2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Facial Micro-Expressions Grand Challenge 2018 Summary

Moi Hoon Yap[1], John See[2], Xiaopeng Hong[3], Su-Jing Wang[4]
[1] School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, UK
[2] Faculty of Computing and Informatics, Multimedia University, Malaysia
[3] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[4] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, 100101, China

***Abstract*— This paper summarises the Facial Micro-Expression Grand Challenge (MEGC 2018) held in conjunction with the 13th IEEE Conference on Automatic Face and Gesture Recognition (FG) 2018. In this workshop, we aim to stimulate new ideas and techniques for facial micro-expression analysis by proposing a new cross-database challenge. Two state-of-the-art datasets, CASME II and SAMM, are used to validate the performance of existing and new algorithms. Also, the challenge advocates the recognition of micro-expressions based on AU-centric objective classes rather than emotional classes. We present a summary and analysis of the baseline results using LBP-TOP, HOOF and 3DHOG, together with results from the challenge submissions.**

I. INTRODUCTION

Facial micro-expressions (MEs) are involuntary movements of the face that occur spontaneously when a person experiences an emotion but attempts to suppress the facial expression, typically found in a high-stakes environment. Computational analysis and automation of tasks on micro-expressions is an emerging area in face research, with a strong interest appearing as recent as 2014. Only a few spontaneously induced facial micro-expression datasets have provided the impetus to advance further from the computational aspect. Particularly comprehensive are two state-of-the-art FACS coded datasets: the Chinese Academy of Sciences Micro-Expression Database II (CASME II) [1] and the Spontaneous Micro-Facial Movement Dataset (SAMM) [2], there has been no attempts to introduce a more rigorous and realistic evaluation. This is the inaugural workshop with the aim of promoting interactions between researchers and scholars from within this niche area of research, and also those from broader, general areas of computer vision and psychology research. The focus of the grand challenge is on the CASME II-SAMM cross-database recognition of micro-expression classes. We hope to solicit original works that address a variety of modern challenges of ME research such as spotting macro-/micro-expressions from long videos, and deep learning techniques.

II. OBJECTIVE CLASSES

The emotion classes in the existing datasets were based on both self-report and Action Units (AUs) classification. Fig 1 and Fig. 2 illustrate the happiness category from CASME II and SAMM, but were coded with different AUs. These inconsistencies adds further justification for the introduction of new classes based on AUs only [3].

This challenge aims to stimulate the micro-expressions researchers in developing new techniques for the AU-centric objective classes. A summary of the objective classes are as illustrated in Table I. A single composite database for this experiment has a total of 253 micro-expressions.

TABLE I
THE TOTAL NUMBER OF MOVEMENTS ASSIGNED TO THE NEW OBJECTIVE CLASSES FOR CASME II AND SAMM.

Class	**CASME II**	**SAMM**	Composite
I	25	24	49
II	15	13	28
III	99	20	119
IV	26	8	34
V	20	3	23
Total	185	68	253

III. CROSS-DATABASE CHALLENGE TASKS

The following newly established protocols are proposed to tests for the robustness in learning salient characteristics of micro-expressions that are universal in nature.

A. Task A: Holdout-database Evaluation (HDE)

Training and testing sets are to be sampled from different micro-expression databases (CASME II and SAMM) and evaluated. The training and testing sets are then swapped and the process is repeated for a second time (similar to a 2-fold cross validation, i.e. Train on CASME II and test on SAMM, vice versa). This protocol mimics a realistic scenario where characteristics of micro-expressions learnt in one system may be transferred to another system (with another group of people enrolled).

