Micro-Expression Recognition Based on Multiple Aggregation Networks

Wenxiang She*†, Zhao Lv*, Jianhua Taoi*†, Bin liu† and Mingyue Niu†
*Anhui University, Heifei, China
E-mail: E18201085@stu.ahu.edu.cn, kjlz@163.com Tel: +86-18855992548
†Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-mail: {jhtao, Liubin,}@nlpr.ia.ac.cn, niumingyue2017@ia.ac.cn

***Abstract*— Micro-expression is a low-intensity, short-term spontaneous facial activity that can reflect people's true feelings. Existing methods mainly extract hand-crafted descriptors from the whole face, which are not enough to capture detailed information of the local regions and are not optimal due to depending on the experience of researcher. Thus, we propose a multiple aggregation networks to explore the impact of local facial regions on micro-expressions recognition in detail. The framework uses multiple different network branches to extract frame-level information about the facial regions of interest, as well as the holistic features of whole face. Finally, the physical meaning of statistical parameters is fully utilized to characterize the average and dynamic changes of frame-level features to generate video-level features. Finally, use video-level features as the input of SVM for micro-expression classification. Experiments are conducted on CASME, CASME II and SMIC databases. The results demonstrate that the proposed method is superior to previous works.**

I. INTRODUCTION

Relevant psychological research [1] have shown that the duration of micro-expressions is only 1/25 to 1/2 seconds, and only minor changes occur in local facial regions. In addition, micro-expression is an unconscious facial activity of people, so it reveals the inner emotions that people may hide [2]. Therefore, this nonverbal facial information has important potential application value in many fields, such as interrogations [3], medical [4] and security field [5].

In recent years, micro-expressions recognition (MER) has received more and more attention from researchers [6] [7]. And how to extract the subtle information from video clips becomes a key issue. Previous studies [6] [7] [8] use hand- crafted descriptors to extract low-level features from micro-expression videos. However, these works depend on individual experience, which cause the results obtained may not be optimal [9]. With the staff of deep learning, some researchers [9] [11] try to use deep neural network models to extract high-level features of the whole face to recognize micro-expressions. However, research by Porter et al. [10] shows that when micro-expressions occur, some special local areas change relatively significantly, named as the regions of interest (RoIs). In other words, when producing micro- expressions, the degree of change in different facial regions is not the same. Therefore, it is necessary to investigate the relatively obvious changes in facial regions to obtain distinguishing information, thereby improving the efficiency of MER.

In response to the above problems, this paper uses the advantages of neural networks to extract micro-expression information. In addition, in order to comprehensively study local information, we extracted RoIs from each frame based on physiological studies. We input the RoIs and the whole face images into the proposed network, which can use the residual network to obtain local or holistic representations, respectively. Since these two types of hierarchical features represent micro-expression in different scales, we concatenate both holistic and local features to generate new features and sent to the neural network for training to modify the parameters, and the features of the first fully connected layer are used as frame-level features. Then, the proposed aggregation strategy is used to aggregate frame-level features into video-level features for classification. And the results show that this method has a positive effect on MER.

The contributions of this paper are as follows：

1) Proposed a multipath aggregation network model that can extract high-level features from the local and holistic facial regions, so as to comprehensively investigate the impact of facial detailed information and global representation on MER.

2) An aggregation strategy is presented, which fully uses the physical meaning of statistical parameters to characterize the average and dynamic changes of frame-level features to generate video-level features for MER.

Rest of this paper is organized as follow: we make a brief review of some existing works on micro-expression recognition in Section II. The details of the works will be specialized in Section III. In Section IV, experiments and result analysis are presented. Finally, we conclude our method in the Section V.1)

II. RELATED WORKS

MER is a challenging task with low intensity and rapid movement of facial muscles. However, due to its wide practical value, the research on MER has gradually improved.

