MEGC2023: ACM Multimedia 2023 ME Grand Challenge

Adrian K. Davison*
Department of Computing and Mathematics, Manchester Metropolitan University
A.Davison@mmu.ac.uk

Jingting Li*
CAS Key Laboratory of Behavioral Science, Institute of Psychology & Department of Psychology, University of the Chinese Academy of Sciences
lijt@psych.ac.cn

Moi Hoon Yap
Department of Computing and Mathematics, Manchester Metropolitan University
m.yap@mmu.ac.uk

John See
School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia
J.See@hw.ac.uk

Wen-Huang Cheng
National Taiwan University
wenhuang@csie.ntu.edu.tw

Xiaobai Li
University of Oulu & Zhejiang University
xiaobai.li@oulu.fi

Xiaopeng Hong
Harbin Institute of Technology
hongxiaopeng@ieee.org

Su-Jing Wang
CAS Key Laboratory of Behavioral Science, Institute of Psychology & Department of Psychology, University of the Chinese Academy of Sciences
wangsujing@psych.ac.cn

ABSTRACT

Facial micro-expressions (MEs) are involuntary movements of the face that occur spontaneously when a person experiences an emotion but attempts to suppress or repress the facial expression, typically found in a high-stakes environment. Unfortunately, the small sample problem severely limits the automation of ME analysis. Furthermore, due to the weak and transient nature of MEs, it is difficult for models to distinguish it from other types of facial actions. Therefore, ME in long videos is a challenging task, and the current performance cannot meet the practical application requirements. Addressing these issues, this challenge focuses on ME and the macro-expression (MaE) spotting task. This year, in order to evaluate algorithms' performance more fairly, based on CAS(ME)^2, SAMM Long Videos, SMIC-E-long, CAS(ME)^3 and 4DME, we build an unseen cross-cultural long-video test set. All participating algorithms are required to run on this test set and submit their results on a leaderboard with a baseline result.

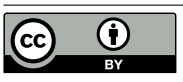
CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Applied computing** → *Psychology*.

KEYWORDS

Micro-expression, Spotting, Long videos

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3613852

ACM Reference Format:
Adrian K. Davison, Jingting Li, Moi Hoon Yap, John See, Wen-Huang Cheng, Xiaobai Li, Xiaopeng Hong, and Su-Jing Wang. 2023. MEGC2023: ACM Multimedia 2023 ME Grand Challenge. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3613852

1 INTRODUCTION

When a person attempts to suppress a facial expression, typically in a high-stakes scenario, there is a possibility of an involuntary movement occurring on the face, namely a facial micro-expression (ME). As such, the duration of a ME is very short, generally being no more than 500 milliseconds (ms), and is the telltale sign that distinguishes them from a normal facial expression. Computational analysis and automation of tasks on MEs is an emerging area in multimedia research. However, only until recently, the availability of a few spontaneously induced facial ME datasets has provided the impetus to advance further from the computational aspect.

Since eliciting and the artificial annotation of MEs is very challenging, the amount of labeled ME samples is limited. So far, there are only around 2500 samples in public spontaneous databases. In addition, it is difficult to unify the standardization of ME labeling for different annotators. To tackle this problem, we expect that the recent advancement in multimedia technologies can help improve the performance of ME spotting and recognition. For example, self-supervised learning is a form of unsupervised learning where the data provides the supervision.

In addition, ME analysis is an interdisciplinary field with multi-modal research capabilities. On the one hand, it is very difficult to analyze MEs solely on RGB images. Multi-modal data, such as time information in video clips and supplementary features of near-infrared images, can help improve the performance of ME spotting

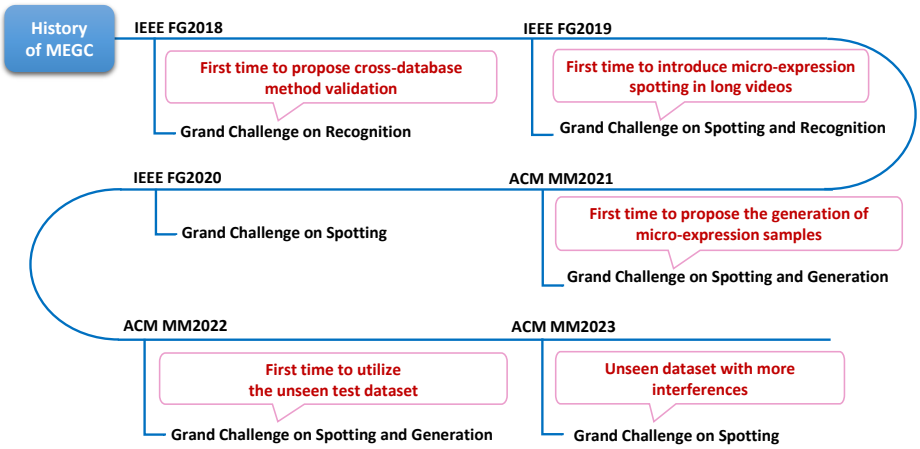


Figure 1: Continuity of Micro-Expression Grand Challenge (MEGC)

