Local Bilinear Convolutional Neural Network for Spotting Macro- and Micro-expression Intervals in Long Video Sequences

Hang Pan, Lun Xie, Zhiliang Wang
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
E-mail: xielun@ustb.edu.cn

***Abstract*—To reduce the impact of low intensity on the spot micro-expressions in long video sequences when facial micro-expressions occur, this paper presents a method based on Local Bilinear Convolutional Neural Network (LBCNN) for spotting macro- and micro-expressions in long videos sequences. Considering the low intensity of facial micro-expressions, and the occurrence of micro-expressions is only related to the local area of the facial, we turn the micro-expression spot in long videos sequences into a fine-grained image recognition. Bilinear Convolutional Neural Network (BCNN) has proven to be effective in fine-grained image recognition. Therefore, we use the BCNN structure to extract the global and local features of the face area of each frame of the image in the long video sequence and obtains the final classification result by fusing the global and local features. The metric F1-scores of our proposed methods were evaluated on the CAS(ME)^2 and SAMM Long Videos dataset of the Third Facial Micro-Expression Grand Challenge (MEGC 2020). For the CAS(ME)^2, the overall F1-scores are 0.0595 for macro- and micro-expressions; for SAMM Long Videos, the overall F1-scores are 0.0813 for macro- and micro-expressions. The experiments show that our method achieved superiorly higher results than the baseline method (MDMD) provided. https://github.com/panhang1023/MEGC2020**

I. INTRODUCTION

With the development of artificial intelligence and affective computing, Facial expression recognition has been applied in various environments [1], [2], [3]. However, in special high-stakes environments, such as criminal investigations [4], communication negotiations [5], and mental illness diagnosis [6], special people often try to restrain or hide their true emotions, and facial macro-expression recognition is not applicable. However, when people try to restrain or hide the inner real emotions, the facial expressions involuntarily revealed are not easily detected, which are called micro-expressions. The micro-expression is almost not transferred by the will of the human being and can reflect the true emotional state of the person [7]. Therefore, Spotting and recognition micro-expression intervals in long video sequences is an emerging area in face research.

The occur of facial expression is manifested in the movement of facial muscles. In order to discover the regularity of muscle movement in facial expressions, Ekman et.al. designed a Facial Action Coding System (FACS) to encode the muscle movements of facial action units (AU) to determine facial expression classification [8]. Pfister et al. [9] first proposed the spontaneous facial micro-expressions recognition framework and that use Local Binary Pattern-three Orthogonal Planes (LBP-TOP) to handle dynamic features and Support Vector Machine (SVM) to perform classification. Wang et al. [10] proposed a Local Binary Pattern with Six Intersections Point (LBP-SIP) to reduce the information redundancy of the LBP-TOP feature for micro-expression classification. Li et al. [11] proposed to use deep convolutional neural network (DCNN) on the apex frame to recognize micro-expression. Huang et al. [12] employed a discriminative spatiotemporal local binary pattern based on an integral projection model for micro-expression recognition. Considering the subtle spatiotemporal changes of micro-expressions, Xia et al. [13] proposed a new type of deep recursive convolutional network to capture the spatiotemporal deformation of micro-expression sequences.