Traditionally, handcraft feature extraction technology is very popular in MER. Liu et al. [6] proposed the Main Directional Mean Optical-flow (MDMO) algorithm for micro-expression recognition. They utilized optical flow technique to compute the subtle movement of RoIs. For each RoI, a polar coordinate including the magnitude and direction of the optical flow vectors are computed. Then put the MDMO features into the SVM with the polynomial kernel function to classify.

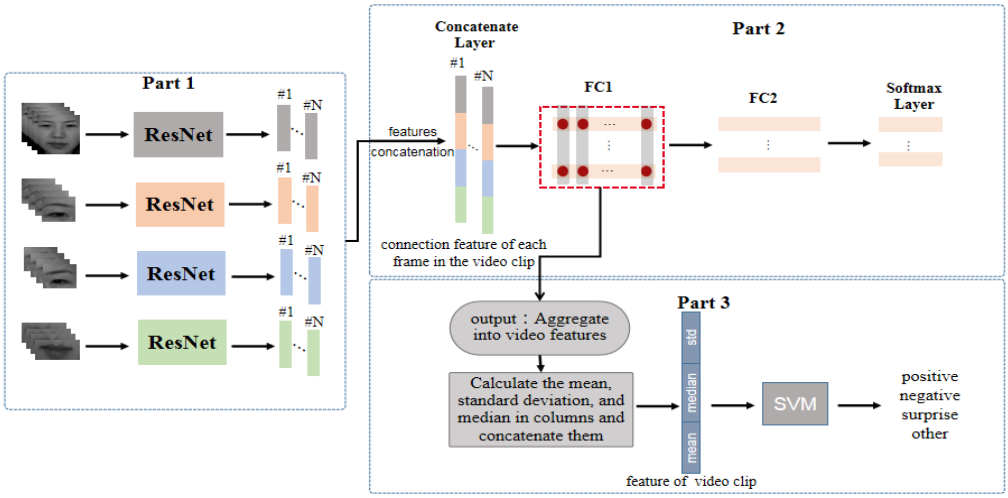

Fig. 1 Micro-expression recognition framework. Part1 extracts local and local representation of micro-expressions through residual networks and then concatenate them. Part 2 trains the concatenate features, and the orange vector in the red box represents the frame-level features. Part 3 aggregates the frame-level futures to get video-level features for classification.

However, they divide the whole face into 36 RoIs, including the regions are inactive when the micro-expressions occur. Moreover, Yu et al. [7] introduced Facial Dynamics Map (FDM) to characterize the movements of micro- expressions at different granularities. They divided each expression sequence into spatio-temporal cuboids in the chosen granularity, and then the principal optical flow direction of each cuboid is calculated. Using these main directions, the obtained facial dynamic map can represent a micro-expression sequence. However, such hand-crafted features have a limitation in the sense that they rely on a prior knowledge and heuristics.

With the advancement in technology, deep neural network leads to a series of breakthroughs in image classification. And deep networks are increasingly used in MER tasks. Kim et al. [11] was the first person to integrate CNN into MER. But CNN mainly focuses on the extraction of spatial features. To analyze temporal information, Ji et al. [19] proposed 3D-CNN to extract features from spatial and temporal dimensions. Then Wang et al. [13] used 3D-CNN to recognize micro- expressions. However, they only focus on extracting the global features while ignoring local information. In fact, studies [10] have shown that both local and holistic representations of the face can help increase the discrimination of MER.

III. PROPOSED METHOD

The micro-expression recognition framework proposed in this paper contains four branch residual networks, as shown in Fig 1. In this section, we first describe the video processing process. Then, the selection of local facial regions is given. Finally, we show the process of obtaining the representation of micro-expression video clips using the proposed method.

A. Video Processing

In order to input video sequence into deep neural network for learning, it is necessary to process the micro-expression videos. Firstly, faces are detected using Haar face detector [14], and cropped the face by using the 68 landmarks from Active Shape Model (ASM) [15]. Pictures containing only facial regions can reduce picture noise and express micro-expressions more efficiently. Secondly, we use time-interpolation model (TIM) [16] to up-sample or down-sample the videos so that all videos are the same length (64 frames). Normalizing the length of the micro-expression sequences that may make the video comparable in the temporal domain. Lastly, all the cropped images are resized to 120×140 pixels. The process of video processing is shown in the Fig. 2.