and recognition. On the other hand, expanding the research on ME analysis can enable more in-depth research on multimedia in fields such as face and emotion analysis, for example, 3D face construction combined with depth information, and personal mental state analysis combined with physiological signals such as heart rate and electroencephalography (EEG).

ME spotting focuses on the identification of whether a ME is present in a video and then to determine the temporal location of that ME (i.e. which frames the movement is located in). This is a necessary step for automated ME analysis in practical applications. Furthermore, accurate ME spotting could improve the reliability of the subsequent ME analysis. Unfortunately, due to the subtlety and fleeting nature of MEs, it is difficult for models to distinguish it from other types of facial actions. Therefore, spotting ME in long videos is a challenging task. Currently, ME spotting methods are gradually turning from traditional unsupervised manual feature comparison methods to those based on deep learning. Yet, the spotting performance is not satisfactory and has large potential for improvement. Moreover, there is no uniform evaluation standard for ME spotting. Through the Micro-Expression Grand Challenge (MEGC), we can promote the research of ME spotting methods and provide a fair comparison of the proposed methods.

This is the inaugural academic activity in this area of research. Our ambition is to conduct ME challenge yearly with continuity. We have held five ME Grand Challenges (MEGC)[1] in conjunction with FG2018 [19], FG2019 [13], FG2020 [4], ACMMM2021 [7] and ACMMM2022 [8] and a ME Recognition Challenge (MER2020)[2] [15] in conjunction with ICIP2020.

This year, 13 teams participated in MEGC 2023[3]. The top three articles were accepted.

2 REVIEW OF PREVIOUS MEGC

As shown in Fig. 1, from 2018, MEGC has always oriented itself to the forefront of ME intelligence analysis research.

Regarding the recognition task, MEGC has witnessed the transition from handcraft feature extraction methods to end-to-end deep learning methods. In addition, MEGC innovatively proposes performance validation based on cross-database and composite databases, which promotes the development of robust ME recognition methods. Due to the limited number of samples in the released databases, the improvement of ME recognition methods based on deep learning is restricted. With the release of new multimodal ME databases in recent years, we may continue to set up ME recognition task in multimodal and complex scenes in the future.

Regarding the generation task, the goal of this task is to generate a specific ME on a given template surface. By assessing the authenticity of the generated MEs through examination by a psychologist, reliable ME generation allows for appropriate data augmentation of the MEs. However, the evaluation of ME generation results currently relies on professional psychologists or FACS-certified experts, which is a subjective and time-consuming task. In addition, there are arguments that MEs are directly related to muscle movements. However, the generation task only generates the external appearance of the human face, and does not reflect the intrinsic connection with muscle actions, which prevents the evaluation of generation quality through Action Units [2]. Therefore, how to treat the ME generation results objectively and rationally remains an undefined issue. Thus, we did not set up the generation task in this year. However, research and discussion on this topic are welcomed in the associated workshop: ACM MM workshop FME2023[4].

Regarding the spotting task, MEGC has gone from working on published databases to setting up the Unseen dataset, in order to facilitate the establishment of robust and transferable ME spotting methods, rather than obtaining a method that is only applicable to a single database through continuous optimization of the parameters. In terms of the spotting methods, traditional handcraft feature

[1] http://www2.docm.mmu.ac.uk/STAFF/m.yap/FG2018Workshop.htm; https://facial-micro-expressiongc.github.io/MEGC2019/; https://megc2020.github.io/; https://megc2021.github.io/; https://megc2022.github.io/

[2] https://2020.ieeeicip.org/challenge/micro-expression-recognition-challenge/

[3] https://megc2023.github.io/

[4] https://megc2023.github.io/workshop.html

(a) (b)

Figure 2: Samples from the Unseen dataset approaching real scenes. Fig. 2(a): illumination fluctuation on the face; Fig. 2(a): body and head movement.

methods and deep learning methods have comparable performance. In the last two years, the optical flow frame difference based method won the first place in the spotting task, and there were also deep learning based methods in the top three. The main reason why deep learning methods failed to outperform traditional methods is the small sample size of micro-expressions.

