Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-expressions

Zhaoqiang Xia, *Member, IEEE,* Xiaopeng Hong, *Member, IEEE,* Xingyu Gao, *Member, IEEE,* Xiaoyi Feng, Guoying Zhao, *Senior Member, IEEE*

***Abstract*—Recently, the recognition task of spontaneous facial micro-expressions has attracted much attention with its various real-world applications. Plenty of handcrafted or learned features have been employed for a variety of classifiers and achieved promising performances for recognizing micro-expressions. However, the micro-expression recognition is still challenging due to the subtle spatiotemporal changes of micro-expressions. To exploit the merits of deep learning, we propose a novel deep recurrent convolutional networks based micro-expression recognition approach, capturing the spatiotemporal deformations of micro-expression sequence. Specifically, the proposed deep model is constituted of several recurrent convolutional layers for extracting visual features and a classificatory layer for recognition. It is optimized by an end-to-end manner and obviates manual feature design. To handle sequential data, we exploit two ways to extend the connectivity of convolutional networks across temporal domain, in which the spatiotemporal deformations are modeled in views of facial appearance and geometry separately. Besides, to overcome the shortcomings of limited and imbalanced training samples, two temporal data augmentation strategies as well as a balanced loss are jointly used for our deep network. By performing the experiments on three spontaneous micro-expression datasets, we verify the effectiveness of our proposed micro-expression recognition approach compared to the state-of-the-art methods.**

***Index Terms*—Micro-Expression Recognition, Spatiotemporal Modeling, Temporal Connectivity, Recurrent Convolutional Networks, Data Augmentation, Balanced Loss**

I. Introduction

MICRO-EXPRESSIONS are very brief and involuntary facial expressions which are shown on the facial regions of humans. Compared to long-duration and obvious changes of normal facial expressions (namely macro-expressions), micro-expressions usually have short duration, i.e., less than 0.2 second, and subtle intensity changes as facial muscle movements caused by micro-expressions only emerge in small and few regions [1], [2]. The spontaneous micro-expressions can reveal the genuine emotions of humans and help understand humans' deceitful behaviors. Thus, it is potential to apply the micro-expressions in diverse fields [3], such as lie detection, police case diagnosis, business negotiation, and psychoanalysis. Whereas, short duration and subtle changes of micro-expressions make it difficult for untrained people to detect and analyze micro-expressions. Even trained by professional micro-expression training tools [4], humans still manually detect and recognize micro-expressions from videos with low accuracy. Consequently, the automatic micro-expression recognition (MER) will be very valuable to promote the performance of analyzing large amounts of video sequences.

To tackle the MER task, several approaches have been presented to model subtle changes of micro-expressions in spatiotemporal domain [5]. Most of these approaches are roughly divided into two main parts. The first part is to extract visual features from facial video clips towards the task of MER. The second part is to choose a classifier for extracted features. Since MER is a typical pattern recognition task, some conventional classifiers, e.g., support vector machine (SVM) [6], [7], [8], [9], [10] and random forest [6], [11], [12], [5], have been used while the feature designing becomes more pivotal for solving MER problem in recent researches. Consequently, plenty of handcrafted features for macro-expressions or new-designed features have been explored in the past decade. For instance, the local binary patterns on three orthogonal planes (LBP-TOP) widely used to describe dynamic textures are firstly applied to recognize micro-expressions [6]. Although LBP-TOP has shown the capacity of discriminability and efficiency, it still suffers the sensitivity problem of global changes. So the second-order Gaussian jet on LBP-TOP [13], LBP six intersection points (LBP-SIP) [14], local spatiotemporal directional features (LSDF) [15], spatiotemporal LBP (STLBP) [8], spatiotemporal completed local quantization patterns (STCLQP) [9], directional mean optical-flow (MDMO) [16], discriminative spatiotemporal LBP (DSLBP) [17] and bi-weighted oriented optical flow (Bi-WOOF) [10] are proposed to improve the robustness of visual descriptors. These handcrafted features are designed to capture temporal differences of micro-expression sequences and achieve an accuracy of more than 50% [5].

However, it is still challenging to extract useful information from subtle changes and achieve high-quality descriptions as handcrafted features cannot well capture the subtle deformations of micro-expressions. Recently, deep convolutional neural networks (CNNs) have shown the great power in various fields and outperformed the handcrafted features as well as shallow classifiers [18], [19], [20], [21], [22]. Deep learning approaches can obviate manual feature design and allow to automatically connect a specific task to the features themselves. Nevertheless, few deep models have been devoted

Z. Xia and X. Feng are with School of Electronics and Information, Northwestern Polytechnical University, 710129 Shaanxi, China. e-mail:xiazhaoqiang@gmail.com, fengxiao@nwpu.edu.cn.

X. Hong is with Xi'an Jiaotong University, Xi'an, P. R. China. e-mail: hongxiaopeng@mail.xjtu.edu.cn.

X. Gao is with Chinese Academy of Sciences, Beijing, China. e-mail: gxy9910@gmail.com.

G. Zhao is with Center for Machine Vision and Signal Analysis, University of Oulu, 90014 Oulu, Finland. e-mail:guoying.zhao@oulu.fi

to the MER problem due to limited video-based (sequence) training samples. Spontaneous micro-expression datasets usually contain insufficient samples, for instance, merely 256 micro-expression sequences for all categories in the largest CASME II dataset [23]. And they also have unbalanced classes, e.g., 26 sequences for category "Happiness" and 99 sequences for category "Other" in CASME II dataset. The limited and imbalanced samples will restrain deep CNNs as the deep network usually needs to learn numerous parameters. For alleviating this problem, the pre-trained CNN [24] has been fine-tuned to recognize image based micro-expressions, in which each image (video frame) is assigned with a micro-expression category. The image based approach can obtain sufficient training samples by using video frames individually (rather than the entire video sequence) while the temporal changes are not considered. In order to leverage limited video-based samples for CNN based deep models, the temporal connectivity of CNNs which consider spatial and temporal changes jointly becomes vitally important for MER problem.

In this paper, we propose spatiotemporal recurrent convolutional networks (STRCN) to automatically recognize micro-expressions by conquering the "limited and imbalanced training samples" problem. To model the spatiotemporal motion deformations, we propose to employ CNNs with recurrent connections (i.e., recurrent convolutional networks) to learn the representation of subtle changes. The convolutional layers with recurrent connections are utilized to learn visual features automatically and a classificatory layer is used to recognize micro-expressions. Towards the micro-expression video frames (clips), we exploit two types of connectivities of STRCN across temporal domain for the network input. Moreover, to facilitate the learning procedure of deep model, we propose two temporal augmentation strategies to greatly enrich the training samples for learning deep model and employ a balanced loss for counterweighing imbalanced classes.

Our main contributions are summarized as follows:

- We propose an STRCN model to explore the powerful representation ability of deep convolutional networks for MER problem by considering the spatiotemporal deformations. To the best of our knowledge, it is the first time that the deep model can be trained from scratch and outperform existing algorithms in MER.
- We propose two ways to extend the connectivity of STRCN across temporal domain for video based micro-expressions, in which not only spatial information but also temporal changes are jointly considered.
- We design two temporal augmentation strategies to greatly enrich the limited training samples for deep framework and employ a balanced loss for facilitating the imbalanced training.

The rest of this paper is organized as follows. Section II reviews the related work briefly and our proposed deep framework for MER is presented in Section III. Then we discuss the experimental results for algorithm evaluation in Section IV. Finally, Section V concludes this work.

II. Related Work

In this section, the researches on micro-expression analysis and deep learning for modeling spatiotemporal information are briefly summarized. The techniques for micro-expression analysis are described to indicate the shifted focus of research community while the deep learning studies are presented briefly to demonstrate the related techniques which can be used for MER problem.

A. Micro-Expression Analysis

To date, some micro-expression datasets (Polikovsky's dataset [25], USF-HD [26], SMIC [27], CASME [28], CASME II [23], CASME^2 [29] and SAMM [30]) have been published in the literatures. Among them, the Polikovskys dataset [25] and USF-HD dataset [26] were constructed by collecting acted micro-expressions of subjects with high-speed cameras and are not available publicly. The remaining databases (SMIC, CASME, CASME II, CASME^2 and SAMM) are spontaneous and obtained by collecting the induced micro-expressions of subjects watching specific videos. The CASME II is the extended version of CASME while CASME^2 focuses on both micro-expressions and macro-expressions in long videos. Thus, in this paper, we study spontaneous MER problem using these three representative micro-expression datasets (i.e., SMIC [27], CASME II [23], and SAMM [30]).

The task of spontaneous micro-expression analysis contains two subtasks: *detection* and *recognition*. The detection task is fundamental to subsequent recognition based on well-segmented video sequences containing micro-expressions while the recognition task aims to distinguish small differences between various kinds of micro-expressions. For the detection task, the geometric features [31], [32], [33] and local textures [34], [35] have been proposed to capture micro-expression frames from videos. For the recognition task, several approaches [3] have been presented by using various features and classifiers, which are discussed in the following.

