DynGeoNet: Fusion network for micro-expression spotting

Thuong-Khanh Tran
Center for Machine Vision and Signal Analysis, University of Oulu
Oulu, Finland
khanh.tran@oulu.fi

Quang-Nhat Vo
Silo.AI
Helsinki, Finland
nhat.vo@silo.ai

Guoying Zhao*
Center for Machine Vision and Signal Analysis, University of Oulu
Oulu, Finland
guoying.zhao@oulu.fi

ABSTRACT

Micro-expressions (MEs) are brief and involuntary facial expressions when people hide their true feelings or conceal their emotions. Based on psychology research, MEs play an important role in understanding genuine emotions, which leads to many potential applications. However, the ME analysis system can still not work well in the real environment because of the challenging performance of ME spotting, which is to spot the images with micro-expressions from long video sequences. To improve the performance of ME spotting, we focus on hybrid feature engineering, which aims to create a robust feature for discriminating tiny movements. The proposed framework consists of two main modules: (1) the feature engineering extracts both geometric features and appearance features based on dynamic image; (2) the new deep neural network inputs the handcrafted feature for the late fusion and ME samples classification. Our experimental results from three baseline datasets demonstrate the promising results.

CCS CONCEPTS

• **Computer systems organization** → **Computer vision**; *Redundancy*; Machine learning; • **Computer Vision** → Affective Computing.

KEYWORDS

micro-expression spotting, multi-model analysis, emotion analysis

ACM Reference Format:
Thuong-Khanh Tran, Quang-Nhat Vo, and Guoying Zhao*. 2022. DynGeoNet: Fusion network for micro-expression spotting. In *Proceedings of* . https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

MEs are brief and involuntary facial expressions that convey the hidden emotions of people. Through psychology research [4, 10], ME analysis has become an attractive topic due to their potential applications [10, 20]. However, building a real system of ME analysis still faces big challenges. One of the defiance is locating the correct positions of ME in long videos that is one of two main tasks of ME analysis in computer vision, called ME spotting.

Through the existing research of ME spotting, there are various techniques involved in this topic, but there still remain several issues [15, 20]. The most critical thing is the poor performance of ME spotting techniques evaluated on long natural video sequences.

In the recent research [20], the authors evaluated diverse techniques for ME spotting tasks but obtained poor results on three commonly used ME spotting datasets. Based on the experimental results of [9, 15, 20], almost all methods returned a high false positive rate, although the state-of-the-art techniques were utilized. One reason is that the number of ME samples is not enough for deep learning methods, which often require huge data samples. In three popular ME spotting datasets, there are only around 300 ME samples in various modals [15, 20, 24]. A new ME spotting database construction is time consuming and often requires well-trained experts to annotate the ME samples. Therefore, we focus on the hybrid models which can discriminate ME samples from other extrinsic movements.

The studies in [12, 16, 18] raise the issue that the combination of handcrafted feature and deep learning technique is a reasonable approach to handle the problems of limited data. In these works, the handcrafted features are utilized as the first step to extract the discriminative information, then the robust tools from deep learning are used to learn the hybrid model for classification. Inspired by this approach, we select dynamic image features due to their impressive performance on the ME recognition task [16]. Furthermore, the output of a dynamic image is the compact representation that captures motion information in a single image which can take advantage of the state-of-the-art image classification architectures. In the research of Niu et al. [12], the geometric feature combining with deep feature obtained promising results for the facial analysis problem. Thus, the geometric feature is a feasible option if we consider the combination.

Overall, we propose a framework for ME spotting, which constructs a new ME feature representation by the dynamic image and deformable feature of facial landmark points. Then, a multi-model deep learning architecture that incorporates the VGG-16 backbone and a multi-layer perception module is utilized to learn the extracted face information end-to-end. The model output is a binary classification of ME and non-ME samples.

The contribution of our works are summarized as follows:

- (1) We propose a new feature engineering fusing dynamic appearance and geometric features for discriminating ME motion samples.
- (2) We learn a specialized deep neural network to take input from the fusion between geometric features and appearance

* Corresponding author.