Performance Metric: Performance of this task is to be measured by two metrics that are typically used in cross-database speech emotion recognition, i.e., unweighted average recall (UAR) and weighted average recall (WAR) [4]. WAR is the normal recognition accuracy (i.e. number of correctly classified samples divided by the total number of samples), while UAR is the "balanced" recognition accuracy (i.e. sum of accuracy of each class divided by the number of classes without considerations of samples per class). It is most desired if a method obtains a high WAR and an equally

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00106

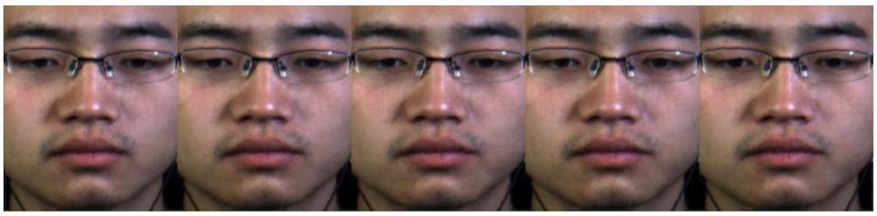

Fig. 1. Sample frames from CASME II database showing Subject 6's micro-expression clip 'EP01_01' that was coded as L14+L15 in the 'happiness' category.

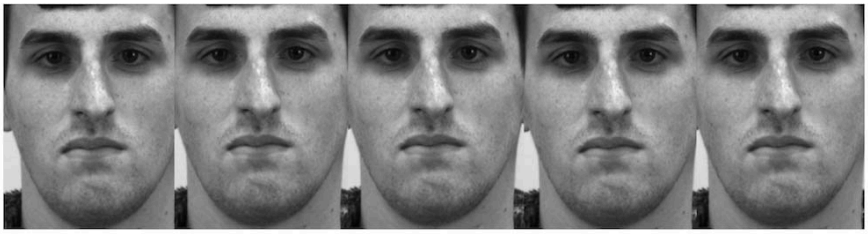

Fig. 2. Sample frames from SAMM database showing Subject 7's micro-expression clip '007_6_3' that was coded as an R6A+R12 in the 'happiness' category.

high UAR. A method is also seen as favourable to certain dominant classes only if the gap between WAR and UAR is large. The WAR and UAR of both folds are averaged to obtain the overall WAR and UAR scores.

B. Task B: Composite database evaluation (CDE)

All samples from both databases (CASME II and SAMM) are combined into a single composite database, on the basis of objective classes. Leave-One-Subject-Out (LOSO) cross-validation is used to determine the training-testing splits (i.e. each subject group is held out as the testing set while all remaining samples are used for training). There are altogether 55 subjects (26 from CASME II and 29 from SAMM) after combining both databases. This protocol mimics a realistic scenario where a more diverse group of people are enrolled to a single system, with subject-independent testing.

Performance Metric: F1-Score (of F-measure) is proposed as the metric for imbalanced databases [5]. Using the conventional accuracy measure may result in a bias towards classes with large number of samples (i.e. naive classification of all samples to a large-sample class would result in much higher result than random chance), hence overestimating the capability of the method. The overall F1-score should micro-averaged across the whole database, i.e. it should be calculated based on the total true positives, false negatives and false positives, across all LOSO folds.

IV. Methodology

This section summarises the methods proposed by three submitting teams.

A. Baseline methods for MEGC 2018 by Merghani et al.

To establish some baselines for the MEGC challenge, Merghani et al. implemented three popular feature descriptors commonly used in micro-expression representation: LBP-TOP, a local spatio-temporal descriptor which was first introduced by Zhao et al. [6]; 3DHOG which had been used in early works by Polikovsky et al. [7], [8] and HOOF, which has been advocated by Lui et al. [9]. For all these descriptors, the authors divided the image frame into 5x5 blocks, a widely used method in micro-expressions analysis [1], [10], [11], [12]. Finally, Sequential Minimal Optimization (SMO) [13] was used for classification. Above the similar methods, the authors also proposed a selective block-based feature fusion representation method for Task B (CDE). Since not all 25 blocks correspond to facial movement, a pre-defined set of salient blocks were used to extract all three descriptors, which are then concatenated into a single histogram.

B. ELRCN by Khor et al.

Khor et al. proposed an enriched version of the Long-term Recurrent Convolutional Network [14] called ELRCN, which comprises of a deep hierarchical spatial feature extractor and a temporal module that characterizes temporal dynamics. Enrichment was achieved through two variety of ways; first, by stacking the input channel with additional low-level information, namely optical flow and optical strain [15], [16]; secondly, by stacking deep spatially-encoded features to enrich the temporal dynamics representation. The VGG-Faces [17] deep CNN model was the authors' choice of the base spatial feature extractor. Temporal interpolation model (TIM) was used to constrain each video sequence to a fixed length of 10 frames due to requirements of the temporal module. Classification was achieved using a basic linear SVM.

