Spontaneous Facial Micro-expression Recognition via Deep Convolutional Network

Zhaoqiang Xia[1], Xiaoyi Feng[1], Xiaopeng Hong[2], and Guoying Zhao[2]
[1] School of Electronics and Information, Northwestern Polytechnical University
e-mail: zxia@nwpu.edu.cn
[2] Center for Machine Vision and Signal Analysis, University of Oulu

***Abstract*— The automatic recognition of spontaneous facial micro-expressions becomes prevalent as it reveals the actual emotion of humans. However, handcrafted features employed for recognizing micro-expressions are designed for general applications and thus cannot well capture the subtle facial deformations of micro-expressions. To address this problem, we propose an end-to-end deep learning framework to suit the particular needs of micro-expression recognition (MER). In the deep model, recurrent convolutional networks are utilized to learn the representation of subtle changes from image sequences. To guarantee the learning of deep model, we present a temporal jittering procedure to greatly enrich the training samples. Through performing the experiments on three spontaneous micro-expression datasets, i.e., SMIC, CASME, and CASME2, we verify the effectiveness of our proposed MER approach.**

***Keywords*— Micro-Expression Recognition, Recurrent Convolutional Networks, Temporal Jittering, Motion Magnification**

I. Introduction

Micro-expressions are repressed and involuntary facial expressions which appear in the facial regions of humans. Compared to normal facial expressions (i.e., macro-expressions), micro-expressions usually have short duration, i.e., less than 0.2 second, and low intensity [1], [2]. Although the macro-expressions can reflect the emotion of humans and have a wide application [3], the people still can pretend the genuine emotions. In contrast, the 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, such as lie detection, police case diagnosis, business negotiation, and psychoanalyzing. 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 tool [4], humans still manually detect and recognize micro-expressions from videos with low accuracy. Consequently, the automatic analysis of micro-expressions will be very valuable to promote the performance of analyzing large amounts of video sequences.

Earlier studies focused on the posed micro-expression analysis differing greatly from the spontaneous ones as they are controlled by different motor pathways [5], [6]. As spontaneous micro-expressions can be observed frequently in real life and reveal more affective information of humans, many works have been devoted to spontaneous micro-expressions recently. 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 [7], [8], local textures [9] and main directional maximal difference [10] have been proposed to capture micro-expression frames from videos. To tackle the recognition task, several handcrafted features have been presented to model subtle changes of micro-expressions. The local binary patterns on three orthogonal planes (LBP-TOP) [11] widely used to describe dynamic textures is applied to recognize micro-expressions. Although LBP-TOP has shown the capacity of discriminability and efficiency, it still suffers the sensitivity problem of global changes. So the spatiotemporal completed local quantization patterns (STCLQP) [12], directional mean optical-flow (MDMO) [13], spatiotemporal local binary pattern [14] and hierarchical spatiotemporal descriptors[15] are proposed to improve the robustness of representation. These handcrafted features are designed to capture temporal differences of micro-expression sequences and achieve the accuracy rate of more than 50%.

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 the task of MER [16], [17]. However, CNNs are used directly on each frame of micro-expression videos without modeling temporal changes. Thus, in this paper, we propose an end-to-end deep framework to automatically recognize micro-expressions by leveraging the temporal changes. In the deep model, the convolutional layers with recurrent connections (i.e., recurrent convolutional neural networks, shorted as R-CNN [18]) are utilized to learn the representation of subtle changes and the last classificatory layer is used to recognize micro-expressions. To guarantee the learning of deep model, we propose a temporal jittering procedure to greatly enrich the training samples for learning deep model. Additionally, before feeding the sequence into the deep network, the motion magnification technique is employed to the entire sequence for enhancing the subtle changes of micro-expressions.

II. Proposed Method

In this section, we present our proposed method based on deep model for micro-expression recognition (MER).

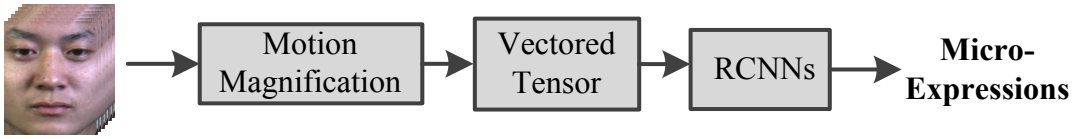

Fig. 1. The framework of our proposed approach for micro-expression recognition.

