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

DOI:
https://arxiv.org/pdf/2105.06340v3

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

Chuin Hong Yap[1], Moi Hoon Yap[1], Adrian K. Davison[2], Ryan Cunningham[1]
[1] Department of Computing and Mathematics, Manchester Metropolitan University, UK
[2] Faculty of Biology, Medicine and Health, The University of Manchester, UK

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. The 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 durations. We show that our solution achieves state-of-the-art performance (F1-score of 0.105) in a dataset of high frame-rate (200 fps) long video sequences (SAMM-LV) and is competitive in a low frame-rate (30 fps) dataset (CAS(ME)^2). In this paper, we document our deep learning model and parameters, including how we use local contrast normalisation, which we show is critical for optimal results. We surpass a limitation in existing methods, and advance the state of deep learning in the domain of facial expression spotting.

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 [26] 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 [6, 7]. As it is an involuntary reaction, the emotional state of a person can be revealed through analysing MEs.

For ME spotting, due to limited dataset availability, early works are based on datasets consisting of short clips containing categorised ME (i.e., SAMM [4], SMIC [14], and CASME II [25]). Spotting with clips containing ME will result in high detection rate regardless. Hence, the recently created long video datasets, SAMM Long Videos (SAMM-LV) [27] and CAS(ME)^2 [18], were created 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.

Most of the previous methods utilise LSTM or optical flow 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 [20]. 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 [23]. 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 correspond to the duration difference of ME and MaE. The two-stream network also possesses shared weights to mitigate overfitting.
- The network architecture is lightweight 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. We also shows that LCN is essential in our network. Our 3-layer network with LCN outperforms deeper network (i.e. 20-layer network).

2. Related Work

Preprocessing The preprocessing steps usually begin with facial alignment. Facial alignment is usually conducted by using facial landmark detection followed by cropping and resizing the detected facial region. Next, it is common to extract the facial region by using masks or facial segmentation. Face masking removes unwanted regions [5]. Facial segmentation consists of converting images into uniform blocks [10] or selecting regions of interest (ROI) such as areas with higher movements such as the eyebrows, nose and mouth [29, 17]. The comparison of preprocessing methods of various approach is shown in Table 1.

Conventional approach Davison et al. [5] performs ME spotting based on 3D Histogram of Oriented Gradients (3D-HOG) by taking Chi-square distance of local regions. This method uses piecewise affine warping to mask off unrelated regions. He et al. [10] uses the magnitude of maximal difference of optical flow features. This method also uses facial blocks which divides the face into a 6×6 region. Zhang et al. [29] uses a spato-temporal feature fusion method using histogram of oriented optical flow of selected ROIs. Instead of facial blocks, this method uses ROI-based method (6 ROIs for eyebrows, 2 ROIs for nose and 4 ROIs for mouth). All of these methods contains feature extraction process. This requires a sequence of steps, such as region of interest (ROI) selection and facial alignment, which are computationally expensive.

Deep learning approach For deep learning based approaches, Tran et al. [22] use conventional features and feed them into a long short-term memory (LSTM) network to detect movement. Verburg et al. [24] implement Histogram of Oriented Optical Flow (HOOF) as input features and use LSTM to learn temporal information. This method uses post-processing, which suppresses the overlapping neighbours of spotted interval. Sun et al. [21] use a spatio-temporal cascaded network that consists of CNN and attention-aware LSTM. However, these methods were trained and evaluated on short ME video clips. Evaluating on a few seconds of short ME clips will have a higher detection rate as each video contains at least one ME. It does not resemble a real-world situation, where ME occurrence is rare and does not happen every few seconds.

Table 1. A comparison of preprocessing steps and input types used in existing ME spotting methods and our proposed method.

Method	Preprocessing	Input
Davison et al. [5]	Face alignment, face masking	3D-HOG features
Tran et al. [22]	facial alignment	LBP-TOP, HOG-TOP HIGO-TOP
Verburg et al. [24]	facial alignment, ROI	Optical flow
Sun et al. [21]	facial alignment, feature matrix processing	Images, optical flow
He et al. [10]	Face alignment, uniform facial blocks	Optical flow of facial blocks
Zhang et al. [29]	Face alignment, ROI	Optical flow, ROI specific pattern
Pan et al. [17]	Face alignment, ROI	Images, ROI
Our approach	Face alignment, LCN	Images

Pan et al. [17] is the only deep learning method evaluated on long video datasets. However, it did not train from scratch using a ME dataset and is not an end-to-end solution. Instead, it uses a pre-trained deep learning model as a feature extractor, optical flow for face detection and ROI extraction. This method classifies each image sequence into either ME, MaE or natural frames. This classification method assumes that ME and MaE are mutually exclusive. On the contrary, ME and MaE occur simultaneously in both datasets and contradicts this assumption. Hence, we address this issue by designing a new network with multi-label design.

