ϕ_G is the feature extracted from the whole face region I. The dimension of ϕ_G corresponds to the number of ME categories. w_G^θ is the global weight for ME category θ.

Given the score $S_G(\theta; I)$ for ME based on global information, the Softmax function is used to compute the probability p_G:

$$p_G(\theta; I) = \frac{exp(S_G(\theta; I))}{\sum_{\theta \in E}(exp(S_G(\theta; I)))}. \tag{7}$$

Thus, the loss function for the GI path is defined as follows:

$$L_G = -\frac{1}{M}\sum_{i=1}^{M}(\log(p_G(\theta = l_i|I_i))), \tag{8}$$

where M represents the batch size and l_i represents the true label of image I_i.

On the other hand, as seen in Figure 3, the LI is proposed to extract the information on the local regions containing ME emotion (*e.g.*, cheek raiser). For the sake of simplicity, τ is defined as a region in I, $R(\tau; I)$ is the set of candidates for the sub-regions in the whole set of regions in I. As the face structure is fixed and symmetrical, LGCcon obtains $R(\tau; I)$ by a sliding window with the height being $1/3$ of the face height and the face width. The step size of the sliding window is $1/6$ of the face height, and six ROIs are obtained in one face image. Then, the ROI pooling layer is used to extract local features ϕ_L for all the $R(\tau; I)$. Following the concept of MIL [38], the set of candidate sub-regions $R(\tau; I)$ can be regarded as a 'bag' of instances in ME recognition. For each ME image, at least one local region contributes emotional information to ME recognition. The most informative region can be seen as the positive instance for the corresponding ME category. The LI path recognizes MEs based on the positive instance contributing the most emotional information through a maximum operation. Therefore, the score and probability of MEs based on the local information are defined in Equations 9 and 10, respectively.

$$S_L(\theta; \tau, I) = \max_{\tau \in R(\tau; I)} w_L^\theta \cdot \phi_L(\tau; I), \tag{9}$$

where ϕ_L is the feature extracted from local face regions $R(\tau; I)$. The dimension of ϕ_L is the number of ME categories. w_L^θ is the local weight for ME category θ.

$$p_L(\theta; \tau, I) = \frac{exp(S_L(\theta; \tau, I))}{\sum_{\theta \in E}(exp(S_L(\theta; \tau, I)))}, \tag{10}$$

where S_L and p_L represent the score and probability of MEs based on the local information, respectively.

Based on $p_L(\theta; \tau, I)$, the loss function for the LI path is defined as L_L:

$$L_L = -\frac{1}{M}\sum_{i=1}^{M}(\log(p_L(\theta = l_i|\tau_i, I_i))). \tag{11}$$

where l_i is the true label of ROI τ_i in facial image I_i.

Finally, the scores based on the global information and local information learning are combined to jointly estimate the final ME probability.

$$S_{LG}(\theta; \tau, I) = S_G(\theta; I) + S_L(\theta; \tau, I), \tag{12}$$

$$p_{LG}(\theta; \tau, I) = \frac{exp(score_{LG}(\theta; \tau, I))}{\sum_{\theta \in E}(exp(score_{LG}(\theta; \tau, I)))}, \tag{13}$$

where S_{LG} and p_{LG} represent the joint score and probability of MEs, respectively. Specifically, the feature representations ϕ_L and ϕ_G, and the weight vectors w_G^θ and w_L^θ in Equations 6 and 9 are learned jointly for all ME categories. The loss

function of the joint local and global information learning is represented as L_{LG},

$$L_{LG} = -\frac{1}{M}\sum_{i=1}^{M}(\log(p_{LG}(\theta = l_i|\tau_i, I_i))). \quad (14)$$

However, based on the previously described framework, the features are not sufficiently discriminative. Due to small ME databases, the possible training identities are very limited and not diversified. For enhancing the discriminant of ME features, Centerloss [23] is employed to strengthen inter-class dispersion and intra-class compactness. Centerloss is defined as:

$$L_C = \frac{1}{2}\sum_{i=1}^{M}\left\|x_i - c_{\theta_i}\right\|_2^2, \quad (15)$$

where x_i represents the sample in the class, while the c_{θ_i} represents the center of samples belonging to ME class θ_i.

During the training process, the L_L and L_G are also used as the constraints to restrict the learning procedure based on local and global information, respectively. They aim to promote the discriminant of local and global representations, Therefore, the final loss function L is formulated as follows:

$$L = L_{LG} + \lambda_C \cdot L_C + \lambda_L \cdot L_L + \lambda_G \cdot L_G, \quad (16)$$

where λ_C, λ_L and λ_G balance the loss functions. λ_C is set as 0.008. λ_L and λ_G are set as 0.7 for faster training convergence. The influences of λ_C, λ_L and λ_G are analyzed in the experiments section.

C. Implementation of LGCcon

LGCcon is built based on VGG and R*CNN [41] and fine-tuned on the VGG-FACE model [19]. In the training stage, the losses L_{LG}, L_C, L_L, and L_G are trained jointly. The learning rate is set as 0.00001 and batch size 64. To avoid over-fitting, the dropout rate is set as 0.8.

As the MEs have low intensity and are difficult recognize, the apex frames are magnified to train the ME classifier. The Eulerian magnification method [42] is used to magnify the subtle motion of apex frames. Here, it enlarges the difference between different ME categories for enhancing the performance of recognition. The level of motion magnification is set as 30 in our framework according to [18]. In addition, due to the small sampling size of the ME database, the new data augmentation strategy is exploited to train a good model. Although the ME is rapid, the neighboring five frames to the apex frame are very similar to the apex frame, especially the magnified one. The apex frame and the two frames before and after the apex frame are chosen, such that the ME database is augmented five times. For the sake of simplicity, the extended database is named the Extended Magnified ME (EMME) database.

IV. Experiments

In this section, experiments regarding apex frame spotting and recognition are conducted on the CASME [43], CASME II [44], SAMM [45], SMIC [28], and composite [46] databases, and the results are quantitatively and qualitatively analyzed. Firstly, the databases are introduced in Subsection A. Subsection B demonstrates the evaluation metrics and performance comparisons on apex frame spotting. Finally, the experiments on micro-expression recognition with the apex frame are elaborated in Subsection C, including evaluation metrics and experimental protocols for ME recognition, comparisons with the single path to validate the effectiveness of joint local and global information learning, ablation study, parameter analysis, performance comparisons with different frames to evaluate the apex frame contribution, comparisons with the state-of-the-art methods and computational time discussion.

A. Databases

The CASME [43] database contains spontaneous ME clips including frames from onset to offset. It contains 195 spontaneous ME clips from 19 subjects, recorded by high-speed camera at 60 fps. Samples in CASME database are categorized into eight ME emotion categories: *happiness*, *disgust*, *sadness*, *surprise*, *fear*, *tenseness*, *repression*, and *contempt*.