A. The Framework

To focus on the problem of spontaneous facial micro-expressions recognition, we crop and normalize face regions from image sequences using the preprocessing method proposed in [11]. This method utilizes conventional eye detector [19] and active shape model (ASM) based algorithm [8] to crop and align face regions. The eye detector determines the starting positions of face shapes and then the accurate locations of face shapes are iteratively fitted by the ASM algorithm. Consequently, the face regions are cropped and normalized in terms of the landmark points of ASM.

Based on the cropped faces, our proposed approach contains several procedures to recognize micro-expressions. The framework of our proposed method is shown in Fig. 1. Firstly, we utilize the motion magnification technique to enhance subtle changes of micro-expressions based on the aligned facial regions. Then, images from an video are resized and flattened to vectors, and then the sequence is concatenated to form a tensor. Lastly, the tensor is fed into the deep RCNN model for recognizing micro-expressions.

B. Motion Magnification

The temporal variations in micro-expression videos are very small and impossible to see with naked eyes of humans. Similarly, it is difficult to automatically learn representations of these subtle changes from noisy content by machine learning techniques. In this context, we use the motion magnification technique to amplify the hidden motion information of adjacent frames and then utilize deep RCNNs to learn the motion information automatically.

To amplify motion deformations and limit magnification distortions, we choose the Eulerian Video Magnification (EVM) method [20] to amplify the temporal motion. More specifically, the Laplacian pyramid method is utilized to decompose the input facial sequence into different spatial frequency bands. Then all bands of images are filtered by temporal filters and achieve the first-order motion information. The magnified temporal motion can be calculated by

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

where $f(x) = I(x,0)$, and $I(x,t)$ denotes the image intensity at position x 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. $\tilde{I}(x,t)$ is the image intensity of tth frame after magnified. 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.

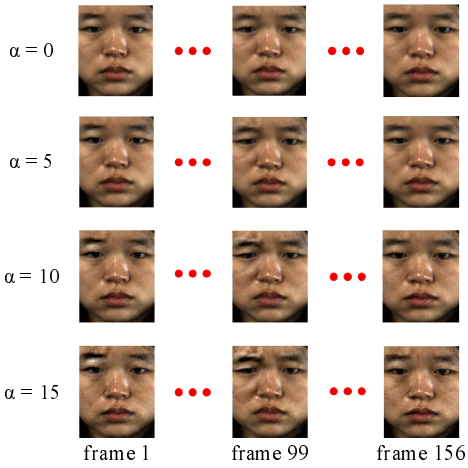

Fig. 2. An example of motion-magnified facial regions with micro-expressions, in which the regions of right eyebrows have more obvious motion changes as well as impulse noises with larger magnification factor.

In order to obtain the subtle changes of facial sequences, an infinite impulse response (IIR) filter with cut-off frequencies of $[0.05, 0.4]$ Hz is chosen as our temporal filter for subsequent motion magnification. In other words, we only use one temporal filter $\delta(t)$ as the bandpass filter. This bandpass filter is more suitable to capture the motion information for MER. Similarly, we use only one magnification factor α to amplify the temporal motion changes of micro-expression sequences for all spatial frequencies.

The motion-magnified facial sequences are shown in Fig. 2. It is noted that the regions of right eyebrows have more obvious motion changes with larger magnification factor α while more impulse noises have been induced with larger α.

C. Recurrent CNNs

Compared to the handcrafted features, CNNs have more powerful ability to describe the subtle changes of micro-expressions. In this paper, we add recurrent connections (i.e., RCNNs [21]) within the feed-forward convolutional layers. The temporal changes of sequences can be captured by multiple-scale receptive fields.

