IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 28, 2019, accepted December 9, 2019, date of publication December 18, 2019,
date of current version December 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960629

Recognizing Spontaneous Micro-Expression Using a Three-Stream Convolutional Neural Network

BAOLIN SONG[1], KE LI[1], YUAN ZONG[2], (Member, IEEE), JIE ZHU[1], WENMING ZHENG[2], (Senior Member, IEEE), JINGANG SHI[3], AND LI ZHAO[1]

[1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, School of Information Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
[3]Center for Machine Vision and Signal Analysis, University of Oulu, 90014 Oulu, Finland

Corresponding authors: Yuan Zong (xhzongyuan@seu.edu.cn) and Li Zhao (101008849@seu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grant 61921004, Grant 61902064, Grant 61572009, Grant 61906094, Grant 61673108, Grant 61571106, and Grant 61703201, in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2016616, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170765, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242018K3DN01 and Grant 2242019K40047.

ABSTRACT Micro-expression recognition (MER) has attracted much attention with various practical applications, particularly in clinical diagnosis and interrogations. In this paper, we propose a three-stream convolutional neural network (TSCNN) to recognize MEs by learning ME-discriminative features in three key frames of ME videos. We design a dynamic-temporal stream, static-spatial stream, and local-spatial stream module for the TSCNN that respectively attempt to learn and integrate temporal, entire facial region, and facial local region cues in ME videos with the goal of recognizing MEs. In addition, to allow the TSCNN to recognize MEs without using the index values of apex frames, we design a reliable apex frame detection algorithm. Extensive experiments are conducted with five public ME databases: CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME. Our proposed TSCNN is shown to achieve more promising recognition results when compared with many other methods.

INDEX TERMS Micro-expression recognition, convolutional neural networks, apex frame location, spatiotemporal information.

I. INTRODUCTION

A special facial expression, a micro-expression (ME) is a rapid facial movement that is not subject to people's conscious recognition and can reveal someone's genuine emotion [1]. Compared with typical macro-expressions, MEs are of short duration (typically only 1/25s to 1/3s) and low intensity (the muscle movements only emerge in small facial regions) [2]. Due to these facts, MER is very difficult for a human to perform, and Ekman suggests that for MER tasks, people without training perform only slightly better than chance on average [3]. Thus, an automatic and reliable MER method should be developed to assist people in recognizing MEs accurately, particularly for application in fields such as clinical diagnosis [4], emotional interfaces [5], and interrogations [6], [7].

Recently, MER became a popular research topic, and extensive effective approaches have been proposed to perform this task. Typical MER approaches have two main components: facial feature extraction, which aims to extract useful information from facial videos to describe ME, and ME classification, which designs a classifier based on features extracted in the first step for MER tasks. Designing reliable facial features that can effectively describe the subtle changes of MEs would improve performance when performing MER tasks [8]. Facial feature extraction has attracted increasing attention from researchers. Among these feature extraction methods, local binary patterns on three orthogonal planes (LBP-TOP) and its variant are widely applied in video-based MER and other computer-vision tasks [9]–[11]. In addition, it is observed by the researchers in studies [5], [12] that the

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

temporal dynamics of video sequences can improve MER performance because they can represent the motion across a sequence of ME frames effectively. Some researchers have employed optical flow (OF) based techniques to extract spatiotemporal motion-dependent information from MEs, and many studies [13]–[15] have demonstrated their effectiveness for MER problems. For ME classification, various types of classifiers are mainly based on machine learning, such as support vector machine (SVM), relaxed K-SVD, and group sparse learning (GSL). Specifically, Zong *et al.* [16] proposed a kernelized GSL to facilitate the process of learning a set of weights from hierarchical spatiotemporal descriptors that can aid the selection of the important blocks from various facial blocks. Zheng *et al.* [17] proposed a relaxed K-SVD that learns a sparse dictionary to distinguish different MEs by minimizing the variance of sparse coefficients.

In recent years, researchers have also investigated deep learning methods to address the MER problem. For example, Kim *et al.* [18] proposed a deep learning method based on LSTM for MER tasks with spatiotemporal information extracted by CNN. Another study that uses a similar deep spatiotemporal structure is ELRCN [19], which uses optical flow features for the VGG-Faces model and then passes them on to recurrent layers. In [20], Xia *et al.* proposed a spatiotemporal extension of RNNs to jointly learn from both spatial and temporal cues of the ME samples to recognize MEs. A recent study [21] by Reddy *et al.* proposed two 3D-CNN-based models (MicroExpSTCNN and MicroExpFuseNet) to recognize MEs by extracting both the spatial and temporal information simultaneously by applying a 3D convolution operation to ME videos. Zhi *et al.* [22] similarly suggested 3D convolutional neural networks (3D-CNNs) architecture for self-learning feature extraction to represent facial MEs. Khor *et al.* [23] proposed a lightweight dual-stream shallow network as a pair of truncated CNNs with heterogeneous input features in MER tasks. Gan *et al.* [24] proposed an OFF-ApexNet to recognize MEs by learning optical flow features from some key frames of a ME video. In [25], Zhou *et al.* proposed a dual-inception network for MER. These deep learning methods perform well with MER tasks and outperform manual features and shallow classifiers.

Inspired by the success of these methods with MER, we propose a novel MER method called the three-stream CNN (TSCNN) in our conference paper [26]. The TSCNN consists of three major convolutional recognition streams that are used to learn the static-spatial, local-spatial, and temporal features from three different cues in ME videos, respectively. Some of recent studies [8], [27], [28] showed that the apex frame contains more ME-aware information, and thus we design a static-spatial stream CNN in the TSCNN to learn the static-spatial feature from the gray image of the apex frame for MER. The main reason of adding the local-spatial stream CNN is mainly inspired by recent findings in [9], [16], [29], [30]. Their studies have proved that the facial local region information has indeed contributions to distinguishing different MEs. Finally, following some studies in [19], [24], [25], [31], a dynamic-temporal stream CNN is also included in TSCNN to learn the temporal features from the optical flow field to deal with MER. Thus, such design explicitly describes facial texture and the subtle changes between ME video frames, reducing the complexity of ME recognition. Decoupling the static-spatial and temporal recognition streams also allows us to exploit the availability of large amounts of annotated image data by pretraining the static-spatial recognition stream using some large facial expression databases such as the FER2013 or ImageNet databases. In addition, our proposed method only analyzes three key frames (the onset, apex, and offset frames) instead of spotting facial micro-movements in all frames in an ME short video. It can avoid the interference of useless frames on the accuracy of MER and reduce redundant data. Thus, the TSCNN achieve parameters fitting in a short time and provide the possibility for real-time application.

This paper is an extended version of our conference paper. We will reinvestigate some problems in MER and extend our conference work. In addition to the contribution of our preliminary work, this paper contains the following main contributions:

1) A reliable apex frame detection algorithm is designed for the TSCNN without using the index values of apex frames given in ME videos from databases. Furthermore, we investigate the influence of parameter λ on the accuracy of TSCNN in MER tasks when locating apex frames from ME videos.
2) More extensive experiments are conducted to evaluate the TSCNN on five public ME databases: CASME II, SAMM, SMIC-HS, CAS(ME)^2, and CASME. In addition, we investigate the networks that have different combinations of three separate recognition streams using the above databases.

The rest of the paper is organized as follows. Our proposed method for MER is presented in Section II. Then, we discuss the experimental results for our method in Section III. Finally, conclusions are drawn in Section IV.

II. PROPOSED METHOD

In this section, we present our proposed method for recognizing MEs in detail. As shown in Fig. 1, our proposed method can be divided into three main parts: apex frames location, spatiotemporal feature extraction, and TSCNN modeling. In the first part, we introduce a reliable apex frame detection algorithm designed for the TSCNN in MER tasks without using the indices of apex frames given in ME videos. In the second part, we introduce the process of extracting spatiotemporal feature from the perspective of feature fusion on the network layer of the TSCNN, as well as its form. In the third part, we present the proposed TSCNN, including its detailed structure and how it deals with MER tasks. The details of each part are described in the following subsections.

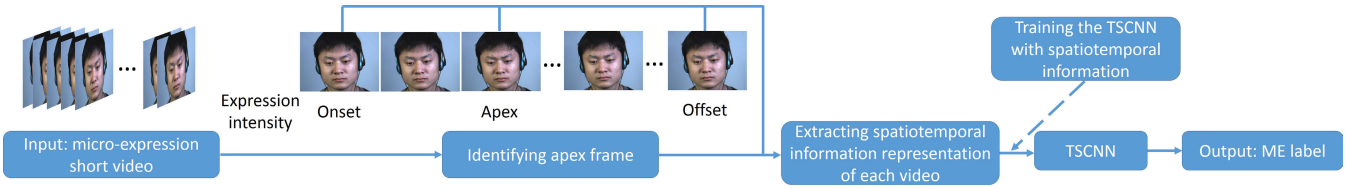


FIGURE 1. The framework of our proposed method for micro-expression recognition.

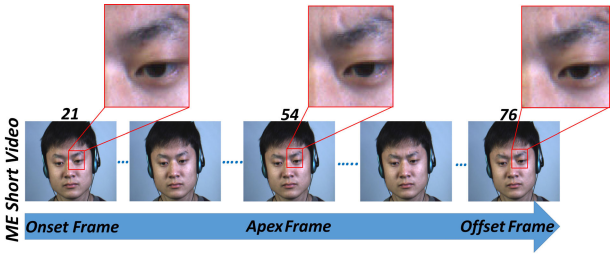


FIGURE 2. A demonstration of an ME short video. The FACS label of this sample is AU4, which indicates angry. The apex frame presents at the 54-th frame of this video. We can easily notice a subtle frowning action on the apex frame when observing each frame of the ME video with the naked eye.