CASME II [44] consists of 247 MEs elicited from 26 participants with a high-speed camera at 200 fps. There are five kinds of ME expressions: *happiness*, *surprise*, *disgust*, *repression*, and *others*.

SAMM [45] collects 159 ME samples from 32 participants of 13 ethnicities using a 200 fps high-speed camera. It includes the ME emotion classes *happy*, *sad*, *surprise*, *angry*, *disgust*, *fear*, *contempt*, and *other*.

SMIC [28] includes three subsets: SMIC-HS, SMIC-VIS and SMIC-NIR. SMIC-VIS and SMIC-NIR are recorded by normal speed cameras with 25 fps of visual (VIS) and near inferred (NIR) light range, respectively. SMIC-HS recorded by 100 fps high-speed cameras is used in our experiments, which contains 164 spontaneous MEs from 16 subjects. These samples are divided into three classes: *positive*, *negative*, and *surprise*.

The composite database [46] collects samples from three spontaneous facial micro-expression databases: CASME II [44], SAMM [45], and SMIC-HS (denoted as SMIC in the following discussions) [28]. Due to various annotations in three databases, the composite database unifies emotion labels in all three databases, in which emotion labels are reannotated as *positive*, *negative*, and *surprise*. Consequently, this consolidation includes 442 samples (145 from CASME II, 133 from SAMM,and 164 from SMIC) from 68 subjects (24 from CASME II, 28 from SAMM, and 16 from SMIC).

In the ME databases, some of the emotion categories have only a few samples, which are not enough for learning. Following the previous works [7], [25], [28], only the emotions with more than 10 samples are considered. The emotion categories in CASME are then classified as: Disgust (**D**), Surprise (**S**), Repression (**R**), and Tense (**T**). The samples in CASME II are classified as: Happiness (**H**), Disgust (**D**), Surprise (**S**), Repression (**R**), and Others (**O**). The samples in SAMM are classified as: Happiness (**H**), Anger (**A**), Surprise (**S**), Contempt (**C**), and Other (**O**). All ME samples in the SMIC database are used for experimentation, classified as: Negative (**N**), Positive (**P**), and Surprise (**S**).

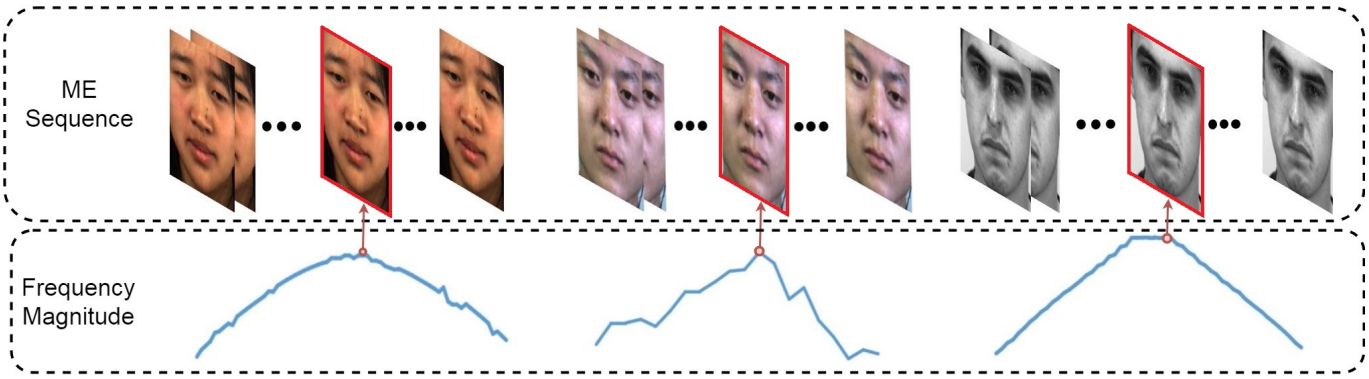


Fig. 4. Examples of frequency amplitude change for CASME, CASME II and SAMM databases. The frame in the red rectangle represents the apex frame in the ME sequence and the red circle represents the highest frequency.

In the experiments, the cropped face images provided by the previous works [7], [25], [28] are employed. For experiments, all the faces are resized into 224×224 pixels.

B. Experiments with apex frame spotting

Because the SMIC database does not provide the annotation for apex frames, the apex frame spotting experiments are conducted with the CASME, CASME II and SAMM databases. As this paper focuses on locating the apex frame in ME clips, all video clips from onset to offset in the CASME, CASME II, and SAMM databases are used. In the experiments, the ME interval T is set at 61, 61, and 19 for CASME II, SAMM, and CASME. The HBF filter threshold D_0 in Equation 2 is set as $\lfloor \frac{T}{2} \rfloor$.

Figure 4 illustrates several examples of the frequency amplitude change for CASME, CASME II, and SAMM. As seen in Figure 4, the apex frame nearly occurs at the position with the highest frequency magnitude. It quantitatively validates that the frequency domain can explicitly explore the micro-expression intensity changes.

1) Evaluation metrics for apex frame spotting: The Normalized Mean Absolute Error (NMAE) and Normalized Standard Error (NSE) are chosen to report the effectiveness of the apex frame spotting method.

NMAE is the average normalized frame distance between the spotted apex frame and the ground-truth:

$$NMAE = \frac{1}{K}\sum_{i=1}^{K} e_i{}', \tag{17}$$

$$e_i{}' = \frac{|e_i|}{len}, \tag{18}$$

where e_i is the frame distance between the spotted apex frame and the ground-truth apex frame of the i-th sample. len is the average length of the samples in the database and K is the number of samples in the databases.

NSE represents the standard deviation of the sample mean distribution:

$$NSE = \frac{\sqrt{(e_i{}' - \overline{e_i{}'})^2}}{\sqrt{K}}, \tag{19}$$

where $\overline{e_i{}'}$ is the average of $e_i{}'$.

TABLE I
THE NMAE (THE LESS THE BETTER) OF APEX FRAME SPOTTING, WHERE 2DF, 3DF-36 AND 3DF-N ARE THE PROPOSED METHODS FOR APEX FRAME SPOTTING. ENTRIES IN BOLD REPRESENT THE BEST PERFORMANCE.

Database	CASME	CASME II	SAMM
LBP [40]	0.3462	0.2037	0.4364
OS-ROI [17]	0.1824	0.1964	0.2550
RHOOF [26]	0.1644	0.1656	N/A
OS-N	0.2037	0.1678	0.2767
2DF	0.1399	0.1954	0.1567
3DF-36	0.1089	0.1687	0.1412
3DF-N	**0.1023**	**0.1471**	**0.1353**

*N/A - no results reported.

TABLE II
THE NSE (THE LESS THE BETTER) OF APEX FRAME SPOTTING, WHERE 2DF, 3DF-36 AND 3DF-N ARE THE PROPOSED METHODS FOR APEX FRAME SPOTTING. ENTRIES IN BOLD REPRESENT THE BEST PERFORMANCE.