, ,
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

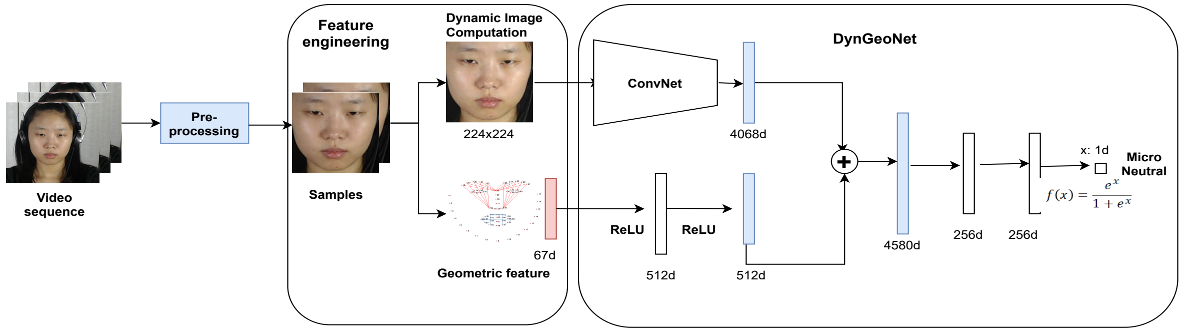

Figure 1: The illustration of our framework for spotting Micro-expression in the long videos. Our framework contains three parts: Pre-processing step, Feature engineering and DynGeoNet. DynGeoNet consists of two modules: the first part relies on ConvNet to extract the feature from a dynamic image. The second part is the multi-layer perceptron to extract the feature from the geometric feature. Then two learning features are concatenated for the binary classification.

features. To our best knowledge, this is the first time geometry feature and appearance feature are fused for spotting ME in long videos.

- (3) The first results of cross-database for ME spotting are reported and show promising results. Our work is the first study that makes the cross-database evaluation to generalization capability.

The remaining sections of this paper are organized as follows: Section 2 reviews the related works; Section 3 introduces the methods used for our proposed framework; Section 4 presents the experimental results and discussions, and Section 5 is the Conclusions.

2 RELATED WORK

When a potential application of ME analysis is implemented in real-life, it needs to detect the temporal locations of ME events before any recognition step can be applied. Therefore, MEs spotting is an indispensable module for a fully automated ME analysis system. Several studies have been involved in this topic. We will briefly go through the existing studies to understand the situation of ME spotting.

Most methods utilized **unsupervised learning** approach to detect ME positions in long videos. Firstly, Moilanen et al. [11] utilized the Chi-Square distance of the Local Binary Pattern (LBP) in fixed-size scanning windows to detect MEs. Patel et al. [14] computed the optical flow vectors for extracting features on small, local regions, then utilizing heuristic techniques to remove non-ME samples. Wang et al. [22] introduced a technique called Main Directional Maximal Differences (MDMD), which utilizes the magnitude of maximal difference in the main direction of optical flow for spotting MEs. In [3], Riesz transforms combining with facial maps have been employed to spot MEs automatically. Kai et al. [1] suggested a ME spotting method in long videos by geometric features in three facial regions. Recently, authors in the study [25] computed the specific patterns (magnitude and angle of motion) on each region of interest.

In the **supervised learning** approach, the authors tried to train the classifiers for discriminating the ME samples from other facial movements that cause the false alarms in detection. The study in [23] is the first study utilizing machine learning based on deformable features and the Adaboost classifier to detect ME samples. Tran et al. [19] proposed using a multi-scale sliding window based on spatial-temporal feature for ME spotting. In [26], Zhang et al. proposed using a Convolutional Neural Network (CNN) to detect the apex frame in two main steps: (1) constructing CNN networks to predict apex frames and neutral frames; (2) introducing a feature engineering technique to merge nearby detected samples. Tran et al. [18] introduced the dense prediction-based technique by fusing spatial-temporal features with Long short-term memory architecture to calculate the apex score of the ME samples in long videos. Recently, Pan et al.[13] proposed the bilinear convolutional neural network (LBCNN) to extract the local and global features of the face area for classifying the ME samples and macro-expression samples.