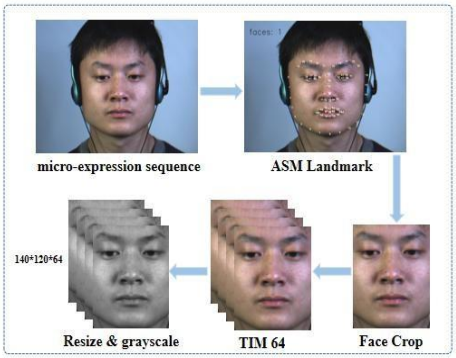

Fig. 2 Video processing.

B. RoIs Selection

Related psychological research has shown that when there are micro expressions, certain partial facial areas (eyes, eyebrows and mouth) change relatively significantly. Although a whole face image can represent the integrity of micro-expression, but RoIs also have their advantages for three reasons:

- Delete regions that are irrelevant or inactive for micro- expressions to better focus on active parts;
- Reduce existing background noise which may affect feature extraction to improve recognition performance;

- Save calculation time and speed up the feature extraction process due to the small input.

TABLE I. EMOTION DESCRIPTION IN TERMS OF FACIAL ACTION UNITS

Emotions	AUs
Happiness	either AU6 or AU12
Disgust	AU9, AU10, AU4 must be present
Repression	AU14, AU15 or AU17 alone or combination
Sadness	AU1 must be present
Surprise	AU1+2, AU25 or AU2 must be present
Fear	AU1+2+4 or AU20 must be present
Others	Other emotion-related facial movement

According to FACS, expressions are encoded by Action Units (AUs) [17]. These AUs represent some tiny but effective facial muscle changes. Therefore, the RoIs is selected empirically according to the frequency of AUs. TABLE I shows the AUs corresponding to the emotions provided in [18] [19]. Before get RoIs, we divide micro- expression pictures into 12 × 14 small pieces with a length and width both of 10 pixels. Then cut out the regions of "Eye + Eyebrow" and "Mouth" from micro-expression images, as shown in Fig. 3. Table II shows the correspondence relationship between the selected RoIs, AU(s) and emotions.

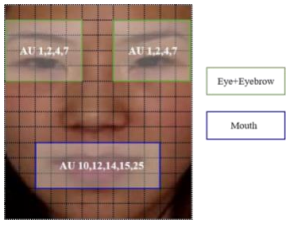

Fig. 3 Correspondence between RoIs and AUs.

TABLE II. EMOTION DESCRIPTION IN TERMS OF AU(S) AND ROIS

RoIs	AU(s)	Emotions
Eye + Eyebrow	1,2,4,7	Sadness, Disgust, Surprise, Tense, Fear
Mouth	10,12,14,15,25	Disgust, Tense, Surprise, Repression, Contempt, Happiness

C. Feature Extraction

Residual network is presented by He et al. [20], and achieve good performance in many tasks related to computer vision in the recent years. In general, Residual network use stacks of residual blocks to build the network. This design can improve network performance with fewer training parameters. By superimposing this residual block, the whole network will be more robust. Formally, residual block can be defined as:

$$y = \mathcal{F}(x, \{\mathrm{W_i}\}) + x \tag{1}$$

where x, y are the input and output vectors of the residual block. In our work, the shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers.