Some primary studies have been devoted to micro-expression recognition [5]. In the earlier stage, scholars attempted to recognize acted micro-expressions. Polikovsky *et al.* [25] proposed a recognition method with descriptors of gradient orientation histogram on Polikovsky's dataset. In [26], authors presented the strain patterns for detecting macro-expressions and micro-expressions through the acted USF-HD dataset. However, the acted micro-expressions are greatly different from the spontaneous facial expressions [36]. Therefore, recent works have further been done on the spontaneous facial micro-expressions, which can be roughly categorized as appearance based and geometric based methods.

1) Appearance Based Methods: In recent studies, some appearance based features have been utilized to recognize micro-expressions. Pfister *et al.* presented a recognition algorithm combing LBP-TOP features with SVMs, multiple kernel classifiers or random forests to recognize negative or positive micro-expressions [6], in which the LBP-TOP features calculate the LBP features from three orthogonal planes and concatenate their histograms. Wang *et al.* extended the LBP-TOP features into the tensor independent color space and

then recognized micro-expressions in this subspace [37], [7]. However, the local textures described by LBP-TOP have the problems of robustness and sparse sampling [14], [9]. Thus, some extended methods [14], [15], [8], [9], [17] have further been proposed to improve the recognition performance. LBP-SIP provided a lightweight representation based on three intersecting lines crossing over the center point of LBP-TOP [14] and trained an SVM classifier. In [8], [9], the STCLBP and STCLQP features extended LBP-TOP features by containing more information (i.e., magnitude and orientation). Their hierarchical versions were reported in [38] and achieved better performance by considering multiple blocks of LBP. In LSDF [15], regions of interest (ROIs) were used to extract LBP-TOP features and further calculate local directional features, which encode the sign feature with magnitude information as weights. Furthermore, DSLBP [17] extracted 1D LBP and LBP-TOP combining with an integral projection and incorporated shape attributes into spatiotemporal texture features.

2) Geometric Based Methods: Geometric based methods extract deformation information from local landmarks or optical flow fields of facial regions without considering facial textures. In [39], facial feature points have been tracked and used to recognize specific micro-expressions (i.e., happiness and disgust). Furthermore, the Delaunay triangulation based on extracted landmarks was used to reveal subtle muscle movements [40] and encoded temporally for dynamical micro-expressions. Besides, based on optical flow estimation, some approaches leverage the magnitude, orientation and other high-order statistics to model the dynamics of micro-expressions. The MDMO [16] features generated histograms from ROIs by counting the oriented optical flow vectors and then those histograms were used to recognize micro-expressions. Facial dynamics map (FDM) features calculated the optical flow direction based on the spatiotemporal cuboids and then were used to characterize a micro-expression sequence [41]. In Bi-WOOF [10], the orientation histogram in each block of facial regions was generated by considering the magnitude and optical strain values as the weights. The weighted histograms were further used for recognizing micro-expressions.

In the earlier conference version of this work [42], the recurrent convolutional networks have been used to recognize micro-expressions. Compared to the handcrafted features, the proposed deep neural network model [42] can capture the subtle temporal changes automatically and obtain good recognition results by simultaneously learning a classifier. Compared to the conference version, this work promotes in three aspects: 1) the appearance based connectivity called STRCN-A in conference version is promoted by a new dimension reduction method with generating masks and selecting the micro-expression-aware areas; 2) this work proposes a new geometric based connectivity method called STRCN-G for modeling the spatiotemporal deformation of micro-expressions; 3) two effective data augmentation strategies and one sample-weighted loss function are presented to benefit the learning of deep models for the challenge of limited and unbalanced samples.

B. Spatiotemporal Deep Learning

In recent years, CNNs have achieved great successes in many computer vision tasks and advanced the state-of-the-art accuracy of image/object recognition [43], [44]. The architecture of CNNs is a type of artificial neural networks and has a purely feed-forward architecture characterized by local connections and weight sharing. Amounts of deep learning approaches based on CNNs have been developed to pursue more excellent performances by using millions of training samples. However, the conventional CNNs only capture the spatial variations without considering temporal information.

To model spatiotemporal variations, several deep approaches have exploited various architectures to be adaptive for video/sequence based data in many fields [45], such as video classification [46], [47], [48] and action recognition [49], [50], [51], to cite only few. In [49], the 3D convolutional filters were used to extract temporal information in every seven frames for action recognition. Similar to [49], the size of filters on the first convolutional layer in the early fusion of [46] was modified to some temporal extent, for example, $11\times11\times3\times10$, which can deal with 10 frames for one time. Similar idea has also been presented in [47], which can deal with full video frames at one time. In the late fusion of [46], two frames separately went through the same networks and merged in the first fully connected layer. These two ways can extract spatiotemporal information progressively in a dense sampling way. Different from the dense sampling strategy [49], [46], [47], a sparse temporal sampling strategy in temporal segment network (TSN) [50] was used to segment long-range video for enabling effective learning using the whole video. And more variants of TSN have been proposed to fuse mid/high-level features [51]. Besides, the recurrent neural network (RNN) based encoder-decoder models [52], [48], [53] combined with CNNs were introduced to learn compact representation for videos. However, the CNN, e.g., VGG16, as an encoder and RNN, e.g., long short-term memory (LSTM) and gated recurrent unit (GRU), as an decoder were usually trained separately and cannot obtain good representations jointly for a specific task. More deep models based on aforementioned architectures have been proposed for modeling spatiotemporal information. These approaches have been proven to be effective for long-range videos and can well capture the significant changes. Whereas, it might not be effective and needs to explore more specific models for the task containing subtle changes, such as micro-expression recognition.

More specifically, the feed-forward architectures can only capture the context in higher layers of CNNs where the units have larger receptive fields and fail to modulate the activities of units in lower layers. It means that only one-scale temporal variation can be captured by CNN based approaches. Inspired by neuroscience, the recurrent connections are adopted to obtain larger receptive fields and reduce the number of learnable parameters. The larger receptive fields are helpful to utilize the context information in lower layers of CNNs, and less parameters are more suitable for tasks without large amounts of samples [54]. In [55] and [54], two different types of convolutional layers with recurrent

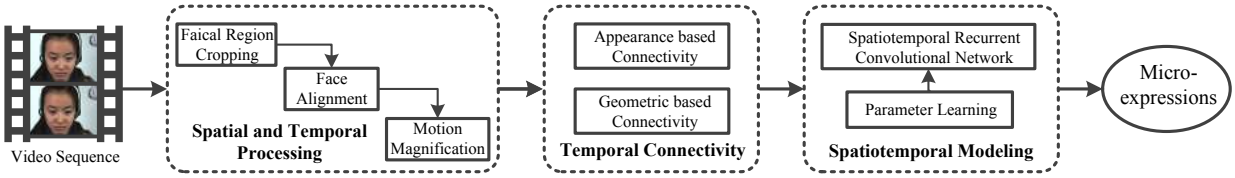


Fig. 1: The framework of our proposed approach for micro-expression recognition.

connections were proposed to label scenes. These recurrent convolutional networks were then applied to model significant changes for pain intensity regression based on static images [56]. In this paper, we propose new spatiotemporal recurrent convolutional networks (STRCN) to model subtle changes for limited-range videos/sequences. The sampling strategies for long-range videos are not suitable for limited-range videos. To address this problem, we propose two types of connectivities of STRCN across temporal domain for jointly modeling the spatial and temporal deformations.

III. Proposed Method

In this section, we will present our proposed deep model (i.e., spatiotemporal recurrent convolutional networks) for recognizing micro-expressions. To explain the details of our proposed STRCN method, total five parts will be successively introduced, including the framework, spatial and temporal processing, temporal connectivity, spatiotemporal modeling and parameter learning. Specifically, our contributions are mainly threefold. Firstly, we apply the RCN in the framework by slightly changing its architecture and learn the deep model without any pre-trained models. Secondly, two ways of temporal connectivies, i.e., appearance based and geometric based connectivity, are proposed to address the problem of sequential network input. In this part, the selection method of micro-expression-aware areas and detection method of apex frames are proposed to generate the network input, respectively. Thirdly, to address the problem of using deep learning techniques for small-size sample problem, we propose a data augmentation method for greatly enriching the limited samples for RCNs in the part of parameter learning.

A. The Framework

The framework of our proposed method is shown in Fig. 1. Our proposed approach contains three key procedures, i.e., spatial and temporal processing, temporal connectivity and spatiotemporal modeling, for recognizing micro-expressions. For training deep models with limited and unbalanced samples, the parameter learning in training procedure is a key component of our proposed approach.

Firstly, we process video sequences of micro-expressions spatially and temporally. In spatial processing, the facial regions are cropped and aligned for each video (image sequence), resulting in the removal of non-facial regions. Then, in temporal processing, the motion deformations of facial regions are augmented to enhance subtle changes of micro-expressions. This will be introduced in Section III-B and a processing example is shown in Section IV-F. Secondly, two types of temporal connectivities are proposed to feed the sequential input into the subsequent deep model. The first type (denoted as **STRCN-A**) is an appearance based way while the second one (denoted as **STRCN-G**) is a geometric based way. This will be introduced in Section III-C. At last, the spatiotemporal modeling with deep recurrent convolutional networks is used to recognize micro-expressions. The architectures and detailed setup will be introduced in Section III-D and the parameter learning of deep models will be introduced in Section III-E.