3 DATASETS

In recent years, with the release of ME databases, the availability of long video samples containing MEs has increased. Simultaneously, the scenarios depicted in these videos have gradually approached real-life situations, commonly referred to as "in the wild" conditions. This means that the subjects in the videos may exhibit head movements, occlusion of hands, and other natural behaviors, in contrast to maintaining a fixed head position and direct gaze at the screen. In particular, several long-video ME databases have been published by the academic community, such as CAS(ME)^2, SMIC-E, SAMM Long Videos, CAS(ME)^3 and 4DME; the last two being the most recently established large-scale datasets. In this challenge, we use these five datasets for the task of ME and macro-expression (MaE) spotting. In addition, in order to evaluate algorithms' performance more fairly, we build an unseen cross-cultural long-video test set and the sample size will be tripled from last year's challenge. This dataset consists of video samples without disclosing the ground truth annotations. All participating algorithms are required to run on this test set and submit their results.

3.1 Recommended Training Datasets

Prior to 2017, ME databases had limited long video samples. Most of the released databases primarily consisted of short videos, focusing only on the ME segments themselves, along with a few frames before and after the start and end frames of those segments. However, starting from 2017, the academic community gradually began to shift its attention towards ME spotting in long videos.

Firstly, CAS(ME)^2 [12] was released. It consists of 22 participants and 98 long videos at 30 fps, including 300 MaEs and 57 MEs. As an improvement compared with MEGC2020 [4], a cropped version with only face region is provided for fair comparison in MEGC2021 [7].

Meanwhile, the authors of the SAMM dataset [1], released their corresponding long videos, i.e. the SAMM Long Videos dataset [17], which consists of 147 long videos at 200fps, including 343 macro-movements and 159 micro-movements in the long videos.

In addition, the authors of the SMIC-E dataset [10] also released their corresponding long videos, i.e. the SMIC-E-long dataset [14], including 162 long videos at 100 fps with 132 MEs.

In 2023, two multi-modal ME databases were released. Specifically, CAS(ME)^3 [5] (RGB & Depth information) offers around 80 hours of videos with over 8,000,000 frames, including 1,109 and 3,490 manually labeled MEs and MaE, respectively.

Shortly thereafter, 4DME [9] (Grayscale, RGB, and Depth) was built by University of Oulu and Imperial College London. The first released version of 4DME data included 270 long clips (with context frames, average length 2.5 seconds, 60 fps) for the spotting test.

3.2 Unseen Test Dataset

The unseen testing set (MEGC2023-testSet) contains 30 long video, including 10 long videos from SAMM Challenge dataset [1, 18] and 20 clips cropped from different videos in CAS(ME)^3 (unreleased before). The frame rate for SAMM Challenge dataset is 200fps and the frame rate for CAS(ME)^3 is 30 fps. The participants should test on this unseen dataset.

Specifically, regarding the clips from CAS(ME)^3, the data collection process followed the same approach as Part A of CAS(ME)^3 . However, we did not impose strict requirements to keep the body and head still during this video recording. Moreover, recordings were made in a natural light environment with no supplemental light to avoid illumination fluctuation. As illustrated in Fig. 2, these samples, which are closer to the real environment, may exhibit head movements, light changes, and other factors that interfere with micro-expression spotting. Overall, we released 20 video samples that include both MaEs and MEs. These samples were obtained from 4 participants (3 males). Additionally, we ensured the stimuli in the videos were balanced.

4 EVALUATION

We use an interval-based evaluation method [4, 6] that was first presented at MEGC2019 [13] and then used for the evaluation of spotting task in subsequent challenges.

Specifically, 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:

$$\frac{W_{spotted} \cap W_{groundTruth}}{W_{spotted} \cup W_{groundTruth}} \geq k \tag{1}$$

where k is set to 0.5, $W_{groundTruth}$ represents the ground truth of the MaE or ME interval (onset-offset). If the condition is not fulfilled, the spotted interval is regarded as a false positive (FP).**We consider that each ground-truth interval corresponds to, at most, one single spotted interval.**

The final evaluation is performed on the entire dataset, based on the overall F1-score of MaE and ME spotting performance. The champion of the challenge will be the best score for overall results in spotting MEs and MaEs. The participants should evaluate their results on the grand challenge system[5], with evaluation codes used by the baseline method [21] to ensure fair comparison.