As shown in Part 1 of Fig. 1, we use the residual network to extract the features of the whole face and RoIs of each frame in the video sequence, so as to obtain local changes and global change. After extracting, the feature of whole face (h_j^i), left eye (l_j^i), right eye (r_j^i) and mouth (m_j^i) will be extracted, where i is the video number and j is the frame number in ith video. In the concatenate layer as shown Part 2 of Fig. 1, the frame-level feature is obtained:

$$\mathcal{L}_j^i = \{h_j^i, l_j^i, r_j^i, m_j^i\} \tag{2}$$

the feature vector $\mathcal{L}_j^i$ will be fed into two fully connected layers for feature fusion. A softmax layer is followed by the fully-connected layer, which maps the output of the previous layer to the expression class. We select the vector of the first fully-connected layers as the frame-level feature vector, So, all the frame-level features in ith video can be represented as $\mathcal{S}_i$:

$$\mathcal{S}_\mathrm{i} = (\mathcal{L}_1^\mathrm{i}; \mathcal{L}_2^\mathrm{i}; \ldots; \mathcal{L}_\mathrm{N}^\mathrm{i})^\mathrm{T} \tag{3}$$

and N represents the length of the video frame. In order to aggregate all the frame-level features $\mathcal{S}_\mathrm{i}$ into video-level features $\mathcal{Z}_\mathrm{i}$. We propose an aggregation strategy, which makes full use of statistics, as shown in formula 4:

$$\mathcal{Z}_\mathrm{i} = \{\mathrm{mean}(\mathcal{S}_\mathrm{i},), \mathrm{std}(\mathcal{S}_\mathrm{i},), \mathrm{median}(\mathcal{S}_\mathrm{i})\} \tag{4}$$

where the function $mean$ represents mean, std means standard deviation, $median$ is a used to find the median.

Finally, we use LIBSVM to classify the $\mathcal{Z}_\mathrm{i}$ and use the result as the recognition result of the ith video clip.

IV. EXPERIMENTS

In this section, we will first introduce the database used in experiment, followed by the experimental setup, and experimental results and discussion are presented at last.

A. Database

The number of samples of spontaneous micro-expression databases is very small. There are three databases used in this work: CASME [18], CASME II [19] and SMIC [21]. The CASME database contains 195 spontaneous micro-expression video clips. Another micro-expression database, CASME II is an extension of CASME and contains a sample of 247 micro-expressions clips from 26 participants. SMIC contains 164 micro-expression clips induced by 16 participants. In TABEL III, we enumerate the number of participants, sample size, and emotion categories contained in the three databases in detail.

TABLE III. DETAILS OF THE THREE DATABASES

database	Size	Emotions	subject
CASME	195	disgust, surprise, tense, contempt, repression, sadness, fear, happiness	35
CASME II	247	disgust, fear, sadness, surprise, happiness, other	26
SMIC	164	repression, positive, negative, surprise	16

B. Experiment Setup

a) Emotion categories: TABLE III shows the details of CASME, CASME II, and SMIC databases, but each database has different emotion categories and the number of individual database is too small to cause overfitting when using deep learning to extract features. Following the recommended strategy [22], we've regrouped the primal emotions into a large database containing four categories, and map the labels to a new label space. The corresponding relationship is: *Positive* = {happiness}, *Surprise* = {surprise}, *Negative* = {disgust, fear, sadness} and those facial movements with unclear emotions are classified as *Others*.

b) Five-folds cross-validation: Leave-one-subject-out (LOSO) cross-validation applies to the database, which has limited subjects. However, the number of subjects in our experiments is too much. Meanwhile, because training deep models is time-consuming, it is difficult to test every sample in our experiment. Therefore, we use a 5-fold cross-validation protocol to evaluate the proposed approach on the new database. In 5-fold cross-validation protocol, we randomly divide the subject into five parts, four of which are used for training and the rest for testing.

c) Recognition framework: The recognition framework of micro-expression is shown in Fig. 1, Part 1 is a network composed of residual network modules. Next is the concatenate layer that can connect the global and local features. Later, two fully-connected layers with 64 and 4 units are used. Each fully-connected layer is followed by a drop layer with a dropout ratio of 0.5. For the activation function, the rectified linear unit (ReLU) is used. The network was trained for 500 epochs with a learning rate of 0.0005. Then as shown in the red box in the Part 2 of Fig. 1, we select the feature of the first fully connected layer as the frame-level feature. In the Part 3, we propose an aggregation strategy which fully utilizes the physical meaning of statistical parameters to characterize the average and dynamic changes of frame-level features to generate video-level features, and uses SVM for micro-expression classification.

