MEGC2021: ACM Multimedia 2021 - Facial Micro-Expression (FME) Challenge

Jingting LI
CAS Key Laboratory of Behavioral Science, Institute of Psychology
Beijing, China
lijt@psych.ac.cn

Moi Hoon Yap
Manchester Metropolitan University
Manchester, UK
m.yap@mmu.ac.uk

Wen-Huang Cheng
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
whcheng@nycu.edu.tw

John See
Heriot-Watt University
Putrajaya, Malaysia
J.See@hw.ac.uk

Xiaopeng Hong
Xi'an Jiaotong University
Xi'an, China
hongxiaopeng@ieee.org

Xiaobai Li
University of Oulu
Oulu, Finland
xiaobai.li@oulu.fi

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

ABSTRACT

Facial micro-expressions (FMEs) are involuntary facial movements that occur spontaneously when a person experiences an emotion but tries to suppress or repress the facial expression and usually occur in high-risk situations. Thus, FMEs are very short in duration, an important feature that distinguishes them from ordinary facial expressions. And MEs are considered to be one of the most valuable cues for complex human emotion understanding and lie detection. Since 2014, the computational analysis and automation of MEs have been an emerging area of face research. The workshop will explore various dimensions of the human mind through emotion understanding and FME analysis, as well as extended research based on multi modal approaches.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Micro expression, Spotting, Generation

ACM Reference Format:
Jingting LI, Moi Hoon Yap, Wen-Huang Cheng, John See, Xiaopeng Hong, Xiaobai Li, and Su-Jing Wang. 2026. MEGC2021: ACM Multimedia 2021 - Facial Micro-Expression (FME) Challenge. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Since the advent of deep learning technology, face recognition has been significantly developed, and its accuracy rate has reached beyond human capabilities. Besides person identification, the broad notion of Emotion understanding has become a hot topic in human face research. Facial expressions could reveal interesting dimensions such as individual personality, personal emotion, etc. Facial micro-expressions (FMEs) are involuntary movements of the face that occur spontaneously when a person experiences an emotion but attempts to suppress or repress the facial expression, a scenario typically found in a high-stakes environment. As such, the duration of FMEs is very short, and it forms a telltale sign that distinguishes them from a normal facial expression. FME is considered one of the most valuable clues for complex human emotion understanding. It can also benefit a wide range of real-world applications, e.g., police interrogation, clinical diagnosis, depression analysis, and business negotiation. Computational analysis and automation of tasks on FMEs are emerging areas in face research, with a strong interest appearing as recently as 2014. The availability of a few spontaneously facial FME datasets has provided the impetus to further advance in the computational aspect.

Since the elicitation and the artificial annotation of FMEs are challenging, the amount of labeled ME samples is limited. So far, there are only around 832 (video) samples across five public spontaneous databases. Besides, it is impossible to unify the standardization of ME labeling for different annotators. To tackle this problem, we expect that the recent advancement in pattern recognition can help improve FME spotting and recognition performance.

Furthermore, FME analysis faces many difficulties and challenges. First, the FME generation mechanism is still not precise, while the valence of FME in lie detection is not sufficiently clear. Second, FME samples with high ecological validity are difficult to induce, and data labeling can be quite time-consuming and labor-intensive. This causes the problems of small sample size and imbalanced distribution in FME tasks. Besides, because the FME itself is relatively short and subtle, it is not as straightforward to generate more samples using popular methods such as GANs.

This is the inaugural workshop in this area of research. Our ambition is to conduct this workshop yearly with continuity. We have held three FME Grand Challenges (MEGC)[1][2][3] in conjunction with FG2018 [14], FG2019 [7] and FG2020 [5], and a FME Recognition Challenge (MER2020)[4] [8] in conjunction with ICIP2020. We aim to promote interactions between researchers and scholars from within this niche research area and include those from broader, general areas of expression and psychology research.

This grand challenge has two main agendas: FME generation, and FME and macro-expression spotting. All the submitted papers received three independent review. The challenge results are published on the FME '21 website[5] for the transparency of the competition. Nine teams participated in the spotting task, and five teams participated in the generation task. The top three articles from both tasks were accepted.

