MEGC 2019 – The Second Facial Micro-Expressions Grand Challenge

John See[1], Moi Hoon Yap[2], Jingting Li[3], Xiaopeng Hong[4], Su-Jing Wang[5]
[1] Faculty of Computing and Informatics, Multimedia University, Malaysia
[2] Faculty of Science and Engineering, Manchester Metropolitan University, UK
[3] CentraleSuplec, CNRS, IETR, UMR 6164, F-35000 Rennes, France
[4] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[5] Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

***Abstract*— Automatic facial micro-expression (ME) analysis is a growing field of research that has gained much attention in the last five years. With many recent works testing on limited data, there is a need to spur better approaches that are both robust and effective. This paper summarises the 2nd Facial Micro-Expression Grand Challenge (MEGC 2019) held in conjunction with the 14th IEEE Conference on Automatic Face and Gesture Recognition (FG) 2019. In this workshop, we proposed challenges for two micro-expression (ME) tasks—spotting and recognition, with the aim of encouraging rigorous evaluation and development of new robust techniques that can accommodate data captured across a variety of settings. In this paper, we outline the evaluation protocols for the two challenge tasks, the datasets involved, and an analysis of the best performing works from the participating teams, together with a summary of results. Finally, we highlight some possible future directions.**

I. Introduction

Facial micro-expressions (MEs) are movements of the face that occur in a spontaneous, involuntary fashion when a person attempts to conceal or hide a particular emotion upon experiencing it. These MEs can typically be found in a high-stakes situations such as criminal interrogations and interviews [1], political debates [2], and poker games [3]. As such, computational analysis and automation of several tasks on micro-expression video has become an emerging area in face processing research, with an increasingly strong interest in the last five years. A number of annotated datasets have emerged: the SMIC dataset [4], and the FACS-coded Chinese Academy of Sciences Micro-Expression Database II [5] and Spontaneous Micro-Facial Movement Dataset (SAMM) [6], giving rise to further advances in this field of study.

Since the inception of the first Micro-Expression Grand Challenge (MEGC) 2018 [7], the second edition of this workshop aims to spur new ideas and approaches with focus on the two primary ME tasks: *spotting* and *recognition*. Spotting ME remains a challenge due to the lack of ME annotated long videos. Most methods proposed either measure feature differences between the frames [8] or use its neutral video as individualised baselines [9]. The spotting challenge introduces two long video benchmark datasets and standardises the performance metrics. As for the ME recognition task, a vast majority of works in literature have mainly focused on building models and approaches that only cater for a specific dataset; if a few datasets are used, models or features are typically constructed individually for each dataset. As such, the rigour and realism in evaluation is most thoroughly lacking in current mode of practices. This edition of the recognition challenge pushes the boundaries of current evaluation schemes towards a comprehensive composite database that comprises of samples collected from different environments, from a diverse range of subjects.

This workshop aims to narrow the gaps found in these tasks and to continuously promote interactions between researchers and scholars from within this area of study, and also those from broader areas of psychology and physiology. Besides the two challenges, we also solicited original works that address a variety of challenges in the computational aspect of ME research, including that of other related fields of neuroscience, psychology and physiology.

II. Spotting Challenge

The goal of this challenge is to spot micro-movement intervals (from onset to offset) in long video sequences. In this challenge, we focus on detecting 57 micro-movements of CAS(ME)^2 database [10] and 159 micro-movements of SAMM database [6].

A. Databases

CAS(ME)^2 [10] and SAMM long videos [6] are among the most recent databases with full annotations including FACS coding, onset, apex, offset frames, and the intensity of the facial movements. CAS(ME)^2 [10] consists of 22 participants with 97 long video sequences, but only 32 of the sequences contained MEs. Meanwhile, SAMM long videos consist of 224 videos in total, i.e. 32 participants with 7 videos each, with MEs existing in only 79 videos. For this challenge, we focus only on sequences with MEs, where the organizers have also provided the cropped version of the SAMM long videos from [9]. For clarity, the summary of the databases is shown in Table I.