Database	CASME	CASME II	SAMM
LBP [40]	0.0223	0.0158	0.0197
OS-ROI [17]	0.0100	0.0118	0.0156
RHOOF [26]	0.0110	0.0159	N/A
OS-N	0.0147	0.0094	0.0178
2DF	0.0137	0.0119	0.0108
3DF-36	0.0094	0.0116	0.0111
3DF-N	**0.0085**	**0.0080**	**0.0107**

*N/A - no results reported.

2) Performance evaluation for apex frame spotting: Tables I and II report the comparative results in terms of NMAE and NSE, respectively. The 2DF method computes the frequency in the X-T and Y-T dimensions, and then sums the frequency magnitudes in X-T and Y-T dimensions up to represent the final change rate. 3DF-36 and 3DF-N represent the proposed apex frame spotting method based on all 36 blocks and maximum N blocks, respectively. OS-N computes optical strain on maximum N blocks.

As shown in Table I, the proposed 3DF-N consistently outperforms LBP [40] by 0.2439, 0.0566, and 0.3011 in

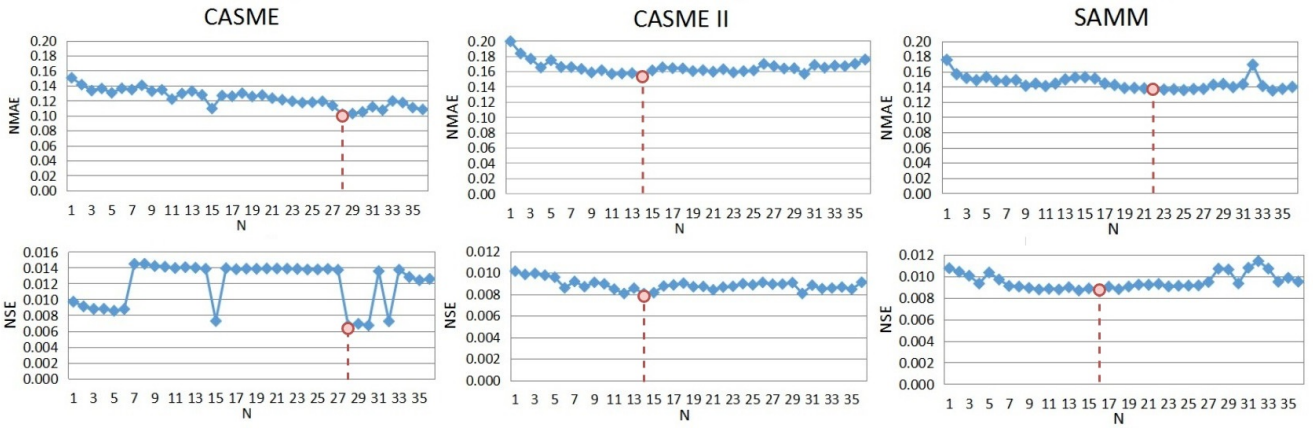


Fig. 5. Apex frame spotting results on the first N blocks with largest frequency amplitude for the CASME, CASME II, and SAMM databases, respectively. The x-axis shows N largest frequency values.

terms of NMAE on CASME, CASME II and SAMM, respectively. 3DF-N improves the OS-N consistently with gains of 0.1014, 0.0207, and 0.1414 in terms of NMAE on CASME, CASME II, and SAMM, respectively. The increasing results demonstrate that our proposed 3DF-N outperforms optical flow on ME apex frame spotting by a large margin. The results indirectly indicate that it is reasonable to spot apex frames in the frequency domain. Furthermore, it can be found that 3DF-36 and 3DF-N work better than 2DF. The results show that both spatial and temporal changes make contributions for apex frame spotting. Moreover, 3DF-N outperforms 3DF-36 by 0.0066, 0.0216, and 0.0059 in terms of NMAE on CASME, CASME II and SAMM, respectively. It shows that reducing the redundant information can improve the apex frame performance. In comparison with the state-of-the-art methods [17], [26], 3DF-N achieves the best performance on all three databases. Compared with RHOOF [26] based on optical flow histogram on ROIs, 3DF-N improves the spotting performances on the CASME and CASME II databases by 37.78% and 11.17% in terms of NMAE, respectively. In Table II, compared with LBP, 3DF-N reduces the NSE by 0.0138, 0.0078, and 0.0090 in the CASME, CASME II, and SAMM databases, respectively. 3DF-N achieves the best robustness compared with LBP, OS-ROI and RHOOF.

In order to see the influence of N, 3DF-N methods are evaluated on different N blocks, in which N blocks correspond to the first N largest frequency amplitudes. The results with various N are illustrated in Figure 5. 3DF-N consistently improves the 3DF-36 when the blocks with lower frequency amplitude are abandoned. It is concluded that the high-frequency signal contributes more valuable information to apex frame spotting. As seen from Figure 5, when N is 28, 14, and 23 for CASME, CASME II, and SAMM, respectively, 3DF-N achieves the best performance by considering NMSE and NSE jointly. Although, the NSE on SAMM is not the lowest when $N = 23$, it slightly decreases the performance by 0.0003 compared with when $N = 14$. The difference of N for the three databases is likely caused by the different properties of the databases including the recording rates and image resolution.

TABLE III
METHODS AND THE CORRESPONDING STRUCTURE (L_G REPRESENTS THE LOSS OF GLOBAL INFORMATION. L_L REPRESENTS THE LOSS OF LOCAL INFORMATION. L_{LG} REPRESENTS THE LOSS OF THE SUM OF GLOBAL AND LOCAL INFORMATION. L_C REPRESENTS THE CENTERLOSS.)

Method	L_G	L_L	L_{LG}	L_C
GI	√	-	-	-
LI	-	√	-	-
LG	-	-	√	-
LGC	-	-	√	√
LGcon	√	√	√	-
LGCcon	√	√	√	√

TABLE IV
MICRO-EXPRESSION RECOGNITION COMPARISONS OF ACCURACY(%) AND F1-SCORE BASED ON THE APEX FRAMES, WHERE THE TEXT IN BOLD IS THE BEST RESULT.

	CASME		CASME II		SAMM		SMIC	
Methods	ACC	F1	ACC	F1	ACC	F1	ACC	F1
LBP	53.21	0.50	40.33	0.11	**41.91**	0.12	46.44	0.46
GI	60.23	0.58	63.21	0.59	36.00	0.25	59.75	0.58
LI	51.46	0.51	60.08	0.54	27.21	0.20	49.39	0.48
LGCcon	**60.82**	**0.60**	**65.02**	**0.64**	40.90	**0.34**	N/A	N/A
LGCconD	57.31	0.54	62.14	0.60	35.29	0.23	**63.41**	**0.62**

*N/A - no results reported.

C. Experiments on micro-expression recognition

1) Evaluation metrics and protocols for ME recognition: This section reports the results of ME recognition on the CASME, CASME II, SAMM, and SMIC databases. As the ground-truth apex frame label in the SMIC database is not available, all the models in the SMIC database are trained with the detected apex frames. In the experiments, the leave-one-subject-out cross validation protocol is used. The recognition accuracy and F1-score are used as performance metrics.