2 SPOTTING CHALLENGE

The 2nd MEGC [7] saw the establishment of the ME spotting in long videos. However, there is only one submission [4] as proposed in [13]. In this 2nd MEGC

The goal of this challenge is to spot macro- and micro-expressions interval in long video sequences. For this challenge, we focus on 98 long videos of CAS(ME)^2 database (300 macro-expressions and 57 micro-expressions) and 147 long videos of SAMM Long Videos dataset (343 macro-expressions and 159 micro-expressions).

2.1 Databases

CAS(ME)^2 consists of 22 participants and 98 long videos, including 300 macro-expressions and 57 micro-expressions. As an improvement compared with MEGC2020 [5], a cropped version with only face region is provided for fair comparison of the challenge. The participant are required to perform the proposed method on this version. Meanwhile, the authors of the SAMM dataset released their corresponding long videos, i.e. the SAMM Long Videos dataset [13], which consists of 147 long videos, including 343 macro-movements and 159 micro-movements in the long videos. Table 1 summarizes the two databases.

2.2 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:

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

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). **We consider that each ground-truth interval corresponds to at most one single spotted interval.**

Table 1: Summary of CAS(ME)^2 and SAMM Long Videos for macro- and micro-expression spotting challenge.

Dataset	CAS(ME)^2-cropped	SAMM Long Videos
Video samples	98	147
Macro-expressions	300	343
Micro-expressions	57	159
Resolution	640×480	2040×1088
FPS	30	200

2. Result evaluation in one video. Supposing there are 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 following metrics:

$$Recall = \frac{a}{m},\ Precision = \frac{a}{n} \tag{2}$$

$$F - score = \frac{2TP}{2TP + FP + FN} = \frac{2a}{m + n} \tag{3}$$

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 Method$_1$ outperforms Method$_2$, if Method$_1$ spots less intervals than Method$_2$.

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.

3. Evaluation for entire database. Supposing in the entire dataset,

- There are V videos including M_1 macro-expressions (MaEs) sequences and M_2 micro-expression (MEs) sequences, where $M_1 = \sum_{i=1}^{V} m_{1i}$ and $M_2 = \sum_{i=1}^{V} m_{2i}$;
- The method spot N_1 MaE intervals and N_2 ME intervals in total, where $N_1 = \sum_{i=1}^{V} n_{1i}$ and $N_2 = \sum_{i=1}^{V} n_{2i}$;
- There are A_1 TPs for MaE and A_2 TPs for ME in total, where $A_1 = \sum_{i=1}^{V} a_{1i}$ and $A_2 = \sum_{i=1}^{V} a_{2i}$.

The dataset could be considered as one long video. The results are firstly evaluated for the MaE spotting and ME spotting separately.

[1] http://www2.docm.mmu.ac.uk/STAFF/m.yap/FG2018Workshop.htm
[2] https://facial-micro-expressiongc.github.io/MEGC2019/
[3] https://megc2020.github.io/
[4] https://2020.ieeeicip.org/challenge/micro-expression-recognition-challenge/
[5] https://megc2021.github.io/

Then the overall result for macro- and micro spotting is evaluated. The *recall* and *precision* for entire dataset can be calculated by following formulas:

- for macro-expression:

$$Recall_{MaE\_D} = \frac{A_1}{M_1} \; Precision_{MaE\_D} = \frac{A_1}{N_1} \tag{4}$$

- for micro-expression:

$$Recall_{ME\_D} = \frac{A_2}{M_2} \; Precision_{ME\_D} = \frac{A_2}{N_2} \tag{5}$$

- for overall evaluation:

$$Recall_D = \frac{A_1 + A_2}{M_1 + M_2} \; Precision_D = \frac{A_1 + A_2}{N_1 + N_2} \tag{6}$$

Then, the values of *F1-score* for all these three evaluations are obtained based on:

$$F1 - score = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{7}$$

The champion of the challenge will be the best score for overall results in spotting micro- and macro-expressions.