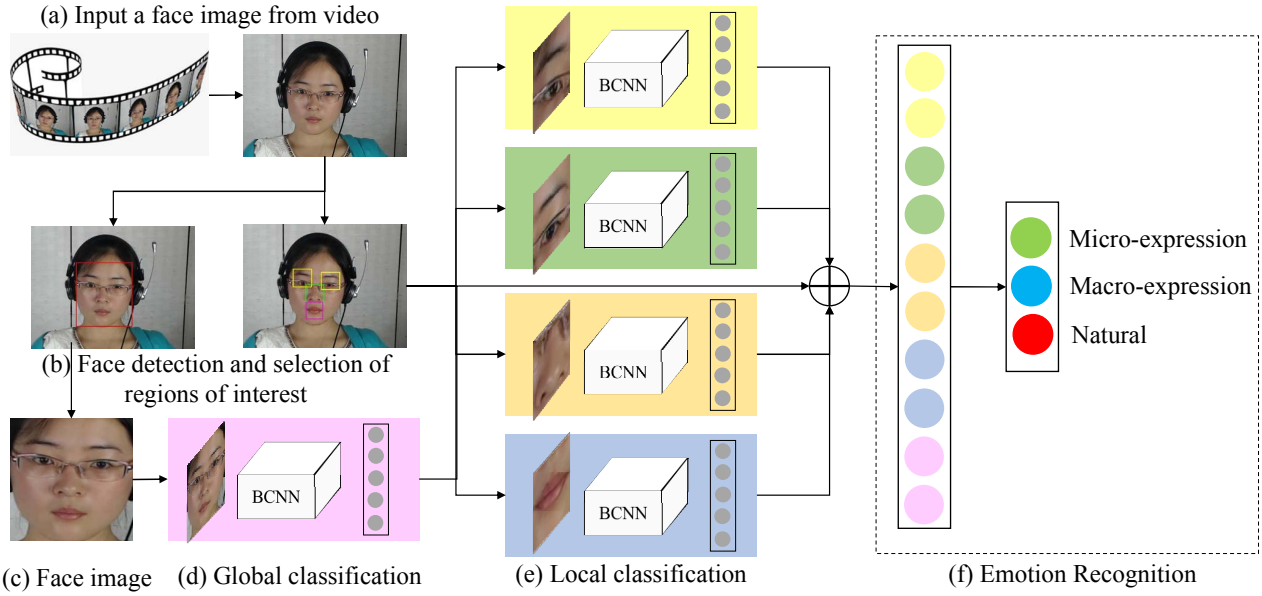
Although some research works have achieved surprising results in improving macro-expression recognition. However, the research on the Spotting micro-expressions is still far behind compared with micro-expression recognition.

When facial expressions images obtained by the imaging device are used for automatic spotting and recognition of micro-expressions. Micro-expression facial muscle movement only occurs in the local area of the face and the problem of low intensity, resulting in a small difference between the micro-expression facial image and the natural facial image. These problems bring great challenges to the spotting and recognition of micro-expressions in long video sequences.

Considering that the micro-expression is only related to the local area of the face and the problem of low intensity, which results in the small inter-class variation of micro-expressions facial image and the natural facial image. We convert the micro-expression spotting problem in long video sequences into fine-grained image classification problems, and propose a local expression based on Local Bilinear Convolutional Neural Network (LBCNN) model for spotting macro- and micro-expression intervals in long video sequences. As illustrated in Fig. 1, given a long video sequence, the face image of each frame is extracted in the preprocessing stage, and the face image and the regions of interest of the micro-expression are input to a bilinear convolutional neural network model. Finally, global and local features of the region of interest are fused to get the final output. The precision, recall, and F1 scores our model obtained in this challenge show that our method outperforms the provided Main Directional Maximal Difference Analysis (MDMD) baseline method.

In general, this paper attempts to propose a Local Bilinear Convolutional Neural Network model for solving the problem

Fig. 1. The complete framework for spotting macro- and micro-expressions based on Local Bilinear Convolutional Neural Network. In a long video, we perform face detection on each frame of image and select regions of interest (a)-(b) to extract regions of interest. The obtained facial region is sent to the BCNN, and the overall feature parameter (c) is output. We obtain local feature parameters (d) through BCNN in the same way. Finally, the global and local features are fused to obtain the emotion type (f).



of low intensity of micro-expressions during the spot micro-expressions in long video sequences. The main contributions of this paper are summarized as follows:

1. The influence of low intensity of micro-expression spotting in long video sequences is analyzed, and a local bilinear convolutional neural network model is proposed.

2. The superior performance of the LBCNN model is verified on two public micro-expression datasets.

The remainder of this paper is organized as follows: In Section 2, a brief review of related research on micro-expressions recognition. Section 3 introduces the datasets and proposed algorithms. Section 4 shows the detailed experimental results and details. Finally, Section 5 presents the conclusions of this research method.

II. Related Work

In 1966, Haggard and Isaacs [14] discovered a fleeting, undetectable facial expression. After 1969, Ekman et al. [15] conducted a series of studies on micro-expressions and proposed the Brief Affect Recognition Test (BART) method.