The architecture of our deep RCNNs is shown in Fig. 3. It contains one feed-forward convolutional layer and several recurrent convolutional layers (RCLs). The layer 1 is the only feed-forward convolutional layer without recurrent connections and used to compute efficiently. Following the standard

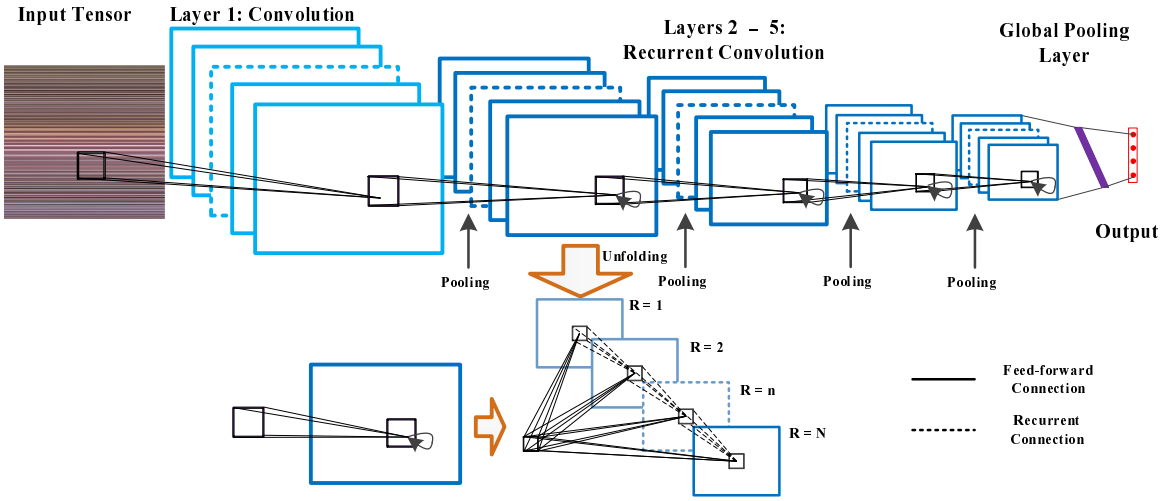

Fig. 3. The architecture of our deep RCNNs for micro-expression recognition.

convolutional layer 1, four RCLs ($RCL2 \sim 5$) 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 to a vector. In the last layer, the Softmax layer is employed to calculate the recognition probabilities with concatenated feature vector.

In each RCL layer, several hidden layers are used to expand the size of receptive fields. As shown in Fig. 3, 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 denotes the depth of one RCL layer, i.e., the number of hidden convolutional layers, and is valued from 1 to N. 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 an RCL layer can be computed as

$$z_{ijk}(n) = \mathbf{w}_k^{f\,T}\mathbf{u}_{ij}(n) + \mathbf{w}_k^{r\,T}\mathbf{v}_{ij}(n-1) + b_k \qquad (2)$$

where $\mathbf{u}_{ij}(n)$ and $\mathbf{v}_{ij}(n-1)$ represent the feed-forward and recurrent input, respectively. In the equation, $\mathbf{w}_k^f$ and $\mathbf{w}_k^r$ 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

$$v_{ijk}(n) = f(z_{ijk}(n)) \qquad (3)$$

where $n = 0, 1, \cdots, N$ and the initial state $v_{ijk}(0) = 0$. $f(\cdot)$ represents the normalized activation function.

Finally, the output of deep network uses the Softmax function to classify feature vectors to C categories and it can be calculated as

$$y_c = \frac{exp(\mathbf{w}_c^T\mathbf{v})}{\sum_{c=1}^{C} exp(\mathbf{w}_c^T\mathbf{v})} \qquad (4)$$

where y_c is the predicted probability of cth category, and $\mathbf{v}$ denotes the output feature vector of last pooling layer.

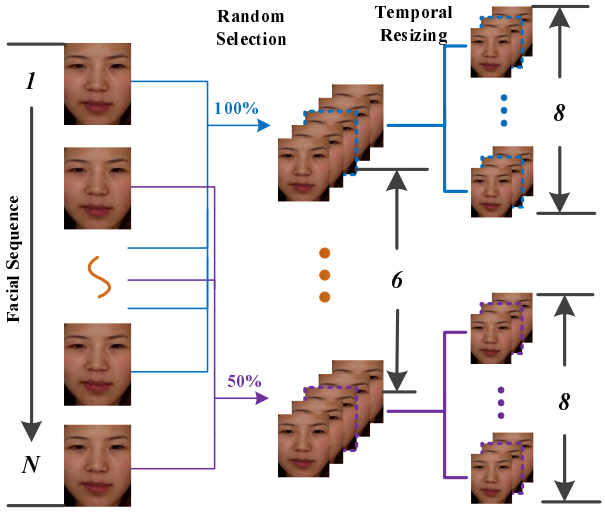

Fig. 4. The illustration of temporal jittering for enriching training samples.