A. IDENTIFYING APEX FRAMES

Index values of some key frames are typically given in ME videos from databases. For example, as shown in Fig. 2, the starting frame when an ME occurs is called an onset frame. The apex frame is the frame where ME intensity reaches its maximum. The ending frame is called the offset frame. The apex frame carries more spatial information about facial muscle micro-movement than other frames because micro-expression intensity reaches its maximum in this frame. The changes of optical flow between these three frames are most obvious in the whole video. So we suggest that the proposed TSCNN network can learn the significant features from spatiotemporal information carried by these three frames. In addition, some studies [8], [28] also suggest that the apex frame is typically the most expressive in an ME video, making it more discriminative and effective for ME recognition. For these reasons, we only need to analyze three frames (the onset, apex, and offset frames) instead of spotting facial micro-movements in all frames in an ME short video. Thus, apex frame location plays a critical role in MER tasks, especially when analyzing ME videos without using the indices of apex frames given by databases. In this subsection, we introduce an approach that can locate apex frames from ME videos for the TSCNN.

To avoid the interference from blank regions without MEs, we use a face-detection method, based on the work of Rowley *et al.* [32], to segment the facial region in each frame of the ME videos and then use the landmark algorithm in [33] to locate 68 facial landmarks using an ensemble of regression trees (ERT). To remove the influence of head posture, we first align facial regions. As shown in Fig. 3, we locate two inner eye corners, (x_1, y_1) and (x_2, y_2), and calculate the rotation

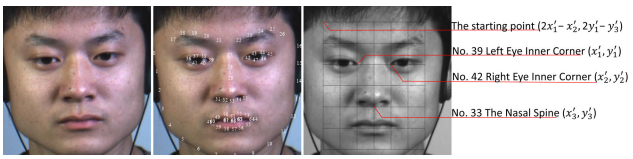


FIGURE 3. An example of 36 facial blocks yielded by 6 × 6 grid on a frame in the ME short video.

matrix R as described below.

$$R=\begin{pmatrix} \dfrac{x_2-x_1}{\sqrt{(y_2-y_1)^2+(x_2-x_1)^2}} & \dfrac{y_1-y_2}{\sqrt{(y_2-y_1)^2+(x_2-x_1)^2}} \\ \dfrac{y_2-y_1}{\sqrt{(y_2-y_1)^2+(x_2-x_1)^2}} & \dfrac{x_2-x_1}{\sqrt{(y_2-y_1)^2+(x_2-x_1)^2}} \end{pmatrix}. \tag{1}$$

In-plane rotation and facial size variations within the facial region are corrected based on

$$\left(x', y'\right) = (x, y)\, R^T. \tag{2}$$

After facial alignment, we obtain the inner eye corners (x_1', y_1'), (x_2', y_2') and the nasal spine point (x_3', y_3'). We then determine the width $\omega = (x_2' - x_1')/2$ and the height $h = (y_3' - y_1')/2$ of every division block and the starting point $(2x_1' - x_2', 2y_1' - y_3')$ based on these three points. Then, the facial region is divided into 6×6 equal-sized blocks, as shown in Fig. 3.

To distinguish the relevant peaks from local magnitude variations in each frame in ME short videos and determine when ME reaches its maximum, we analyze facial texture and shape appearance. In many studies, LBP and its variant are preferred when analyzing facial texture and shape appearance [9]. Ojala *et al.* [34] proposed a uniform pattern LBP (UP-LBP) to reduce the sparse conditions caused by feature dimensions and improve the statistical properties of facial features. Thus, we calculate the UP-LBP histogram from each block in the facial area of each frame to describe facial texture and shape appearance to determine which can yield the best performance. For each frame of the input video, we calculate the UP-LBP histogram with P = 8 & R = 3 for each of its 36 blocks as $H_{i,0}, H_{i,1}, \ldots, H_{i,35}$. The dimension of each histogram is 10; thus, each frame in videos will correspond to 36 10-dimensional vectors as $H_i : H_{i,0}, H_{i,1}, \ldots, H_{i,35}$.

Feature difference (FD) analysis compares the differences in the appearance-based features of sequential video frames within a specified interval and provides information about

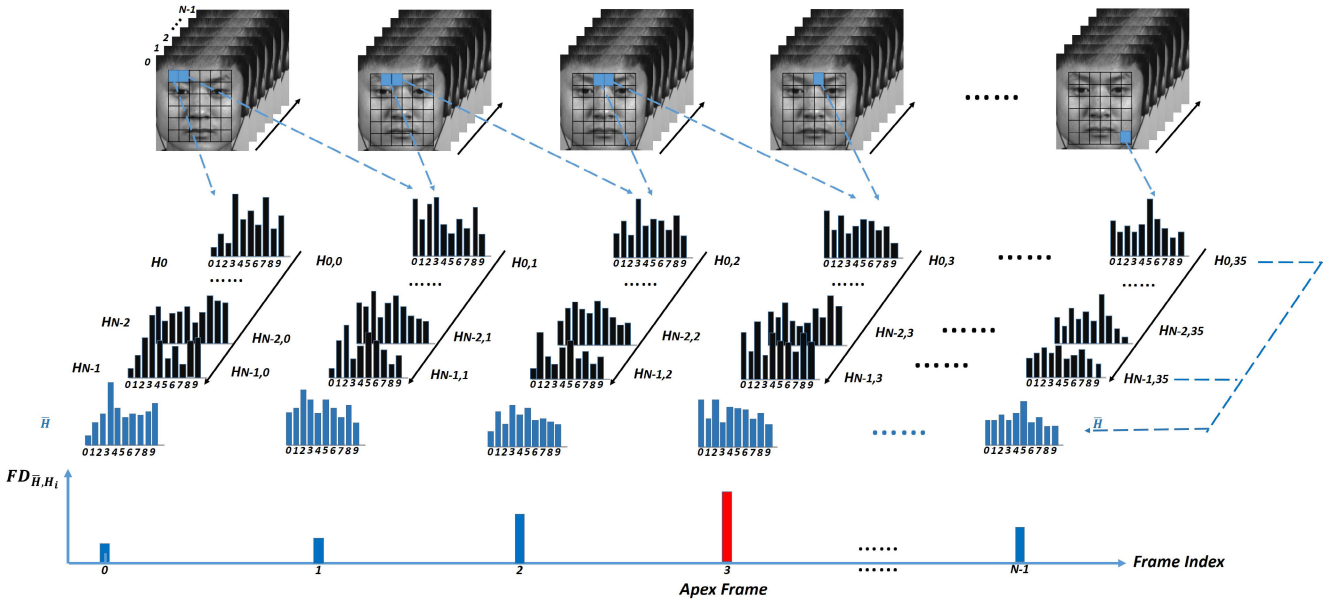


FIGURE 4. An illustration of how to identify apex frames from ME short videos. The top describes the UP-LBP histogram of each block in facial regions, which describes the facial texture and shape appearance of each frame. The bottom presents that FD is calculated to locate the apex frame (i.e., the highest intensity frame of rapid facial movements).

spatial location in identified facial movements. To capture the greatest changes, we use FD values to roughly locate the apex frame (i.e., the highest intensity frame of rapid facial movements). The FD value between H_i of the i-th frame and H_j of the j-th frame is given by:

$$\begin{aligned} d_k &= \frac{1}{10}\sum_{\alpha=1}^{10}\frac{(H_{i,k}^{(\alpha)}-H_{j,k}^{(\alpha)})^2}{H_{i,k}^{(\alpha)}+H_{j,k}^{(\alpha)}}, \quad k=0,1,2,\ldots,35 \\ &\{d_0', d_1', \ldots, d_{35}'\} \\ &\quad = \{d_0, d_1, \ldots, d_{35}\}, \quad d_0' > d_1' > \cdots > d_{35}' \\ FD_{H_iH_j} &= \frac{1}{\lambda}\sum_{\beta=0}^{\lambda-1} d_\beta'. \end{aligned} \tag{3}$$

where $H_{i,k}^{(\alpha)}$ and $H_{j,k}^{(\alpha)}$ respectively represent the values of the same dimension α in the k-th UP-LBP histogram corresponding to H_i and H_j. Only the largest λ values among the 36 distances are used in calculations, because the occurrence of an ME will result in larger d_i values in some (but not all) blocks between two adjacent frames. More details using different values of λ are presented and discussed in Section III.D.

Finally, we calculate H_{onset} of the onset frame and H_{offset} of the offset frame in an ME short video respectively, and then calculate the average value $\overline{H}$ between H_{onset} and H_{offset}. If the $FD_{H_k,\overline{H}}$ value between H_k of the k-th frame and $\overline{H}$ exists the greatest value, the k-th frame is seen as the apex frame in the whole video. As shown in Fig. 4, a higher value of $FD_{H_k,\overline{H}}$ indicates that a muscle movement with a larger amplitude exists in the facial area of the frame.

B. SPATIOTEMPORAL FEATURE EXTRACTED BY THE TSCNN FOR MER

Spatiotemporal feature is characterized by the type of information encoded in space and time, which can describe MEs in videos and allows it to represent subtle expressions in videos more efficiently. In this paper, our proposed spatiotemporal feature consists of three components: static-spatial, local-spatial, and temporal components. Details of each component are described below.

1) STATIC-SPATIAL COMPONENT

Static-spatial information, especially some appearance and overall outline information, has gained increasing attention in facial image analysis and has been shown to be effective in tackling the MER problems [9], [35], [36]. The static appearance and overall outline of a whole face is the most intuitive since some facial expressions are strongly associated with particular facial muscle contractions. In an ME video, the apex frame carries more spatial information because facial muscle micro-movement of this frame is more obvious than that of other frames. Thus, we consider the gray image of the whole face in the apex frame as the input of the static-spatial recognition stream in the TSCNN, which is cropped to 48 × 48 pixels. Finally, the static-spatial feature extracted from the whole face is fused together with two other feature vectors from the other two recognition streams at the second fully connected layer of the TSCNN network.