In 2017, Qu et al. [16] considered that spotting micro-expression in long video sequences has shown great potential as a promising cue for deception detection, and present a new database, Chinese Academy of Sciences Macro-Expressions and Micro-Expressions (CAS(ME)2), which used to spotting and recognition macro and micro-expressions in long video sequences. Yap et al. [17] released a richer spontaneous micro-expression dataset, Spontaneous Actions and Micro-Movements (SAMM) for micro-expression spotting and recognition, and subsequently released an updated database version SAMM Long Videos [18].

In order to solve the problem of evaluation standards for the micro-expression spotting method, Tran et al. [19] constructed a multi-scale evaluation benchmark based on a sliding window for fairly and better evaluate the micro-expression spotting approaches. Li et al. [20] proposed the first automatic ME analysis system (MESR), which can spot and recognize MEs from spontaneous video data. In The Second Facial micro-expressions Grand Challenge (MEGC2019), the micro-expression spotting challenge task in long video sequences was performed for the first time in two databases of CAS (ME) 2 and SAMM. Li et al. [21] used local temporal patterns (LTP-ML) [22] for spontaneous micro-expression spotting, which achieved better experimental results than the state of the art LBP-χ2-distance (LBP-χ2) method [23]. These datasets and challenges established the foundation for spotting macro-expressions and micro-expressions intervals in long video sequences.

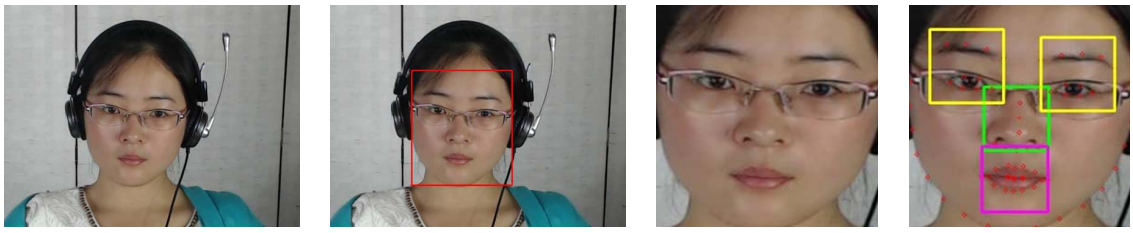
III. Methodology

A. Datasets

Two micro-expression database CAS(ME)2 [16] and SAMM Long Videos [18] based on the induction method are used for the third facial micro-expressions grand challenge (MEGC 2020). Due to the particularity of micro-expressions, there are two notable elements in the micro-expressions dataset: PFS and resolution.

Because the duration of micro-expressions is short, most micro-expression datasets collected by a high-speed camera. But it is worth noting that the CAS(ME)2 database records at 30 frames per second, and the SAMM records at 200 frames per second. The higher the image resolution, the easier it is to capture detailed information about facial expressions. CAS(ME)2v is a low-resolution database and SAMM is a high-resolution database. These inconsistencies make it

Fig. 2. Face detection and selection of regions of interest. (a) The original image; (b) Face detection; (c) Face cropping; (d) Selection of regions of interest.



(a) (b) (c) (d)

difficult to extract features. Table. I list the inconsistencies in frame rate and resolution between these databases that explain the MEGC 2020 challenge.

TABLE I. FPS AND RESOLUTION COMPARISON OF TWO MICRO-EXPRESSION DATASETS CAS(ME)2 AND SAMM LONG VIDEOS.

Dataset	FPS	Resolution
CAS(ME)2	30	640×480
SAMM Long Videos	200	2040×1088

The goal of this challenge is to spot macro- and micro-expressions intervals in long video sequences. For this challenge, we focus on two databases CAS(ME)2 and SAMM Long Videos. In part A of CAS(ME)2 database, there are 22 subjects and 98 long videos, including 300 macro-expressions and 57 micro-expressions. The facial movements are classified as macro- and micro-expressions. The video samples may contain multiple macro or micro facial expressions. The onset, apex, offset index for these expressions are given in the excel file. In addition, eye blinks are labeled with the onset and offset time. The original SAMM dataset [17] with 159 micro-expressions. In the SAMM Long Videos dataset [18], there are 32 subjects and 147 videos, including 343 macro-expressions and 159 micro-expressions. The index of onset, apex and offset frames of micro-movements are outlined in the ground truth excel file. The micro-movements interval is from the onset frame to the offset frame. In this database, all the micro-movements are labeled. The sample distribution of the two databases is shown in Table. II.