TABLE I: Summary of CAS(ME)^2 and SAMM long videos for ME Spotting Challenge.

Database	ME Sequences	Resolution	Frame rate (fps)
CAS(ME)^2	32	640×480	30
SAMM	79	2040×1088	200

B. Performance Metrics

To evaluate the spotting performance, we first calculate several standard measures: *True Positive* (*TP*), *False Positive* (*FP*) and *False Negative* (*FN*). A spotted video interval (detected onset frame to detected offset frame), denoted as $W_{spotted}$, is categorised as *TP* if the following condition is fulfilled:

$$\frac{W_{spotted} \cap W_{gt}}{W_{spotted} \cup W_{gt}} \geq 0.5 \tag{1}$$

where W_{gt} represents the ground truth ME interval, i.e. [onset, offset]. Otherwise, the detected interval is regarded as *FP*. *FN* is counted when the algorithm failed to detect the ground truth interval. To obtain an overall result, we use *F1-Score* as denoted by:

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

C. Methods

The lone submission for the spotting challenge was by Li et al. [11], which proposed the use of local temporal patterns [12] (LTP-ML). The long video is divided into short clips by a sliding window. Then, 12 facial regions-of-interests (ROIs) are selected, and the sequences of these ROIs are processed by PCA across the time axis to conserve the main temporal movement. After distance calculation and normalization, the local motion pattern is analyzed according to the average ME duration (300 ms). Based on the general LTP for ME sequences, local movements on the ROIs are classified by a SVM classifier. Finally, the global spotting result is obtained by a fusion process, which includes temporal qualification, spatial selection and a merge process.

D. Results and Analysis

Table II shows the results of LTP-ML when compared to the state of the art LBP-χ^2-distance (LBP-χ^2) method [8]. SAMM^f represents the full frame (whole image) SAMM while SAMM^c represents the cropped face version. The proposed LTP-ML performs better than the baseline LBP-χ^2 method. LBP-χ^2 method spots the maximal movement in the video sequence, and it works well in short videos which contain only a single micro-expression. However, in SAMM and CAS(ME)^2, video sequences are longer and may contain irrelevant facial movements or macro-expressions. Meanwhile, LTP-ML spots micro-expression by recognizing the local temporal pattern, i.e. facial movements of varying motion patterns are better distinguished by this algorithm. However, the high number of *FP* has deteriorated the overall performance. This is due to the fact that the raw video sequences are noisy and consists of macro-movements and eye blinks.

III. Recognition Challenge

The 1st MEGC [7] saw the establishment of the cross-database challenge, which used a combination of two datasets (CASME II and SAMM), with objective class labels as proposed in [13]. In this 2nd MEGC, the cross-database challenge increases its coverage to include the classic SMIC [4] dataset, which is one of the earliest spontaneous micro-expression dataset to be created. The motivation behind this challenge is to mimic a more realistic scenario by:

- Increasing the number of subjects considered in the system, particularly with subjects, captured from different environment and settings. This also increases the overall number of video samples, which facilitates the use of more contemporary machine learning or deep learning techniques that are data-driven in nature;
- Using a reduced set of general emotion classes to better accommodate contrasting types of emotions which have been elicited from different stimuli and environment setup. This also reduces the ambiguity in the elicited emotions caused by such differences.

To enable all three datasets to be used together, a reduced set of common emotion classes are used. The original emotion classes are mapped to three distinct classes (original classes in parentheses):

- **Negative** (i.e. 'Repression', 'Anger', 'Contempt', 'Disgust', 'Fear' and 'Sadness')
- **Positive** ('Happiness'), and
- **Surprise** ('Surprise')

Videos containing other unrelated and undefined emotions are omitted, particularly the 'Others' class from CASME II, which consists of a diverse mix of all kinds of emotions. These samples are likely to cause confusion to model training if included together. Table III shows the summary of the distribution of samples for all three datasets.