3. 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 few 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 will be critical for generalisation, as there is likely more variation in skin tone and brightness. Therefore, we apply LCN to all images before presented to our network.

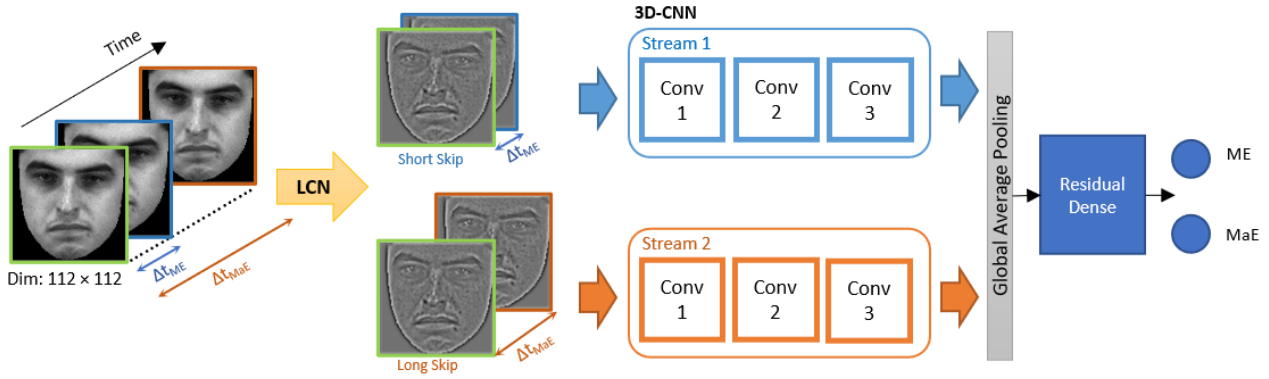

Figure 1. Network architecture of our two-stream 3D-CNN. It is lightweight as it has only 3 layers (4 layers if you include LCN). Temporal oriented frame skip based on the duration differences of ME and MaE (where $\Delta t_{ME} < \Delta t_{MaE}$). LCN is applied using a convolutions kernel which performs 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.

3.1. 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) [28] 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 [15] 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).

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.

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.

3.2. 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. [16], 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 weighing factors are used to address the imbalance dataset. *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 imbalance labels of ME and MaE. The model performance of different weighing factors is shown in Table 8.

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 consisting of a residual dense layer. This layer consists of two fully connected layers with skip connections inspired by ResNet [9].

Output Layer The output layer consists of two dense nodes representing the presence of ME and MaE. A sigmoid activation function is used as the output, and is in the range of 0 and 1.

Table 2. Duration analysis of SAMM-LV and CAS(ME)^2.

Dataset	**SAMM-LV**		**CAS(ME)^2**	
Type of Expression	ME	MaE	ME	MaE
Minimum (s)	0.15	0.51	0.27	0.10
Mean (s)	0.37	2.17	0.42	1.25
Maximum (s)	0.51	25.88	0.53	3.90

4. Experiment

Our experiment involves an end-to-end 3D-convolutional network using leave-one-subject-out (LOSO) cross validation. This section provides dataset information and introduces a novel move-to-neutral ratio which estimates the movements of a subject in a dataset. Training details of our experiment is also included.

4.1. Datasets

The datasets used are SAMM Long Videos (SAMM-LV) [27] with 147 long videos containing 343 MaEs and 159 MEs; and CAS(ME)^2 [18] with 87 long videos containing 300 MaEs and 57 MEs. The duration analysis of MEs and MaEs in the long videos are shown in Table 2. It is noted that the ME duration of SAMM-LV is shorter than CAS(ME)^2, but the MaE duration is longer. 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.