TABLE II. SAMPLE DISTRIBUTION OF TWO MICRO-EXPRESSION DATASETS CAS(ME)2 AND SAMM LONG VIDEOS.

Dataset	CAS(ME)^2	SAMM Long Videos
Participants	22	32
Video samples	98	147
Macro-expressions	300	343
Micro-expressions	57	159

B. Local Bilinear Convolutional Neural Network

So far, the literatures [24], [25], [26], [27], [28], [29] have verified the feasibility of deep learning in micro-expression recognition. However, the work of micro-expression through deep learning is still relatively small. On the one hand, there are fewer samples in the data set, and on the other hand, the low-intensity characteristics of micro-expressions. Considering these situations, we transform the micro-expression localization problem into a fine-grained image classification problem. Each frame of the long video is input to a local bilinear convolutional neural network for classification, and the emotion category of each frame of image is obtained. Perform micro-expression positioning.

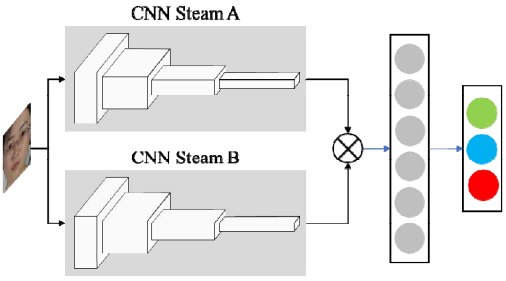
1) Preprocess: This competition used the CAS (ME) 2 and SAMM Long Videos databases for experimental verification. We first uniformly sample the micro-expression, macro-expression, and natural images in the long video according to the provided tag file as training samples. For clarity, the notations used in this paper are defined and explained in the following sections. The image sequence of the long video in the database is represented as:

$$V=\{f_1, f_2, \ldots, f_n\} \quad (1)$$

Where n is the number of frames of the long video. Each long video sample is expressed as:

$$V=\left\{\begin{array}{l}\left(f_{micro_1}, f_{micro_2}, \ldots, f_{micro_k}\right)\\ \left(f_{macro_1}, f_{macro_2}, \ldots, f_{macro_k}\right)\\ \left(f_{natural_1}, f_{natural_2}, \ldots, f_{natural_k}\right)\end{array}\right\} \quad (2)$$

Fig. 3. The framework of Bilinear Convolutional Neural Network.



Due to the micro-expression localization, more attention is paid to the small changes in the local area of the face. Therefore, we use the face detection method in MDMD [30] to

obtain the facial area and the micro-expression region of interest ("left eyebrows", "right eyebrows", "nose" and "mouth"). Face detection and selection of regions of interest are shown in Fig. 2.

2) Bilinear Convolutional Neural Network: Bilinear model is a fine-grained image classification model proposed by Lin et al. [31]. This model uses two parallel Convolutional Neural Network (CNN) models. This CNN model uses AlexNet [32] or VGGNet [33] to remove the last fully connected layer and the softmax layer. This is used as a feature extractor, and then SVM is used as the final one. Linear classifier. This architecture can model local pairwise feature interactions in a translation-invariant manner and is suitable for fine-grained classification [34].

Considering the low intensity of micro-expressions, we use a bilinear model to classify the entire facial area and local regions of interest in each frame of the long video sequence. The structure of the bilinear model is shown in Figure 3. In order to reduce overfitting during the training process, we use a shallower convolutional neural network structure to modify the AlexNet network structure. After inputting the image, three convolutional layers and Two downsampling layers and two fully connected layers. We adjust the image input to the model to 224 * 224 * 3 each time, and the output is micro-expression, macro-expression, and natural.