B. Spatial and Temporal Processing

1) Spatial Processing: To avoid the pollution of regions without containing micro-expressions, facial regions need to be cropped and aligned from image sequences. In this context, the facial landmark points are detected firstly and then used to crop and align facial regions.

We employ an eye detector [57] and active shape model (ASM) algorithm [32] to detect landmark points. The eye detector [57] can accurately locate two centering points of eye regions and determine the starting positions for face shapes described by ASM. And then the accurate locations of face shapes are iteratively fitted by the ASM algorithm [32]. Following [6], we use 68 landmark points to crop the facial regions. Given the centering points of eyes in frame i, i.e., (x_i^l, y_i^l) for left eye and (x_i^r, y_i^r) for right eye, the cropping coordinates of facial regions are calculated as

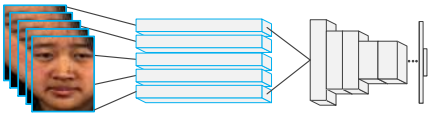
$$\begin{aligned} topleft &= ((x_i^l, y_i^l)) + \delta_1(0, y_i^l - y_i^r) - \delta_2(x_i^r - x_i^l, 0) \\ height &= \delta_3\sqrt{(x_i^l - x_i^r)^2 + (y_i^l - y_i^r)^2} \\ width &= \delta_4\sqrt{(x_i^l - x_i^r)^2 + (y_i^l - y_i^r)^2} \end{aligned} \tag{1}$$

Based on the 68 landmark points, a local weighted mean (LWM) transformation [6] of any frame for sequence i is used for aligning cropped facial regions. The transformed value of an arbitrary point is set to

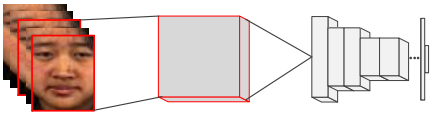
$$f(x,y) = \frac{\sum_{i=1}^{N} W(\sqrt{(x-x_i)^2 + (y-y_i)^2}/D_n) S_i(x,y)}{\sum_{i=1}^{N} W(\sqrt{(x-x_i)^2 + (y-y_i)^2}/D_n)} \tag{2}$$

where W is the weight, D_n is the distance of control point (x_i, y_i) from its $(n-1)$th nearest control point, and $S_i(x,y)$ is the polynomial with n parameters passing through a measure for (x_i, y_i) and $n-1$ other measurements nearest to it. Using LWM transformation, all images in one sequence can be aligned frame by frame.

2) Temporal Processing: The temporal changes of aligned facial regions are very small and (almost) impossible to see with naked eyes of humans. Moreover, it is difficult to automatically learn representations of these subtle changes from



(a) Type 1: STRCN-A



(b) Type 2: STRCN-G

Fig. 2: The schematic diagram of temporal connectivity. (a) Type 1: Appearance based connectivity (STRCN-A); (b) Type 2: Geometric based connectivity (STRCN-G).

noisy content by machine learning techniques. In this context, we utilize the motion magnification technique to amplify the hidden motion information of adjacent frames.

In this context, we choose the Eulerian Video Magnification (EVM) method [58] to amplify the temporal motion. The magnified temporal motion can be calculated by

$$\tilde{I}(x,y,t) = f(x,y) + \sum_{k}(1+\alpha_k)\delta_k(t)(\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}) \tag{3}$$

where $f(x,y) = I(x,y,0)$ and $I(x,y,t)$ denotes the image intensity at position (x,y) and time t. $\delta_k(t)$ is a displacement function and can be obtained by the temporal bandpass filter with respect to the kth frequency. α_k is a frequency-dependent motion magnification factor. $\tilde{I}(x,y,t)$ is the image intensity of tth frame after magnified.

In order to amplify the subtle changes of facial sequences, an infinite impulse response (IIR) filter is chosen as our temporal filter. In other words, we only use one temporal filter $\delta(t)$ ($k=1$) as the bandpass filter. After the temporal filtering, all images of each band are amplified with a fixed magnification factor α_k. Finally, all bands of Laplacian pyramid are used to reconstruct the motion-magnified facial sequences.

C. Temporal Connectivity

We investigate two types of temporal connectivities for fusing information across temporal domain. Type 1 concatenates all images in one sequence by vectorizing one channel of an image into a column of matrix. The spatiotemporal variations are reserved in a composite matrix and then appearance features can be learned by STRCN. So we call it as appearance based connectivity (abbreviated as **STRCN-A**). In type 2, the geometric motion is described by optical flow fields and then used to learn spatiotemporal features by STRCN. So we call it as geometric based connectivity (abbreviated as **STRCN-G**). The schematic diagram of these two types of temporal connectivities is shown in Fig. 2.

1) Type 1: Appearance based Connectivity: To obtain the appropriate input for STRCN, all pixels of an entire image can be directly vectorized into a column of one matrix, however, this way may induce the lengthy size of matrix column. In STRCN-A, we propose a better way to reserve spatiotemporal information in one matrix. Usually, not all pixels in facial regions are helpful for learning representations of micro-expressions. Observed from the magnified video sequences, micro-expressions are usually the fragments of normal expressions and only occur in particular areas of face, such as eyes, brows and mouth, which are called as *micro-expression-aware areas* in this paper. Other facial areas, e.g., chin and hairs, do not reveal micro-expressions, which are *micro-expression-unaware areas*. In addition, the short-duration micro-expressions do not involve wide-range facial changes. So, in the context, we choose to eliminate the effects of micro-expression-unaware facial areas and select out the micro-expression-aware ones. In other words, some partial regions sensitive to micro-expressions, rather than an entire image, are selected out to be concatenated into a matrix. In order to obtain micro-expression-aware areas, we propose a mask generation method and choose those areas with the generated mask, which has an example in Fig. 8(a). The mask is generated by thresholding a difference heat map, which is calculated by accumulating temporal differences of video frames on entire datasets. In the following paragraph, we will introduce how to calculate the heat map and threshold it.

The difference heat map $E(x,y)$ of magnified video frames is calculated as follows:

$$\begin{aligned} E(x,y) &= \sum_{i}\sum_{t} d_i(x,y,t) \\ d_i(x,y,t) &= |\tilde{I}_i(x,y,t) - \tilde{I}_i(x,y,0)| \end{aligned} \tag{4}$$

where $\tilde{I}_i(x,y,t)$ denotes the magnified image intensity at spatial position x,y and time t for ith facial sequence. $d_i(x,y,t)$ represents the difference of two frames with temporal interval t and is accumulated to generate the heat map $E(x,y)$. Through accumulating temporal differences, the micro-expression-aware facial areas become active in the difference heat map. We further design an efficient thresholding strategy to generate the mask. Firstly, we sort all values of difference heat map in a descending order. Then the top $p\%$ percentiles of sorted values are chosen and others are abandoned. The binary mask is finally generated by setting the chosen values to 1 and others to 0, which has an example in Fig. 8(a).

With the binary mask, the micro-expression-aware areas can be selected out and continue to be flattened to a tensor. Supposed that the magnified data $\tilde{I}(x,y,t) \in R^{d_1 \times T}$ ($d_1 = W \times H$), the data after selection is denoted as $\hat{I}(x,y,t) \in R^{d_2 \times T}$ (d_2 equals to the number of pixels in selected areas). Usually, the dimensionality d_2 after selection is greatly smaller than the original dimensionality d_1, i.e., $d_2 \ll d_1$. So, the entire video in STRCN-A can be denoted as a tensor $\mathcal{V}_1 \in \mathbb{R}^{d_2 \times T \times C}$, where C represents the image channels.

2) Type 2: Geometric based Connectivity: In this type (STRCN-G), we investigate the motion information extracted from the entire video sequence by optical flow method. The dense sampling method used in [49], [46], [47] can model the motion variations of micro-expressions, however, it computes inefficiently and still need to further fuse the dense sampling. As the micro-expression clips are different from common video clips having multiple scenes and usually have one scene,

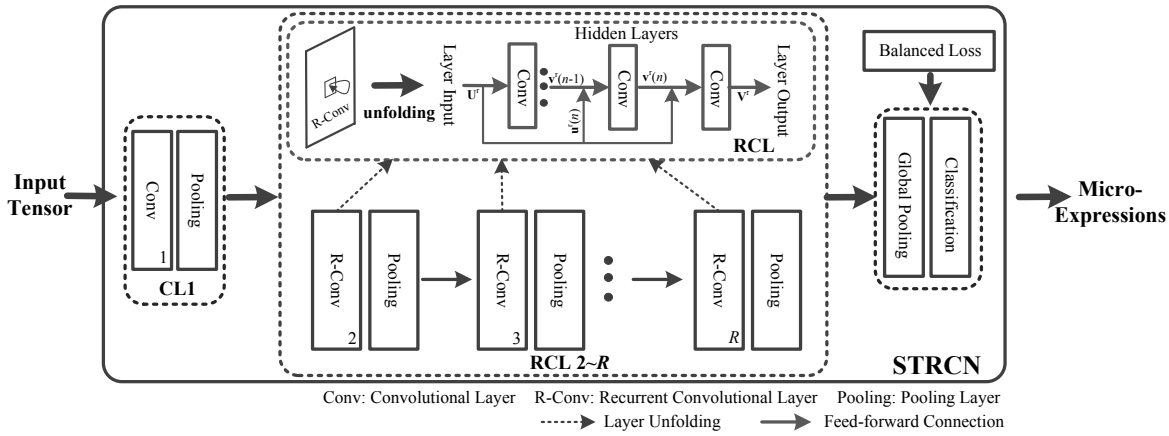


Fig. 3: The architecture of deep RCNs for micro-expression recognition and the detailed configuration is shown in Table I.

we choose to use the sparse sampling to compute the optical flows. Inspired by [59], [10], only the apex and onset frames are used to compute the optical flow images. Then the optical flow images are fed into STRCN for modeling geometric deformations.