Table III lists the proposed framework and its corresponding combination of loss functions. The basic framework is LG with the L_{LG} loss. LGcon represents the basic framework with L_G and L_L. LGC contains the Centerloss L_C. The final training framework combining L_{LG}, L_G, L_L and L_C is represented as

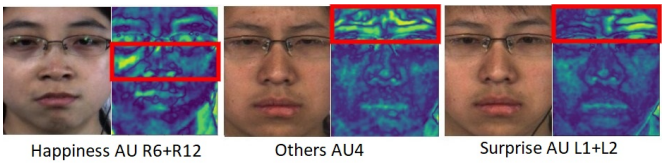


Fig. 6. Visualization of the examples for subjects with eyeglasses. The feature map is shown on the right. The red rectangle implies the most informative region.

TABLE V
MICRO-EXPRESSION RECOGNITION COMPARISONS OF ACCURACY(%) ON THE SUBJECTS WITH AND WITHOUT EYEGLASSES, WHERE THE BEST RESULTS ARE SHOWN IN BOLD. NG REPRESENTS THE SUBJECTS WITHOUT EYEGLASSES AND WG REPRESENTS THE SUBJECTS WITH EYEGLASSES.

	CASME		CASME II		SAMM		SMIC*	
Methods	WG	NG	WG	NG	WG	NG	WG	NG
VGGMag [18]	62.96	**56.45**	61.50	64.41	31.30	36.99	48.61	**67.39**
LGCcon	**63.88**	**56.45**	**64.40**	**65.29**	**32.17**	**43.19**	**59.72**	66.30

*The results on SMIC are based on the detected apex frame

LGCcon.

2) Performance comparisons with the single-path network and LBP: This section compares the two-path LGCcon with the single-path network based on sole local information (LI), sole global information of the whole face (GI), and LBP. The LBP features are extracted on 6×6 blocks and the radii is set at $(3, 3, 3)$. Linear SVM is used. Table IV reports the comparison results in terms of accuracy and F1-score. It is seen that LGCcon improves the single-path architecture significantly. Comparing with sole learning based on local and global information independently, the two-path LGCcon framework achieves improvements of 0.09 and 0.02 on the CASME database, 0.10 and 0.05 on the CASME II database, 0.14 and 0.09 on the SAMM database, and 0.14 and 0.04 on the SMIC database in terms of F1-score, respectively. The increasing results demonstrate that both the local and global information make contributions to ME recognition. Joint learning the local and global information can also improve the ME recognition performance. With the SAMM database, the LBP method is inferior to the LGCcon though its accuracy is relatively high. This is explained by the class-imbalance of the SAMM database, which makes the SVM classifier falsely classify most of the samples to the class that has the largest number of samples.

Furthermore, Table IV reports the results of LGCcon with the detected apex frame based on 3DF-N, namely LGCconD. Compared with LGCcon, LGCconD slightly degrades the performance by 0.06, 0.04 and 0.11 in terms of F1-score on the CASME, CASME II, and SAMM databases, respectively, which suggests that our proposed apex frame spotting method is reliable. Furthermore, from Figure 7, it can be seen that LGCconD can achieve good performances on most ME categories. For 'Repression' in the CAME database and 'Happiness' in the SAMM database, the accuracies are lower, it may be caused by the data imbalance.

In order to evaluate the effectiveness of LGCcon with outliers, LGCcon is further studied on participants with eyeglasses. Table V reports the comparisons between LGCcon and the previous work [18] on subjects with and without eyeglasses. In [18], the local information on ME recognition is not considered. In Table V, it is seen that LGCcon outperforms VGGMag [18] with 2.9% when they recognize the ME of the subjects with eyeglasses on the CASME II database. It narrows the performance gap between subjects with and without eyeglasses by 2.02%. Moreover, LGCcon outperforms VGGMag by 0.92% and 0.87% on subjects with eyeglasses on the CASME and SAMM databases, respectively. When comparing with VGGMag, LGCcon slightly decreases the accuracy on subjects without eyeglasses by 1.09% on the SMIC database, but despite this, LGCcon consistently outperforms the VGGMag by a large margin (11.11%) on subjects with eyeglasses. Figure 6 illustrates some visualization examples for subjects with eyeglasses. It can be seen that LGCcon can extract the emotion information on informative facial regions, even though the face is occluded by eyeglasses to some degree. The results indicate that joint learning local and global information not only improves the discrimination of the ME feature, but also reduces the influence of outliers to some extent.

TABLE VI
ABLATION STUDY ON MICRO-EXPRESSION RECOGNITION ACCURACY(%) AND F1-SCORE OF THE PROPOSED METHODS, WHERE THE TEXT IN BOLD REPRESENTS THE BEST RESULT.

	CASME		CASME II		SAMM		SMIC	
Methods	ACC	F1	ACC	F1	ACC	F1	ACC	F1
LG	44.44	0.50	61.73	0.62	34.71	0.22	53.05	0.52
LGcon	53.21	0.53	63.79	0.63	33.38	0.26	57.93	0.57
LGC	48.54	0.49	61.08	0.61	35.52	0.29	54.88	0.54
LGCcon	**60.82**	**0.60**	**65.02**	**0.64**	**40.90**	**0.34**	N/A	N/A
LGCconD	57.31	0.54	62.14	0.60	35.29	0.23	**63.41**	**0.62**

*N/A - no results reported.

3) Ablation study: To reveal the contribution of each module, the accuracy and F1-score of LGCcon with different configurations are evaluated. Table VI reports the comparison results. The proposed backbone LG obtains 0.50, 0.62, 0.22, and 0.52 in terms of F1-score on the CASME, CASME II, SAMM, and SMIC databases, respectively. Figure 7 illustrates confusion matrices of LG, the variations of LG, and LBP.

(1) LGcon includes constraints L_L and L_G on local and global information learning. The experiments show that LGcon increases the F1-score by 0.03 on average on all four databases in comparison to LG. It indicates that constraints on local and global information learning can control the learning and lead to a better fusion result.

(2) The LGC consists of the basic LG with the Centerloss L_C. As seen in Table VI, LGC performs better than basic LG on the CASME, SAMM, and SMIC databases. LGC also achieves comparable results on the CASME II database. From the confusion matrices in Figure 7, it can be seen that LGC outperforms LG by 0.01, 0.05, and 0.22 for the *disgust*, *surprise*, and *repression* categories, respectively. Based on the results, it is inferred that the Centerloss improves the discriminative ability of the ME feature.