In this paper, we contribute a technique from supervised learning approach to spot ME in long videos.

3 PROPOSED METHOD

In order to enhance the performance of ME spotting, we propose a combination of appearance features and geometric features. The proposed framework is illustrated in Fig. 1, which consists of three main parts: (1) Several sub-steps are employed to pre-process the video and extract the Region of Interest of ME samples. (2) we apply feature engineering on the interested samples to extract the dynamic appearance features that summarize the frames sequence information, and we compute the geometric features that model the changes of specific facial landmark points. (3) the fusion network is constructed to learn the appearance and geometric features for classifying ME samples and non-ME samples.

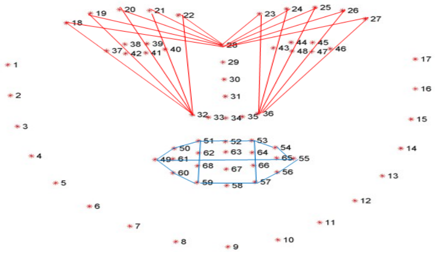

Figure 2: The illustration of geometric feature on single frame. The lines indicate the distance of specific landmark points.

3.1 Pre-processing

Pre-processing step carries out face detection and face-alignment for image sequences for which DLIB toolbox is utilized. Then, we apply the technique in [10] to extract the ROI of ME samples for next steps.

3.2 Dynamic image features

In order to take the advantages of deep convolutional neural network, we explore the features which summarize the information of frame sequence into one image. Dynamic image which have got the impressive results for the micro-expression recognition [7, 21] is selected here as appearance features.

Dynamic image summarizes video content by remodeling video structure to a typical static image which represents all information of the whole video. This feature resembling is basically based on the idea of a ranking function proposed in [2], where each frame in a given input video is ranked and ultimately mapped to a real vector that carries distinctive appearance elements from original image sequence. In [2], authors proposed the fast computation version for the dynamic image by equation 1:

$$p(I_1, I_2, ..., I_T; \psi) = \sum_{t=1}^{T} \alpha_t \psi(I_t) \tag{1}$$

where $p(I_1, I_2, ..., I_T; \psi)$ represents the dynamic image and ψ is the video frame. $\alpha_t = 2t - T + 1$ where T is number of frame in sample. For the improvement, we apply dynamic image computation on the particular regions of the face. We only focus on the eye, nose, and mouth area because of the action of ME samples.

3.3 Geometric features

In this paper, we propose the geometric feature computation between the onset frame and apex frame, which are the first frame and middle frame of the sample, respectively, to obtain the changes of these two frames.

First, we calculate the single-frame geometric features by the Euclidean distance between specific points. We utilize the DLIB toolbox [6] to extract 68 landmark points on the face. Then, we select points on the eyebrow and mouth regions to calculate the distance between points. Fig. 2 shows the particular landmark points which are utilized. The lines between two points describe the distance feature, which are: (1) the lines between 10 eyebrow landmark points linked to markers: 28^{th}, 32^{th} and 36^{th} points; (2) the lines of (P_i, P_{i+1}) where i is from 49 to 60. Additionally, we calculate the distances of couples: (P_{51}, P_{59}), (P_{53}, P_{557}) and (P_{49}, P_{55}) to explore the motion of mouth. In total, $N = 45$ lines are obtained.

To extract the motion information between apex frame and onset frame, we suggest combining two features: ratios of distance $Ratio$ and the deformable feature $Deform$. The $Ratio$ is calculated by the ratios between onset frame and apex frame as formula 2:

$$Ratio = [\frac{d_1^1}{d_T^1}, \frac{d_1^2}{d_T^2}, ..., \frac{d_1^N}{d_T^N}] \tag{2}$$

where d_1^i is the i^{th} Euclidean distance of onset frame, while d_T^i is the i^{th} Euclidean distance of apex frame.