2) LOCAL-SPATIAL COMPONENT

However, it is not sufficient to represent all characteristics of ME videos if only static-spatial components are considered

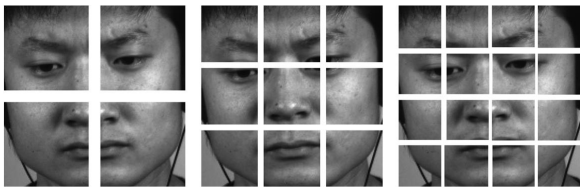


FIGURE 5. Images split into 2 × 2 blocks (left), 3 × 3 blocks (intermediate), and 4 × 4 blocks (right).

when performing ME recognition. Since ME muscle movements only appear in facial local regions (e.g., mouth, cheek, eyebrows, and eyes), motion changes at these regions and conveys meaningful information from different MEs. Block-based segmentation of a face to extract facial local features is a common practice when extracting facial local features, that can be assigned to regions that contain key facial features with the goal of enhancing recognition power [16]. Some studies [9], [16], [29], [30] have proved that the facial local region information has indeed contributions to distinguishing different MEs.

However, many methods use block-based segmentation of a face without considering the effects of block size. Ideally, the contribution from all blocks in a face should be varied greatly from different grid divisions of a face. Thus, we use spatial grids with multiple sizes $\{n \times n | 2 \times 2, 3 \times 3, 4 \times 4\}$ to divide the grayscale image of the apex frame into several facial blocks and then stack them up to obtain a facial block sequence to serve as the input of local-spatial stream CNN in the TSCNN, where the division detail is shown in Fig. 5. Specifically, the gray image of the apex frame in an ME video sample is scaled to $48n \times 48n$ before image segmentation. The facial block sequence as input is an n^2-channel gray image, and the size of each channel is 48×48 pixels. Finally, we obtain the local-spatial component at the last fully connected layer of the local-spatial stream in the TSCNN.

3) TEMPORAL COMPONENT

Compared to still image classification, videos provide data augmentation for single image classification. The temporal components of videos provide an additional information for MER. Many muscle movements emerge in facial regions and can be reliably recognized based on the motion information [31]. For example, we select the onset frame F_{21}, the apex frame F_{54} and the offset frame F_{76} in the sub01/EP04 02.avi sample of the CASME II database, to calculate the horizontal and vertical optical flow field and visualize it (see Fig. 6). The FACS label of this sample (AU4) indicates a frowning action. Using this image of the optical flow field, we can observe the muscle movements in the subject's eyebrows from the occurrence to the disappearance of an angry micro-expression, although the amplitude of the facial muscle motion between adjacent frames is very small. Thus, using only the spatial component does not capture the motion well in ME videos. In this section, we describe the process of extracting the temporal component from ME videos using dynamic-temporal stream in the TSCNN.

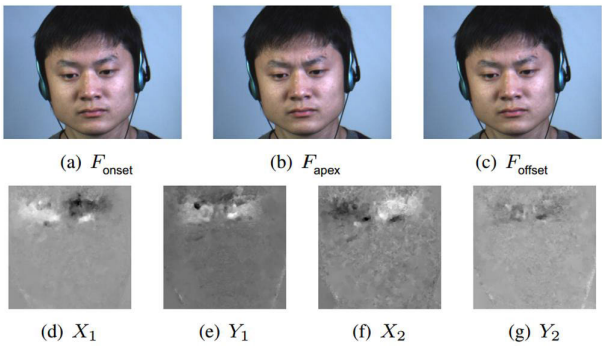


FIGURE 6. The horizontal and vertical optic flow fields and visualization.

Many studies have used methods based on optic-flow (OF) [37]–[42] to characterize the local dynamics of a temporal texture and detect motion information between adjacent frames. The optical flow field is a set of displacement vector fields between pairs of consecutive frames. The horizontal and vertical components of the vector field can be thought of as image channels. Thus, it is suitable for deep networks to learn advanced features.

Optical flow fields between three frames (the onset, apex, and offset frames) are calculated by the approach in [43], in which the function $flow(F_1, F_2)$ takes two frames as inputs and a horizontal optical flow field X and a vertical optical flow field Y as outputs, as described below,

$$\begin{aligned} X_1, Y_1 &= flow\left(F_{onset}, F_{apex}\right), \\ X_2, Y_2 &= flow\left(F_{apex}, F_{offset}\right). \end{aligned} \tag{4}$$

where F_{onset}, F_{apex}, and F_{offset} represent the onset frame, the apex frame and the offset frame in an ME video, respectively. Two sets of optical flow fields are obtained via the formula above. Each set contains two optical flow fields (horizontal and vertical) that move pixels in the x- and y-directions, respectively. Thus, the two sets of optical flow fields can completely represent ME movements from occurrence to peak and then from peak to termination.

Since data in optical flow fields is represented as float64 values, we must normalize the optical flow matrix via min-max normalization as follows:

$$H_{norm} = \frac{H_{org} - \min(H_{org})}{\max(H_{org}) - \min(H_{org})}. \tag{5}$$

where H_{org} and H_{norm} are the matrix before and after normalization, respectively. By transforming the original matrix linearly, all elements fall into the [0,1] interval. Thus, we obtained two sets of normalized optical flow fields for each ME video in a given database and then stack them in the same way as processing the local-spatial component, which can be considered a 4-channel image of size 48×48 pixels. Finally, we take the 4-channel image as the input of the dynamic-temporal recognition stream and obtain the temporal component that is a 1024-dimension vector.

TABLE 1. The configuration of the TSCNN network.

Model	TSCNN								
Stream	S			L			T		
Conv1	Filter size = 5 × 5	Stride = 1	N = 64	Filter size = 5 × 5	Stride = 1	N = 64	Filter size = 5 × 5	Stride = 1	N = 64
Maxpool1	Filter size = 5 × 5	Stride = 2	N = 64	Filter size = 5 × 5	Stride = 2	N = 64	Filter size = 5 × 5	Stride = 2	N = 64
Conv2	Filter size = 3 × 3	Stride = 1	N = 64	Filter size = 3 × 3	Stride = 1	N = 64	Filter size = 3 × 3	Stride = 1	N = 64
Conv3	Filter size = 3 × 3	Stride = 1	N = 64	Filter size = 3 × 3	Stride = 1	N = 64	Filter size = 3 × 3	Stride = 1	N = 64
Avepool1	Filter size = 3 × 3	Stride = 2	N = 64	Filter size = 3 × 3	Stride = 2	N = 64	Filter size = 3 × 3	Stride = 2	N = 64
Conv4	Filter size = 3 × 3	Stride = 1	N = 128	Filter size = 3 × 3	Stride = 1	N = 128	Filter size = 3 × 3	Stride = 1	N = 128
Conv5	Filter size = 3 × 3	Stride = 1	N = 128	Filter size = 3 × 3	Stride = 1	N = 128	Filter size = 3 × 3	Stride = 1	N = 128
Avepool2	Filter size = 3 × 3	Stride = 2	N = 128	Filter size = 3 × 3	Stride = 2	N = 128	Filter size = 3 × 3	Stride = 2	N = 128
FC1	1024			1024			1024		
FC2	3072								
Output	3/4/5								

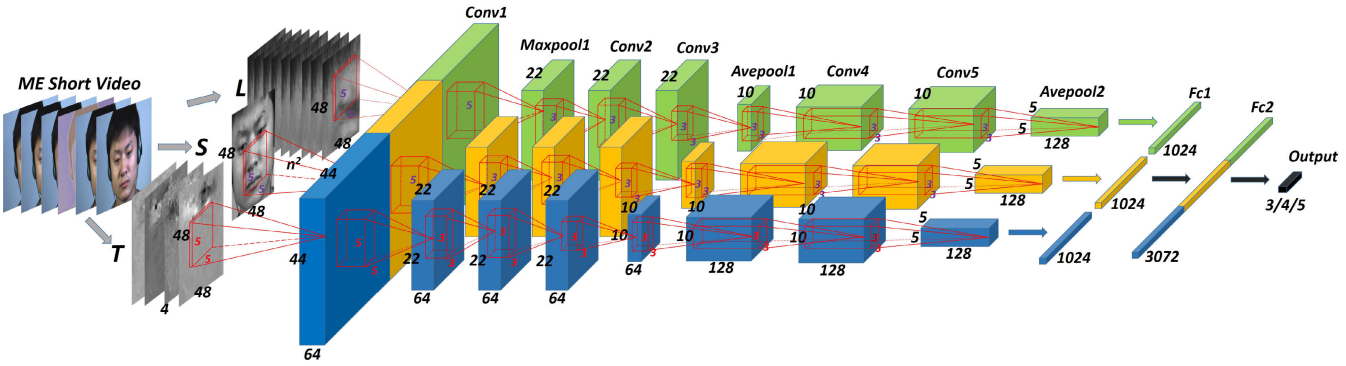


FIGURE 7. The architecture of the TSCNN network.

C. TSCNN MODEL FOR ME RECOGNITION

The proposed TSCNN is based on the research on CNN networks. It is composed of multiple processing layers in a multi-stream architecture and can learn representations of data using multiple levels of abstraction. The TSCNN consists of three-stream CNNs (i.e., the static-spatial stream (S), the local-spatial stream (L), and the dynamic-temporal stream (T)), which learn discriminative features for recognizing ME from three different clues in three key frames from ME videos. Its detailed structure and how it deals with MER tasks, is shown in Fig. 7.