(3) Our framework LGCcon is designed by combining Centerloss and constraints with local and global information learn-

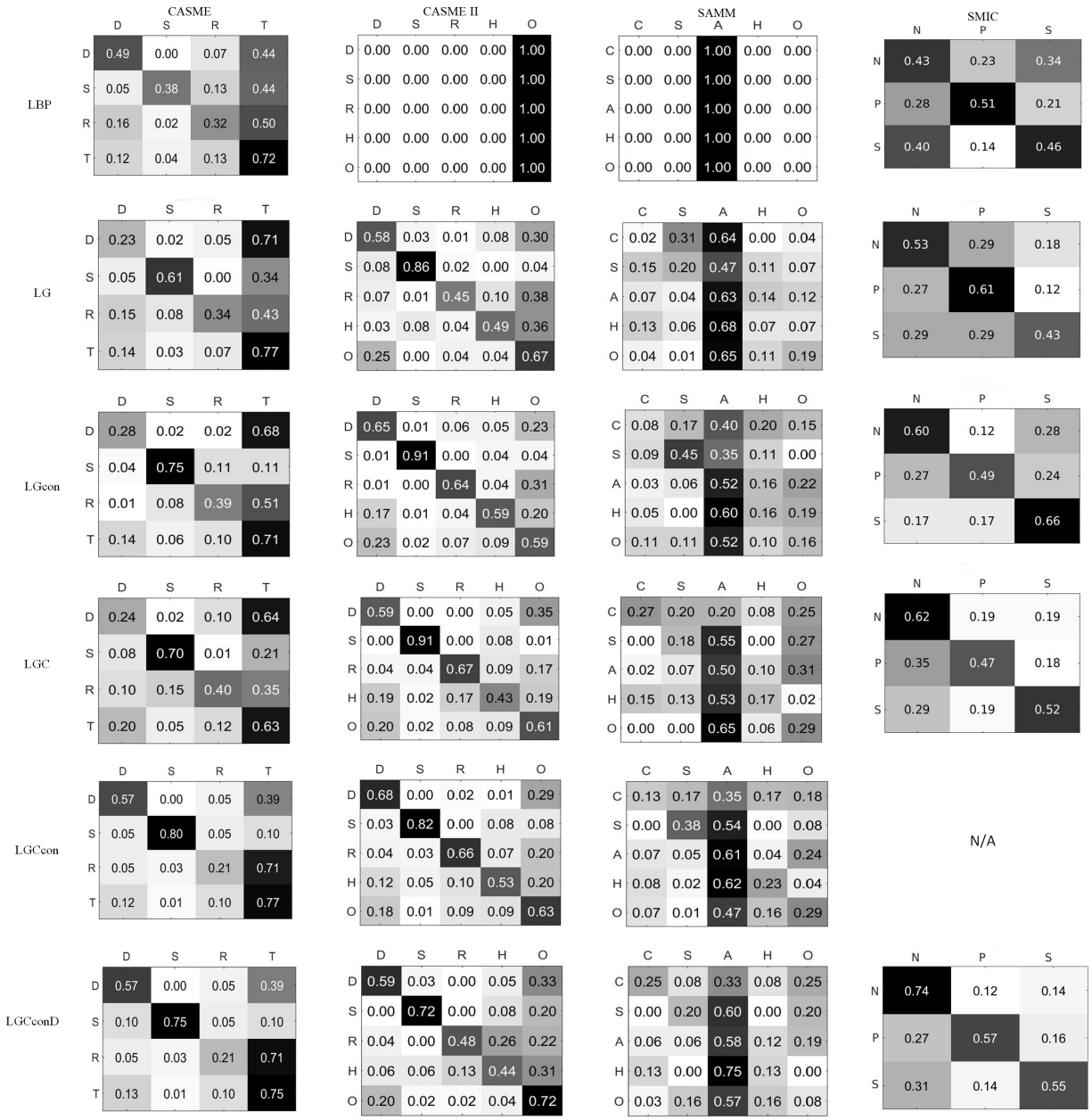


Fig. 7. The confusion matrices of the proposed architectures and the comparative method (LBP) on the CASME, CASME II, SAMM, and SMIC databases.

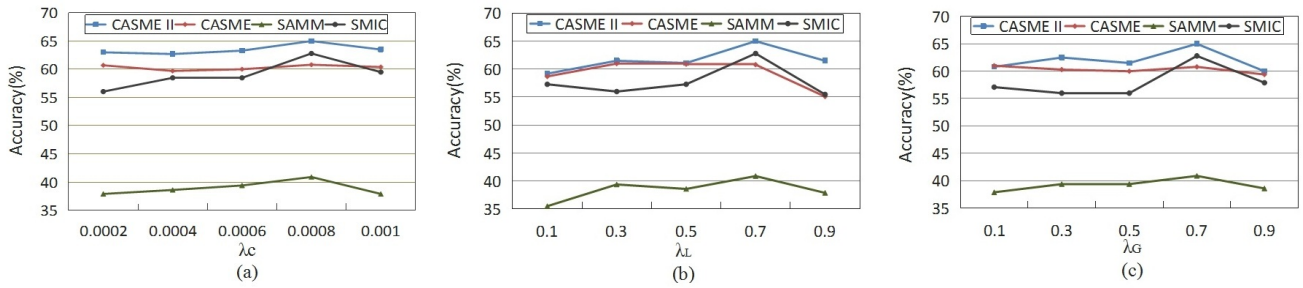


Fig. 8. Micro-expression recognition accuracies on the CASME, CASME II, SAMM, and SMIC databases, respectively, achieved by (a) models with different λ_C and fixed λ_L = λ_G=0.7. (b) models with different λ_L, fixed λ_C=0.0008 and λ_G=0.7. (c) models with different λ_G, fixed λ_C=0.0008 and λ_L=0.7.

TABLE VII
MICRO-EXPRESSION RECOGNITION COMPARISONS OF ACCURACY(%) ON APEX FRAMES AND RANDOM FRAMES. APEX AND RANDOM REPRESENT THE RESULTS BASED ON APEX FRAMES AND RANDOM FRAMES.

	CASME		CASME II		SAMM	
Methods	Apex	Random	Apex	Random	Apex	Random
VGGMag [18]	60.23	53.40	63.21	55.43	36.00	33.82
LGCcon	**60.82**	54.60	**65.02**	57.56	**40.90**	35.75

ing. LGCcon achieves an accuracy of 60.82%, 65.02%, and 40.90% on the CASME, CASME II, and SAMM databases, respectively. Compared with LG, the addition of L_G, L_L, and L_C losses improves the recognition accuracy by 16.38%, 3.29%, and 6.19% on the CASME, CASME II, and SAMM databases, respectively. It validates the effectiveness of multi-constraints and Centerloss.

4) Parameter analysis: For LGCcon, the hyper parameter λ_C limits the influence of intra-class variations, and λ_L and λ_G constrain the learning of local and global information individually. Here, three experiments are conducted to investigate the influence of the three parameters. In order to validate the effectiveness of λ_C, λ_L and λ_G are both fixed at 0.7. λ_C is varied from 0.0001 to 0.001. Figure 8(a) shows the accuracy corresponding to various λ_C on the CASME, CASME II, SAMM, and SMIC databases. LGCcon achieves the best performance on the CASME II, SAMM, and SMIC databases when $\lambda_C = 0.0008$, while on the CASME database, $\lambda_C = 0.0002$. This difference is due to the intra-class variations of each database.

