3D-CNN for Facial Micro- and Macro-expression Spotting on Long Video Sequences using Temporal Oriented Reference Frame

Chuin Hong Yap
Centre for Advanced Computational Science, Manchester Metropolitan University, Manchester, UK
chuin.h.yap@stu.mmu.ac.uk

Moi Hoon Yap
Centre for Advanced Computational Science, Manchester Metropolitan University, Manchester, UK
M.Yap@mmu.ac.uk

Adrian Davison
University of Manchester
Manchester, UK

Connah Kendrick
Centre for Advanced Computational Science, Manchester Metropolitan University, Manchester, UK

Jingting Li, Su-Jing Wang
CAS Key Laboratory of Behavioral Science, Institute of Psychology
Beijing, China

Ryan Cunningham
Centre for Advanced Computational Science, Manchester Metropolitan University, Manchester, UK
R.Cunningham@mmu.ac.uk

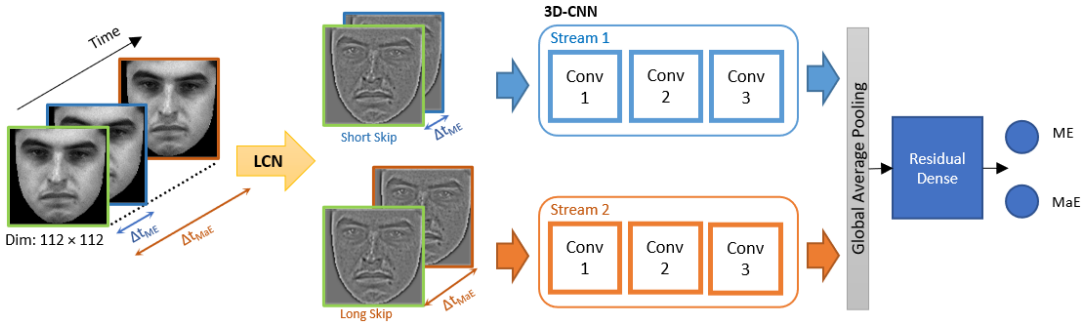

Figure 1: Network architecture of our two-stream 3D-CNN. Our network has only 3 layers (4 layers included LCN). Temporal oriented frame skip based on the duration differences of ME and MaE (where $\Delta t_{ME} < \Delta t_{MaE}$). LCN is applied using a convolutions kernel which performs local contrast normalisation as described in Equation 1. Each convolutional block consists of depthwise separable convolution, batch normalisation and dropout. The residual dense layer possesses the skip connections that shares weights. Two dense nodes were used at the end to resemble the presence of ME and MaE.

ABSTRACT

Facial expression spotting is the preliminary step for micro- and macro-expression analysis. The task of reliably spotting such expressions in video sequences is currently unsolved. Current best systems depend upon optical flow methods to extract regional motion features, before categorisation of that motion into a specific class of facial movement. Optical flow is susceptible to drift error, which introduces a serious problem for motions with long-term dependencies, such as high frame-rate macro-expression. We propose a purely deep learning solution which, rather than tracking frame differential motion, compares via a convolutional model, each frame with two temporally local reference frames. Reference frames are sampled according to calculated micro- and macro-expression duration. As baseline for MEGC2021 using leave-one-subject-out evaluation method, we show that our solution performed better in a high frame-rate (200 fps) SAMM long videos dataset (SAMM-LV) than a low frame-rate (30 fps) (CAS(ME)^2) dataset. We introduce a new unseen dataset for MEGC2022 challenge (MEGC2022-testSet) and achieves F1-Score of 0.1531 as baseline result.

CCS CONCEPTS

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

KEYWORDS

Facial micro-expressions, baseline result, spotting, deep learning

1 INTRODUCTION

Facial expression is the main way people convey visual information of human emotion. It can predict a person's current state of emotion. Facial expressions can be classified into two groups: macro-expression (MaE) and micro-expression (ME). These classifications are based on their relative duration and intensity, where MaE (also known as a regular facial expression) lasts from 0.5 to 4.0s [27] and has higher intensity; ME occurs in less than 0.5s and has lower intensity. ME occurs more frequently in high-stake and stressful circumstances [7, 8]. As it is an involuntary reaction, the emotional state of a person can be revealed through analysing MEs.