2.3 Methods

[16]. The optical flow method is applied to estimate the motion trend of the facial regions. Because the head shaking is an essential reason for the high false-positive rate of micro- expression spotting, a reliable face alignment method becomes crucial. According to the optical flow of the nose tip region, the cutting box was adjusted several times to optimize the relative position between the face and the cutting box stable. On this basis, the optical flow features from the 14 regions of interest on the face are used to build a feature matrix, and a wave peak location technology is proposed to accurately locate the moment when the micro-expression occurs on the time-domain curve of the features.

[15] In this paper, we propose an efficient two-stream network named location suppression spotting network (LSSNet), which includes three parts. First, the optical flow is extracted using the traditional TV-L1 algorithm which captures subtle facial movements while adding temporal information to alleviate the problem of insufficient samples. Then, fixed length features are extracted from the sampled optical flow and raw images by an I3D model, which is used to set sliding windows. Finally, location suppression modules are added to the pyramidal full convolutional neural network to reduce the proposals with longer and shorter intervals. In addition, we use two different validation methods, named top_k and top_threshold, respectively.

[12] The proposed framework focuses on individual components of facial muscle movement rather than processing the whole image, which eliminates the influence of image change caused by noises, such as body or head movement. Compared with existing models deploying deep learning methods with classical Convolutional Neural Network (CNN) models, the proposed framework utilizes Gated Recurrent Unit (GRU) or Long Short-term Memory (LSTM) or our proposed Concat-CNN models to learn the characteristic correlation between AUs of distinctive frames. The Concat-CNN uses three convolutional kernels with different sizes to observe features of different duration, and emphasizes both local and global mutation features by changing dimensionality (max-pooling size) of the output space.

2.4 Results & Analysis

Table 2: Spotting results reported in the submitted papers.

ID	CAS(ME)$\hat{2}$-cropped			SAMM-LV		
	MaE	ME	all	MaE	ME	all
3288	**0.377**	0.042	0.325	0.281	0.131	0.238
3289	0.2515	0.2275	0.2466	0.2395	0.1969	0.2213
3291	0.125	0.025	0.1168	0.1469	0.0125	0.1257
3304			0.1763			0.136
3308	0.2505	0.0153	0.2019	0.3553	0.1155	0.2736
3314	0.3782	0.1965	0.3436	0.4149	0.2162	0.3638
3325	0.2608	0.1419	0.2269	0.2176	0.1526	0.1884
3326	0.0864	0.0175	0.0754	0.2011	0.3042	0.1316
3344	0.3252	0.0333	0.278	0.2936	0.0506	0.1667

Table 3: Results of the combined datasets calculated from the submitted log files.

CAS(ME)$\hat{2}$-cropped + SAMM-LV					
ID	TP sum	Positive Sum	Recall	Precision	F1
3314	**367**	**1218**	**0.4272**	**0.3013**	**0.3534**
3288	**202**	**628**	**0.2352**	**0.3217**	**0.2717**
3308	**196**	**740**	**0.2282**	**0.2649**	**0.2452**
3289	245	1283	0.2852	0.191	0.2288
3325	225	1327	0.2619	0.1696	0.2059
3344	229	1394	0.2666	0.1643	0.2033
3326	303	2849	0.3527	0.1064	0.1634
3304	132	870	0.1537	0.1517	0.1527
3291	121	1136	0.1409	0.1065	0.1213

3 GENERATION CHALLENGE

The goal of this challenge is to generate specific micro-expression (source) on the given template faces (target). For this challenge, we focus on three commonly used micro-expression databases: the Chinese Academy of Sciences Micro-Expression Database II (CASME II) with 247 FMEs at 200 fps, SMIC-E with 157 FMEs at 100 fps and Spontaneous Facial Micro-Movement Dataset (SAMM) with 159 FMEs at 200 fps. By evaluating the authenticity and strength of the generated micro-expression via psychologists' inspection, reliable micro-expression generation can enable proper data augmentation of FMEs, thereby promoting the further development of micro-expression analysis.