To reduce the redundant parameters and realize parameters sharing, each stream module in the TSCNN has the same structure. This design aim to make the TSCNN achieve parameters fitting in a short time and reduce the amount of training. As shown in Table 1, each recognition stream is a simplified network that uses a 2D convolution kernel and pooling cell to automatically represent the properties of subtle facial movements. The three recognition streams are then combined by late fusion in a fully connected layer.

Among the three recognition streams in our TSCNN, the static-spatial recognition stream (S) operates on individual video frames (e.g., the apex frame), effectively performing action recognition using still images. We consider the gray image of the apex frame as the input of this recognition stream. The local-spatial recognition stream (L) operates on the n^2-channel gray image after stacking $n \times n$ blocks of the gray image of the apex frame. The input to the dynamic-temporal stream (T) contains optical flow displacement fields between three frames (the onset, apex, and offset frames), whose center frame is the apex frame. We use the dynamic-temporal stream with optical flow sequences to ensure that the TSCNN networks can further acquire higher-level features. Such inputs explicitly describes the motion between video frames, which significantly improves accuracy and makes ME recognition easier.

Each recognition stream is compacted with only 9 layers: 5 convolutional layers, 3 pooling layers and 1 fully connected layer. For the first convolution layer in each recognition stream, the kernel size is set equal to 5×5, the stride size is set equal to 1, and zero padding is set equal to "valid". For other four convolutional layers in every recognition stream, we use a kernel size of 3×3 with a stride of S = 1, and zero padding is set equal to 1. The number of kernels (N) for each layer is 64, 64, 64, 128, and 128 respectively. The N value of the last two convolutional layers is much larger than that of the other layers and will increase the computational complexity of the network. Many studies [9], [44] have demonstrated

that a large N can cover more abstract features of certain important facial regions, such as the eye or mouth region, and thus improve the performance of MER.

Three pooling layers of every stream are used to down-sample the spatial dimensions of the input, which contains a max pooling layer with a window size of 5×5, and two mean pooling layers with a window size of 3×3. The stride of each pooling layer is 2, and the number of kernels (N) is set equal to 64, 64, and 128, respectively. This design is important in real applications because there is no agreed standard frame rate so far for recoding the micro-expressions (i.e., the ME video could be recorded in various frame rate). The design of different network streams can adapt to different frame rates, which may make the whole network robust to the frame rate of the input data.

The final layer of every recognition stream is a fully connected layer that has the same configuration. Their output dimension are all set equal to 1024 to reduce the number of parameters in the model, and prevents overfitting. Then, the output of three recognition streams are merged into a 3072-dimensional feature vector. In the final layer of the TSCNN, we transform the feature vector to one having the same dimension as the ME class number in MER tasks. Thus, the output dimension of the final layer is different for different databases.

All hidden layers are equipped with the Parametric Rectified Linear Unit (PRELU) function, which is defined as follows:

$$PReLU(y_i) = \begin{cases} y_i, & y_i > 0 \\ a_i y_i, & y_i \leqslant 0 \end{cases} \tag{6}$$

where i denotes the channel, and a_i is a parameter obtained during training. Compared with the traditional activation function (sigmoid, tanh, ReLU, etc.), the PReLU can improve classification of the CNN model at a cost of overfitting and computational complexity.

Cross entropy is used to calculate the loss function of TSCNN, which can be defined as:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{Y} \tau\left(y_n, j\right) \times \log P_{n,j}, \tag{7}$$

where N denotes the number of the training samples, Y is the number of emotion types, y_n is the label of n-th training sample and $P_{n,j}$ represents the value of the prediction that the n-th training sample is predicted to be the j-th class. We use the backpropagation (BP) algorithm to minimize the loss function of the TSCNN and update the weight parameters. The training optimizer is the stochastic gradient descent (SGD) algorithm with Nesterov Momentum.

The iterative process is as follows:

$$\begin{aligned} \upsilon_t &= \gamma\upsilon_{t-1} + \alpha\nabla_\theta J\left(\theta - \gamma\upsilon_{t-1}\right), \\ \theta &\leftarrow \theta - \upsilon_t. \end{aligned} \tag{8}$$

where α represents the learning rate. The attenuation of the weight parameters is set equal to 10^{-5}, and the correction factor is set equal to 0.9.

III. EXPERIMENTS

In this section, we present experimental results of our proposed method in detail, including the datasets we used, the implementation details, and the comparison of experimental results, etc.

A. DATABASES AND EXPERIMENT SETTING

In this section, we conduct extensive MER experiments to evaluate our proposed TSCNN method. The CASME II [45], SMIC-HS [46], SAMM [47], CAS(ME)^2 [48], and CASME databases [49] are used in our experiments as they are widely used spontaneous ME databases. Details of the five spontaneous ME databases used in this paper are listed below.

1) The CASME II database was collected by Yan *et al.* from the Institute of Psychology, Chinese Academy of Science. The database includes 247 ME samples with high spatial and temporal resolutions from 26 subjects. The face videos were recorded at 200 fps, with an average face size of 280 × 340. These samples are categorized into 5 ME classes: happiness (32), surprise (25), disgust (64), repression (27), and others (99), where the number in the brackets are the number of corresponding MEs present in the database. We pick all ME samples in the CASME II database for experimentation.
2) The SMIC-HS database was collected by Li *et al.* from the University of Oulu. The database includes 164 ME samples from 16 subjects, which are recorded at 100 fps with an average face size of 280 × 340. These samples are divided into 3 classes: positive (51), negative (70), and surprise (43), where the number in the brackets are the number of corresponding MEs present in the database. We use all ME samples in the SMIC-HS database for experimentation.
3) The SAMM database was collected by Davison *et al.* from Manchester City University. The database includes 159 ME samples from 29 subjects. These samples are divided into 8 ME classes. The face videos were recorded at 200 fps with an average face size of 650 × 960. Note that since the sample number of several MEs in SAMM is small, we only use ME samples whose number is larger than 10 for experimentation; these include anger (57), contempt (12), happiness (26), surprise (15), and others (26), where the number in the brackets are the number of corresponding MEs present in the database.
4) The CAS(ME)^2 database (Chinese Academy of Science Macro- and Micro-expression) was established by the Chinese Academy of Science. The database includes both spontaneous macro (300) and micro (57) expression video sequences of 22 subjects (13 females and 9 males). These videos have been captured by a

camera with a 500-ms shutter speed, and the recorder's resolution was set equal to 640×480 pixels at 30 frames per second. By extracting more than 600 AUs, these image sequences are categorized into three emotion classes: anger, happy and disgust. In our experimental setup, we selected 341 image sequences, anger (102), happy (151) and disgust (88) of macro- and micro-expressions, where the number in the brackets is the number of corresponding expressions present in the database. To ensure fair comparisons and following other methods, such as 3D CNN based techniques in the literature [21], we also report the recognition results under the same conditions as the literature.

5) The CASME database was built by the Chinese Academy of Sciences. The database contains two datasets A and B with 195 ME samples from 19 subjects; videos were recorded at 60 fps. The video clips in dataset A of the database were recorded with the resolution of 1280×720 pixels in natural light. The samples in dataset B were recorded with the resolution of 640×480 pixels under LED illumination. All samples were coded with onset, apex and offset frames with action units (AUs) marked and emotions labeled. There are 8 classes of the micro-expressions in this database: tense, disgust, repression, surprise, happiness, fear, sadness, and contempt. Since the three classes of happiness, fear and sadness contain very few samples, we chose the remaining four classes in our experiment: tense (69), disgust (44), repression (38), and surprise (20).

For all experiments in the above five public databases, the leave-one-subject-out (LOSO) protocol is used to calculate the recognition accuracy and mean F_1-score to report the performance of the MER methods. In each fold, the samples of one subject are used as the test set, while the remaining samples are used for training. This method can eliminate appearances of samples from the same subject in the training and verification sets, thus ensuring the reliability of the experimental results.

The accuracy rate can be calculated as follows:

$$Accuracy = \frac{\sum_{i=1}^{S} T_i}{\sum_{i=1}^{S} N_i} \times 100\%. \tag{9}$$

where T_i and N_i are the number of correct predictions and the number of testing samples, respectively, when the samples of the i-th subject is used as the test set. The accuracy rate shows the average ''hit rate'' across all classes and does not evaluate the performance of the algorithm objectively.

The CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME databases are highly imbalanced [19], [48]–[52], which means that the number of one type of micro-expression samples is significantly more or less abundant than other types of ME samples. The performance of the classifier that deals with each emotion class is not revealed. Thus, we calculate an F_1-score to describe the classification effect for each class, and use it as a criterion to measure the network performance along with the accuracy rate. The F_1-score can be defined as:

$$F = \frac{1}{c} \sum_{i=1}^{c} \frac{2p_i \times r_i}{p_i + r_i}. \tag{10}$$

where p_i and r_i are the precision and recall of the i-th micro-expression, respectively, and c is the number of micro-expressions.

FIGURE 8. Expansion of training samples.

B. IMPLEMENTATION DETAILS

We set the input image size of each recognition stream in the TSCNN equal to 48×48, and the facial block number in the local-spatial stream is set equal to 3×3. The base learning rate is set equal to 10^{-3} in the experiment due to difficulties related to the subtlety of MEs. The attenuation of weight parameters is set equal to 10^{-5}, and the correction factor is set equal to 0.9 in the experiment. Dropout is used on all fully connected layers in the TSCNN model to avoid overfitting problem. The λ values of the carrying experiments with the CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME databases equal 25, 21, 23, 25, and 20, respectively, when locating apex frames from ME videos.