The onset frame is defined as the first frame of video sequence containing micro-expressions while the apex frame is the frame containing the strongest-intensity expressions, which have the greatest changes from the first frame. To capture the greatest changes, we propose to use the difference deviation for roughly locating the apex frame. The index t_a of apex frame can be calculated as

$$\begin{aligned} t_a &= \arg\max_t \sigma_i(t) \\ \sigma_i(t) &= std\big(\tilde{I}_i(x,y,t) - \tilde{I}_i(x,y,0)\big) \end{aligned} \tag{5}$$

where $std(\cdot)$ computes the standard deviation of input matrix.

With the sparse sampling frames (i.e., onset and apex frames), we employ the accurate optical flow method [60] for optical flow approximation. To compute more efficiently, the Lorentzian penalty function with improved model is chosen as the estimation method. The estimated displacement vector at position (x, y) for sequence i is denoted as $(\mathbf{u}_{x,y}, \mathbf{v}_{x,y})$. So the entire video in STRCN-G can be denoted as a tensor $\mathcal{V}_2 \in \mathbb{R}^{W\times H\times 2}$.

D. Spatiotemporal Modeling

Compared to the handcrafted features, CNNs have more powerful ability to describe subtle changes of micro-expressions. In this paper, we add recurrent connections (i.e., RCNs [54]) within the feed-forward convolutional layers by using multiple-scale receptive fields. Based on two types of temporal connectivities, the spatiotemporal information of sequences can be further extracted by RCNs.

The architecture of our deep RCNs is shown in Fig. 3. It contains one feed-forward convolutional layer (CL) and several recurrent convolutional layers (RCLs). The layer 1 ($CL1$) is the only feed-forward convolutional layer without recurrent connections and used to compute efficiently. Following the standard convolutional layer 1, several RCLs ($RCL2 \sim R$) are employed to extract visual features for recognition task. Between each convolutional layers (feed-forward and recurrent), max pooling operations are adopted to reduce dimensionality and save computation. Following the RCLs, a global average pooling layer is adopted to concatenate all feature maps into a vector. In the last layer, the classificatory layer with Softmax function is employed to calculate the recognition probabilities with concatenated feature vector.

With the tensor $\mathcal{V}$ of network input, the convolutional layer $CL1$ with a pooling layer outputs the feature maps denoted by V^c. In each RCL layer, several hidden layers are used to expand the size of receptive fields and one RCL layer can be unfolded into several convolutional layers. The layer latter in the subnetwork has larger receptive field in the same RCL layer. r represents the index of RCL layers ($r = 2, \cdots, R$) and n denotes the depth index of rth RCL layer, i.e., the index of hidden convolutional layers. For every convolutional layer, fixed-size feature maps are used to obtain the consistent connections. The input of an unit located at (i, j) on the kth feature map in rth RCL layer can be computed as

$$z^r_{ijk}(n) = \mathbf{w}^{f^T}\mathbf{u}^r_{ij}(n) + \mathbf{w}^{r^T}_k\mathbf{v}^r_{ij}(n-1) + b_k \tag{6}$$

where $\mathbf{u}^r_{ij}(n)$ and $\mathbf{v}^r_{ij}(n-1)$ represent the feed-forward and recurrent input, respectively. In the equation, $\mathbf{w}^f_k$ and $\mathbf{w}^r_k$ denote the feed-forward and recurrent weight vectors for kth feature map. b_k is the bias of kth feature map. The output of an unit located at (i, j) on the kth feature map is given by

$$v^r_{ijk}(n) = f(z^r_{ijk}(n)) \tag{7}$$

where $n = 0, 1, \cdots, N$ and the initial state $v^r_{ijk}(0) = v^{r-1}_{ijk}(N)$. $f(\cdot)$ represents the normalized activation function. In a summary, the rth RCL layer can transform a tensor $U^r \in \mathbb{R}^{W'\times H'\times K'}$ of layer input into a tensor $V^r \in \mathbb{R}^{W''\times H''\times K''}$. $K(K', K'')$ represents the number of feature maps in rth RCL layer. After R RCL layers, the output V^c of $CL1$ is transformed to a tensor V^R.

Finally, the output of deep network uses the Softmax function to classify the C-categories micro-expressions based

on the feature vector from the global pooling layer. The probability of micro-expressions can be calculated as

$$\mathbf{p}_c = \frac{exp(\mathbf{W}^T\mathbf{v})}{\sum_{c=1}^{C} exp(\mathbf{W}^T\mathbf{v})} \tag{8}$$

where $\mathbf{p}_c$ is the predicted probability vector of all categories, $\mathbf{v}$ denotes the output feature vector of global pooling layer, which computes from V^R, and $\mathbf{W}$ denotes the weight matrix between last global pooling layer and classification layer.

E. Parameter Learning

The parameter learning can be performed by minimizing the cross entropy loss function using the back propagation through time (BPTT) algorithm [54]. However, it is noted that two challenging issues for deep learning methods exist in current micro-expression datasets, which will make the parameter learning ineffective. The first one is having slightly imbalanced classes while the second one is having limited samples. Take CASME II [23] for example. The category of "Other" have almost 4 times more samples than "happiness". Besides, no more than 250 original samples in CASME II [23] can be used to train deep RCNs. This will cause the problem of over-fitting and limit the recognition performance. To address these two problems, we employ two operations, namely, multi-class balanced loss and multi-scale data augmentation to train the deep RCNs.

We extend the binary balanced loss [61] into MER problem with multiple categories. We define a quantity $\mathbf{p}_t$ as

$$\mathbf{p}_t = \mathbf{p}_c^{\mathbf{y}}(1-\mathbf{p}_c)^{1-\mathbf{y}} \tag{9}$$

where $\mathbf{y}$ is the label vector and its arbitrary element $y_i \in \{0,1\}$. Then the multi-class balanced loss is computed as follows

$$\mathcal{L} = \sum_i \Big(-\beta_i log(\mathbf{p}_t^i) \Big) \tag{10}$$

where β_i is the weighting factor of class for sample i and inversely proportional to the sample's class ratio in batch data. With the balanced loss function, the imbalance classes are balanced by β_i.

On the other side, we propose two multi-scale data augmentation strategies to enrich training samples and further restrain the problem of small-size samples. Firstly, we use multiple-scale amplification factors α_k for training samples. All categories use multi-scale amplification factors $\alpha_k = [5, 14]$. With different factors, the size of samples are augmented by 10 times. Secondly, some frames are randomly selected out from one sequence with a percentage. Totally, five levels of percentages are adopted for random selection, i.e., 100%, 90%, 80%, 70%, and 60%. So, with five random selections, the data can be augmented by 5 times and these data contain different-sized sequences. Performing these two strategies jointly, the original data can be augmented by 50 times. These augmented data can make it more sufficient for training deep architecture.

TABLE I: The detailed configuration of our deep STRCNs.

Layers	Configurations	
	STRCN-A	STRCN-G
Input	Tensor:$d_1 \times T \times 3$	Tensor: $W \times H \times 2$
Conv1	$k: 5 \times 5$, $p: 0$, $s: 1 \times 1$	
Pool1	MAX, $k: 4 \times 1$, $s: 4 \times 1$	MAX, $k: 4 \times 4$, $s: 4 \times 4$
RCL2	1 $feed\text{-}forward$: $k: 1 \times 1$, $p: 0$, $s: 1 \times 1$ 3 $recurrents$: $k: 3 \times 3$, $p: 1 \times 1$, $s: 1 \times 1$	
Pool2	MAX, $k: 4 \times 1$, $s: 4 \times 1$	MAX, $k: 4 \times 4$, $s: 4 \times 4$
RCL3	1 $feed\text{-}forward$: $k: 1 \times 1$, $p: 0$, $s: 1 \times 1$ 3 $recurrents$: $k: 3 \times 3$, $p: 1 \times 1$, $s: 1 \times 1$	
Pool3	MAX, $k: 4 \times 4$, $s: 4 \times 4$	MAX, $k: 4 \times 4$, $s: 4 \times 4$
RCL4	1 $feed\text{-}forward$: $k: 1 \times 1$, $p: 0$, $s: 1 \times 1$ 3 $recurrents$: $k: 3 \times 3$, $p: 1 \times 1$, $s: 1 \times 1$	
Pool4	MAX, $k: 2 \times 2$, $s: 2 \times 2$	MAX, $k: 2 \times 2$, $s: 2 \times 2$
RCL5	1 $feed\text{-}forward$: $k: 1 \times 1$, $p: 0$, $s: 1 \times 1$ 3 $recurrents$: $k: 3 \times 3$, $p: 1 \times 1$, $s: 1 \times 1$	
Pool5	Global AVG with alterable size	
Output	C categories	

All the convolutional layers contain M feature maps.
k - filter or pooling size, p - padding size, s - stride size.