The parameter learning is performed by minimizing the cross entropy loss function using the back propagation through time (BPTT) algorithm [22].

D. Temporal Jittering for Model Training

Currently, the training samples with spontaneous micro-expressions are not sufficient for learning numerous parameters of deep RCNNs. It is still necessary to train deep RCNNs with at least thousands of samples even though RCNNs need less samples than the standard CNNs. However, no more than 200 original samples in spontaneous micro-expression datasets can be used to train the deep RCNNs. This will cause the problem of over-fitting and limit the recognition performance. In order to reduce over-fitting, we propose a temporal jittering strategy to train our deep model, which is shown in Fig. 4.

Firstly, we randomly select some frames from sequences with a percentage. Totally, five levels of percentages are adopt-

ed for random selection, i.e., 100%, 90%, 80%, 70%, 60% and 50%. So, with five random selection, the original data can be augmented by six times while these data contain different-sized sequences. Secondly, eight resizing methods (down-sampling or up-sampling) are utilized to generate sequences with fixed size of 30 frames. In this context, we use eight sampling approaches, and these methods involve the *nearest-neighbor approach, bilinear approach, bicubic approach* and *kernel based approaches* (i.e., *box-shaped kernel, triangular kernel, cubic kernel, Lanczos-2 kernel* and *Lanczos-3 kernel*). So every sample can generate eight similar samples but having different temporal deformations. Thus, the original data are augmented by eight times. Performing these two procedures jointly, namely temporal jittering, the original data can be augmented by 48 times. These augmented data can make it sufficient for training deep architecture.

III. Experiments

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

A. Implementation Details

To amplify the motion changes and avoid inducing excessive noises, the magnification factor is fixed to $\alpha = 10$ as a tradeofff between the magnification and noises. Whereas, some image intensities are not satisfied with the assumptions in Wu's method [20]. According to [20], the bound for factor α in any frame is adopted as follows

$$\alpha_c = \frac{\lambda}{8\delta(t)} - 1 \tag{5}$$

where λ denotes the spatial wavelength and is set to $\lambda = 16$ in this context. Therefore, the magnification factor can be finally used as $\alpha = min(10, \alpha_c)$ for subsequent procedures.

Before feeding the magnified images into the vector, images are resized by selecting core region of facial images. The 80% pixels along the width and height are reserved and remaining pixels close to boundary are removed. The number of elements for one channel of color image can be reduced from 3072 to 1966. In the next step, the temporal normalization operations are performed to obtain fixed-size input tensor. Here, we utilize different kinds of resizing methods to temporally down-sample or up-sample the sequences and obtain fixed number of frames. In this context, we choose 30 frames to feed the deep model. So the input tensor is with size of $1966 \times 30 \times 3$.

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 [21], [18]. The detailed configurations are shown in Table 1.

In stochastic gradient decent (SGD) procedure of BPTT for parameter learning, the momentum is set to 0.9 and weight decay 0.0005. The stopping criterion for SGD is set to 10^{-4} for iterations. The learning rate is set to 10^{-3} in the beginning and will be multiplied with damping factor 0.5 when all mini-batches are traversed and re-allocated randomly. To accelerate the parameter learning, we employ the library MatConvNet [23] to accomplish our proposed model. The mini-batch size for training model is set to 64 as it is limited by the memory of GPUs (One Geforce TiTan X).

Table 1. The detailed configuration of our deep RCNNs.