A. Composite Database Evaluation (CDE) Protocol

In contrary to the dual protocol used in MEGC 2018, this challenge only adopts the Composite Database Evaluation (CDE), i.e. samples from all datasets are combined into a single composite database based on the reduced emotion classes. Leave-one-subject-out (LOSO) cross-validation is used to determine the training-test splits. With a total of 68 subjects (16 from SMIC, 24 from CASME II, 28 from SAMM), evaluation is repeated 68 times by holding out test samples of each subject group while the remaining samples are used for training. This protocol mimics a realistic scenario where people from diverse backgrounds (ethnicity, gender, emotional sensitivities) are "enrolled" separately in different environment and settings, in a single recognition system. The LOSO cross-validation also ensures subject-independent evaluation.

B. Performance Metrics

The composite database is clearly imbalanced in terms of its class distribution, i.e. the distribution for sur-

TABLE II: F1-Score of LTP-ML and LBP-χ^2 for ME spotting from long videos.

Database	SAMM^c	SAMM^f	CAS(ME)^2
LTP-ML	**0.0316**	0.0229	**0.0179**
LBP-χ^2	0.0055	N/A†	0.0035

†This method requires cropped faces, so SAMM^f is not applicable.

TABLE III: 3-class sample distribution of all datasets for CDE

Emotion Class	SMIC	CASME II	SAMM	3DB-combined
Negative	70	88†	92‡	250
Positive	51	32	26	109
Surprise	43	25	15	83
TOTAL	164	145	133	442

†Negative class of CASME II: Disgust and Repression.
‡Negative class of SAMM: Anger, Contempt, Disgust, Fear and Sadness.

prise:positive:negative classes are in the ratio of $1 : 1.3 : 3$ [1]. To properly handle such class imbalances [14], the performance is to be reported with two balanced metrics:

1) Unweighted F1-score (UF1): This metric is also known as the *macro-averaged F1-score*. This flavour of F1-score is a good choice in imbalanced multi-class settings for providing equal emphasis on rare classes[2]. To compute this, first obtain all the True Positives (TP), False Positives (FP) and False Negatives (FN) over all k folds of LOSO for each class c (of C classes), and proceed to compute their respective F1-scores. The unweighted F1-score (UF1) is determined by averaging the per-class F1-scores, $F1_c$:

$$\text{F1}_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{3}$$

$$\text{UF1} = \frac{F1_c}{C} \tag{4}$$

2) Unweighted Average Recall (UAR): This metric is also known as the *balanced accuracy* of the system. This is a more reasonable metric in place of the standard Accuracy (or Weighted Average Recall) metric which may be bias towards classifiers that predict the larger classes well. In similar manner, the per-class accuracy scores are first computed, before averaging by the number of classes:

$$\text{UAR} = \frac{1}{C} \sum_c \frac{TP_c}{n_c} \tag{5}$$

where n_c is the number of samples of the c-th class.

Both these metrics provide a balanced judgement whether an approach can predict all classes equally well, hence reducing the possibility that an approach could be well-fitted to only work for certain classes.

C. Methods

This section summarises the methods proposed by the top four submitting teams [16], [17], [18], [19], in terms of recognition performance. There were a total of seven submissions; four were accepted into the workshop proceedings, the remaining three unaccepted submissions are not described in this summary paper.

It is interesting to note that all four approaches opted to utilize the apex frame as the choice of input as inspired by the work of Liong et al. [20]. In the case of [18], a mid-position frame between the onset and offset is used as the approximated apex frame. We briefly summarize these four approaches as follows:

1) Expression Magnification and Reduction (EMR) with Adversarial Training [16]: The authors proposed a part-based deep neural network approach with two domain adaptation techniques – adversarial domain adaptation and motion magnification and reduction, which help to enrich the available training samples. Motivated by domain-adversarial framework proposed in [21], the authors made use of macro-expression samples from CK+ dataset [22] together with micro-expression samples to minimize the combined loss function. The authors did not disclose how the apex frame for the SMIC samples were obtained as they were not annotated.