IV. Experiments

In this section, we present the experimental details, including the implementation details, the datasets we used, the protocols, the approaches for comparison and experimental results.

A. Implementation Details

In the spatial processing procedure, the constants $\delta_1, \delta_2, \delta_3, \delta_4$ for cropping facial regions are set to $\{0.4, 0.6, 2.2, 1.8\}$, following [6]. For the temporal processing, the cut-off frequencies (filter interval) of IIR filter are chosen as $[0.05, 0.4]Hz$. And the magnification factor in testing procedure is fixed to $\alpha = 8$. According to [58], the bound for factor α in any frame is adopted as $\alpha_c = \frac{\lambda}{8\delta(t)} - 1$. λ denotes the spatial wavelength and is set to $\lambda = 16$ in this context. Therefore, the magnification factor can be finally used as $\alpha = min(8, \alpha_c)$.

In temporal connectivity procedure, the facial images are resized to fixed sizes for subsequent procedures. The fixed sizes are set to 64×48 for STRCN-A (type 1) and 300×245 for STRCN-G (type 2). For the mask generation in STRCN-A, considering the trade-off between the dimensionality reduction and information preserving, we choose the percentile value $p = 30\%$ as the threshold, which is further investigated in Section IV-E. The corresponding value of 70% percentile of maximum is denoted as E_p so that the positions with values $E(x) > E_p$ are selected as active elements. In the next step of STRCN-A, the temporal normalization operations are performed to obtain fixed-size input tensor for RCNs. Here, we choose 30 frames to feed the deep model. For STRCN-G, the fixed-size (300×245) frames are used to generate optical flow map and then fed to RCNs.

The detailed configurations of RCNs are shown in Table I. Two types of temporal connectivities share the same architecture of RCNs but have different parameter setups. For each convolutional layer (feed-forward and recurrent), the batch normalization is used for scaling the activation and then a rectified linear unit is followed as the activation

function. Since many parameters in our deep architecture may affect the performance of micro-expression recognition, we fix some parameters (e.g., filter sizes and stride size) with prior values in [54], [56]. For other some important parameters (e.g., filter interval of IIR filter, number of feature maps and recurrent layers), we explore their values by grid searching and investigate their impact in Section IV-E.

For learning parameters, the momentum is set to 0.9 and weight decay 0.0005 in stochastic gradient decent (SGD) procedure of BPTT. The stopping criterion for SGD is set to 10^{-3} for iterations. The learning rate is set to 10^{-3} in the beginning and will be multiplied with damping factor 0.8 when all mini-batches are traversed and re-allocated randomly. To accelerate the parameter learning, we employ the library MatConvNet [62] to accomplish our proposed model. The mini-batch size for training model is set to 20 as it is limited by the memory of GPUs (One Geforce TiTan X).

B. Micro-expression Datasets and Setups

Three representative micro-expression datasets are used to evaluate the performance of our proposed approach in the experiments: SMIC dataset [27], CASME II dataset [23] and SAMM [30]. All of them are specially designed to detect and recognize spontaneous micro-expressions, which are constructed by inducing subjects' micro-expressions. These three corpora have following characteristics:

- The SMIC dataset contains 164 spontaneous micro-expressions from 16 subjects. It is recorded by 100 fps high-speed cameras. These participants undergo high emotional arousal and suppress their facial expressions in an interrogation room setting with a punishment threat and highly emotional clips.
- The CASME II dataset has 256 micro-expressions from 26 subjects. It has higher video quality and larger image size compared with SMIC. The recording rate of cameras in CASME II is 200 fps. Thus the video sequences of micro-expressions in CASME II have more frames than SMIC.
- The SAMM dataset contains 159 micro-expressions from 29 subjects at 200 fps. It uses similar procedures like CASME II but has a higher image resolution and employs an array of LEDs to avoid flickering. Because of the creators with professional rating skills, these expressions are obtained from stricter lab situations and labeled more accurately. And it has a wide ethnicity compared to other datasets.

To keep three datasets consistent with each other, we merge seven more categories in CASME II and SAMM into four classes. Following [37], [16], [38], the happy micro-expressions in CASME II and SAMM are classified into "Positive" class as they indicate good emotions of subjects. In contrast, the disgust, sadness, fear, contempt and anger micro-expressions are classified into "Negative" class as they are usually considered as bad emotions. Surprise usually occurs when there is a difference between expectations and reality and can be neutral/moderate, pleasant, unpleasant, positive, or negative. Tense and repression are classified into the "Other" class as they indicate the ambiguous feelings of subjects and require further inference. In SMIC dataset, only the first three classes (i.e., positive, negative and surprise) are used to annotate the micro-expressions.

For all experiments on three datasets, both *leave-one-subject-out* (LOSO) and *leave-one-video-out* (LOVO) protocols are used to evaluate our proposed methods. To save the training time of deep models, we leave 5% videos for testing, which will reduce the testing time. Based on two protocols, we use both accuracy and F1-score to evaluate the performance of our proposed STRCNs (STRCN-A and STRCN-G), avoiding the impact of imbalanced class problem for three datasets. Assume that TP, FP and FN are the true positive, false positive and false negative, respectively. The accuracy is calculated as $Acc = \frac{TP}{N}$ where N is the number of testing samples. The F1-score is computed as $F = 2\times \frac{P\times R}{P+R}$, where $P = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FP_i)}$ and $R = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FN_i)}$.

C. Method Comparison in LOSO Protocol

Since the subject-independent evaluation protocol, i.e., LOSO, is becoming the main-stream for evaluating MER problem, we report the comparison results of our two deep models (STRCN-A and STRCN-G) in LOSO protocol with all state-of-the-art approaches, including the conventional appearance based methods [6], [14], [7], [8], [38], geometric based methods [16], [41], [10] and deep methods [24], [] in Table II.

1) Comparison Results to Appearance Based Methods: The appearance based methods are based on LBP and have LBP-TOP [6], LBP-SIP [14], LBP-TICS [7], STLBP-IP [8] and Hierarchical STLBP-IP [38]. The LBP parameters of these methods are set as $R = 3$ and $P = 8$, which achieve the best performance in all configurations [14], [8], [38]. Except Hierarchical STLBP-IP, all LBP-based methods employ SVMs as classifiers, in which the polynomial kernel $k(x_i, x_j) = (\gamma_1 x_i^T x_j + \gamma_2)^{\gamma_3}$ is used and the optimal values are set to $\gamma_1 = 0.22, \gamma_2 = 0, \gamma_3 = 2$. All these methods are implemented according to their descriptions and retrained over three datasets. In contrast, the results with the optimal parameter set of kernelized group sparse learning are obtained from [38] for hierarchical STLBP-IP and only reported on two datasets, i.e., SMIC and CASME II.

Compared with the results of appearance based methods in Table II, our proposed methods (STRCN-A and STRCN-G) achieve better results than them in most configurations and datasets. Especial for STRCN-G, its results are better than all the other appearance based methods (including STRCN-A) under the LOSO protocol while STRCN-A achieves better performance only in LOVO protocol. In LOSO protocol, both our proposed appearance based method (STRCN-A) and image-based CNN cannot achieve better performance than hierarchical STLBP-IP. That might be because the appearance based deep models would be affected by the intra-class variations of each subject (person) as these deep models would learn certain appearance from each subject's samples.

2) Comparison Results to Geometric Based Methods: The geometric based methods mainly utilize the optical flow fields

TABLE II: The recognition accuracy and F1-score of different methods under the **LOSO** protocol on three datasets.

Approaches	SMIC		CASME II		SAMM	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
LBP-TOP [6]	0.457	0.461	0.409	0.369	0.415	0.406
LBP-SIP [14]	0.421	0.422	0.457	0.425	0.417	0.402
LBP-TICS [7]	0.439	0.384	0.405	0.378	0.395	0.374
STLBP-IP [8]	0.543	0.547	0.551	0.497	0.568	0.527
Hierarchical STLBP-IP [38]	0.601	0.613	0.638	0.611	$N\backslash A$	$N\backslash A$
MDMO [16]	0.615	0.406	0.510	0.418	$N\backslash A$	$N\backslash A$
FDM [41]	0.714	0.540	0.417	0.297	$N\backslash A$	$N\backslash A$
Bi-WOOF[10]	0.593	0.620	0.589	0.610	0.598	0.591
Image-based CNN [24]	0.312	0.305	0.444	0.428	0.436	0.429
CNN-LSTM [48]	0.376	0.357	0.482	0.455	0.448	0.437
CNN-GRU [52]	0.385	0.368	0.493	0.467	0.452	0.441
STRCN-A(Ours)	0.531	0.514	0.560	0.542	0.545	0.492
STRCN-G(Ours)	**0.723**	**0.695**	**0.803**	**0.747**	**0.786**	**0.741**

*$N\backslash A$ - no results reported.

to obtain the geometric information of facial movements and have MDMO [16], FDM [41] and Bi-WOOF [10]. The LOSO results originally from MDMO [16] and FDM [41] are used directly while the Bi-WOOF is implemented mainly according to [10]. In Bi-WOOF, the block size for local weight and optical strain is set to 8×8. Then SVM classifiers are retrained on three datasets with the same parameters of LBP-based methods. Different from [10], the optical flow estimation method [60] used in STRCN-G is also used to implement Bi-WOOF.