Earlier works of ME are based on datasets of short clips containing categorised ME (i.e., SAMM [5, 6], SMIC [16], and CASME II [26]). These were used to facilitate ME expressions recognition [21, 30]. With recent interest in ME and MaE spotting, researchers created long video datasets, SAMM Long Videos (SAMM-LV) [28, 29] and CAS(ME)^2 [20], to better represent spontaneous emotion for ME and MaE spotting. This paper focuses on automated spotting of MaE and ME on SAMM-LV and CAS(ME)^2. We produce the baseline results for two Facial Micro-expressions Grand Challenge (MEGC), i.e. MEGC2021 [15] and MEGC2022. To increase the level of challenge, we introduce a new unseen dataset.

Most of the previous methods utilise long short-term memory (LSTM) [4, 25] or optical flow [10, 23, 25, 32] to detect temporal correlation of video sequences. LSTM is a recurrent neural network that computes sequential time steps with a new element of the input sequence being added to the network at each time step [22]. Optical flow computes the differences of two image frames every time when it is applied within a video sequence. Both LSTM and optical flow are computationally expensive. In addition, optical flow has weaknesses such as drifting over frames [2] and is very susceptible to illumination changes [24]. We also noticed that previous attempts lack duration centred analysis. We take advantage of the major difference between ME and MaE (they occur for different duration, where ME occurs less than 0.5s while MaE occurs in 0.5s or longer) and propose a two-stream network with a different frame skip based on the duration differences for ME and MaE spotting.

The main contributions are:

- Our approach is the first end-to-end deep learning ME and MaE spotting method trained from scratch using long video datasets.
- Our method uses a two-stream network with temporal oriented reference frame. The reference frames are two frame pairs corresponding to the duration difference of ME and MaE. The two-stream network also possesses shared weights to mitigate overfitting.
- The network architecture consists of only 3 convolutional layers with the capability of detecting co-occurrence of ME and MaE using a multi-label system. This method has the potential to be used on lightweight devices (e.g., smartphones) in real-time.
- To make the network less susceptible to uneven illuminations, Local Contrast Normalisation (LCN) is included into our network architecture. LCN drastically improves the overall network performance across a range of configurations and parameters.

Figure 2: Preprocessing: (Top) Face alignment and data augmentation (randomised brightness and contrast change) on a subject of SAMM-LV; and (Bottom) Image normalised using LCN. Despite the brightness and contrast differences, the facial features remain well-preserved.

2 PROPOSED METHOD

Our goal is to detect ME and MaE within long video sequences. By using the duration difference of ME and MaE, we propose a two-stream 3D-Convolutional Neural Network (3D-CNN) with temporal oriented frame skips. We define the two streams as ME and MaE pathways, as illustrated in Fig. 1. They are structurally identical networks with shared weights, but differ in frame skips. We use 3 convolutional layers and pool all the spatial dimensions before the dense layers using global average pooling. This design constrains the network to focus on regional features, rather than global facial features. Next, we further propose that normalising the brightness and/or contrast of the images. This is important for generalisation and real world applications, as there is likely more variation in skin tone and brightness between different individuals, and lighting conditions. Therefore, we apply LCN to all images before presented to our network.

2.1 Preprocessing

Facial Alignment OpenFace 2.0 [1] is used for facial alignment. It is a general-purpose toolbox for facial analysis. OpenFace uses Convolutional Experts Constrained Local Model (CE-CLM) [31] of 84-points for facial landmark tracking and detection. Based on the detected facial landmarks, the face in each frame of a video sequence is aligned and extracted. In our experiment, image resolution is 112×112 pixels, which is the default output resolution of OpenFace.
Local Contrast Normalisation (LCN) LCN [12] was inspired by computational neuroscience models that mimic human visual perception [17] by mainly enhancing low contrast regions of images. LCN normalises the contrast of an image by conducting local subtractive and divisive normalisations [12]. It performs normalisation on local patches (per pixel basis) by comparing a central pixel value with its neighbours. The unique feature of LCN is its divisive normalisation, which consists of the maximum of local variance or the mean of global variance. If an area of image has very low variance (approximately 0), dividing with a small value will form a bright spot. Dividing using the mean of global variance mitigates this issue. The main advantage of this method is robustness towards the change in brightness or contrast (shown in Figure 2). The facial

features are well preserved despite the random changes in brightness and contrast. This can be a solution to address the weakness of overused conventional optical flow method of dealing with uneven lighting. In our implementation, Gaussian convolutions are used to obtain the local mean and standard deviation. Gaussian convolution acts as a low pass filter which reduces noise. It also speeds up the local normalisation process as it is a separable filter (where 2-dimensional data can be calculated using 2 independent 1-dimensional functions).

The general equation of LCN can be described as

$$g(x,y) = \frac{f(x,y) - m_f(x,y)}{max(\sigma_f(x,y), c)} \quad (1)$$

where $f(x,y)$ is the input image, $m_f(x,y)$ is the local mean estimation, $\sigma_f(x,y)$ is the local variance estimation, c is the mean of local variance estimation and $g(x,y)$ is the output image.