2) Shallow Triple Stream Three-dimensional CNN (STST-Net) [17]: Motivated by the observation that very deep CNN architectures do not perform well under limited ME data, the authors proposed a shallow 3-D CNN which comprises of three parallel streams, each with a different number of feature maps to curb underfitting. Optical flow guided features (optical strain, and the horizontal and vertical flows) form the input cube for further network learning. The authors also re-implemented a few notable approaches (e.g. [20]) and some vanilla deep CNNs for the purpose of performance benchmarking.

3) Dual-Inception Network [18]: The authors proposed a two-stream two-block variant of the Inception network [23] to learn a robust feature representation from the horizontal and vertical components of TV-L1 optical flow information. Their pipeline first takes the onset and mid-position (instead of apex) frames to compute optical flow before allowing the proposed network to learn from two identical streams consisting of two stacked Inception blocks. Their corresponding convolutional feature maps are merged, and appended with one fully-connected layer and final softmax layer.

4) CapsuleNet [19]: With standard CNNs limited by their weakness in representing part-whole relationships, the authors proposed the use of Capsule Networks (CapsuleNet) [24] which has been successful in general object recognition. This is likely the first known use of Capsule Networks for ME recognition. CapsuleNet takes in the ME apex frame To avoid resizing the input frames to fit the original CapsuleNet architecture, the authors used a ResNet18 to obtain reduced-size local features for the primary capsule layer. Another output capsule layer captures the learned weight matrix which is refined via a dynamic routing process. They also proved their method to be better than a few full-fledged CNN models.

D. Results & Analysis

In this section, we report the top four submissions of the challenge and provide some analysis and observations. In total, we received submissions from seven participating teams. The results reported in this section were reported by the respective submitted papers, and verified again using the submitted output logs. We will not be reporting results

[1]Accuracy of the system is 0.565 simply by making a random naive Negative class prediction

[2]See the work of Forman & Scholz [15] which advocates this as the most unbiased way of calculating F1-score in a k-fold cross-validation setting. It caters well for cases of strong class imbalance.

TABLE IV: Results of the top four submissions to the Recognition Challenge against various baseline and recent methods (first 3 rows).

Method	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [25]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
Bi-WOOF [20]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
OFF-ApexNet [26]	0.7196	0.7096	0.6817	0.6695	**0.8764**	0.8681	0.5409	0.5392
Quang et al. [19]	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
Zhou et al. [18]	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
Liong et al. [17]	0.7353	0.7605	0.6801	0.7013	0.8382	**0.8686**	0.6588	0.6810
Liu et al. [16]	**0.7885**	**0.7824**	**0.7461**	**0.7530**	0.8293	0.8209	**0.7754**	**0.7152**

from the other three unaccepted papers, all of which, fared poorer overall compared to the four reported here. Prior to submission, all participants have also been requested to provide their code implementations in GitHub. This would encourage reproducibility of the proposed methods, and also invite future researchers to contribute to this field of study.

Table IV summarizes the performance (in both UF1 and UAR) of the four submitting methods, against the hand-crafted LBP-TOP [25] and Bi-WOOF [20] baselines, and also a recent neural network method named OFF-ApexNet [26] proposed for ME recognition on composite (combined) databases.

Overall, the method by Liu et al. [16] emerged as the method with the best overall result on the full composite database, outperforming the other competing submissions on the SMIC and SAMM subsets. The work of Liong et al. [17] is notably strong on the CASME II, obtaining the highest balanced accuracy (UAR) across the board.

From these results, we highlight a number of interesting observations.

1) The EMR method with adversarial training [16] perform exceedingly well on the SAMM database, more than 0.11 higher (UF1) than the closest competitor. Their domain adaptation using expression-reduced CK+ samples appears to work well in the SMIC and SAMM, but not so for CASME II, which contain predominantly Chinese subjects.
2) The method of Zhou et al. [18] chalked up reasonably strong scores, by opting to use the mid-position frames as a reasonable substitute for the "apex" frame (they observed that most apex frames are located in the middle part of the sequence) rather than the more precisely annotated apex frames. This actually circumvents the lack of apex information, particularly for the case of SMIC. Of course, this is somewhat inconclusive until further studies are conducted on fixed methods.
3) The top 3 works all used optical flow as their choice of input data, rather than relying on pixel intensities. This points towards the advantages of employing more discriminative information for cases of extremely subtle facial changes.
4) SMIC and SAMM remained the more challenging datasets, as compared to CASME II. There could be various possible reasons: the SMIC dataset was captured at a slower frame rate and lower resolution whereas the SAMM dataset contains quite a diverse range of age and ethnicity. These factors are likely to contribute towards limited recognition capability in these two datasets.