To validate the influence of λ_L, λ_C and λ_G are fixed at 0.0008 and 0.7, respectively. λ_L is varied from 0.1 to 0.9. Figure 8(b) illustrates the accuracy corresponding to various λ_L on these databases. As seen from Figure 8(b), the increasing λ_L boosts the performance of model. It is seen that when $\lambda_L = 0.7$, the LGCcon achieves the best results.

Furthermore, to validate the influence of λ_G, the same analysis to λ_L is conducted. Here, λ_C and λ_L are fixed at 0.0008 and 0.7, respectively. Figure 8(c) illustrates the accuracy corresponding to various λ_G on these databases. As seen from Figure 8(c), the conclusion is the same as the analysis of λ_L. It is explained by the fact that the final ME category is obtained based on the joint probability of MEs p_{LG}. Large λ_L and λ_G on independent local and global information learning will distract the learning of the joint local and global information. Small λ_L and λ_G are not enough to constrain the learning of local and global information, respectively.

According to Figure 8, the performance on the SMIC database remains relatively unstable across the range of λ_C, λ_L, and λ_G. This is perhaps because of the diversity of samples on the SMIC database, which leads LGCcon to be more sensitive to these hyper parameters on the SMIC database.

5) Performance comparisons with different frames: In order to validate the importance of the apex frame, a comparison of ME recognition performances based on the apex frame with the other frames is conducted. One frame from the ME clip between the onset and apex frames is randomly selected. Table VII reports the recognition accuracy based on the random frame selection approaches and apex frame approaches. Several interesting finds are observed: (1) LGCcon still obtains the competitive results by using other frames. (2) More importantly, the utilization of the apex frame yields better recognition results for VGGMag and LGCcon methods than using random frames. Furthermore, Figure 9 shows ME recognition accuracy with a function of training frames on the CASME database. When the selected frame is closer to the apex frame, LBP and LGCcon gain improvements by around 10% in terms of accuracy. These results indicate that the apex frame in ME clips contributes more important information to ME recognition, compared with the other frames.

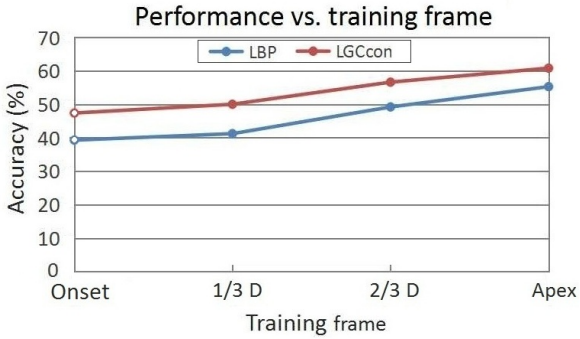


Fig. 9. Performance vs. training frame on the CASME database. D represents the distance between the onset frame and apex frame.

6) Comparison with the state-of-the-art algorithms: This section compares the proposed LGCcon, LGCconD with state-of-the-art methods [3], [7], [15], [35]. Table VIII summarizes the compared results. It is seen that LGCcon surpasses the existing deep learning methods based on whole ME sequence [4], [15]. Furthermore, our proposed LGCcon with LOSO protocol shows superior performance on CASME II, compared with 3D-FCNN [33] and TIM-DCNN [9] which employ the leave one-video-out protocol (LOVO), while LGCcon achieves competitive results on SMIC. Although these results cannot be compared directly, they still indicate the effectiveness of LGCcon. Compared with LOSO, the LOVO protocol is easier to obtain better performance, as the LOVO can include more training data and the test data can come from the same subject with training data. The results demonstrate that the proposed deep method can achieve good performance with the apex frame. Besides, LGCcon and LGCconD achieve promising results compared with the hand-crafted methods using ME sequences [7], [28], [47] on the CASME, CASME II, SAMM, and SMIC databases.

Furthermore, STRCN [34] has two types of input. The first type (denoted as STRCN-A) is a masked ME sequence and the second one (denoted as STRCN-G) is optical flow between onset and apex frames. As STRCN [34] used different training protocol to the LGCcon, STRCN-G and STRCN-A are re-implemented in PyTorch with the same training protocol and data augmentation (magnification ratio set at 8 by following [34]) to LGCcon. The results of STRCN-A and STRCN-G are shown in Table VIII.

TABLE VIII
MICRO-EXPRESSION RECOGNITION ACCURACY AND F1-SCORE OF THE PROPOSED METHODS AND THE STATE-OF-THE-ART METHODS. THE BEST RESULTS WITH LOSO PROTOCOL ARE SHOWN IN BOLD AND BRACKETS. THE SECOND BEST RESULTS WITH LOSO PROTOCOL ARE SHOWN IN BOLD, AND THE THIRD BEST RESULTS ARE SHOWN IN BRACKETS.

	CASME		CASME II		SAMM		SMIC	
Methods	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Baseline [28]	40.35	0.26	40.65	0.33	34.56	0.25	45.70	0.46
LBP-SIP [29]	36.84	0.33	46.56	0.45	36.76	0.21	42.12	0.42
FHOFO [47]	**65.99**	0.54	55.86	0.52	N/A	N/A	51.83	0.52
STCLQP [7]	57.31	0.50	58.39	0.58	N/A	N/A	64.02	**0.63**
Bi-WOOF [3]	N/A	N/A	59.67	N/A	N/A	N/A	62.80	N/A
HIGOMag [25]	N/A	N/A	**67.21**	N/A	[41.91]	N/A	**68.29**	N/A
3D-FCNN† [33]	54.44	N/A	59.11	N/A	N/A	N/A	55.49	N/A
TIM-DCNN† [9]	N/A	N/A	64.90	N/A	N/A	N/A	65.85	N/A
VGGMag [18]	60.23	[0.58]	63.21	0.59	36.00	0.25	59.75	0.58
CNNLSTM [4]	N/A	N/A	60.96	N/A	N/A	N/A	N/A	N/A
Selective [15]	N/A	N/A	47.30	N/A	N/A	N/A	53.60	N/A
STRCN-A‡ [34]	40.93	0.35	45.26	0.38	32.85	0.24	49.39	0.47
STRCN-G‡ [34]	59.65	0.57	63.37	[0.62]	**53.48**	**0.36**	[64.63]	**0.63**
TSCNN [35]	[**73.88**]	[**0.72**]	[**80.97**]	[**0.81**]	[**71.76**]	[**0.69**]	[**72.74**]	[**0.72**]
LGCcon	[60.82]	**0.60**	[65.02]	**0.64**	40.90	[0.34]	N/A	N/A
LGCconD	57.31	0.54	62.14	0.60	35.29	0.23	63.41	[0.62]

*N/A - no results reported.
† employing LOVO which leaves one video out.
‡ re-implemented in PyTorch with the same training protocol and data augmentation as LGCcon.