To train the TSCNN model to distinguish MEs, large amounts of training data is needed. However, only a few key expression frames can be selected for the training in an ME video. We thus expand the number of training samples by taking the original samples and applying a horizontal flip and clockwise/counterclockwise rotation in 5 or 10 degree increments a total of 10 times, as shown in Fig. 8. This process yields 2470, 1360, 1640, 3410, and 1710 samples from the CASME II, SAMM, SMIC-HS, CAS(ME)^2 and CASME databases, respectively. When the training data is ready, we begin to train the TSCNN network according to our purposes.

We pretrain the static-spatial recognition stream using the large facial expression database FER2013, where obtained weights are used for initialization. The weights of the local-spatial recognition stream and dynamic temporal recognition stream are randomly initialized. Mini-batch is not applied in the experiment due to the small sample size. Early stopping is used to train our TSCNN model over 500 iterated epochs in each fold. When the validation loss curve is generally stable, training for each fold will stop, and our TSCNN model will output the emotion classification label.

TABLE 2. Comparison between our method with some state-of-the-art methods on CASME II database.

Method	Class	Accuracy(%)	F1-score
LBP-TOP + AdaBoost [50]	5	43.78	0.3337
EVM + HIGO [5]	5	67.21	N\A[a]
STCLQP [53]	5	58.39	0.5836
FDM [54]	5	41.96	0.4700
Riesz Wavelet [55]	5	46.15	0.4307
SIP + MOP [56]	5	45.75	N\A
Hierarchical STLBP-IP + KGSL [16]	5	65.18	0.6254
LBP-TOP [57]	5	51.91	N\A
LBP-TOP [52]	5	51.00	0.4700
DMDSP + LBP-TOP [51]	5	49.00	0.5100
CNN + LSTM [18]	5	60.98	N\A
FMBH [63]	5	69.11	N\A
STLBP-IIP [64]	5	62.75	N\A
DiSTLBP-IIP [64]	5	64.78	N\A
AlexNet [66]	5	62.96	0.6675
OF + CNN [71]	5	56.94	N\A
ELRCN [19]	5	52.44	0.5000
SSSN [23]	5	71.19	0.7151
CNN + SFS [70]	5	47.30	N\A
DSSN [23]	5	70.78	0.7297
TIM+DCNN+SVM [75]	5	64.90	N\A
3D-CNNs (with transfer learning) [22]	5	65.90	N\A
TSCNN-I(Ours)[b]	**5**	**74.05**	**0.7327**
TSCNN-II(Ours)[c]	**5**	**80.97**	**0.8070**

[a]N\A - no results reported.

[b]TSCNN-I - The TSCNN with using the apex frame identified in Section II.A.

[c]TSCNN-II- The TSCNN with using the apex frame given by the CASME II database.

C. COMPARISON WITH THE STATE-OF-THE-ART METHODS

In this subsection, we compare the best result achieved by our method with those of other state-of-the-art methods [5], [9], [16], [18]–[25], [29], [50]–[84] using the five public ME databases (CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME). The LOSO protocol was used for all the methods. In Tables 2 through 6, TSCNN-II represents the results achieved by TSCNN when using the apex frame given by databases. TSCNN-I represents the results achieved by TSCNN when using the apex frame located by our proposed approach in Section II.A.

From Tables 2 through 6, our TSCNN is shown to yield an accuracy of 80.97% and an F_1-score of 0.8070 with the CASME-II database; 71.76% and 0.6942 with the SAMM database; 75.41% and 0.7463 with the CAS(ME)^2 database; and 73.88% and 0.7270 with the CASME database when we use index values of some key frames given by these databases in MER tasks. Thus, our TSCNN shows significant improvement in recognition compared to other methods. Additionally, our TSCNN model achieves improved classification results in MER tasks, especially when assuming that these databases do not provide us with index values and that apex frames must be located. In this case, the accuracies and F_1-score with the CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME databases are 74.05% and 0.7327; 72.74% and 0.7236; 63.53% and 0.6065; 71.62% and 0.7129; and 70.73% and 0.6736, respectively.

As described above, the experimental performance of the TSCNN-I in MER is worse than that of the TSCNN-II. This result agrees with our expectations, because accurately locating an apex frame in an ME video is difficult and may decrease the performance of the deep learning method. Additionally, many other methods [8], [28], [85], [86] only locate apex-feature time intervals roughly.

Next, to analyze the recognition performance of our TSCNN in MER tasks, we only compare the results of the TSCNN-I and other methods when simulating the MER problem without using true indices given by the databases.

1) Comparison of results using the CASME II database: As shown in Table 2, we report the performance of the TSCNN-I for MER tasks using the CASME II database and compare it with that of other methods [5], [16], [18], [19], [22], [23], [50]–[57], [63], [64], [66], [70], [71], [75] using the LOSO protocol. Our TSCNN-I yielded an accuracy of 74.05% and a mean F_1-score of 0.7327 using the CASME II database. Compared with state-of-the-art methods (FMBH [63], OF+CNN [71], ELRCN [19], SSSN [23], DSSN [23], and 3D-CNNs [22]), our method exhibits an improvement of 4.94%, 17.11%, 21.61%, 2.86%, 3.27%, and 8.15% in accuracy, respectively. Thus, our TSCNN-I yielded improved recognition, especially in the absence of index values given by the databases.
2) Comparison of results using the SMIC-HS database: As shown in Table 3, our TSCNN-I yields the highest recognition accuracy (72.74%) and F_1-score (0.7236) among other state-of-the-art approaches [16], [22]–[25], [50], [51], [53], [54], [56], [58]–[60], [63]–[69], [72]–[75]. Compared with the best results of other methods (Bi-WOOF+Phase [67], TIM+DCNN+SVM [75], Dual-Inception Network [25], SSSN [23], DSSN [23], 3D-CNNs [22], OFF-ApexNet [24], and 3D-FCNN [74]), our method yields 4.45%, 6.84%, 6.74%, 9.33%, 9.33%, 6.44%, 4.96%, and 17.25% better recognition accuracy, respectively.
3) Comparison of results using the SAMM database: As shown in Table 4, our TSCNN-I yields a recognition accuracy of 63.53% and an F_1-score of 0.6065, which are considerably better than the other methods [5], [9], [20], [23], [61], [62], [66]. Xia *et al.* [20]'s STRCN-G yields a 78.60% recognition accuracy in 4 ME classes. However, the results achieved by our TSCNN-I (5 ME classes) are better than that of STRCN-A (4 ME classes).
4) Comparison of results using the CAS(ME)^2 database: Since the CAS(ME)^2 database is a mixed database of spontaneous micro- and macro-expressions, few methods for micro-expression recognition [21], [65], [76], [77] have been designed and tested using this database.

TABLE 3. Comparison between our method with some state-of-the-art methods on SMIC-HS database.

Method	Class	Accuracy(%)	F1-score
OSW-LBP-TOP [58]	3	53.66	N\A[a]
LBP-TOP + TIM [59]	3	53.56	N\A
FDM [54]	3	54.88	0.5380
LBP-TOP + AdaBoost [50]	3	44.34	0.4731
SIP + MOP [56]	3	51.83	N\A
Hierarchical STLBP-IP+KGSL [16]	3	60.37	0.6125
STCLQP [53]	3	64.02	0.6381
2Standmap [60]	3	57.90	N\A
DMDSP + LBP-TOP [51]	3	58.00	0.6000
FMBH [63]	3	71.95	N\A
Bi-WOOF + Phase [67]	3	68.29	0.6730
STLBP-IIP [64]	3	60.37	N\A
DiSTLBP-IIP [64]	3	63.41	N\A
VGG-11 [65]	3	34.61	0.3558
VGG-16 [65]	3	58.00	0.5964
ResNet18 [73]	3	35.76	0.3602
GoogLeNet [68]	3	51.23	0.5511
SqueezeNet [69]	3	53.81	0.5603
SSSN [23]	3	63.41	0.6329
TIM+DCNN+SVM [75]	3	65.90	N\A
DSSN [23]	3	63.41	0.6462
Dual-Inception Network [25]	3	66.00	0.6700
AlexNet [66]	3	59.76	0.6013
OFF-ApexNet [24]	3	67.68	0.6709
CapsuleNet [72]	3	58.00	0.5900
3D-FCNN (G-5×5,XF,YF) [74]	3	55.49	N\A
3D-CNNs (with transfer learning) [22]	3	66.30	N\A
TSCNN-I(Ours)[b]	**3**	**72.74**	**0.7236**

[a]N\A - no results reported.

[b]In this table, only TSCNN-I is reported while TSCNN-II is not available because SMIC-HS database do not provide index values of apex frames.

TABLE 4. Comparison between our method with some state-of-the-art methods on SAMM database.

Method	Class	Accuracy(%)	F1-score
LBP-TOP [9]	5	34.56	0.2892
LBP-SIP [61]	5	36.03	0.3133
HOG-TOP [5]	5	36.03	0.3403
HIGO-TOP [5]	5	41.18	0.3920
LPQ-TOP [62]	5	38.97	0.2468
STRCN-A [20][a]	4	54.50	0.4920
STRCN-G [20]	4	78.60	0.7410
AlexNet [66]	5	52.94	0.4260
SSSN [23]	5	56.62	0.4513
DSSN [23]	5	57.35	0.4644
TSCNN-I(Ours)[b]	**5**	**63.53**	**0.6065**
TSCNN-II(Ours)[c]	**5**	**71.76**	**0.6942**

[a]In Reference [20], the author only reports the experimental results on 4 classes of MEs.

[b]TSCNN-I - The TSCNN with using the apex frame identified in Section II.A.