[5]https://codalab.lisn.upsaclay.fr/competitions/14254

Table 1: Top-3 spotting results of the spotting tasks for MEGC2023. [21] provides the baseline results.

Participants	F1 Score			Precision			Recall		
	Overall	SAMM	CAS	Overall	SAMM	CAS	Overall	SAMM	CAS
Xu et al. [16]	0.22	0.37	0.21	0.28	0.37	0.27	0.19	0.36	0.17
Qin et al. [11]	0.19	0.29	0.18	0.25	0.29	0.24	0.16	0.29	0.15
Yu et al. [20]	0.18	0.31	0.17	0.18	0.28	0.16	0.19	0.35	0.17
Baseline [21]	0.03	0.00	0.03	0.04	0.00	0.04	0.02	0.00	0.03

5 METHODS

Baseline method: Based on the results of the previous two years' challenges, it has been observed that traditional frame difference methods can achieve spotting performance comparable to that of deep learning methods. Additionally, training deep learning models requires a substantial amount of time and data. Therefore, we have decided to adopt the spatiotemporal fusion method [21], which won the first-place in the MEGC2020 [4].

Specifically, Zhang et al. separate the local movement vector from the overall optical flow field through the estimation of mean optical flow in the nose region. Following preprocessing, the completed specific pattern of MEs within each region of interest are extracted. Then, considering the impact of frame rate and varying intensities of ME and MaE, they propose the utilization of a multi-scale filter to enhance the spotting capability.

For MEGC2023, the top three teams all used optical flow features for the main task of the challenge: MaE and ME spotting in long videos.

Third place method: The third place team proposed an efficient spotting method using Main Directional Mean Optical Flow features [20]. The overall framework consists of three main modules: Face Cropping and Alignment Module (FCAM), optical flow Feature Extraction Module (FEM), and ROI-based expression Proposal Generation Module (PGM).

Second place method: In second place, Qin et al. [11] uses traditional optical flow to represent the motion between frames, and also use it to distinguish between MaE and ME types. As optical flow can be prone to missing small movements, the method defines ROIs based on FACS [3] and uses expression segment generation, where the extracted optical flow undergoes low-pass filtering to eliminate noise.

First place method: The winner of MEGC2023 was Xu et al. [16]. Their method utilised a pre-trained masked autoencoder, namely VideoMAE, to spot MaE and MEs in a data-efficient way. The method uses different "fine-grain sizes" during training to focus on the characteristics found in MaE and MEs. Optical flow is used to output expression intervals, along with the output from the vision transformer (ViT), as a post-processing step for frame interval refinement.

6 RESULTS AND ANALYSIS

Baseline result: Considering that the parameters in the baseline method were tailored according to the specifications of CAS(ME)^2, it is anticipated that the same method can identify a few MEs and MaEs within CAS(ME)^3 videos. This expectation arises from the shared characteristics of similar video frame rates and collection scenarios between the two datasets. However, when examining the SAMM dataset, characterized by a significantly higher frame rate of 200fps and grayscale samples, the performance of the baseline method is suboptimal, as it does not successfully detect any true positives (TP).

Third place result [20]: For the third place team, their overall F1 score was 0.18, but the method performed particularly well in the SAMM only scores, achieving an F1 score of 0.31.

Second place result [11]: The second place team marginally outperformed the third place team, achieving an overall F1 score of 0.19. Similar to the other results for the challenge, including those outside of the top-3, the team's SAMM only scores performed much better than CAS only. The larger sample of CAS data within the unseen test dataset will likely have contributed to this.

First place result [16]: The highest performing team achieved an overall F1 score of 0.22. They also achieved an F1 score of 0.37 for SAMM only, showing how the historically low scores of ME/MaE spotting are gradually increasing each year.

Overall, through effective preprocessing, filtering and appropriate threshold settings, the traditional frame difference method can remove the interference of head movements and other external factors, and thus achieve good ME spotting performance. Meanwhile, the deep learning network is able to better learn the ME features, so as to effectively discriminate MEs, MaEs and other head movements. Furthermore, constructing pre-trained models on large-scale datasets by means of migration learning, self-supervised learning, etc. is proven to be effective for ME spotting performance enhancement.