Layers	Configurations
Input	Tensor:$1966 \times 30 \times 3$ Width×Height×Channel
Conv1	$k : 5 \times 5$, $p : 0$, $s : 1 \times 1$
Pool1	MAX, $k : 4 \times 1$, $s : 4 \times 1$
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, $k : 4 \times 1$, $s : 4 \times 1$
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, $k : 4 \times 4$, $s : 4 \times 4$
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, $k : 4 \times 2$, $s : 4 \times 2$
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	AVG
Output	C categories

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

B. Micro-expression datasets

Three spontaneous micro-expression datasets are used to evaluate the performance of our proposed approach in our experiments: SMIC dataset [24], CASME dataset [25] and CASME2 dataset [26]. All of them are designed to detect and recognize 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. These participants undergo high emotional arousal and suppress their facial expressions in an interrogation room setting with a punishment threat and highly emotional clips. Half data are recorded by low-speed cameras.
- The CASME dataset has 195 micro-expressions from 19 subjects. It uses similar procedures like SMIC to elicit micro-expressions from subjects. Because of the creators with psychological background, these expressions are obtained from stricter lab situations and labeled more accurately. More emotions and action units (AUs) are labeled by psychologists.
- The CASME2 dataset has 256 micro-expressions from 26 subjects. It has higher video quality and image size compared with CASME. The recording rate of cameras in CASME2 is 200 fps. Thus the video sequences of micro-expressions in CASME2 have more frames than other corpus.

To keep three datasets consistent with each other, we merge seven categories in CASME and CASME2 into four classes.

Table 2. The recognition accuracy of different methods on three datasets.

Approaches	Evaluation		
	SMIC	CASME	CASME2
LBP-TOP[11]	0.537	0.577	0.592
LBP-SIP [29]	0.445	0.368	0.466
TICS[28]	0.561	0.618	0.623
MDMO [13]	**0.640**	0.573	0.584
Pre-trained CNNs [30]	0.301	0.376	0.304
CNNs[17]	0.325	0.471	0.491
MER-RCNN (Ours)	0.571	**0.632**	**0.658**

Following [27], [13], the happy micro-expressions in CASME and CASME2 are classified into "Positive" class as they indicate good emotions of subjects. In contrast, the disgust, sadness and fear 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, the first three classes (i.e., positive, negative and surprise) are used to annotate the micro-expressions.

C. Experimental Setup and Protocols

In previous works [28], [27], the *leave-one-sample-out* protocols are utilized to explore the limited samples of aforementioned datasets for handcrafted features. However, it is intractable to test every sample in our experiments as the training of deep models is time-consuming. Instead of exhaustive testing of leave-one-sample-out protocols, we use *5-fold* evaluation protocol to evaluate our proposed method on all datasets. In *5-fold* protocol, we randomly split the samples into five parts, in which four of them are used for training and the rest for testing.

Following [11], [12], [28], we utilize the mean recognition accuracy to evaluate the performance of our deep Micro-Expression Recognition algorithm using Recurrent CNNs (abbreviated as MER-RCNN). We compare our MER-RCNN method with the state-of-the-art methods on all datasets, i.e., LBP-TOP [11], LBP-SIP [29], TICS [28] and MDMO [13]. To observe the effect of recurrent connections in deep model, we also compare our proposed method with standard CNNs for single images [17].

D. Experimental Results

All methods are compared on three datasets and the performances are reported in Table 2. Observed from Table 2, our proposed method achieves the best accuracies in most configurations. Compared to the handcrafted features, the deep features from automatic learning in our proposed method are competitive even with limited training samples.

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. The TICS method [28] can improve the performance of LBP-based features in new color space while the subtle changes can still not be well captured. Based on optical flow maps, the MDMO features [13] can extract subtle changes of micro-expressions but depend on the accurate partitions of facial regions. And the MDMO features can easily be influenced by the illuminations as the optical flow can be affected by lighting changes. Our proposed method can obtain the descriptions of subtle changes and outperform these state-of-the-art methods.

It is worth noticing that MDMO features achieve better results than our proposed method on SMIC datasets. The SMIC dataset has less subjects and micro-expression samples than other datasets. Besides, half of data in SMIC dataset are recorded by using a 25fps camera. These low-speed micro-expression videos are less apparent to learn the subtle changes from them compared the high-speed videos in other datasets. So the insufficient and incomplete samples can limit our proposed method.