2.2 Network Architecture

We propose a two-stream network using a 3D-CNN (network architecture shown in Figure 1). Our network takes advantage of the duration differences of ME and MaE and encouraging one network to be more sensitive to ME and the other to MaE. This is made possible by using a different number of skipped frames in each respective stream (using the maximum duration of a ME, 0.5s, as the threshold for the duration difference). Our network consists of depthwise separable convolutions, which has about 10% less parameters compared to regular convolution counterpart.

Input Layer The input of this network consists of 4 images. The frame pair in the first stream has a shorter frame skip compared to the latter pair. The frame skips are determined based on the k-th frame. The k-th frame, described by Moilanen et al. [18], is the average mid-point of odd-numbered facial expression interval of the whole dataset. These pairs are then fed into two separate but identical neural networks with shared weights.

Weighted loss function To the best of our knowledge, we are the first in ME spotting to weight imbalanced datasets using a loss function. The datasets used in our experiment are imbalanced, and there are more neutral frames relative to frames containing ME or MaE. We also weighted the loss based on ME and MaE, as ME occurs less than MaE. The loss can be described as

$$Loss = -\sum_{i=1}^{C'} M_i \cdot [W \cdot t_i \cdot log(s_i) - (1 - t_i) \cdot log(1 - s_i)] \quad (2)$$

where t_i is ground truth labels, s_i are the predictions, C' is the number of expression types (C'=2 in our case, for ME and MaE), W is the weighting factor that functions to penalise more when the network predicts ME/MaE wrongly as neutral and M_i is the weighting factor for expression (ME or MaE).

We only apply weighted loss function when training SAMM-LV as we found out model trained with SAMM-LV improves with weighted loss function. The effects in CAS(ME)^2 is negligible. We used C' = 2, M_0 = 0.9 (for ME), M_1 = 0.1 (for MaE). Coefficient W used is 3. All the weighting factors are used to address the dataset imbalance. W is used to address different number of ground truth labels of ME/MaE and neutral; M_0 and M_1 is used to address the imbalanced labels of ME and MaE.

Depthwise Separable Convolution We use depthwise separable convolution of MobileNet [11] that reduces total trainable parameters with minimal performance impact. It consists of depthwise and pointwise convolution. Depthwise convolution is convolution applied on individual channels instead of all channel at once (as in regular convolutional). Pointwise convolution is convolution that uses a 1×1 kernel with a third dimension of d (where d is the number of channels) on the feature maps.

GAP and Residual Dense Layer A global average pooling (GAP) layer is used to flatten the convolution output and enforce modelling of localised facial movements. It is followed by the final hidden layer consists of a residual dense layer. This layer is two fully connected layers with skip connections inspired by ResNet [9].

Output Layer The output layer consists of two dense nodes with sigmoid activation representing the presence of ME and MaE.

3 EXPERIMENT

This section provides datasets information, training details and performance metrics of our experiment.

3.1 Datasets

We evaluate our method on two datasets, i.e., MEGC2021 Spotting Datasets and introduce MEGC2022 unseen test set.

MEGC2021 Spotting Datasets. The datasets used are SAMM Long Videos (SAMM-LV) [29] with 147 long videos containing 343 MaEs and 159 MEs; and CAS(ME)^2 [20] with 87 long videos containing 300 MaEs and 57 MEs. The original ground truth of these datasets consist of onset, apex, and offset frame labels of each facial expression. We label the ground truth of movement from the onset frame to the offset frame, inclusively. Our ground truth consists of two labels of binaries where 0 represents absence while 1 represents presence of ME or/and MaE.

MEGC2022 Unseen Test Set. In MEGC2022, we introduce an unseen test set with 10 long videos, which consists of 5 long videos from SAMM [5] (SAMM Challenge dataset) and 5 clips cropped from different videos in CAS(ME)^3 [14]. The frame rate for SAMM Challenge dataset is 200 fps and the frame rate for CAS(ME)^3 is 30 fps. The participants can use SAMM-LV and CAS(ME)^2 as training set, and test on this unseen dataset. For facilitate the spotting challenge and to enable fair assessment, we do not release the ground truth for this dataset. The participants will submit their results to our grand challenge system (https://megc2022.grand-challenge.org).