Compared with the results of geometric based methods in Table II, our proposed geometric based method (STRCN-G) achieves better results than them in all configurations and datasets. Merely in SMIC dataset, FDM reports the accuracy near to our STRCN-G while FDM achieves very poor performance in CASME II dataset. For other approaches, STRCN-G achieves very promising performance improvement on all datasets. On the other side, in LOSO protocol, our proposed appearance based method (STRCN-A) only achieves better performance in some configurations and datasets. For instance, in SMIC dataset, all geometric based approaches achieve better accuracy and F1-score than the proposed STRCN-A under the LOSO protocol. Furthermore, the geometric based methods achieve better performance than almost all appearance based methods, not just our proposed STRCN-A. Compared to appearance based methods, geometric based methods may eliminate the intra-class information of each subject as only geometric information of subjects are reserved.

3) Comparison Results to Deep Methods: The image-based CNN [24] recognizes the category of each frame in a micro-expression sequence. To compare with our proposed deep models fairly, we recognize one sequence as the category c_i when half frames in the sequence are recognized as the category c_i. Strictly following the image-based CNN, the VGGFace architecture is utilized to pre-train the CNN model[1] [62] and then fine-tuned in MER datasets. In [24], the temporal changes are not utilized for recognizing micro-expressions. Besides, the LSTM [48] and GRU [52] combined with CNN (denoted as CNN-LSTM and CNN-GRU) are also used as baseline methods. Both CNN-LSTM and CNN-GRU use the VGGFace model [2] to extract the visual features and then train RNNs (LSTM and GRU) for extracting temporal information. The CNN cannot be trained with LSTM and GRU jointly in CNN-LSTM and CNN-GRU for the task of MER as the network input of RNN uses sequential data while the CNN input needs one individual image. In contrast, the temporal changes are modeled by two types of connectivities in our proposed methods and can be trained from scratch for MER.

Observed from Table II, our proposed methods outperform the image-based CNN, CNN-LSTM and CNN-GRU obviously under the LOSO protocol. Even for the handcrafted features, the image based CNN cannot outperform them as those manually-designed features consider the temporal information of sequences. And merely using the pre-trained model for CNN-LSTM and CNN-GRU limits the representation ability of CNNs for MER. So it is indicated that the joint spatiotemporal modeling of deep models and its fine-tuning are prerequisite for MER problem.

D. Method Comparison in LOVO Protocol

To complement the LOSO protocol, we also report the comparison results under the LOVO protocol. We compare our proposed methods with several conventional and deep methods on three datasets, including LBP-TOP [6], LBP-SIP [14], LBP-TICS [7], STLBP-IP [7] Bi-WOOF [10], image-based CNN [63], CNN-LSTM [48] and CNN-GRU [52]. All these comparison methods are compared on three datasets and the performances are reported in Table III. Among these state-of-the-art approaches, the variants of LBP method are based on appearance features while Bi-WOOF is a geometric based method. To compare deep models in LOVO protocol, we also compare our proposed method with deep models (image based CNN [24], CNN-LSTM [48] and CNN-GRU [52]). Since we use 5% of all samples, rather than only one sample, as the testing sample for each evaluation, all methods in Table III are reevaluated over three datasets with implementing their algorithms, rather than citing their reported results directly. The parameters of all methods are same to the configurations in LOSO protocol.

[1] http://www.vlfeat.org/matconvnet/pretrained/

[2] https://github.com/dangz90/Deep-Learning-for-Expression-Recognition-in-Image-Sequences

TABLE III: The recognition accuracy and F1-score of different methods under the **LOVO** protocol on three datasets.

Approaches	SMIC		CASME II		SAMM	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
LBP-TOP [6]	0.618	0.577	0.591	0.537	0.593	0.542
LBP-SIP [14]	0.645	0.598	0.657	0.622	0.645	0.592
LBP-TICS [7]	0.637	0.586	0.643	0.607	0.647	0.604
STLBP-IP [8]	0.701	0.669	0.723	0.677	0.712	0.663
Bi-WOOF [10]	0.745	0.706	0.725	0.689	0.718	0.672
Image-based CNN [24]	0.531	0.476	0.564	0.507	0.572	0.518
CNN-LSTM [48]	0.559	0.478	0.581	0.534	0.569	0.521
CNN-GRU [52]	0.562	0.509	0.599	0.541	0.583	0.536
STRCN-A(Ours)	**0.758**	**0.714**	**0.841**	0.784	**0.836**	**0.792**
STRCN-G(Ours)	0.749	0.710	0.833	**0.807**	0.827	0.781

Observed from Table III, our proposed methods (STRCN-A and STRCN-G) achieve the best performance in all configurations under the LOVO protocol. Compared to the handcrafted features, the deep features from automatic learning in our proposed method are competitive to other methods under the LOVO protocol. We can see that LBP-based features cannot outperform other methods as they are more suitable for the description of obvious changes of macro-expressions and the subtle changes can still not be well captured. Based on optical flow maps, the Bi-WOOF feature [10] extracts subtle changes of micro-expressions but has limited ability of describing micro-expressions. Our proposed methods can obtain the descriptions of subtle changes and outperform these state-of-the-art methods.

Moreover, the evaluation under the LOVO protocol is easier than LOSO protocol as all subjects can occur in training samples. Compared results in Table II and III, our proposed methods, i.e., STRCN-A and STRCN-G, achieve different performances in both protocols. Especially, STRCN-A achieve better performance than STRCN-G in most configurations under the LOVO protocol. As STRCN-A is an appearance based method, which learns certain information of subjects, it can outperform STRCN-G in LOVO protocol while achieving poor performance in LOSO protocol.

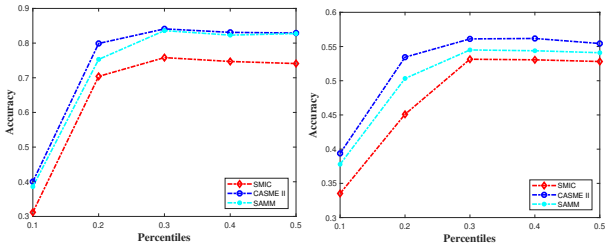
E. Parameter Analysis

In this section, we analyze the parameters of the proposed methods and evaluate the impact of these parameters individually. Totally, the accuracy performances of the data augmentation, the balanced loss, the percentage threshold p, the size of feature maps, the number of recurrent layers, the interval of IIR filter and the training size are reported and discussed in this context.

1) The Impact of Data Augmentation and Balanced Loss: To observe the impact of temporal data augmentation and balanced loss, we replace the proposed methods with removing data augmentation (DA) procedure and balanced loss (BL). For removing the procedure of temporal data augmentation (DA), only the original samples are used by keeping model parameters consistent to proposed models. To fairly compare with final deep models (STRCN-A and STRCN-G), the same iterations are used in all deep models. For using imbalanced loss, we set all weights of samples to $\beta = 1$ for removing the balancing weights. These methods are denoted as **STRCN-A without DA**, **STRCN-G without DA**, **STRCN-A without BL** and **STRCN-G without BL** in Table IV.

TABLE IV: The recognition accuracy of different methods without using data augmentation and balanced loss under the LOVO and LOSO protocol on three datasets.

LOVO			
Approaches	SMIC	CASME II	SAMM
STRCN-A without DA	0.583	0.593	0.617
STRCN-G without DA	0.605	0.605	0.641
STRCN-A without BL	0.737	0.829	0.831
STRCN-G without BL	0.741	0.818	0.822
STRCN-A (Ours)	**0.758**	**0.841**	**0.836**
STRCN-G (Ours)	0.749	0.833	0.827
LOSO			
Approaches	SMIC	CASME II	SAMM
STRCN-A without DA	0.481	0.471	0.488
STRCN-G without DA	0.576	0.621	0.642
STRCN-A without BL	0.541	0.557	0.538
STRCN-G without BL	0.718	0.802	0.775
STRCN-A (Ours)	0.531	0.560	0.545
STRCN-G (Ours)	**0.723**	**0.803**	**0.786**

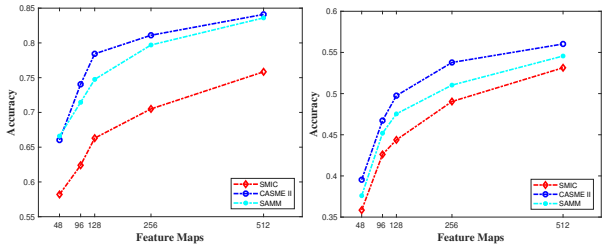


(a) LOVO (b) LOSO

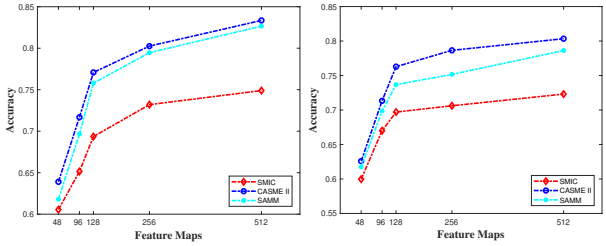
Fig. 4: The accuracy performance of various percentiles p by STRCN-A on three datasets.