Besides, the architecture of deep model without recurrent connections are also investigated in Table 2. In the experiments, the prominent architecture of CNNs (i.e., VGG-Face) used for face recognition [17]is employed to micro-expression recognition. The CNNs are trained merely with the augmented data and recognize micro-expressions for single images. Meanwhile, the pre-trained model based on VGG-Face is used to leverage the additional data. The pre-trained model is achieved on a large-scale dataset of facial images [30], i.e., 982,803 images for 2,622 identities. It is observed that the architecture without recurrent connections cannot achieve good performance on micro-expression recognition as it neglects the temporal information. Moreover, the supplementary data are not helpful to learn subtle changes of micro-expressions even if we use the facial images.

IV. Conclusion

In this paper, we proposed an end-to-end framework comprised of recurrent convolutional networks to recognize micro-expressions. In the deep framework, the RCNNs were utilized to learn the representation of subtle changes and recognize micro-expressions. For employing the RCNN model, resizing and vectorization methods were used to transfer one image into a tensor. Besides, the temporal jittering was utilized to enrich the training samples to facilitate the learning procedure. Through performing the experiments on three spontaneous micro-expression datasets, we verified the effectiveness of our proposed micro-expression recognition approach.

Acknowledgment

This work is partly supported by the National Nature Science Foundation of China (No.61702419), and the Natural Science Basic Research Plan in Shaanxi Province of China (No.2018JQ6090).

References

[1] Madhumita Takalkar, Min Xu, Qiang Wu, and Zenon Chaczko, "A survey: facial micro-expression recognition," *Multimedia Tools & Applications*, pp. 1–25, 2017.

[2] Xunbing Shen, Qi Wu, and Xiaolan 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] Xianlin Peng, Zhaoqiang Xia, Lei Li, and Xiaoyi Feng, "Towards facial expression recognition in the wild: A new database and deep recognition system," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1544–1550.

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

[5] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas 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.

[6] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, 2017.

[7] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep 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.

[8] Zhaoqiang Xia, Xiaoyi Feng, Jinye Peng, Xianlin Peng, and Guoying Zhao, "Spontaneous micro-expression spotting via geometric deformation modeling," *Computer Vision & Image Understanding*, vol. 147, pp. 87–94, 2016.

[9] Antti Moilanen, Guoying Zhao, and Matti 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.

[10] Su Jing Wang, Shuhang Wu, Xingsheng Qian, Jingxiu Li, and Xiaolan Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, 2017.

[11] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikainen, "Recognising spontaneous facial micro-expressions," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 1449–1456.

[12] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2015.

[13] Yong Jin Liu, Jin Kai Zhang, Wen Jing Yan, Su Jing Wang, Guoying Zhao, and Xiaolan 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.

[14] Xiaohua Huang, Su Jing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng, and Matti Pietikainen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 8, pp. 1–15, 2017.

[15] Yuan Zong, Xiaohua Huang, Wenming Zheng, Zhen Cui, and Guoying Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–13, 2018.

[16] Devangini Patel, Xiaopeng Hong, and Guoying Zhao, "Selective deep features for micro-expression recognition," in *International Conference on Pattern Recognition (ICPR)*, 2017, pp. 2258–2263.

[17] Madhumita A. Takalkar and Min 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.

[18] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying 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.

[19] Zhaoqiang Xia, Wenhao Zhang, Fang Tan, Xiaoyi Feng, and Abdenour Hadid, "An accurate eye localization approach for smart embedded system," in *International Conference on Image Processing Theory Tools and Applications*, 2016, pp. 1–5.

[20] Hao Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Fredo Durand, and William Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 13–15, 2012.

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

[22] Paul J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

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

[24] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti 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.

[25] Wen Jing Yan, Qi Wu, Yong-Jin Liu, Su Jing Wang, and Xiaolan 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.

[26] Wen Jing Yan, Xiaobai Li, Su Jing Wang, Guoying Zhao, Yong Jin Liu, Yu-Hsin Chen, and Xiaolan Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, pp. e86041, 2014.

[27] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 4678–4683.

[28] Su Jing Wang, Wen Jing Yan, Xiaobai Li, and Guoying Zhao, "Micro-expression recognition using color spaces," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6034, 2015.

[29] Yandan Wang, John See, Raphael C. W. Phan, and Yee Hui 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.

[30] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, 2015, pp. 1–12.