Table 1: Training configuration. Stream 1 is designed to be more sensitive to ME, while Stream 2 is more sensitive to MaE by using different range of frame skips based on the duration differences of ME and MaE. The k-th frame is the average mid-point of facial expression interval. (Note: * used in training and validation, † used in testing)

Dataset	SAMM-LV	CAS(ME)^2
Random frame skip* (Stream 1 & 2)	25~75 & 200~400	3~9 & 16~50
k-th frame skip† (Stream 1 & 2)	37 & 217	6 & 19

3.2 Training

Randomised frame skips are used in training and validation. This creates a more realistic scenario as the duration of each facial expression is unknown in real life. For model testing, we used a frame skip based on the k-th frame of ME and MaE of each respective dataset shown in Table 1. The visual differences of frames calculated using this interval (frames skipped) is larger, making the facial movements more distinct for the algorithm to spot.

Regularisation Random augmentations (i.e., contrast, gamma intensity, and gamma gain) on the input images are performed with a range of 0.5 to 1.5. Other augmentations include 50% probability of horizontal flip and $\pm 10^{\circ}$ of image rotation. Other regularisations include adding dropout layers and random frame skips during training and validation.

Training Configuration As shown in Table 1, the results are evaluated using leave-one-subject-out (LOSO) cross-validation.

3.3 Performance Metrics

We apply the Intersection over Union (IoU) method used in Micro-Expression Grand Challenge (MEGC) III [10, 13] to compare with other methods. The interval is then evaluated using the following IoU method

$$\frac{Predicted \cap GT}{Predicted \cup GT} \geq J \tag{3}$$

where J is the minimum overlapping to be classified as true positive, GT represents the ground truth expression interval (onset-offset), $Predicted$ represents the detected expression interval. In our experiment, J is set to 0.5.

For MEGC2022, we create an automated evaluation system, which is available at grand challenge system[1]. Our evaluation code is used to standardised the evaluation method and tested on MEGC2022-testset (unseen dataset). To facilitate future research in ME spotting, we provide a live leaderboard, where the authors can continue to use our MEGC2022-testset (with agreement in placed) and evaluate their results online.

4 RESULTS

4.1 Baseline Result for MEGC2021 Spotting Task

We convert our results into intervals using automated thresholding based on ROC evaluation. First, the test results are normalised and smoothed using a Butterworth filter [3], which is a low-pass filter that cuts off high frequency noises while retaining low frequency signals. The main advantage of this filter is it has a flat magnitude filter whereby signals with frequency below cut-off frequency do not undergo attenuation. Next, the onset and offset of both ground truth and the predictions are obtained. Finally, the overlapping was analysed using the IoU method (where TP must fulfill the criteria in Equation 3). At the time of producing this baseline result, Pan et al. [19] is the only deep learning method evaluated on long video datasets for ME and MaE spotting and without using any post-processing algorithm. Therefore, we only compare to their result, as shown in Table 2.

[1]https://megc2022.grand-challenge.org/

Table 2: F1-score of ME and MaE spotting using our Automated IoU Method on SAMM-LV and CAS(ME)^2-cropped dataset.

Method	SAMM-LV			CAS(ME)^2-cropped		
	MaE	ME	Overall	MaE	ME	Overall
Pan [19]	-	-	0.0813	-	-	0.0595
Ours	0.1863	0.0409	0.1193	0.0401	0.0118	0.0304

It is noted that the result in Table 2 is CAS(ME)^2-cropped. When we aligned the face with OpenFace 2.0 [1], we achieved F1-score of 0.0686, 0.1190, 0.0497 on MaE, ME, and overall, respectively. Our results show better spotting performance in SAMM-LV compared to CAS(ME)^2. One possibility is SAMM-LV has higher frame rate (200 fps) and the randomised frame skipping used in our training pipeline has more variety of input data to be learnt compared to CAS(ME)^2 (30 fps). Hence, our model is able to learn data with more variation in SAMM-LV and show better performance. ME which occur in less than 0.5s, has a small window of detection. A lower ME detection rate in CAS(ME)^2 might also be a consequence of the lower frame rate.

4.2 Baseline Result for MEGC2022 Spotting Task

Table 3 shows the baseline result to facilitate MEGC2022 Spotting Task. We achieved an overall F1-score of 0.1351, with 0.1176 and 0.1739 on SAMM Challenge Dataset and CAS(ME)^3, respectively. For evaluation on SAMM Challenge, we train our network using SAMM-LV; for CAS(ME)^3, the network was trained on CAS(ME)^2. It is noted that on unseen dataset, our method performed better in detecting ME of CAS(ME)^3.