IV. Conclusion and Future Challenges

This summary paper highlights the second Facial Micro-Expressions Grand Challenge (MEGC) workshop and the two sub-challenges for ME spotting and recognition. With a total of eight submissions for both sub-challenges, the highest ever obtained, we observe an encouraging momentum in ME research with further interesting avenues worth investigating.

This challenge focuses on the tasks of ME spotting and recognition; the latter still obviously more popular than the former. As ME spotting is still an important and practical problem, we urge the research community to give more attention to novel propositions and ground-breaking ideas. Robust and accurate spotting of ME occurrences from long videos or unconstrained "in-the-wild' settings could be beneficial to advance this field further. Further to this workshop, the idea of domain transfer or adaptation is one that could be viable for harnessing additional micro-expression cues from various sources such as gestures, body language and physiological signals.

Acknowledgement

We would like to thank Image Metrics Ltd for sponsoring the prizes for the challenge winners. The workshop chairs would like to thank their funders: National Natural Science Foundation of China (61772511, 61472138, 61572205), The UK Royal Society Industry Fellowship (IF160006), MOHE Malaysia Grant No. FRGS/1/2016/ICT02/MMU/02/2, Shanghai 'The Belt and Road' Young Scholar Exchange Grant (17510740100), Academic of Finland, Tekes, Infotech OuluChina scholarship council and ANR Reflet.

References

[1] D. Luciew, J. Mulkern, and R. Punako, "Finding the truth: interview and interrogation training simulations," in *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2011.

[2] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–4.

[3] P. Husák, J. Cech, and J. Matas, "Spotting facial micro-expressions in the wild," in *22nd Computer Vision Winter Workshop (Retz)*, 2017.

[4] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[5] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, 2014.

[6] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, Jan 2018.

[7] M. H. Yap, J. See, X. Hong, and S.-J. Wang, "Facial micro-expressions grand challenge 2018 summary," in *13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 675–678.

[8] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 1722–1727.

[9] A. K. Davison, M. H. Yap, and C. Lansley, "Micro-facial movement detection using individualised baselines and histogram-based descriptors," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, 2015, pp. 1864–1869.

[10] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)ˆ2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. on Affective Computing*, 2017.

[11] J. Li, C. Soladié, R. Séguier, S. Wang, and M. H. Yap, "Spotting micro-expressions on long videos sequences," *arXiv preprint arXiv:1812.10306*, 2018.

[12] J. Li, C. Soladié, and R. Séguier, "LTP-ML: Micro-expression detection by recognition of local temporal pattern of facial movements," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, 2018, pp. 634–641.

[13] A. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *Journal of Imaging*, vol. 4, no. 10, p. 119, 2018.

[14] A. C. Le Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Computer Vision–ACCV 2014*. Springer, 2014, pp. 33–48.

[15] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, 2010.

[16] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Automatic Face & Gesture Recognition (FG 2019), 2019 14th IEEE International Conference on*, 2019.

[17] 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," in *Automatic Face & Gesture Recognition (FG 2019), 2019 14th IEEE International Conference on*, 2019.

[18] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Automatic Face & Gesture Recognition (FG 2019), 2019 14th IEEE International Conference on*, 2019.

[19] N. V. Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *Automatic Face & Gesture Recognition (FG 2019), 2019 14th IEEE International Conference on*, 2019.

[20] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94–101.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[24] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in NIPS*, 2017, pp. 3856–3866.

[25] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[26] Y. 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*, 2019.