[c]TSCNN-II- The TSCNN with using the apex frame given by the SAMM database.

TABLE 5. Comparison between our method with some state-of-the-art methods on CAS(ME)^2 database.

Method	Test types	Class	Accuracy(%)	F1-score
VGG-16 [65]	macro- and micro-expressions	3	44.29	N\A[a]
VGG-19 [65]	macro- and micro-expressions	3	44.28	N\A
ResNet [76]	macro- and micro-expressions	3	74.48	N\A
LEARNet [77]	macro- and micro-expressions	3	**76.33**	N\A
TSCNN-I(Ours) [b]	macro- and micro-expressions	3	71.62	0.7129
TSCNN-II(Ours) [c]	macro- and micro-expressions	3	75.41	0.7463
MicroExpFuseNet (Late) [21]	only micro-expressions	3	79.31	N\A
MicroExpFuseNet (Intermediate) [21]	only micro-expressions	3	83.25	N\A
MicroExpSTCNN [21]	only micro-expressions	3	**87.80**	N\A
TSCNN-I(Ours)	only micro-expressions	3	84.47	0.8421
TSCNN-II(Ours)	only micro-expressions	3	86.22	0.8618

[a]N\A - no results reported.

[b]TSCNN-I - The TSCNN with using the apex frame identified in Section II.A.

[c]TSCNN-II-The TSCNN with using the apex frame given by the CAS(ME)^2 database.

Additionally, the number of samples and test types selected for testing with this database are different from each other in these studies. Therefore, to ensure a fair comparison, our TSCNN was tested under two different experimental conditions. One is the same as that used in [65], [76], [77], which contains 341 image sequences with macro- and micro-expressions selected by the authors. The other is the same as that used in [21], which only contains micro-expression videos that have the same samples as the literature.

As shown in Table 5, our TSCNN-I yields a recognition accuracy of 71.62% and an F_1-score of 0.7129 when a total of 341 image sequences of macro- and micro-expressions are selected. Compared with the results of other state-of-the-art approaches [65], [76], [77], our method is very competitive using this database. We also compare the TSCNN with two 3D-CNN methods (MicroExpSTCNN and MicroExpFuseNet) that were proposed in [21] using only micro-expression videos that have the same number of samples as in the literature. The recognition accuracy (84.47%) of our TSCNN-I outperforms that (79.31%) of the MicroExpFuseNet (Late) and 83.25% of MicroExpFuseNet (Intermediate) methods. The performance of the MicroExpSTCNN method outperforms that of our TSCNN-I. The above experimental results show that our TSCNN is very competitive, compared with the two 3D-CNN based methods in [21].

5) Comparison of results using the CASME database: As shown in Table 6, we report the performance of the TSCNN-I in MER tasks using the CASME database and compare it with other methods [9], [29], [53], [54], [61], [78]–[84] using the LOSO protocol. Our TSCNN-I yielded an accuracy of 70.73% and a mean F_1-score of 0.6736. Compared with other methods

TABLE 6. Comparison between our method with some state-of-the-art methods on CASME database.

Method	Class	Accuracy(%)	F1-score
LBP-TOP [9]	4	37.43	0.3296
LOCP-TOP [79]	4	31.58	N\A[a]
MDMO [29]	4	56.29	0.5551
CLBP-TOP (S+M) [80]	4	45.31	N\A
LBP-SIP [61]	4	36.84	N\A
STCLQP [53]	4	57.31	0.5618
MPCA [84]	4	41.01	N\A
DiSTLBP-RIP [83]	4	64.33	N\A
FDM [54]	4	56.14	0.5499
Cuboids [81]	4	33.33	N\A
STLMBP [78]	4	46.20	N\A
LTOGP(without FS) [82]	4	61.07	N\A
LTOGP(with FS) [82]	4	68.64	N\A
TSCNN-I(Ours)[b]	**4**	**70.73**	**0.6736**
TSCNN-II(Ours)[c]	**4**	**73.88**	**0.7270**

[a]N\A - no results reported.

[b]TSCNN-I - The TSCNN with using the apex frame identified in Section II.A.

[c]TSCNN-II -The TSCNN with using the apex frame given by CASME database.

(LTOGP(with FS) [82], LTOGP(without FS) [82], FDM [54], DiSTLBP-RIP [83], and STCLQP [53]), our method exhibits an improvement of 2.09%, 9.66%, 14.59%, 6.4%, and 13.42% in accuracy. Thus, our TSCNN-I yields better recognition and outperforms other methods, especially in the absence of index values given by the databases.

We also calculate the confusion matrix for each of the five databases to determine the recognition of the TSCNN for each emotion label, as shown in Fig. 9. For the CASME-II database, the TSCNN yielded an improved recognition, especially on the ''surprise'' and ''others'' labels. However, the method still encountered a bottleneck with the ''repression'' label because repression emotions have a relatively small range of muscle motion and is thus more difficult to detect and classify correctly. For the SMIC-HS database, our TSCNN yielded an improved recognition result for all labels. For the SAMM database, the network also performed well on two labels (anger and others) but did not perform well on the ''contempt'' and ''surprise'' labels. This result agrees with our expectations because all poorly performing labels have a small sample size, which hinders such deep-learning methods. For the CAS(ME)^2 database, the network performed well on two labels (anger and disgust) but did not perform well on the ''happy'' labels. For the CASME database, the TSCNN performed well on three labels (disgust, surprise, and tense) but did not perform well on the ''Repression'' labels. The above results show that for MER tasks, the recognition of our TSCNN is superior to that of the image feature extraction method used by most previous researchers.

TABLE 7. Experimental results on CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME databases under different division scheme.

Database	Block pattern	Net structure	Protocal	Accuracy(%)	F1-score
CASME II	2×2	L	LOSO	52.13	0.5064
		L+S	LOSO	54.68	0.5238
		L+T	LOSO	67.34	0.6249
		L+S+T	LOSO	71.92	0.7162
	3×3	L	LOSO	50.04	0.4612
		L+S	LOSO	59.23	0.5750
		L+T	LOSO	70.20	0.6876
		L+S+T	**LOSO**	**74.05**	**0.7327**
	4×4	L	LOSO	48.56	0.4729
		L+S	LOSO	57.93	0.5306
		L+T	LOSO	62.15	0.5817
		L+S+T	LOSO	68.77	0.6546
SMIC-HS	2×2	L	LOSO	50.29	0.4798
		L+S	LOSO	51.93	0.4987
		L+T	LOSO	66.32	0.6572
		L+S+T	LOSO	69.71	0.6713
	3×3	L	LOSO	58.60	0.5480
		L+S	LOSO	49.21	0.4858
		L+T	LOSO	70.00	0.6981
		L+S+T	**LOSO**	**72.74**	**0.7236**
	4×4	L	LOSO	47.63	0.4296
		L+S	LOSO	50.82	0.4762
		L+T	LOSO	64.89	0.5991
		L+S+T	LOSO	67.13	0.6653
SAMM	2×2	L	LOSO	55.26	0.4708
		L+S	LOSO	52.01	0.4192
		L+T	LOSO	60.64	0.4991
		L+S+T	LOSO	61.51	0.5062
	3×3	L	LOSO	55.37	0.4727
		L+S	LOSO	51.18	0.4321
		L+T	LOSO	53.68	0.4755
		L+S+T	**LOSO**	**63.53**	**0.6065**
	4×4	L	LOSO	51.32	0.3818
		L+S	LOSO	54.63	0.4676
		L+T	LOSO	55.47	0.4552
		L+S+T	LOSO	59.35	0.5267
CAS(ME)^2	2×2	L	LOSO	60.79	0.5900
		L+S	LOSO	64.88	0.6365
		L+T	LOSO	69.44	0.6852
		L+S+T	LOSO	70.35	0.7020
	3×3	L	LOSO	65.00	0.6423
		L+S	LOSO	64.74	0.6316
		L+T	LOSO	70.62	0.7027
		L+S+T	**LOSO**	**71.62**	**0.7129**
	4×4	L	LOSO	60.53	0.5886
		L+S	LOSO	61.88	0.5993
		L+T	LOSO	62.38	0.6097
		L+S+T	LOSO	69.29	0.6865
CASME	2×2	L	LOSO	63.82	0.5808
		L+S	LOSO	60.67	0.5216
		L+T	LOSO	64.72	0.6309
		L+S+T	LOSO	68.26	0.6435
	3×3	L	LOSO	63.31	0.5946
		L+S	LOSO	60.17	0.5357
		L+T	LOSO	67.92	0.6385
		L+S+T	**LOSO**	**70.73**	**0.6736**
	4×4	L	LOSO	61.29	0.5405
		L+S	LOSO	60.84	0.5149
		L+T	LOSO	63.93	0.5981
		L+S+T	LOSO	68.43	0.6232

D. PARAMETER ANALYSIS

In this subsection, we analyze the parameters of the proposed methods and evaluate the impact of these parameters individually. The block pattern of the input image for the local-spatial stream, the number of the network's recognition stream, and the effect of λ on the TSCNN are reported and

TABLE 8. The influence of parameter λ on TSCNN model.