C. Macro to micro transfer learning by Peng et al.

The third participating work by Peng et al. employed a transfer-learning-based approach to applying deep CNN

on small ME datasets. Specifically, the authors first fine-tune an ImageNet-pretrained ResNet10 [18] on four macro-expression datasets – the Extended Cohn-Kanade (CK+) [19], Oulu CASIA NIR & VIS [20], JAFFE [21] and MUGFE [22]. Using a large number of images (>10k) from these macro-expression datasets, a resampling technique was applied to provide balance training with further alterations by colour shift, rotation and smoothing applied. Then, the model is further fine-tuned on the CASME II and SAMM database using the apex frame of each sample.

V. Challenge Results

In this section, we report and analyze the results of the challenge. We have received submissions from three teams of participants. All three teams have submitted the results for Task B: Composite Database Evaluation (CDE). Meanwhile for Task A: Holdout-Database Evaluation (HDE), results of two methods were received. The results reported in this section are computed using the output result logs submitted by the participants.

A. Task A: Holdout-Database Evaluation (HDE)

Table II summarizes the results of HDE. The overall result shows the method based on the transferred ResNet10 model proposed by Peng et al. performs the best with an average WAR of 0.561 and an average UAR of 0.389. The enriched long-term recurrent convolutional network with spatial dimension enrichment (SE) proposed by Khor et al. achieves the second-best performance. The three methods using manually designed features, namely LBP-TOP, 3DHOG, and HOOF, shows inferior performance. Nevertheless, HOOF is the best among the three, with a WAR of 0.353 and a UAR of 0.348. Moreover, HOOF surprisingly obtains the best UAR out of all methods when training on SAMM and testing on CASME II.

Moreover, there are two important observations. Firstly, deep learning methods outperform handcrafted features substantially in the HDE test; Secondly, as Table II indicates, the values of WAR and UAR may vary even for the same method. For example, when testing on the CASMEII dataset using SAMM as training data, the 3DHOG is ranked in the third position in terms of WAR, while it becomes the last one in terms of UAR. It is therefore suggested to use both measures jointly for comprehensive comparisons.

B. Task B: Composite Database Evaluation (CDE)

Table III lists the results of the CDE task (Task B) in terms of F1-score[1] and the weighted F1-score[2]. LBP-TOP, 3DHOG, HOOF are the baseline methods using a block size of 5×5. The other three methods are computed using the output result logs submitted by the participants.

[1]The F1-score computed here is an average of the class-specific F1-scores across the five classes (or *macro-averaging*) by stacking all results of the LOSO evaluation.

[2]The class-specific F1-scores are weighted by the number of samples in the corresponding classes before averaging.

The macro-to-micro transferred ResNet10 model proposed by Peng et al. outperforms the other competing methods substantially, with an F1-score of 0.639 and a weighted F1-score of 0.733. This suggests the great potential of using modern deep learning methods for cross-dataset subject-independent micro-expression recognition. The selective block-based feature fusion method by Merghani et al. achieves the second best performance. When comparing its F1-score and weighted F1-score against that of the three baseline methods, it can be easily observed that by appropriate feature fusion techniques, methods using handcrafted features are still able to improve their performance and obtain highly competitive results.

Among the three baseline methods, HOOF and LBP-TOP outperform, by a large margin, the 3DHOG. Moreover, it is interesting to find that their performance is marginally better than the enriched long-term recurrent convolutional network with spatial dimension enrichment (SE) proposed by Khor et al., which only has an F1-score of 0.393 and a weighted F1-score of 0.523. The inferior performance of the recurrent network is likely due to the inability of deep architecture in learning from small datasets alone, and possibly also, a significantly different face pre-processing procedure[3]. Moreover, unlike the method by Peng et al., where a pre-trained model from a large-scale dataset is available, there is no such pre-train process for the temporal recurrent LSTM model in the method by Khor et al. Thus, this suggests that there is room for the recurrent network to improve in the CDE task.