The accuracy of proposed methods without DA and BL are reported in Table IV. From the results, we can see that the temporal data augmentation can improve the performance of both two proposed deep models in both protocols. Obviously, the recognition ability of deep STRCN can be enhanced by leveraging more training samples. On the other side, the balanced loss can improve the performance slightly with considering the imbalanced classes. Overall, more sufficient training samples can promote the performances of deep models.

2) The Impact of Percentiles: The values of p in STRCN-A are evaluated by the accuracy of all datasets in both LOVO

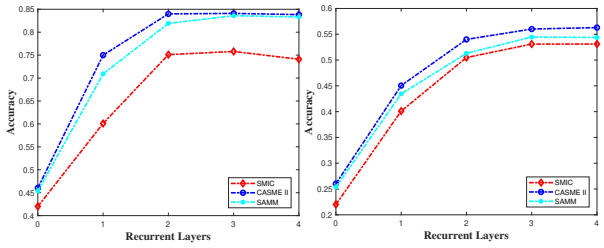


(a) LOVO, STRCN-A (b) LOSO, STRCN-A

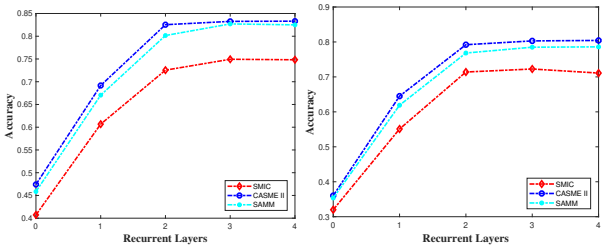


(c) LOVO, STRCN-G (d) LOSO, STRCN-G

Fig. 5: The accuracy performance with different feature maps by Type-1 and Type-2 connections on three datasets.



(a) LOVO, STRCN-A (b) LOSO, STRCN-A



(c) LOVO, STRCN-G (d) LOSO, STRCN-G

Fig. 6: The accuracy performance with different recurrent layers by STRCN-A and STRCN-G on three datasets.

and LOSO protocols, which are shown in Fig. 4. According to the results, the performance will be degraded when too many pixels are abandoned by using lower values of p. In contrast, extra pixels may not be helpful for recognizing micro-expressions as some regions may contain noises. So the micro-expression-aware areas can be optimally determined by choosing appropriate values of p.

3) The Impact of Feature Maps: Figure 5 shows the results of LOVO and LOSO evaluation for different feature maps on the SMIC, CASME II and SAMM datasets. It is noted that the performance can be steadily improved with more features maps on all datasets. More feature maps in each recurrent layer have more representation ability for extracting spatiotemporal information. However, the improvement will become less with the increase of feature maps. Besides, more feature maps occupy more memory and need more time to learn.

4) The Impact of Recurrent Layers: Different recurrent layers are evaluated on the SMIC, CASME II and SAMM datasets in Fig. 6. According to Table I, the RCL2 $\sim$ 5 can be removed in sequence to obtain less recurrent layers (i.e., 1 $\sim$ 4 layers) while all recurrent layers can be replaced by convolutional layers to obtain zero recurrent layer (i.e., 0 layer). The experimental results show that more recurrent layers can improve the recognition performance. It is worth noticing that the performance may decrease in some cases when there are too many recurrent layers, e.g., more than four layers in SMIC dataset. This might because the insufficient samples limit the performance as more recurrent layers need more training samples.

5) The Impact of IIR Filter Interval: Table V shows the recognition accuracy of our proposed methods with different-interval IIR filters and a learned filter under the LOSO protocol on the SMIC, CASME II and SAMM datasets. Firstly. we observe the effect of motion magnification by removing the temporal filter. Without the magnification, the data augmentation cannot further be performed. So only the original samples are used to train the deep RCNs. From the results, we can see that the performances of RCNs decrease greatly without using motion magnification (including the data augmentation). Furthermore, it is noted that the performance of recognizing micro-expressions can be affected by choosing different filter intervals. Specifically, the choice of filter interval would further affect the recognition performance. High intervals for filtering more obvious deformations cannot be used to describe the micro-expression deformations as the micro-expressions contain short-duration and subtle changes. Besides, the learning based method [64] is also used to verify the filter interval. The learned temporal filter can improve the recognition performance of MER but fail to outperform the EVM method. The reason is that the learned filter needs to be trained on synthetic images while ME images are limited and cannot be used to fine-tune the learned filter for MER.

6) The Impact of Larger Training Size: To observe the impact of larger training size, i.e., using a larger dataset, we construct a larger dataset having 55 subjects and 415 samples, rather than using the individual dataset, to evaluate our proposed methods. The composite dataset contains two individual datasets, i.e., CASME II and SAMM, as these two datasets have four same recognition tasks, i.e., positive, negative, surprise and other (the SMIC dataset only has three tasks). The experimental results are shown in Table VI. Our proposed methods still achieve very competitive performance compared to other approaches when using a larger dataset to train the deep model. It is worth noting that the performance degradation of all methods is mainly induced by the domain shift of combining two datasets. For instance, the "negative" task in CASME II contains "disgust", "sadness" and "fear" subtasks (three fine-grained micro-expression classes) while

TABLE V: The recognition accuracy of our proposed methods with different-interval IIR filters and learned filter under the LOSO protocol.

STRCN-A			
Filter Interval	SMIC	CASME II	SAMM
removed	0.432	0.419	0.435
$[0.0, 0.05]$	0.517	0.552	0.531
$[0.05, 0.4]$	**0.531**	**0.560**	**0.545**
$[0.4, 1.0]$	0.509	0.522	0.513
$[1.0, 5.0]$	0.414	0.451	0.429
learned [64]	0.471	0.518	0.496
STRCN-G			
Filter Interval	SMIC	CASME II	SAMM
removed	0.492	0.504	0.527
$[0.0, 0.05]$	0.698	0.785	0.753
$[0.05, 0.4]$	**0.723**	**0.803**	**0.786**
$[0.4, 1.0]$	0.675	0.759	0.742
$[1.0, 5.0]$	0.596	0.678	0.648
learned [64]	0.647	0.716	0.701

TABLE VI: The recognition accuracy of different methods under the LOSO and LOVO protocols over the composite dataset.

Composite Dataset (CASME II + SAMM)		
Method	LOSO	LOVO
LBP-TOP [6]	0.384	0.426
Bi-WOOF [10]	0.453	0.591
Image-based CNN [24]	0.365	0.495
CNN-LSTM [48]	0.413	0.503
CNN-GRU [52]	0.417	0.508
STRCN-A(Ours)	0.495	**0.698**
STRCN-G(Ours)	**0.629**	0.693

the "negative" task in SAMM has "disgust", "sadness", "fear", "contempt" and "anger" subtasks (five classes). That would greatly decrease the performance of all approaches when compositing two datasets.

F. Visual Investigation

An example of spatial and temporal processing is shown in Fig. 7, in which the magnification factor α is set to 8. It is shown that the aligned facial regions can be obtained with spatial processing and the temporal changes are amplified by temporal processing. It is worth noticing that the temporal changes of micro-expressions cannot be observed easily by naked eyes without temporal processing (i.e., motion magnification), which also increases the difficulty of automatic learning. The temporal processing amplifies the temporal changes and helps to learn deep models.

The generated mask in STRCN-A is shown in Fig. 8 (a). From the figure, it can be seen that some areas are not active in the difference heat map and might not reveal micro-expressions. Oppositely, the areas around eyes, nose and mouth are mostly active for micro-expressions and can be chosen by the binary mask. With the selected areas, the dimension of learning space for deep model can be drastically reduced and thus be helpful for the learning of deep models. The quantificational validation for selecting areas is investigated in Section IV-E2. The calculated optical flow fields are shown in Fig. 8 (b), in which the fields are normalized for visualization.

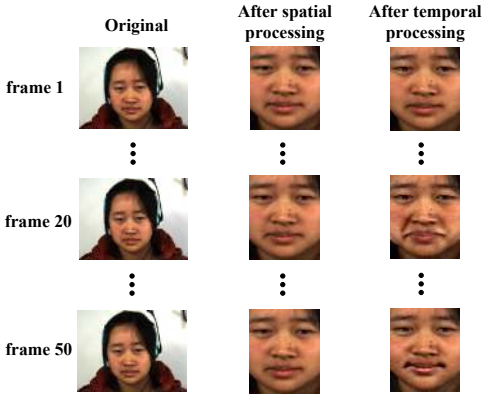
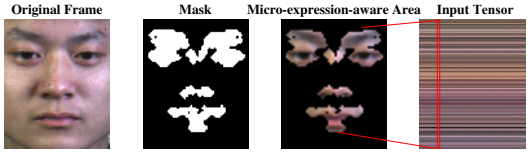
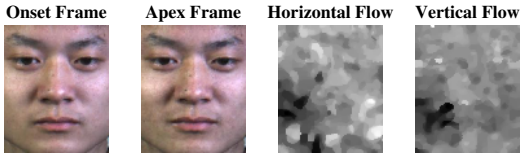


Fig. 7: An example of using spatial and temporal processing (indexed "sub02\EP01_11f" in CASME II).