	CASME-II		SAMM		SMIC-HS		CAS(ME)^2		CASME	
λ	Accuracy(%)	F1-score	Accuracy(%)	F1-score	Accuracy(%)	F1-score	Accuracy(%)	F1-score	Accuracy(%)	F1-score
10	72.55	0.7231	61.91	0.5757	70.06	0.6994	70.76	0.7050	69.83	0.6666
15	72.51	0.7197	61.69	0.5702	70.55	0.7031	71.06	0.7077	70.22	0.6695
20	72.83	0.7237	63.16	0.5884	72.38	0.7221	71.47	0.7115	70.73	0.6736
25	74.05	0.7327	62.43	0.5807	71.77	0.7129	71.62	0.7129	70.45	0.6713
30	71.94	0.7091	61.61	0.5794	71.22	0.7098	71.26	0.7096	70.06	0.6681

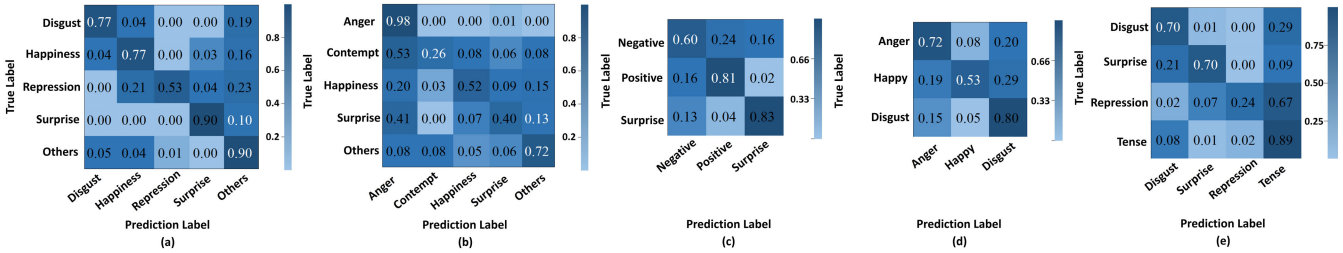


FIGURE 9. Confusion matrices on CASME II, SAMM, SMIC-HS, CAS(ME)^2, and CASME. The (a), (b), (c), (d), and (e) denote the results of CASME II, SAMM, SMIC-HS, CAS(ME)^2, and CASME, respectively.

discussed in this section. The following experimental results are obtained with the assumption that the five databases do not provide apex frame indices.

1) EVALUATION OF DIFFERENT BLOCK PATTERNS

The facial block sequence served as the input of the local spatial recognition stream in the TSCNN is different when we choose spatial grids with multiple sizes (2×2, 3×3, and 4×4) to divide the gray image of the apex frame. To test which block pattern is optimal, we compared the performance of the TSCNN under three above cases. Experiment results are shown in Table 7, which shows that the TSCNN yields the best results (74.05% for the CASME II database, 72.74% for the SMIC-HS database, 63.53% for the SAMM database, 71.62% for the CAS(ME)^2 database, and 70.73% for the CASME database) when the gray image of the apex frame is divided into 3×3 image blocks and used as the input of the local-spatial recognition stream in the TSCNN.

2) EVALUATION OF THE TSCNN ARCHITECTURE

To analyze our network's structure in depth and find the most prominent module, we compare the results between the TSCNN with that of the network that retains two recognition streams and that of the network that only retains a single stream. We set the block pattern equal to 3×3 for the local-spatial stream.

Results using five databases are shown in Table 9. An accuracy of 74.05% using the CASME II database is achieved by the TSCNN (L(3×3)+S+T), which is higher than all single-stream networks (S: 60.08%, L(3×3): 50.04% and T: 71.53%) and outperforms all two-stream networks (L(3×3)+S: 59.23%, L(3×3)+T: 70.20% and S+T: 73.20%). Our TSCNN also shows better

TABLE 9. Experimental results on CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME databases under different combination of recognition streams.

Database	Net structure	Protocal	Accuracy(%)	F1-score
CASME II	S	LOSO	60.08	0.5867
	L	LOSO	50.04	0.4612
	T	LOSO	71.53	0.7135
	L+S	LOSO	59.23	0.5750
	L+T	LOSO	70.20	0.6876
	S+T	LOSO	73.20	0.7292
	L+S+T	**LOSO**	**74.05**	**0.7327**
SMIC-HS	S	LOSO	49.63	0.4824
	L	LOSO	58.60	0.5480
	T	LOSO	71.52	0.7146
	L+S	LOSO	49.21	0.4858
	L+T	LOSO	70.00	0.6981
	S+T	LOSO	68.90	0.6878
	L+S+T	**LOSO**	**72.74**	**0.7236**
SAMM	S	LOSO	48.38	0.3741
	L	LOSO	55.37	0.4727
	T	LOSO	60.66	0.5723
	L+S	LOSO	51.18	0.4312
	L+T	LOSO	53.68	0.4755
	S+T	LOSO	56.62	0.5078
	L+S+T	**LOSO**	**63.53**	**0.6065**
CAS(ME)^2	S	LOSO	65.56	0.7435
	L	LOSO	65.00	0.6423
	T	LOSO	70.32	0.6949
	L+S	LOSO	64.74	0.6316
	L+T	LOSO	70.62	0.7027
	S+T	LOSO	70.03	0.6908
	L+S+T	**LOSO**	**71.62**	**0.7129**
CASME	S	LOSO	60.22	0.5301
	L	LOSO	63.31	0.5946
	T	LOSO	66.57	0.6577
	L+S	LOSO	60.17	0.5357
	L+T	LOSO	67.92	0.6385
	S+T	LOSO	67.47	0.6454
	L+S+T	**LOSO**	**70.73**	**0.6736**

performance than the single-stream or two-stream networks when using the SMIC-HS, SAMM, CAS(ME)^2, and CASME databases. Thus, our proposed TSCNN yields the best performance in both accuracy and F_1-score (CASME II: 74.05%/0.7327, SMIC-HS: 72.74%/0.7236, SAMM: 63.53%/0.6065, CAS(ME)^2: 71.62%/0.7129, and CASME: 70.73%/0.6736).

The performance of the TSCNN is significantly better than those of single-stream and two-stream networks, particularly for the dynamic-temporal stream. These results agree with our assumptions, because the calculated image of the optical flow field can describe the two-dimensional projection of an ME movement intuitively and make it easy to distinguish ME emotion categories. Additionally, the results show that the three streams of our TSCNN can better utilize various forms of effective characteristics for ME recognition, yielding better performance for MER tasks than single characteristics.

3) THE IMPACT OF PARAMETER λ ON THE TSCNN

In this subsection, we analyze how λ (see Section II.B) affects the proposed TSCNN model for MER tasks. Its value is evaluated using the CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME databases. Specifically, we change the value of λ to observe the recognition results of the TSCNN with the five databases, as shown in Table 8. The MER accuracy is shown to be stable even when λ varies within a given range. As shown in Table 8, we can see that the occurrence of MEs is a process of gradual change in facial expression intensity. If the apex frame located by our method falls on adjacent frames of the real apex frame, the classification performance of the TSCNN in MER tasks is stable and satisfactory when these location frames are applied to the TSCNN.

IV. CONCLUSION

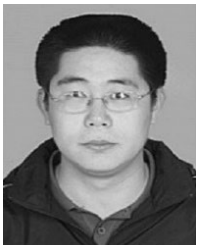
In this paper, we propose a three-stream convolutional neural network (TSCNN) for ME recognition. Experiments are conducted on five public spontaneous ME databases, (CASME II, SMIC-HS, SAMM, CAS(ME)^2, and CASME) to evaluate the proposed method. The experimental results show that our method can effectively improve recognition accuracy in MER tasks compared with the results of other methods using the same five databases. Additionally, this paper also summarizes the problems that have not received sufficient attention in research to date but are crucial for feasible MER interpretations. Incorporating static-spatial, local-spatial and temporal information associated with MEs is shown to be important when describing MEs and aides distinguishing MEs. In our method, the dynamic-temporal recognition stream plays a critical role and, depends on the calculation of optical flow. However, this calculation has a high computational cost and thus must occur offline; this is the key bottleneck to the application of this method. In the future, we plan to study faster optical flow calculation methods to facilitate using the proposed method in real-time identification. Additionally, we plan to design a simpler network structure with multiple recognition tubes to handle ME details and use different datasets of spontaneous MEs with various kinds of metrics.

REFERENCES

[1] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers Psychol.*, vol. 8, p. 1745, Oct. 2017.
[2] 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.
[3] P. Ekman, M. O'Sullivan, and M. G. Frank, "A few can catch a liar," *Psychol. Sci.*, vol. 10, no. 3, pp. 263–266, 1999.
[4] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *Proc. Annu. Meeting Int. Commun. Assoc. Sheraton*, New York, NY, USA, 2009.
[5] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "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./Dec. 2017.
[6] M. G. Frank, C. J. Maccario, and V. Govindaraju, "Behavior and security," in *Protecting Airline Passengers in the Age of Terrorism*. Santa Barbara, CA, USA: ABC-CLIO, LLC, 2009, pp. 86–106.
[7] M. O'Sullivan, M. G. Frank, C. M. Hurley, and J. Tiwana, "Police lie detection accuracy: The effect of lie scenario," *Law Hum. Behav.*, vol. 33, no. 6, p. 530, 2009.
[8] 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.
[9] G. Zhao and M. Pietikäinen, "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, Jan. 2007.
[10] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognit. Lett.*, vol. 33, no. 16, pp. 2181–2191, 2012.
[11] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: Databases, methods, and challenges," *Frontiers Psychol.*, vol. 9, p. 1128, Jul. 2018.
[12] R. Péteri and D. Chetverikov, "Dynamic texture recognition using normal flow and texture regularity," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.* Estoril, Portugal: Springer, Jun. 2005, pp. 223–230.
[13] D. Chetverikov and R. Péteri, "A brief survey of dynamic texture description and recognition," *Proc. Int. Conf. mboxCORES Comput. Recognit. Syst. 2*. Springer, 2005, pp. 17–26.
[14] R. C. Nelson and R. Polana, "Qualitative recognition of motion using temporal texture," *CVGIP, Image Understand.*, vol. 56, no. 1, pp. 78–89, 1992.
[15] P. Bouthemy and R. Fablet, "Motion characterization from temporal cooccurrences of local motion-based measures for video indexing," in *Proc. 14th Int. Conf. Pattern Recognit.*, vol. 1, 1998, pp. 905–908.
[16] 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.
[17] H. Zheng, X. Geng, and Z. Yang, "A relaxed K-SVD algorithm for spontaneous micro-expression recognition," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Springer, 2016, pp. 692–699.
[18] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 382–386.
[19] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 667–674.
[20] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," 2019, *arXiv:1901.04656*. [Online]. Available: https://arxiv.org/abs/1901.04656
[21] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," 2019, *arXiv:1904.01390*. [Online]. Available: https://arxiv.org/abs/1904.01390
[22] R. Zhi, H. Xu, M. Wan, and T. Li, "Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," *IEICE Trans. Inf. Syst.*, vol. 102, no. 5, pp. 1054–1064, 2019.