IV. Result and Discussion

A. Performance metrics

1) True positive in one video definition: The true positive (TP) per interval in one video is first defined based on the intersection between the spotted interval and the ground-truth interval. The spotted interval $W_{spotted}$ is considered as TP if it fits the following condition:

$$\frac{W_{spotted} \cap W_{groundTruth}}{W_{spotted} \cup W_{groundTruth}} \geq k \quad . \tag{3}$$

where k is set to 0.5, $W_{groundTruth}$ represents the ground truth of the macro- or micro-expression interval (onset-offset). If the condition is not fulfilled, the spotted interval is regarded as false positive (FP).

2) Result evaluation in one video: Supposing there is m ground truth interval in the video, and n intervals are spotted. According to the overlap evaluation, the TP amount in one video is counted as a ($a \leq m$ and $a \leq n$), therefore FP = n - a, FN = m - a. The spotting performance in one video can be evaluated by the following metrics:

$$recall = \frac{a}{m}, precision = \frac{a}{n} \quad . \tag{4}$$

$$F1-score = \frac{2TP}{2TP + FP + FN} = \frac{2a}{m+n} \tag{5}$$

Yet, the videos in real life have some complicated situations which influences the evaluation per single video:

- There might be no macro- nor micro-expression in the test video. In this case, m = 0, the denominator of recall would be zeros.
- If there is no spotted intervals in the video, the denominator of precision would be zeros since n = 0.
- It is impossible to compare two spotting methods when both TP amounts are zero. The metric (recall, precision or F1-score) values both equal to zeros. However, the Method1 outperforms Method2, if Method1 spots less intervals than Method2. Thus, to avoid these situations, we propose for single video spotting result evaluation, we just note the amount of TP, FP, and FN. Other metrics are not considered for one video.

B. Performance metrics

For the experimentally verified databases CAS(ME)^2 and SAMM Long Videos, we use Leave-One-Subject-Out (LOSO) cross-validation to train the LBCNN model, predicting macro- and micro-expression intervals in long video sequences that each subject in the database contains macro- and micro-expressions and macro expression fragments of long video sequences for each subject in the database. The details of the final baseline results for spotting macro- and micro-expressions are shown in Table III. For CAS(ME)^2, the overall F1-scores are 0.0595 for macro- and micro-expressions. For SAMM Long Videos, the overall F1-scores are 0.0813 for macro- and micro-expressions. More details about the number of true labels, TP, FP, FN, precision, recall, and F1-score for various situations are shown in Table III.

V. Conclusions

This paper focuses on the low intensity of micro-expressions, and proposes a method based on Local Bilinear Convolutional Neural Network (LBCNN) for spotting macro- and micro-expressions in long video sequences for the Third Facial Micro Expression Spotting Challenge (MEGC 2020). The BCNN structure is used to extract the global and local features of the facial area of each frame of the image is a long video sequence, and to obtain the emotion type of the image by fusing the global and local features. The results have shown

TABLE III. Sample Distribution of Two Micro-expression Datasets CAS(ME)2 and SAMM Long Videos..

Method	TP	FP	FN	Precision	Recall	F1-score
MDMD (CAS(ME)^2)	130	6428	335	0.0198	0.3641	0.0376
LBCNN (CAS(ME)^2)	35	872	234	0.0386	0.1301	0.0595
MDMD (SAMM Long Videos)	51	1741	451	0.0285	0.1016	0.0445
LBCNN (SAMM Long Videos)	29	397	258	0.0681	0.1010	0.0813

that the experimental results of our method overstepped the provided baseline method (MDMD).

Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2018YFC 2001700), and the National Natural Science Foundation of China (No. 61672093), and Advanced Innovation Center for Intelligent Robots and Systems Open Research Project (No.2018IRS01).

References

[1] A. M. Shabat, J. R. Tapamo. Angled Local Directional Pattern for Texture Analysis with an Application to Facial Expression Recognition. *IET Computer Vision*, vol. 12, no. 5, pp. 603-608, 2018.

[2] M. Liu, S. Shan, R. Wang, and X. Chen. Learning Expressionlets on Spatio-Temporal Manifold for Dynamic Facial Expression Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1749-1756.

[3] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2983-2991.

[4] P. Ekman. Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition). *WW Norton & Company*, 2009

[5] T. A. Russell, E. Chu, M. L. Phillips. A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. *British journal of clinical psychology*, vol. 45, no. 4, pp. 579-583, 2006.