The deformable feature $Deform$ is computed by the distance between selected points of two onset and apex frames: $Deform = [D(P_1^1, P_T^1); D(P_1^2, P_T^2); ..., D(P_1^M, P_T^M)]$, where $D(P_1^i, P_T^i)$ is the distance between i^{th} point of onset frame and i^{th} point of apex frame. We select $M = 22$ points from the two regions: 18^{th} point to 27^{th} point for eyebrow, and 49^{th} point to 60^{th} point for mouth region.

Finally, we concatenate $Deform$ and $Ratio$ to form a feature vector f with dimension $M + N = 67$.

3.4 DynGeoNet

After feature extraction, two kinds of features are inputted to the network for predicting ME or non-ME classification. For the purpose of fusing geometric information and dynamic image, we construct a new network, namely DynGeoNet.

In Fig 1, we illustrate our network which consists of two parts: the first part is the VGG-16-based network which extracts the deep appearance feature from dynamic image. The second part is a simple two-layers network to learn the facial movements.

After learning the deep feature from each module, features are concatenated for a fused feature and go through two Fully-connected (FC) layers for the binary classification of ME samples.

The objective function is the Binary Cross Entropy loss as the in Eq. 3:

$$Loss = -\frac{1}{N} \sum_{i=0}^{N} y_i log(y_i^*) + (1 - y_i).log(1 - y_i^*) \tag{3}$$

where y_i^* and y_i are predicted value and ground truth of each sample. In our research, $y \in [0, 1]$.

4 EXPERIMENTS

4.1 Experimental Setting

We evaluate the proposed method on three spontaneous databases: CAS(ME)^2, SAMM-Long and SMIC-E-Long [8, 15, 20]. These datasets have long videos which are suitable for ME spotting. In CAS(ME)^2, there are 57 ME samples in 97 videos from 22 subjects. SAMM-Long has 159 ME samples recorded in 147 videos from 32 subjects. There are 166 MEs in SMIC-E-Long with 162 videos from 16 subjects.

We set the parameters for each dataset as follows: for SMIC-E-long with $FPS = 100$, we set the window size as $L = 35$; for CAS(ME)^2 with $FPS = 30$, we set size as 11 for the detected window; for SAMM-long with $FPS = 200$, we set the value as 65. Through the setting of window size for each dataset, we determine the apex frame of ME sample by the middle frame of window ($L/2$).

As well, the parameters for learning are set as: the number of epoch is 100, the learning rate is 0.0001. The training and testing sets follow the protocol Leave-one-subject-out (LOSO) of [8, 20]. For the evaluation metric, we utilize the $F1 - score$ metric which is widely used in ME spotting [8, 20]. Intersection over Union (IoU) is set by 0.5 to determine the true positive samples.

Additionally, we conduct the cross-database evaluation; thus we denote three experimental setups: A, B, and C. Set A contains the training samples from SMIC-E-long and CAS(ME)^2 while testing from SAMM-Long; setup B has the samples from CAS(ME)^2 and SAMM-Long and evaluates in SMIC-E-long; setup C using the samples from SMIC-E-long and SAMM-Long for training while testing in CAS(ME)^2. To compare the cross-database evaluation with our method, we re-implement the CNN-based idea from the studies [20, 26]. The VGG-16 architecture [17] is selected as the backbone of ConvNet. The learning rate is set as 0.0001, and we optimize the network by Stochastic gradient descent (SGD). Finally, we utilize the evaluation metric introduced in [20] to compare the methods.