C. Experimental Results and Discussion

Based on the above settings, we evaluated the aggregation strategy, and conducted an experimental comparison between the global image and each RoI. In addition, we compared the proposed method with the previous methods.

1) Comparison of Different Aggregation Results

To further evaluate the effectiveness of the proposed aggregation strategy, we use different functions to fuse frame-level features into video-level features, such as mean, standard deviation (std), median, and possible combinations between them. The results are shown in Table IV.

From this experimental result, it can be seen that the video-level futures obtained by standard deviation are poor, and the recognition results obtained by the mean and median are similar. This may be because the labels of each frame in the same video are consistent, so that the difference in frame- level features is relatively small. However, it is very obvious that when the video features obtained by using the standard deviation and mean are concatenated together, the experimental result is further improved, which shows that calculating the standard deviation can reflect the dynamic changes in one video and promote to improve the recognition accuracy. The concatenation of median and standard deviation is similar. Further, we stitched together the three statistical results to obtain the best recognition performance. This is mainly because it not only contains the average level of each frame feature, but also captures the dynamic changes between frames.

TABLE IV. MICRO-EXPRESSION RECOGNITION ACCURACY OF AGGREGATION STRATEGY BETWEEN FRAMES

Terms	Recognition accuracy (%)
mean	60.24
std	57.09
median	60.14
mean + std	61.51
mean + median	62.14
std + median	61.54
mean + std + median	**63.51**

2) Results of Local and Global Features

In Fig. 4, we list the recognition accuracy using each RoI feature, the whole face feature, and aggregation feature. When the recognition result of the global feature is taken as the baseline, the results of "Left eye + Eyebrows", "Right eye + Eyebrows" and "Mouth" is lower than the baseline. This is because the baseline uses the whole face to extract features, containing multiple RoIs features. However, the recognition rate after combining the features of whole face and each RoIs is 5.81% higher than the baseline. This result proves that some distinguishing details can be captured through RoIs, which is complementary to the overall characteristics. In other words, the features extracted from RoIs have positive prospects in MER, and the combination of local information and holistic information can achieve better recognition results.

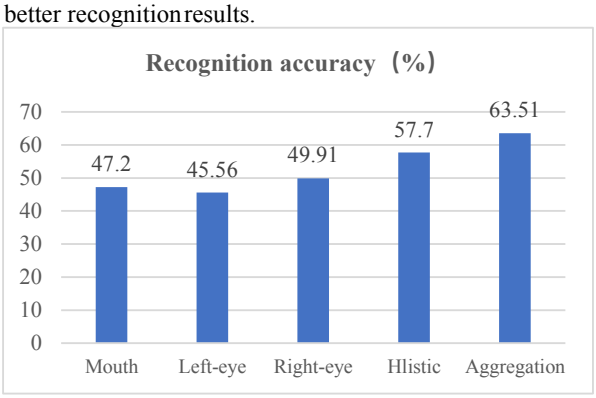

Fig. 4 Detection accuracy of Global feature, local feature and fusion feature

3) Comparison with Other Methods

Our method has been compared with may of the state-of-the-art methods, and the result is illustrated in TABLE V. The methods mentioned in the table use the same database collections and verification method.

As can be seen from TABLE V, our method is 0.86% and 16.90% higher than MDMO and FDM, respectively. Both methods use the hand-craft features. Although they cut the face into multiple RoIs, they use each RoI instead of focusing on specific parts. Different from them, our method is not only to extract high-level features using deep learning, but also to use specific RoIs to focus on local changes. Therefore, on the one hand, the features extracted by the neural networks have a positive effect on micro-expression recognition tasks. On the other hand, the RoIs that we choose are helpful for MER tasks, and concatenate them together with the holistic features to form a complement to further improve the recognition accuracy.