Compared with STRCN-G based on the optical flow between onset and apex frames, LGCcon based on apex frame achieves comparable performance, *i.e.*, 0.60 vs. 0.57 on CASME, 0.64 vs. 0.62 on CASME II, and 0.34 vs. 0.36 on SAMM in terms of F1-score. On the SMIC database, LGCconD slightly decreases the F1-score by 0.01 with the detected apex frame, compared to STRCN-G.

LGCcon outperforms STRCN-A by 19.89%, 19.76%, and 2.44% in terms of accuracy on the CASME, CASME II, and SAMM databases, respectively. In addition, LGCconD improves the accuracy on the SMIC database by 14.02%, in comparison with STRCN-A. The results demonstrate the effectiveness of LGCcon structure and the important apex frame contribution to ME recognition.

TSCNN [35] utilized the dynamic temporal information of optical flow between onset, apex, and offset frames, and the static spatial information of apex frames. Although the accuracy of LGCcon is a bit lower than TSCNN, LGCcon, based on only static apex frame information, can deal with the situation when the onset frame, offset frame, and temporal information are missing. Even though many methods [34], [35] indicated that optical flow-based methods always outperform the appearance-based methods, our proposed LGCcon, which is considered an appearance-based method, achieves competitive performance compared with optical flow-based methods. The results further verify the effectiveness of LGCcon.

7) Performance on the composite database: Table IX reports the results on the composite database. Following the evaluation metrics provided by MEGC2019 [46], UF1 and UAR are used to measure the performance of various methods. As seen from Table IX, the proposed method only based on the apex frame outperforms the LBP-TOP and ResNet [48] employing the temporal information. The methods [49]–[51] utilizing the optical flow between onset and apex frames outperform LGCcon. It is explained by the fact that (1) optical flow eliminates subject diversity across databases to some extent, and (2) EMR [51] also employed domain adaptation using expression-reduced CK+ samples [52]. However, because it is different from these methods, LGCcon can handle situations without temporal information and auxiliary databases. On the other hand, experimental results indicate that the apex frame significantly contributes to ME recognition.

TABLE IX
MICRO-EXPRESSION RECOGNITION RESULTS OF THE PROPOSED METHODS AND THE STATE-OF-THE-ART METHODS ON THE COMPOSITE DATABASE. THE BEST RESULTS ARE SHOWN IN BOLD.

	CASME II		SAMM		SMIC	
Methods	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [28]	0.7026	0.7429	0.3954	0.4102	0.2000	0.5280
Bi-WOOF [3]	0.7805	0.8026	0.5211	0.5139	0.5727	0.5829
ResNet [48]	0.6248	0.6136	0.4726	0.4359	0.4609	0.4327
OFF-Apex [49]	**0.8764**	**0.8681**	0.5409	0.5392	0.6817	0.6695
DualInp [50]	0.8621	0.8560	0.5868	0.5663	0.6645	0.6726
EMR [51]	0.8293	0.8209	**0.7754**	**0.7152**	**0.7461**	**0.7530**
LGCcon	0.7929	0.7639	0.5248	0.4955	N/A	N/A
LGCconD	0.7762	0.7499	0.4924	0.4711	0.6195	0.6066

*N/A - no results reported.

8) Computational time: Our proposed apex frame detection is implemented with Matlab on CPU (Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz). The apex frame detection time varies from 0.2s to 3s as the length of the ME sequence varies from 9 to 100 frames. Thus, the average apex frame detection duration is about 1.26s. Our proposed 3DF-N method employs the sliding window. In other words, the detection time is influenced by the ME sequence length. We conducted the experiments about the LGCcon network using a NVIDIA Tesla K80c GPU with 12 GB memory. The average testing time for ME image (224×224) is 0.18s. The LGCcon is based on the framework of fast RCNN. And the ROI proposal is designed according to facial structure without extra computation. With the powerful GPU and further development, the ME system could be realized in real-time in the future.

V. Conclusion

In ME research, the apex frames are very important, as they convey the representative information in micro-expression videos. This paper studies the contribution of apex frames to ME recognition. A complete pipeline is proposed to firstly locate the apex frame through analyzing MEs in the frequency domain and furthermore recognize MEs by a joint local and global information learning architecture. In contrast to existing spotting methods in the spatio-temporal domain, the proposed 3DF-N spots the apex frame in the frequency domain, which is more powerful for describing rapid changes. Different from the publicly available deep learning methods considering all the regions on face equally, LGCcon learns discriminative representation from the region containing the most emotional information automatically. It is found that LGCcon can focus on emotion learning and reduce the influences of eyeglasses and other negative effects.

The proposed approach is evaluated on the CASME, CASME II, SAMM, SMIC and composite databases. The experiments demonstrate that frequency analysis is suitable for describing ME change. LGCcon employing apex frames achieves comparable results when compared with the state-of-the-art methods employing the information from the whole ME sequence. The conclusion that both local and global information contributes to ME recognition can be drawn. Joint learning local and global information can reduce the side effects of outliers to some degree and improve the performance of ME recognition. The proposed LGCcon approach achieves promising ME recognition performance with the apex frame. Furthermore, the performance comparison of different frames in ME clips demonstrate that the apex frame contributes more important emotion information to ME recognition compared with other frames. In future work, we will explore the local and global information contributions to ME recognition based on video clips.

VI. ACKNOWLEDGMENT

This work was supported by Infotech Oulu, National Natural Science Foundation of China (Grant Nos: 61772419, 62076122), the Academy of Finland for project MiGA (grant 316765), and ICT 2023 project (grant 328115), the Jiangsu Specially-Appointed Professor Program (Grant No. 3051107219003), and the Talent Startup project of NJIT (No. YKJ201982). As well, the authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

[1] P. Ekman and W.V. Friesen, "Constants across cultures in the face and emotion," *Personal. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[2] P. Ekman, "Lie catching and microexpressions," *Phil. Decept.*, pp. 118–133, 2009.

[3] ST. Liong, J. See, K. S. 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, 2018.

[4] D.H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 24th ACM Int. Conf. on Multimedia*, 2016, pp. 382–386.

[5] F. Xu, J. Zhang, and J. Wang, "Micro-expression identification and categorization using a facial dynamics map," *IEEE Transact. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, 2017.

[6] S.-J. Wang et al., "Micro-expression recognition using color spaces," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, 2015.

[7] X. Huang, G. Zhao, X. Hong, and W. Zheng, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2016.

[8] 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, 2007.

[9] V. Mayya, R.M. Pai, and M.M. Pai, "Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences," in *2016 Int. Conf. on Advances in Comput., Commun. and Informat. (ICACCI)*. IEEE, 2016, pp. 699–703.

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

[11] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[12] ST. Liong, J. See, K. Wong, and R.CW. Phan, "Automatic micro-expression recognition from long video using a single spotted apex," in *Proc. IEEE Asian Conf. Comput. Vis.* Springer, 2016, pp. 345–360.

[13] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3359–3368.