3.1 Databases

CASME II [10]: CASME II contains 26 subjects and 255 ME sequences. All videos are at 200 fps to retain more facial information and the resolution is 640 × 480. The onset, apex, offset index for these expressions are given in the excel file. In addition, the eye blinks are labeled with onset and offset time.

SAMM[1] The original SAMM dataset [1] with 159 micro-expressions. In SAMM Long Videos dataset [13], there are 147 videos. 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 onset frame to offset frame. In this database, all the micro-movements are labeled. Thus, the spotted frames can indicate not only micro-expression but also other facial movements, such as eye blinks.

SMIC[6]: includes three subsets: SMIC-HS, SMIC-VIS and SMIC-NIR. SMIC-VIS and SMIC-NIR contrains 71 samples recorded by normal speed cameras with 25 fps of visual (VIS) and near inferred (NIR) light range, respectively. SMIC-HS recorded by 100 fps high-speed cameras contains 164 spontaneous micro-expression clips from 16 subjects. These micro-expression clips are divided into three classes: positive, negative, and surprise.

3.2 Metrics

3.2.1 Submission Video Format. Each database specifies three kinds of emotion samples (positive, negative, and surprised), i.e. source samples, as listed in Table **??**.

Table 4: Sequence names of the assigned emotion samples for the generation task (Source).

Databases	CASME II	SAMM	SMIC-HS
Positive	EP01_01f	022_3_3	s3_po_05
Negative	EP19_06f	018_3_1	s11_ne_02
Surprise	EP01_13	007_7_1	s20_sur_01

The participants should generate the specified expressions on the two provided template faces, which are Asian female (from CASME I [11]) and western male (from SMIC-VIS) respectively (as shown in Fig. 1), i.e. target samples.

Figure 1: Normalized template faces for generation (Target).

The expected output video should generated the facial micro expression from source samples videos on the target face, as illustrated in Fig. 2.

The total number of videos for submission is 18, i.e. 2 templates × 3 emotions for each database. All submitted videos should be unified at 100fps, with a resolution of 256 × 256. The length of the generated video is based on the specified emotion sample and does not need to be normalized.

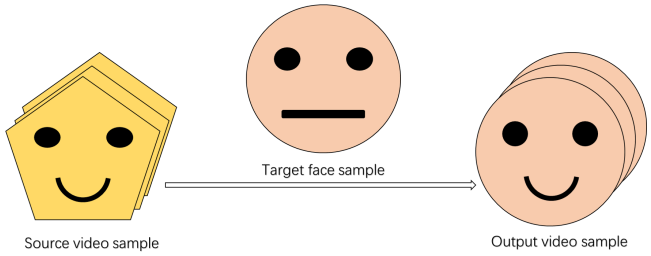

Figure 2: An example of generation task output video.

3.2.2 Evaluation Protocol. Each generated image will be evaluated based on the quality and action units. Specifically, the facial region will be divided into upper and lower parts (see Fig. 3) and evaluated separately. By separating the face into two parts, evaluations can take into account partial facial movements that may occur.

The quality and action unit of each block will be scored 0-3 by experts who have Facial Action Coding System (FACS) certification [3]. The following details the score categories:

- Score 0: Completely incorrect
- Score 1: Poor
- Score 2: Good
- Score 3: Excellent

In addition, there will be a 'noise' category, which judges the overall generation's image quality and is also score 0-3. For example, if a generation has background artifacts, this would reduce the noise score.

The maximum available score will be 9. Three experts will evaluate the generated images individually and the final score will be the average of scores provided by three experts.

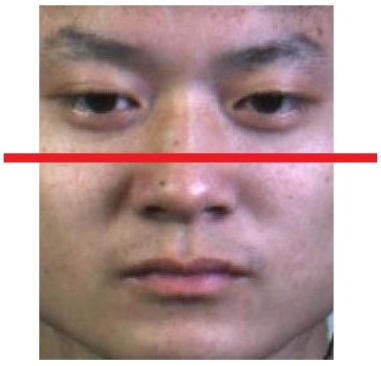

Figure 3: A basic representation of the upper and lower face used for assessment.