VI. Conclusion and Future Challenge

This challenge is the first Grand Challenge on facial micro-expressions. Three teams have taken part in the challenge. In view of the results, the submission by Peng et al. used a simple deep learning approach to obtain the best performances on both Task A and Task B. Although the number of micro-expression samples is very small and probably not suited to use deep learning techniques, the use of transfer learning from macro-expression samples offered a potential solution to achieve better performances for micro-expression recognition. Particularly, the broad availability of large-scale facial expression databases promotes such possibilities.

This challenge focuses on the task of micro-expression recognition. However, in practice, *micro-expression spotting* is also an essentially important problem, if not more important. The task of micro-expression spotting is to find the period of time which overlaps with the duration between onset frame and offset frame of a micro-expression. If we are able to spot the occurrence of micro-expressions from long videos or live video streams in an automated manner, we can provide assessment of whether a person is telling lies at that opportune moment. Future challenges will focus on micro-expression spotting. CAS(ME)^2 [23] database contains 87

[3]Khor et al. applied TIM10 with face-cropping by Face++ API, and alignment with DLib library for their SAMM samples. Peng et al. used AAM for face area segmentation. Merghani et al. used the original pre-cropped faces.

TABLE II

THE RESULTS OF HOLDOUT-DATABASE EVALUATION (TASK A).

Method	WAR			UAR		
	@SAMM	@CASME II	Average	@SAMM	@CASME II	Average
LBP-TOP	0.338	0.232	0.285	0.327	0.316	0.322
3DHOG	0.353	0.373	0.363	0.269	0.187	0.228
HOOF	0.441	0.265	0.353	0.349	**0.346**	0.348
Peng et al.	**0.544**	**0.578**	**0.561**	**0.440**	0.337	**0.389**
Khor et al.	0.485	0.384	0.435	0.382	0.322	0.352

TABLE III

THE RESULTS OF COMPOSITE DATABASE EVALUATION (TASK B) BASED ON LOSO CROSS VALIDATION.

Method	F1-Score	Weighted F1-score
LBP-TOP	0.400	0.524
3DHOG	0.271	0.436
HOOF	0.404	0.527
Peng at al.	**0.639**	**0.733**
Merghani et al.	0.454	0.579
Khor et al.	0.393	0.523

long videos that contain both macro- and micro-expressions. This provides an interesting challenge for future work in computational analysis of facial micro-expression.

ACKNOWLEDGEMENT

The 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, and Infotech Oulu.

REFERENCES

[1] 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.

[2] 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.

[3] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *arXiv preprint arXiv:1708.07549*, 2017.

[4] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[5] 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.

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

[7] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*. IET, 2009, pp. 1–6.

[8] S. Polikovsky and Y. Kameda, "Facial micro-expression detection in hi-speed video based on facial action coding system (facs)," *IEICE transactions on information and systems*, vol. 96, no. 1, pp. 81–92, 2013.

[9] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transaction of Affective Computing*, 2015.

[10] A. K. Davison, M. H. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: An investigation on spatio-temporal descriptors," in *Computer Vision-ECCV 2014 Workshops*. Springer, 2014, pp. 111–123.

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

[12] 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*. IEEE, 2015, pp. 1864–1869.

[13] J. Platt *et al.*, "Fast training of support vector machines using sequential minimal optimization," 1999.

[14] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. of IEEE CVPR*, 2015, pp. 2625–2634.

[15] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," in *Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE, 2009, pp. 1–6.

[16] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing: Image Communication*, vol. 47, pp. 170–182, 2016.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[18] M. Simon, E. Rodner, and J. Denzler, "Imagenet pre-trained models with batch normalization," *arXiv preprint arXiv:1612.01452*, 2016.

[19] 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*. IEEE, 2010, pp. 94–101.

[20] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[21] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 200–205.

[22] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in *Image analysis for multimedia interactive services (WIAMIS), 2010 11th international workshop on*. IEEE, 2010, pp. 1–4.

[23] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas (me)ˆ 2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, 2017.