7 CONCLUSION AND FUTURE CHALLENGES

This year, it is clear that the performance of MaE and ME spotting in long videos has improved substantially, demonstrating the growing success of the field. The challenges faced this year included creating a new base to host the leaderboard and analysis code, and so only a single challenge was focused on. In the year leading up to the next challenge, we will investigate emerging methods and how we can utilise these in a rewarding way for future participants.

ACKNOWLEDGMENTS

We would like to thank the ACM MM '23 conference organisers for agreeing to host our Grand Challenge and for their support. We would also like to thank the CodaLab open-source competition platform for hosting our leaderboard submissions. This work is supported by grants from the National Natural Science Foundation of China (62106256, U19B2032, 62276252, 62076195), the Academy of Finland (Grant 323287), and Ministry of Science and Technology of Taiwan (MOST-109-2223-E-009-002-MY3, MOST-110-2634-F-007-015).

REFERENCES

[1] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2018. [SAMM]: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing* 9, 1 (2018), 116–129.

[2] Paul Ekman, Wallace V Friesen, and Silvan S Tomkins. 1971. Facial affect scoring technique: A first validity study. (1971).

[3] W. V Friesen and P. Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3 (1978).

[4] LI Jingting, Su-Jing Wang, Moi Hoon Yap, John See, Xiaopeng Hong, and Xiaobai Li. 2020. MEGC2020-the third facial micro-expression grand challenge. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 777–780.

[5] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. 2022. CAS(ME)^3: A Third Generation Facial Spontaneous Micro-Expression Database with Depth Information and High Ecological Validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 2782–2800.

[6] Jingting Li, Catherine Soladie, Renaud Seguier, 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.

[7] Jingting Li, Moi Hoon Yap, Wen-Huang Cheng, John See, Xiaopeng Hong, Xiaobai Li, and Su-Jing Wang. 2021. FME'21: 1st Workshop on Facial Micro-Expression: Advanced Techniques for Facial Expressions Generation and Spotting. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5700–5701.

[8] Jingting Li, Moi Hoon Yap, Wen-Huang Cheng, John See, Xiaopeng Hong, Xiaobai Li, Su-Jing Wang, Adrian K Davison, Yante Li, and Zizhao Dong. 2022. MEGC2022: ACM Multimedia 2022 Micro-Expression Grand Challenge. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7170–7174.

[9] Xiaobai Li, Shiyang Cheng, Yante Li, Muzammil Behzad, Jie Shen, Stefanos Zafeiriou, Maja Pantic, and Guoying Zhao. 2022. 4DME: A Spontaneous 4D Micro-Expression Dataset With Multimodalities. *IEEE Transactions on Affective Computing* (2022), 1–18. https://doi.org/10.1109/TAFFC.2022.3182342

[10] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen. 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] Wenfeng Qin, Bochao Zou, Xin Li, Weiping Wang, and HUimin Ma. 2023. Micro-Expression Spotting with Face Alignment and Optical Flow. In *Proceedings of the 31st ACM International Conference on Multimedia*. Association for Computing Machinery.

[12] 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.

[13] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. 2019. Megc 2019–the second facial micro-expressions grand challenge. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.

[14] 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.

[15] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2020. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2871–2880.

[16] Ke Xu, Kang Chen, Licai Sun, Zheng Lian, Bin Liu, Gong Chen, Haiyang Sun, Mingyu Xu, and Jianhua Tao. 2023. Integrating VideoMAE based model and Optical Flow for Micro- and Macro-expressions Spotting. In *Proceedings of the 31st ACM International Conference on Multimedia*. Association for Computing Machinery.

[17] Chuin Hong Yap, Connah Kendrick, and Moi Hoon Yap. 2020. Samm long videos: A spontaneous facial micro-and macro-expressions dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 771–776.

[18] Chuin Hong Yap, Moi Hoon Yap, Adrian Davison, Connah Kendrick, Jingting Li, Su-Jing Wang, and Ryan Cunningham. 2022. 3d-cnn for facial micro-and macro-expression spotting on long video sequences using temporal oriented reference frame. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7016–7020.

[19] Moi Hoon Yap, John See, Xiaopeng Hong, and Su-Jing Wang. 2018. Facial micro-expressions grand challenge 2018 summary. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 675–678.

[20] Jun Yu, Zhongpeng Cai, Shenshen Du, Xiaxin Shen, Lei Wang, and Fang Gao. 2023. Efficient Micro-Expression Spotting Based on Main Directional Mean Optical Flow Feature. In *Proceedings of the 31st ACM International Conference on Multimedia*. Association for Computing Machinery.

[21] 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)*. IEEE, 734–741.