[6] F. Salter, K. Grammer, A. Rikowski. Sex differences in negotiating with powerful males. *Human Nature*, vol. 16, no. 3, pp. 579-583, 2005.

[7] M. Frank, M. Herbasz, K. Sinuk. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. *Annual Meeting of the International Communication Association*. 2009.

[8] P. Ekman. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). *Oxford University Press*, 1997.

[9] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen. Differentiating Spontaneous from Posed Facial Expressions within a Generic Facial Expression Recognition Framework,. *IEEE International Conference on Computer Vision (CVPR) Workshops*, Nov 2011, pp. 868-875.

[10] Y. Wang, J. See, R. C. W. Phan, and H. Yee. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. *Asian Conference on Computer vision (ACCV),* 2014, pp: 525-537.

[11] Y. Li, X. Huang, G. Zhao. Can micro-expression be recognized based on single apex frame? *IEEE International Conference on Image Processing (ICIP)*, 2018: 3094-3098.

[12] X. Huang, S. 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. 10, no. 1, pp. 32-47, 2019.

[13] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia,* 2019.

[14] E. A. Haggard, S. I. Kenneth. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. *Methods of research in psychotherapy*. 1966. pp:154-165.

[15] P. Ekman, V. F. Wallace. Nonverbal leakage and clues to deception. *Psychiatry*, vol. 32, no, 1, pp: 88-106, 1969.

[16] 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. 9, no, 4, pp: 424-436, 2017.

[17] K. D. Adrian, L. Cliff, C. Nicholas, T. Kevin, and H. Y. Moi. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, vol, 9, no, 1, pp:116-129, 2018.

[18] H. Y. Chuin, K. Connah, H. Y. Moi.. Samm long videos: A spontaneous facial micro- and macro-expressions dataset. *arXiv preprint arXiv*:1911.01519, 2019.

[19] T. K. Tran, X. Hong, G. Zhao. Sliding window based micro-expression spotting: a benchmark. *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*. 2017, pp: 542-553.

[20] X. Li, X. Hong, A. Moilanen A, X. Huang, T. Pfister, G. Zhao. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, vol, 9, no, 4, pp: 563-577, 2017.

[21] J. Li, C. Soladie, R. Seguier, S-. J. Wang and H. Y. Moi. Spotting micro-expressions on long videos sequences. *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 2019, pp: 1-5.

[22] J. Li, C. Soladie, R. Seguier. LTP-ML: Micro-expression detection by recognition of local temporal pattern of facial movements. *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018, pp: 634–641.

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

[24] Y. S. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, vol, 74, pp: 129-139, 2019.

[25] Q. Li, S. Zhan, L. Xu, C. Wu. Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. *Multimedia Tools and Applications*, vol, 78, no, 20, pp: 29307-29322, 2019.

[26] Y. Liu, H. Du, L. Zheng T. Gedeon. A neural micro-expression recognizer," *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2019.

[27] S. T. Liong, Y. Gan, J. See, H. Q. Khor, and Y.-C. Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. *IEEE International Conference on Automatic Face & Gesture Recognition (FG),* 2019.

[28] L. Zhou, Q. Mao, L. Xue. Dual-inception network for crossdatabase micro-expression recognition. *IEEE International Conference on Automatic Face & Gesture Recognition (FG),* 2019.

[29] N. V. Quang, J. Chun, T. Tokuyama. Capsulenet for microexpression recognition. *IEEE International Conference on Automatic Face & Gesture Recognition (FG),* 2019.

[30] Y. He, S. J. Wang, J. Li and H. Y. Moi. Spotting Macro-and Micro-expression Intervals in Long Video Sequences. *arXiv preprint arXiv*:1912.11985, 2019.

[31] T. Y. Lin, A. RoyChowdhury, S. Maji. Bilinear cnn models for fine-grained visual recognition. *IEEE International Conference on Computer Vision (ICCV)*. 2015, pp: 1449-1457.

[32] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. *International Conference on Neural Information Processing Systems (NIPS)*. 2012, pp:1097-1105.

[33] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556, 2014.

[34] T. Y. Lin, A. RoyChowdhury, S. Maji. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, vol, 40, no, 6, pp: 1309-1322, 2017.