[14] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, 2019.

[15] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2017, pp. 2258–2263.

[16] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in Psychology*, vol. 8, pp. 1–12, 2017.

[17] ST. Liong, J. See, K. Wong, and N. Le, "Automatic apex frame spotting in micro-expression database," in *Proc. IEEE Asian Conf. Pattern Recognit.* IEEE, 2015, pp. 665–669.

[18] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?," in *IEEE Int. Conf. Image Process.* IEEE, 2018, pp. 3094–3098.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[20] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognit. Letters*, vol. 33, no. 16, pp. 2181–2191, 2012.

[21] Z. Wang and G. Peng, "Weakly supervised dual learning for facial action unit recognition," *IEEE Trans. Multimedia*, vol. PP, pp. 1–1, 2019.

[22] S. Huang, W. Gao, and Z. Zhou, "Fast multi-instance multi-label learning," *IEEE Transact. Pattern Anal. Mach. Intell.*, vol. PP, pp. 1–1, 2018.

[23] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 499–515.

[24] D. Patel, G. Zhao, and M. Pietikäinen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *Int. Conf. Adv. Concepts Intell. Vis. Syst.* Springer, 2015, pp. 369–380.

[25] 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, 2018.

[26] H. Ma, G. An, S. Wu, and F. Yang, "A region histogram of oriented optical flow (RHOOF) feature for apex frame spotting in micro-expression," in *2017 Int. Symposium on Intell. Signal Process. and Commun. Systems (ISPACS)*. IEEE, 2017, pp. 281–286.

[27] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Int. Conf. Comput. Vis.* 2011, pp. 1449–1456, Springer.

[28] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. IEEE Conf. Workshops on Autom. Face Gesture Recognit.*, 2013, pp. 1–6.

[29] Y. Wang, J. See, R. CW. Phan, and Y. Oh, "LBP with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Proc. Asian Conf. on Comput. Vis.* Springer, 2014, pp. 525–537.

[30] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, 2016.

[31] S. Wang, W. Yan, X. Li, and G. Zhao, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2014, pp. 4678–4683.

[32] D. Alexey, F. Philipp, I. H. Philip, H. Caner, G. Vladimir, V. Patrick, C. Daniel, and B. Thomas, "Flownet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis*, 2015, pp. 2758–2766.

[33] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3d flow convolutional neural network," *Pattern Anal. and Applications*, pp. 1–9, 2018.

[34] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, vol. PP, pp. 1–1, 2019.

[35] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184537–184551, 2019.

[36] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Adv. Neural Inf. Process. Syst.*, 1998, pp. 570–576.

[37] O. Maxime, B. Léon, L. Ivan, S. Josef, et al., "Weakly supervised object recognition with convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014. hal-01015140.

[38] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1080–1088.

[39] R. Dikpal, R. Ramamoorthi, and B. Curless, "Frequency-space decomposition and acquisition of light transport under spatially varying illumination," in *European Conf. on Comput. Vis.* 2012, pp. 596–610, Springer.

[40] W. J. Yan, S. J. Wang, Y. H. Chen, G. Zhao, and X. Fu, "Quantifying micro-expressions with constraint local model and local binary pattern," in *Proc. IEEE Eur. Conf. Comput. Vis.* Springer, 2014, pp. 296–305.

[41] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Adv. Neural Inf. Process. Syst.*, 2006, pp. 1417–1424.

[42] H. Wu, E. Shih, E. Shih, J. Guttag, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graphics*, vol. 31, no. 4, pp. 1–8, 2012.

[43] W. Yan, Q. Wu, Y. Liu, S. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. IEEE Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–7.

[44] 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, pp. e86041, 2014.

[45] 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, 2018.

[46] J. See, Y.M. Hoon, J. Li, X. Hong, and S. Wang, "MEGC 2019–the second facial micro-expressions grand challenge," in *2019 14th IEEE Int. Conf. on Autom. Face & Gesture Recognit. (FG 2019)*. IEEE, 2019, pp. 1–5.

[47] S. L. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 394–406, 2017.

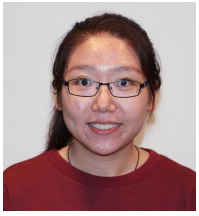
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2016, pp. 770–778.

[49] 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, 2019.

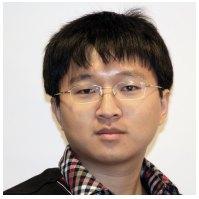
[50] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *2019 14th IEEE Int. Conf. on Autom. Face & Gesture Recognit. (FG 2019)*. IEEE, 2019, pp. 1–5.

[51] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *2019 14th IEEE Int. Conf. on Autom. Face & Gesture Recognit. (FG 2019)*. IEEE, 2019, pp. 1–4.

[52] 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 *2010 IEEE comput. society conf. on comput. vis. and pattern recognit. workshops*. IEEE, 2010, pp. 94–101.



Yante Li received the B.S. degree in communication engineering from China University of Petroleum (East China), Shandong, China in 2014. She received her master degree in Computer Science and Engineering from China University of Petroleum (East China), Shandong, China in 2017. She is currently a Ph.D student in University of Oulu, Oulu, Finland. Her current research interests include micro-expression analysis and facial action unit detection.



Xiaohua Huang received the B.S. degree in communication engineering from Huaqiao University, Quanzhou, China in 2006. He received his Ph.D degree in Computer Science and Engineering from University of Oulu, Oulu, Finland in 2014. He was a research assistant in Southeast University since 2006. He has been a scientist researcher in the Center for Machine Vision and Signal Analysis at University of Oulu since 2015. He is currently a Professor in School of Computer Engineering, Nanjing Institute of Technology. He is also a distinguished Professor of Jiangsu province. He has authored or co-authored more than 40 papers in journals and conferences, and has served as a reviewer for journals and conference. His current research interests include facial expression recognition, micro-expression analysis, group-level emotion recognition, multi-modal emotion recognition and texture classification.

Guoying Zhao (IEEE Senior member 2012) is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where she has been a senior researcher since 2005 and an Associate Professor since 2014. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She has authored or co-authored more than 230 papers in journals and conferences. Her papers have currently over 13530 citations in Google Scholar (h-index 53). She is co-program chair for ACM International Conference on Multimodal Interaction (ICMI 2021), was co-publicity chair for FG2018, General chair of 3rd International Conference on Biometric Engineering and Applications (ICBEA 2019), and Late Breaking Results Co-Chairs of 21st ACM International Conference on Multimodal Interaction (ICMI 2019), has served as area chairs for several conferences and is associate editor for Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, and Image and Vision Computing Journals. She has lectured tutorials at ICPR 2006, ICCV 2009, SCIA 2013 and FG 2018, authored/edited three books and nine special issues in journals. Dr. Zhao was a Co-Chair of many International Workshops at ICCV, CVPR, ECCV, ACCV and BMVC. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics. Her research has been reported by Finnish TV programs, newspapers and MIT Technology Review.