At the same time, in work [9], Wang et al. use 3D-CNN to extract high-level features from micro-expression sequences, but these features are not enough to capture the subtle local changes in the face. Our method pays attention to local features and fuse them, and finally achieves better results. This proves that it is necessary to pay attention to relatively significant local facial regions when exploring the MER task.

TABLE V. PERFORMANCE COMPARISON WITH EXISTING METHODS

Methods	Recognition accuracy (%)
3D-CNN	62.68
MDMO	62.65
FDM	46.61
Our method	**63.51**

V. CONCLUSION

In this paper, a novel hybrid framework which combines the heuristic and automatic approaches is proposed. The framework of the framework contains four branches. One of the branches performs holistic feature extraction on the whole micro-expression frames, while the other three branches extract the local features from each RoI. The choice of RoIs is determined by the frequency of occurrence of AU in all expressions. We combine the holistic features and local features to generate frame-level features, and aggregate the frame-level features of each video to obtain video-level features for classification. Experiments were conducted on three spontaneous micro-expression databases to prove the effectiveness of our method. However, when integrating the temporal information, we used statistical methods. In the future, we hope to use the neural network to directly obtain temporal information to achieve end-to-end micro-expression recognition.

ACKNOWLEDGMENT

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002804), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473).

REFERENCES

[1] P. Ekman and W. Friesen, "Nonverbal leakage and clues to deception," Psychiatry, vol. 32, no. 1, pp. 88–106, 1969.

[2] P. Ekman, "Darwin, deception, and facial expression," Annals of the New York Academy of Sciences, vol. 1000, no. 1, pp. 205–221, 2003.

[3] M. G. Frank, C. J. Maccario, and V. Govindaraju, "Behavior and security," Protecting Airline Passengers in the Age of Terrorism, Greenwood Pub Group, Santa Barbara, California, pp.86–106, 2009.

[4] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: training laypeople and professionals to recognize fleeting emotions," in the Annual Meeting of the International Communication Association, 2009.

[5] P. Seidenstat and F. X Splane, "Protecting airline passengers in the age of terrorism," ABC-CLIO, 2009.

[6] 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, pp. 1-1, 2015.

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

[8] M. Niu, J. Tao, Y. Li, J. Huang and Z. Lian, "Discriminative video representation with temporal order for micro-expression recognition," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, pp. 2112-2116, 2019.

[9] S. J. Wang, B. J. Li, Y. J. Liu, et al., "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," Neurocomputing, 312(OCT.27), pp. 251-262, 2018.

[10] S. Porter, and L.T. Brinke, "Reading between the lies identifying concealed and falsified emotions in universal facial expressions," Psychological Science, vol. 19, no. 5, pp. 508-514, 2008.

[11] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in Proc. of the ACM MM. ACM, pp. 382–386, 2016.

[12] P. Ekman, W. V. Friesen, and J. C. Hager, "FACS manual," Salt Lake City, UT, USA: A Human Face, May 2002.

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

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Computer Vision and Pattern Recognition, pp. 511-518, 2001.

[15] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," Computer vision and image understanding, vol. 61, no. 1, pp. 38–59, 1995.

[16] X. Li, T. Pfifister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in IEEE FG, pp. 1–6, 2013.

[17] P. Ekman, W. Friesen, "Facial Action Coding System," 1977.

[18] W. Yan, Q. Wu, Y. Liu, S. Wang, and X. Fu, "CASME Database: a dataset of spontaneous micro-expressions collected from neutralized faces", 10th IEEE conference on automatic face and gesture recognition, Shanghai IEEE, pp. 1–7, 2013.

[19] W.J. Yan, X. Li, S.J. Wang, et al., "CASME II: an improved spontaneous micro-expression database and the baseline evaluation," PLOS ONE, vol. 9, no. 1, pp. 1-8, 2014.

[20] K. He, X. Zhang, S. Ren, et al, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[21] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," international conference on computer vision, 2011.

[22] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition, " Frontiers in Psychology, pp. 1745-1745, 2017