[23] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 36–40.
[24] Y. 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.
[25] 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.
[26] K. Li, Y. Zong, B. Song, J. Zhu, J. Shi, W. Zheng, and L. Zhao, "Three-stream convolutional neural network for micro-expression recognition," in *Proc. 26th Int. Conf. Neural Inf. Process. (ICONIP)*, 2019.
[27] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?" in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3094–3098.
[28] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, "SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos," *IEEE Access*, vol. 6, pp. 71143–71151, 2018.
[29] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. 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, Oct./Dec. 2015.
[30] 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.
[31] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[32] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.
[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001, vol. 1, nos. 511–518, p. 3.
[34] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[35] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 509–516.
[36] S. E. Kahou, P. Froumenty, and C. Pal, "Facial expression analysis based on high dimensional binary features," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, Sep. 2014, pp. 135–147.
[37] R. Fablet and P. Bouthemy, "Motion recognition using spatio-temporal random walks in sequence of 2D motion-related measurements," in *Proc. Int. Conf. Image Process.*, vol. 3, 2001, pp. 652–655.
[38] R. Fablet and P. Bouthemy, "Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1619–1624, Dec. 2003.
[39] Z. Lu, W. Xie, J. Pei, and J. Huang, "Dynamic texture recognition by spatio-temporal multiresolution histograms," in *Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION)*, vols. 1–2, Jan. 2005, pp. 241–246.
[40] C. Peh and L. Cheong, "Exploring video content in extended spatio-temporal textures," in *Proc. 1st Eur. Workshop Content-Based Multimedia Indexing*, 1999, pp. 147–153.
[41] C.-H. Peh and L.-F. Cheong, "Synergizing spatial and temporal texture," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1179–1191, Oct. 2002.
[42] R. Péteri and D. Chetverikov, "Qualitative characterization of dynamic textures for video retrieval," in *Computer Vision and Graphics*. Warsaw, Poland: Springer, 2006, pp. 33–38.
[43] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.
[44] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 435–442.
[45] 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, 2014, Art. no. e86041.
[46] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "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.
[47] 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./Mar. 2016.
[48] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 424–436, Oct./Dec. 2017.
[49] 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.
[50] A. C. Le Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Proc. Asian Conf. Comput. Vis.* Singapore: Springer, 2014, pp. 33–48.
[51] A. C. Le Ngo, J. See, and R. C.-W. Phan, "Sparsity in dynamics of spontaneous subtle emotions: Analysis and application," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 396–411, Jul./Sep. 2017.
[52] 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.
[53] X. Huang and G. Zhao, "Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern," in *Proc. Int. Conf. Frontiers Adv. Data Sci. (FADS)*, 2017, pp. 159–164.
[54] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, Apr./Jun. 2017.
[55] Y.-H. Oh, A. C. Le Ngo, J. See, S.-T. Liong, R. C.-W. Phan, and H.-C. Ling, "Monogenic Riesz wavelet representation for micro-expression recognition," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 1237–1241.
[56] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition," *PLoS ONE*, vol. 10, no. 5, 2015, Art. no. e0124674.
[57] S. Y. Park, S. H. Lee, and Y. M. Ro, "Subtle facial expression recognition using adaptive magnification of discriminative facial motion," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 911–914.
[58] S.-T. Liong, J. See, R. C.-W. Phan, A. C. Le Ngo, Y.-H. Oh, and K. Wong, "Subtle expression recognition using optical strain weighted features," in *Proc. Asian Conf. Comput. Vis.* Singapore: Springer, Nov. 2014, pp. 644–657.
[59] S.-T. Liong, R. C.-W. Phan, J. See, Y.-H. Oh, and K. Wong, "Optical strain based recognition of subtle emotions," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, 2014, pp. 180–184.
[60] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, and M. Pietikäinen, "Capturing correlations of local features for image representation," *Neurocomputing*, vol. 184, pp. 99–106, Apr. 2016.
[61] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Proc. Asian Conf. Comput. Vis.* Springer, 2014, pp. 525–537.
[62] J. Päivärinta, E. Rahtu, and J. Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Proc. Scand. Conf. Image Anal.* Ystad, Sweden: Springer, May 2011, pp. 360–369.
[63] H. Lu, K. Kpalma, and J. Ronsin, "Motion descriptors for micro-expression recognition," *Signal Process., Image Commun.*, vol. 67, pp. 108–117, Sep. 2018.
[64] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection," 2016, *arXiv:1608.02255*. [Online]. Available: https://arxiv.org/abs/1608.02255
[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[67] S.-T. Liong and K. Wong, "Micro-expression recognition using apex frame with phase information," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2017, pp. 534–537.
[68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[69] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50X fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360
[70] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 34, no. 3, pp. 1588–1595, Jun. 2004.

[71] Q. Li, J. Yu, T. Kurihara, and S. Zhan, "Micro-expression analysis by fusing deep convolutional neural network and optical flow," in *Proc. 5th Int. Conf. Control, Decis. Inf. Technol. (CoDIT)*, 2018, pp. 265–270.
[72] N. Van 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.
[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[74] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, pp. 1331–1339, Nov. 2018.
[75] V. Mayya, R. M. Pai, and M. M. Pai, "Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences," in *Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, 2016, pp. 699–703.
[76] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image steganalysis," *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 10437–10453, 2018.
[77] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "Learnet dynamic imaging network for micro expression recognition," 2019, *arXiv:1904.09410*. [Online]. Available: https://arxiv.org/abs/1904.09410
[78] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Spatiotemporal local monogenic binary patterns for facial expression recognition," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 243–246, May 2012.
[79] C. H. Chan, B. Goswami, J. Kittler, and W. Christmas, "Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 602–612, Apr. 2012.
[80] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 868–875.
[81] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. VS-PETS*, Beijing, China, Oct. 2005, pp. 65–72.
[82] M. Niu, J. Tao, Y. Li, J. Huang, and Z. Lian, "Discriminative video representation with temporal order for micro-expression recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2112–2116.
[83] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 32–47, Jan./Mar. 2017.
[84] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, Jan. 2008.
[85] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Automatic micro-expression recognition from long video using a single spotted apex," in *Proc. Asian Conf. Comput. Vis.* Taipei, Taiwan: Springer, Nov. 2016, pp. 345–360.
[86] X. Li, J. Yu, and S. Zhan, "Spontaneous facial micro-expression detection based on deep learning," in *Proc. IEEE 13th Int. Conf. Signal Process. (ICSP)*, Nov. 2016, pp. 1130–1134.



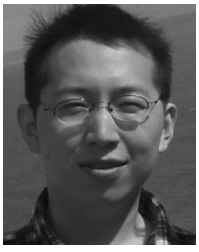
BAOLIN SONG received the B.S. and M.S. degrees from Qingdao University (QDU) and Shandong Normal University (SNU), in 2010 and 2014, respectively. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Southeast University, China. His research interests include computer vision, pattern recognition, and micro-expression recognition.

KE LI received the B.S. degree in information engineering from Southeast University, Nanjing, China, in 2019, where he is currently pursuing the M.E. degree with the Engineering Centre of Signal and Information Processing, School of Information Science and Engineering. His research interests include image processing, affective computing, and pattern recognition.

YUAN ZONG (M'18) received the B.S. and M.S. degrees in electronics engineering from Nanjing Normal University, Nanjing, China, in 2011 and 2014, respectively, and the Ph.D. degree in medical engineering from Southeast University, Nanjing, in 2019. From 2016 to 2017, he was a Visiting Student with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland. He is currently a Lecturer with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University. His research interests include affective computing, pattern recognition, and computer vision.

JIE ZHU received the B.S. degree in communication engineering from Hohai University, Nanjing, China, in 2018. She is currently pursuing the Ph.D. degree with the Key Laboratory of Signal and Information Processing, School of Information Science and Engineering, Southeast University, Nanjing. Her research interests include speech recognition and computer vision.

WENMING ZHENG (SM'18) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004. Since 2004, he has been with the Research Center for Learning Science, Southeast University. He is currently a Professor with the School of Biological Science and Medical Engineering and the Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University. His research interests include neural computation, pattern recognition, machine learning, and computer vision. He is an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, *Neurocomputing*, and *The Visual Computer*.



JINGANG SHI received the B.S. and Ph.D. degrees from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. Since 2017, he has been a Post-Doctoral Researcher with the Center for Machine Vision Research and Signal Analysis, University of Oulu, Finland. His current research interests mainly include image restoration and face analysis.

LI ZHAO received the B.E. degree from the Nanjing University of Aeronautics and Astronautics, China, in 1982, the M.S. degree from Suzhou University, China, in 1988, and the Ph.D. degree from the Kyoto Institute of Technology, Japan, in 1998. He is currently a Professor with Southeast University, China. His research interests include speech signal processing and pattern recognition.