Table 3: F1-score of ME and MaE spotting on unseen test set of MEGC2022 that uses SAMM Challenge and CAS(ME)^3

Method	SAMM Challenge			CAS(ME)^3			Overall
	MaE	ME	Overall	MaE	ME	Overall	
Ours	0.1739	0.0714	0.1176	0.1622	0.2222	0.1739	0.1351

5 CONCLUSION

We presented a temporal oriented two-stream 3D-CNN model that shows promising results in ME and MaE spotting in long video sequences. Our method took advantage of the duration difference of ME and MaE by making a two-stream network that is sensitive to each expression type. Despite only having 3 convolutional layers, our model showed state-of-the-art performance in SAMM-LV and remained competitive in CAS(ME)^2. LCN has proven to have significant improvement in our model and the ability to address uneven illumination, which is a major weakness of optical flow.

ACKNOWLEDGMENTS

This work is supported by grants from The Manchester Metropolitan University VC PhD Studentship and the National Natural Science Foundation of China (U19B2032, 62106256).

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[2] Mario Bertero, Tomaso A Poggio, and Vincent Torre. 1988. Ill-posed problems in early vision. *Proc. IEEE* 76, 8 (1988), 869–889.

[3] Stephen Butterworth et al. 1930. On the theory of filter amplifiers. *Wireless Engineer* 7, 6 (1930), 536–541.

[4] Dawood Al Chanti and Alice Caplier. 2019. ADS-ME: Anomaly Detection System for Micro-expression Spotting. *arXiv preprint arXiv:1903.04354* (2019).

[5] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* 9, 1 (Jan 2018), 116–129. https://doi.org/10.1109/TAFFC.2016.2573832

[6] Adrian K Davison, Walied Merghani, and Moi Hoon Yap. 2018. Objective classes for micro-facial expression recognition. *Journal of Imaging* 4, 10 (2018), 119.

[7] Paul Ekman. 2003. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences* 1000, 1 (2003), 205–221.

[8] Paul Ekman and Gavin Yamey. 2004. Emotions revealed: recognising facial expressions: in the first of two articles on how recognising faces and feelings can help you communicate, Paul Ekman discusses how recognising emotions can benefit you in your professional life. *Student BMJ* 12 (2004), 140–142.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[12] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. 2009. What is the best multi-stage architecture for object recognition?. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2146–2153.

[13] 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.

[14] 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* (2022).

[15] 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*. Association for Computing Machinery, New York, NY, USA, 5700–5701. https://doi.org/10.1145/3474085.3478579

[16] Xiaobai Li, Thorsten Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen. 2013. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–6.

[17] Siwei Lyu and Eero P Simoncelli. 2008. Nonlinear image representation using divisive normalization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[18] Antti Moilanen, Guoying Zhao, and Matti Pietikainen. 2014. Spotting Rapid Facial Movements from Videos Using Appearance-Based Feature Difference Analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. 1722–1727. https://doi.org/10.1109/ICPR.2014.303

[19] Hang Pan, Lun Xie, and Zhiliang Wang. 2020. Local Bilinear Convolutional Neural Network for Spotting Macro- and Micro-expression Intervals in Long Video Sequences. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. 343–347.

[20] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. 2017. CAS (ME)ˆ 2: A Database for Spontaneous Macro-expression and Micro-expression Spotting and Recognition. *IEEE Transactions on Affective Computing* (2017).

[21] 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.

[22] Sanchari Sen and Anand Raghunathan. 2018. Approximate computing for long short term memory (LSTM) neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 37, 11 (2018), 2266–2276.

[23] Bo Sun, Siming Cao, Jun He, and Lejun Yu. 2019. Two-stream Attention-aware Network for Spontaneous Micro-expression Movement Spotting. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 702–705.

[24] Pavan Turaga, Rama Chellappa, and Ashok Veeraraghavan. 2010. Advances in video-based human activity analysis: challenges and approaches. In *Advances in Computers*. Vol. 80. Elsevier, 237–290.

[25] Michiel Verburg and Vlado Menkovski. 2019. Micro-expression detection in long videos using optical flow and recurrent neural networks. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–6.

[26] 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).

[27] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior* 37, 4 (2013), 217–230.

[28] Chuin Hong Yap, Ryan Cunningham, Adrian K Davison, and Moi Hoon Yap. 2021. Synthesising Facial Macro-and Micro-Expressions Using Reference Guided Style Transfer. *Journal of Imaging* 7, 8 (2021), 142.

[29] 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) (FG)*. IEEE Computer Society, Los Alamitos, CA, USA, 194–199. https://doi.org/10.1109/FG47880.2020.00029

[30] 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.

[31] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2519–2528.

[32] L-w Zhang, Jingting Li, S Wang, X Duan, W Yan, H Xie, and S Huang. 2020. Spatio-Temporal Fusion for Macro-and Micro-Expression Spotting in Long Video Sequences. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. 245–252.