(a) STRCN-A



(b) STRCN-G

Fig. 8: The network input in STRCN-A and STRCN-G. (a) The original frame, generated binary mask, micro-expression-aware area ($p = 0.3$) and input tensor in STRCN-A on entire datasets; (b) The onset image, apex image, and optical flow fields (normalized for illustration).

Finally, three (two successful and one failed) recognition examples from different datasets are shown in Fig. 9 in which "GT" represents the ground-truth label. From the first example (first row), it is observed that the augmented changes indicate a negative expression compared the apex frame with onset frame. The second example (second row) shows that the smile faces of subject in the entire sequence exhibit the positive emotion when the subtle changes around the mouth are obviously augmented. In the third example (third row), the person has not shown any facial deformations, which cannot be observed by naked eyes, even the motion augmentation has been applied. The exemplar subject shows low-level emotions by facial expressions. It further demonstrates that the personality of humans will make the micro-expression recognition being challenging.

V. Conclusion and Future Work

In this paper, we proposed a novel micro-expression recognition approach based on spatiotemporal recurrent convolutional networks (STRCNs). The proposed STRCNs modeled the spatiotemporal deformations of micro-expression sequence

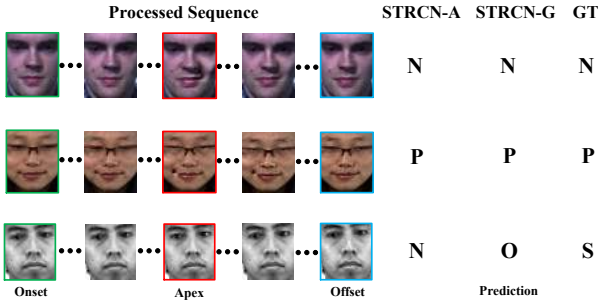


Fig. 9: Recognition examples from three datasets by our proposed STRCN methods ("P" - Positive, "N" - Negative, "S" - Surprise, "O" - Other).

by two types of connectivities (STRCN-A and STRCN-G), in which the connectivity of RCNs was extended across temporal domain for sequential data. The STRCN-A was an appearance based method, which transformed one sequence into a matrix by concatenating frames to reserve the appearance of facial regions. In contrast, the STRCN-G was a geometric based method, which transformed one sequence into a matrix by computing the optical flow fields of onset and apex frames, to obtain the geometric information of facial movements. Furthermore, to overcome the shortcomings of limited and imbalanced training samples, two temporal data augmentation strategies were designed for network input and a balanced loss was integrated to this recognition task. Through performing the experiments in LOVO and LOSO protocols on three spontaneous micro-expression datasets, i.e., SMIC, CASME II and SAMM, we verified the effectiveness of our proposed micro-expression recognition approach compared to the state-of-the-art methods. The STRCN-A achieved the best performance under the LOVO protocol while STRCN-G achieved the best performance under the LOSO protocol.

In the future work, we would further explore a more effective processing framework in an end-to-end way. The learnable deep models will be employed for searching the optimal processing, such as face cropping and motion magnification, reducing the setup of hyper-parameters. Besides, another type of connectivity will be further exploited to model the spatiotemporal deformation for micro-expression recognition in larger datasets.

ACKNOWLEDGMENT

We would like to thank the reviewers for their valuable and constructive comments, which greatly helped us to improve this work. This work is partly supported by the National Natural Science Foundation of China (Nos. 61702419, 61702491), and the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JQ6090).

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] X. Shen, Q. Wu, and X. Fu, "Effects of the duration of expressions on the recognition of microexpressions," *Journal of Zhejiang University SCIENCE B*, vol. 13, no. 3, pp. 221–230, 2012.

[3] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: facial micro-expression recognition," *Multimedia Tools & Applications*, vol. 77, no. 15, p. 1930119325, 2018.

[4] P. Ekman, "The micro-expression training tool, v. 2. (mett2)," www.mettonline.com, 2007.

[5] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, pp. 1–1, 2017.

[6] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 1449–1456.

[7] S. J. Wang, W. J. Yan, X. Li, and G. Zhao, "Micro-expression recognition using color spaces," *IEEE Transactions on Image Processing*, vol. 24, no. 12, p. 6034, 2015.

[8] X. Huang, S. J. Wang, G. Zhao, and M. Pietikinen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *ICCV Workshop on Computer Vision for Affective Computing*, 2015, pp. 1–9.

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

[10] S. T. Liong, J. See, C. W. Phan, and K. S. Wong, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[11] A. K. Davison, M. H. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: An investigation on spatio-temporal descriptors," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 111–123.

[12] X. Duan, Q. Dai, X. Wang, Y. Wang, and Z. Hua, "Recognizing spontaneous micro-expression from eye region," *Neurocomputing*, vol. 217, pp. 27–36, 2016.

[13] J. A. Ruizhernandez and M. Pietikainen, "Encoding local binary patterns using the re-parametrization of the second order gaussian jet," in *International Conference and Workshops on Automatic Face and Gesture Recognition (FG Workshops)*, 2013, pp. 1–6.

[14] 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 *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2014, pp. 21–23.

[15] S. J. Wang, W. J. Yan, G. Zhao, X. Fu, and C. G. Zhou, "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *European Conference on Computer Vision Workshops (ECCV Workshops)*, 2015, pp. 325–338.

[16] 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 Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2016.

[17] X. Huang, S. J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikainen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[18] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1404–1416, 2015.

[19] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[20] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+3d facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[21] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[22] H. Kim, T. Mei, H. Byun, and T. Yao, "Exploiting web images for video highlight detection with triplet deep ranking," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2415–2426, 2018.

[23] 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, p. e86041, 2014.

[24] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017, pp. 1–7.

[25] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in *International Conference on Crime Detection and Prevention*. IET, 2009, pp. 1–6.

[26] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," in *Workshop on Applications of Computer Vision (WACV)*. IEEE, 2009, pp. 1–6.

[27] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[28] 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 *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[29] 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 Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[30] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2018.

[31] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *International Conference on Automatic Face & Gesture Recognition (FG) and Workshops*. IEEE, 2011, pp. 51–56.

[32] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao, "Spontaneous micro-expression spotting via geometric deformation modeling," *Computer Vision & Image Understanding*, vol. 147, pp. 87–94, 2016.

[33] S. J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, 2017.

[34] A. Moilanen, G. Zhao, and M. Pietikainen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1722–1727.

[35] S. T. Liong, J. See, C. W. Phan, Y. H. Oh, A. C. L. Ngo, K. S. Wong, and S. W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing Image Communication*, vol. 47, no. C, pp. 170–182, 2016.

[36] S. Afzal and P. Robinson, "Natural affect datacollection & annotation in a learning context," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.

[37] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 4678–4683.

[38] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2018.

[39] S. Yao, N. He, H. Zhang, and O. Yoshie, "Micro-expression recognition by feature points tracking," in *International Conference on Communications*. IEEE, 2014, pp. 1–4.

[40] Z. Lu, Z. Luo, H. Zheng, J. Chen, and W. Li, "A delaunay-based temporal coding model for micro-expression recognition," in *Asian Conference on Computer Vision Workshops (ACCV Workshops)*, 2015, pp. 698–711.

[41] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[42] Z. Xia, X. Feng, X. Hong, and G. Zhao, "Spontaneous facial micro-expression recognition via deep convolutional network," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2018, pp. 1–6.

[43] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose, and M. Richardson, "Do deep convolutional nets really need to be deep and convolutional?" in *International Conference on Learning Representations (ICLR)*, 2017, pp. 1–13.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[45] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image & Vision Computing*, vol. 60, pp. 4–21, 2017.

[46] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1725–1732.

[47] T. Du, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 4489–4497.

[48] Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694–4702.

[49] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[50] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.

[51] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 1–17.

[52] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv*, pp. 1–9, 2014.

[53] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning (ICML)*, 2015, pp. 843–852.

[54] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3367–3375.

[55] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene parsing," in *International Conference on Machine Learning (ICML)*, 2014, pp. 82–90.

[56] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2016, pp. 84–92.

[57] Z. Xia, W. Zhang, F. Tan, X. Feng, and A. Hadid, "An accurate eye localization approach for smart embedded system," in *International Conference on Image Processing Theory Tools and Applications (IPTA)*, 2016, pp. 1–5.

[58] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 13–15, 2012.

[59] A. C. L. Ngo, J. See, and C. W. Phan, "Sparsity in dynamics of spontaneous subtle emotion: Analysis & application," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 396–411, 2017.

[60] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2432–2439.

[61] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. online, pp. 1–1, 2018.

[62] A. Vedaldi and K. Lenc, "Matconvnet - convolutional neural networks for matlab," in *ACM Multimedia*. ACM, 2015, pp. 689–692.

[63] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, 2015, pp. 1–12.

[64] T. H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, and W. Matusik, "Learning-based video motion magnification," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 663–679.