Table 1: The experimental results on three ME spotting datasets

Method	CAS(ME)^2	SAMM	SMIC
LBP-X2 [8, 20]	0.0055	0.0111	0.0666
HOGTOP-LSTM [20]	0.0111	0.03202	0.0535
HIGOTOP-LSTM [20]	0.00902	0.03428	0.0835
MDMD [5, 20]	0.0082	0.0364	0.0268
LBCNN [13]	**0.0595**	0.0813	-
DynNet (Ours)	0.00921	0.0208	0.0438
GeoNet (Ours)	0.0081	0.0198	0.0278
DynGeoNet (Ours)	0.05012	**0.0974**	**0.1035**

4.2 Results

4.2.1 Comparison on each dataset. In this Section, we describe the experimental results when we conducted on three datasets by comparing our methods (denoted as *DynGeoNet*) with existing techniques from the studies [5, 8, 13, 20].

Ablation study: we also conducted an independent experiment for each component in our network to explore the performance of each module. As shown in Table 1, *DynNet* and *GeoNet* are two subnets of the *DynGeoNet* that process only dynamic image feature or geometric feature, respectively. We can see that the independent modules have lower performance than the fusion network.

In Table 1, we can see that our method is better than the baseline results from [5, 8, 20]. The proposed method obtains F1-score as 0.05012, 0.0974 and 0.1035 for the datasets CAS(ME)^2, SAMM-long and SMIC-E-Long, respectively. In the CAS(ME)^2 result, our method gets the comparable result to the best one from [13] with F1-score 0.0595.

4.2.2 Comparison on cross-database evaluation. The experimental results from DynGeoNet are reported on Table 2. The best result with F1-socre 0.06512 is obtained by the SMIC-E-long when testing . The dataset CAS(ME)^2 is still very challenging with the F1-score 0.01521. It is probably because that CAS(ME)^2 has only 30FPS; thus, it can not extract robust features when applying models trained on data with high-speed camera images from SAMM-Long ($FPS = 200$) and SMIC-E-Long ($FPS = 100$).

In Table 2, we compare our technique with one existing method from the study of Tran et al. [20] which is denoted as $VGG16$. As shown in Table 2, the proposed technique outperforms $VGG16$ method on all setups. We obtain F1-scores by 0.0481, 0.06512 and 0.01521 on set A, B and C, respectively. The proposed technique also reduces the number of false positive effectively than $VGG16$. The results show that our proposed technique is promising with better generalization capability.

Table 2: The cross-database evaluation performance of DynGeoNet.

Method	Setup	TP	FP	FN	F1-score
VGG16 [20]	A	45	2157	114	0.0381
DynGeoNet	A	49	1830	110	**0.0481**
VGG16 [20]	B	58	2340	108	0.0452
DynGeoNet	B	48	1260	118	**0.06512**
VGG16 [20]	C	11	2521	46	0.0085
DynGeoNet	C	12	1509	45	**0.01521**

4.3 Discussion

High false positive rate is still an issue. One reason is that the false detections caused by the subtle macro-expressions are still one of the biggest challenges for the ME spotting task. In later works, multi-scale window detection should be considered to discriminating the macro-expression from ME samples.

Furthermore, the current method depends on the handcrafted features for the input of the deep network. In the next work, different end-to-end architectures robust to extract valuable features with limited data will be explored.

5 CONCLUSION

This paper introduced a new method for ME spotting based on the fusion of geometric features and appearance features. The experimental results show that our proposed network model achieves state-of-the-art performance with limited data. Furthermore, our research also provides the baseline results for the cross-database evaluation. However, there is still room for further improvement while the performance is not good enough for real-world applications.

ACKNOWLEDGMENTS

This work was supported by Infotech Oulu, Ministry of Education and Culture of Finland for AI forum project, and Academy of Finland for ICT 2023 project (grant 328115). As well, the authors wish to acknowledge CSC IT Center for Science, Finland, for computational resources.

REFERENCES

[1] Kai Xin Beh and Kam Meng Goh. 2019. Micro-Expression Spotting Using Facial Landmarks. In *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, 192–197.

[2] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic image networks for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3034–3042.