3.3 Methods

[17] We try to formulate micro-expression generation task and then propose a novel method using first order motion model based on facial prior. In our method, Region-Focusing Module obtains our designed facial prior feature. Motion Prediction Module estimates facial motions using key points and local affine transformations.

Finally in Expression Generation module, generative adversarial network is applied, driving the target face to generate FME videos.

[2]. To solve them, we propose to use a deep learning based method for facial micro-expression generation. First, to be able to capture subtle changes, we use a deep motion re-targeting network that can learn landmarks in a self-supervised manner and generate dense motion between reference and desired images. Second, to alleviate the problem of lack of training data, we apply deep transfer learning by borrowing knowledge from macro-expression generation.

[9] As a result, current data-driven machine learning methods are easy to be over-fitting and hard to obtain discriminative feature representations of MEs. To address this challenge and inspired by the current face generation technology, in this paper we introduce Generative Adversarial Network based on fine-grained AUs modulation to generate MEs sequence (FAMGAN) with different magnitudes of AUs. In particular, to guarantee the continuity of motion of the generated MEs sequence, we propose a fine-grained combination method to fine-tune the AUs.

3.4 Results & Analysis

Table 5: Overall evaluation of ME generation task

	Original score				Normalized score			
ID	E1	E2	E3	Total	E1	E2	E3	Total
3295	139	101	76	316	0.99	0.94	1.00	2.94
3282	140	107	56	303	1.00	1.00	0.74	2.74
3320	104	66	66	236	0.74	0.62	0.87	2.23
3311	85	51	37	173	0.61	0.48	0.49	1.57

4 CONCLUSION AND FUTURE CHALLENGES

5 ACKNOWLEDGMENTS

We would like to thank the ACM MM '21 conference organizers for agreeing to host our workshop and for their support, and all reviewers for their time and helpful contributions. This work is supported by grants from the National Natural Science Foundation of China (61772511, U19B2032), the Academy of Finland (Grant 323287), The Royal Society (INF/PHD/180007), 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] Xinqi Fan, Ali Raza Shahid, and Hong Yan. 2021. Facial Micro-Expression Generation based on Deep Motion Retargeting and Transfer Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4735–4739.

[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] 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.

[5] Jingting Li, Su-Jing 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)*. IEEE Computer Society, 234–237.

[6] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. 2013. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (FG)*. IEEE, 1–6.

[7] 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.

[8] 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.

[9] Yifan Xu, Sirui Zhao, Huaying Tang, Xinglong Mao, Tong Xu, and Enhong Chen. 2021. FAMGAN: Fine-grained AUs Modulation based Generative Adversarial Network for Micro-Expression Generation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4813–4817.

[10] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* 9, 1 (2014), e86041.

[11] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. 2013. *CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces*. IEEE, 1–7.

[12] Bo Yang, Jianming Wu, Zhiguang Zhou, Megumi Komiya, Koki Kishimoto, Jianfeng Xu, Keisuke Nonaka, Toshiharu Horiuchi, Satoshi Komorita, Gen Hattori, et al. 2021. Facial Action Unit-based Deep Learning Framework for Spotting Macro-and Micro-expressions in Long Video Sequences. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4794–4798.

[13] 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.

[14] 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.

[15] Wang-Wang Yu, Jingwen Jiang, and Yong-Jie Li. 2021. LSSNet: A Two-stream Convolutional Neural Network for Spotting Macro-and Micro-expression in Long Videos. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4745–4749.

[16] He Yuhong. 2021. Research on Micro-Expression Spotting Method Based on Optical Flow Features. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4803–4807.

[17] Yi Zhang, Youjun Zhao, Yuhang Wen, Zixuan Tang, Xinhua Xu, and Mengyuan Liu. 2021. Facial Prior Based First Order Motion Model for Micro-expression Generation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4755–4759.