[3] Carlos Arango Duque, Olivier Alata, Remi Emonet, Anne-Claire Legrand, and Hubert Konik. 2018. Micro-expression spotting using the Riesz pyramid. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 66–74.

[4] Paul Ekman. 2009. Lie catching and microexpressions. *The philosophy of deception* 1, 2 (2009), 5.

[5] Ying He, Su-Jing Wang, Jingting Li, and Moi Hoon Yap. 2020. Spotting macro-and micro-expression intervals in long video sequences. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 742–748.

[6] Dae Hoe Kim, Wissam J Baddar, and Yong Man Ro. 2016. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *Proceedings of the 24th ACM international conference on Multimedia*. 382–386.

[7] Trang Thanh Quynh Le, Thuong-Khanh Tran, and Manjeet Rege. 2020. Dynamic image for micro-expression recognition on region-based framework. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 75–81.

[8] Jingting Li, Catherine Soladié, Renaud Séguier, Su-Jing Wang, and Moi Hoon Yap. 2019. Spotting micro-expressions on long videos sequences. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.

[9] J Li, S Wang, Moi Hoon Yap, John See, Xiaopeng Hong, and Xiaobai Li. [n.d.]. MEGC2020-The Third Facial Micro-Expression Grand Challenge. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. 234–237.

[10] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. 2017. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing* 9, 4 (2017), 563–577.

[11] A. Moilanen, G. Zhao, and M. Pietikäinen. 2014. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *Proc. ICPR*.

[12] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. 2019. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11917–11926.

[13] Hang Pan, Lun Xie, and Zhiliang Wang. 2020. Local bilinear convolutional neural network for spotting macro-and micro-expression intervals in long video sequences. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. IEEE Computer Society, 343–347.

[14] D. Patel, G. Zhao, and M. Pietikäinen. 2015. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *Proc. ACIVS*.

[15] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. 2017. CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing* 9, 4 (2017), 424–436.

[16] Hadas Shahar and Hagit Hel-Or. 2019. Micro Expression Classification using Facial Color and Deep Learning Methods. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[17] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[18] TK Tran, Nhat VQ, X. Hong, and G. Zhao. 2019. Dense prediction for micro-expression spotting based on deep sequence model. In *Proc. Electronic Imaging*.

[19] Thuong-Khanh Tran, Xiaopeng Hong, and Guoying Zhao. 2017. Sliding window based micro-expression spotting: a benchmark. In *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 542–553.

[20] Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, Xiaobai Li, and Guoying Zhao. 2021. Micro-expression spotting: A new benchmark. *Neurocomputing* 443 (2021), 356–368.

[21] Monu Verma, Santosh Kumar Vipparthi, Girdhari Singh, and Subrahmanyam Murala. 2019. LEARNet: Dynamic imaging network for micro expression recognition. *IEEE Transactions on Image Processing* 29 (2019), 1618–1627.

[22] S. Wang, S. Wu, and X. Fu. 2016. A main directional maximal difference analysis for spotting micro-expressions. In *ACCV 2016*. Springer, 449–461.

[23] Z. Xia, X. Feng, J. Peng, X. Peng, X. Fu, and G. Zhao. 2016. Spontaneous micro-expression spotting via geometric deformation modeling. *CVIU* 147 (2016), 87–94.

[24] Chuin Hong Yap, Connah Kendrick, and Moi Hoon Yap. 2019. Samm long videos: A spontaneous facial micro-and macro-expressions dataset. *arXiv preprint arXiv:1911.01519* (2019).

[25] Li-Wei Zhang, Jingting Li, Su-Jing Wang, Xian-Hua Duan, Wen-Jing Yan, Hai-Yong Xie, and Shu-Cheng Huang. 2020. Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. IEEE Computer Society, 245–252.

[26] Zhihao Zhang, Tong Chen, Hongying Meng, Guangyuan Liu, and Xiaolan Fu. 2018. SMEConvnet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos. *IEEE Access* 6 (2018), 71143–71151.