4.2. Move-to-neutral ratio

We introduce a new metric named move-to-neutral ratio to analyse the subject of dataset used. In LOSO cross validation, by knowing the amount of movements (ME or MaE) of each subject, we can estimate the amount ME or MaE in each subject. As each subject of the dataset used has different numbers of frames with movement (ME or MaE) and duration of recorded videos, the proportion of movements to video duration of each subject is different. This can result in easier predictions on some subjects as they have more movements and vice versa. The move-to-neutral ratio of each subject is shown in Figure 3. The average move-to-neutral ratio for both datasets are approximately 0.40 (SAMM-LV) and 0.05 (CAS(ME)^2), which shows that both are imbalanced as most of the videos consist of neutral frames. However, this metric is solely based on the movements labelled in the ground truth, which consists of MaE and ME. Other movements such as head movements might

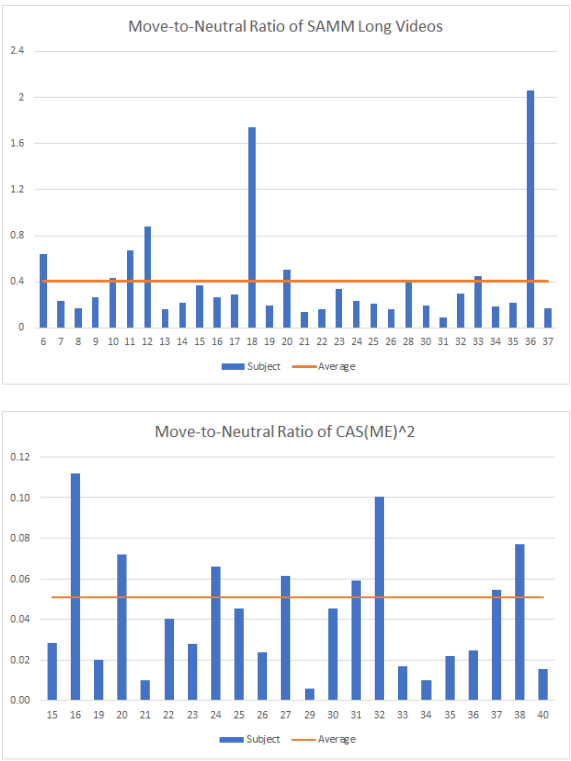

Figure 3. Move-to-Neutral Ratio for each subject of SAMM-LV and CAS(ME)^2. It shows that every subject has different relative number of movement (ME or MaE) to neutral frames. From the average move-to-neutral ratio, SAMM-LV is higher compared to CAS(ME)^2. *x-axis labels are the subject indices of the datasets.

not be included.

4.3. 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. It can also act as a regularisation process by adding variations and perturbations to the input. 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 3. 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 3, the results are evaluated using leave-one-subject-out (LOSO) cross-validation. Early stopping is used during training, ending when the loss does not improve for 5 consecutive epochs.

Table 3. 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**
Batch Size	16	
Learning Rate	0.007	0.005
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
Manual frame skip† (Stream 1 & 2)	30 & 310	10 & 33

Table 4. Results (Raw Output) of macro- and micro-expression spotting of our method

	MaE		**ME**	
	F1-score	AUC	F1-score	AUC
SAMM-LV	0.3872	0.6780	0.0720	0.5687
CAS(ME)^2	0.1369	0.6925	0.0174	0.5762

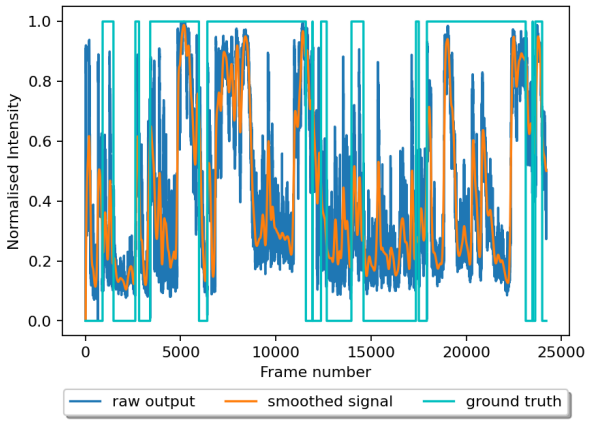

Figure 4. Real long video testing data of a subject smoothed using Butterworth filter with ground truth comparison

5. Results

Our network predicts the presence of facial expression on a per frame basis. F1-score and area under the curve (AUC) of receiver operating characteristic (ROC) of our raw output are reported in Table 4. We compare each frame using normalised results filtered using threshold based on ROC curve. From the F1-scores, our model performs better on SAMM-LV. However, CAS(ME)^2 performs better for the AUC. Both F1-score and AUC indicate that our model performs better in MaE, which is expected as MaE occurs more frequently and has longer duration.

We apply the Intersection over Union (IoU) method used

Table 5. F1-score of ME and MaE spotting using Automated IoU Method, where Ours* represents our proposed method with k-th frame skip and Ours** represents our proposed method with manual frame skip. Manual frame skip is performed by first taking k-th frame as a reference, proceeded by increasing or decreasing the frame skips until the results improve.

Method	SAMM-LV			CAS(ME)^2		
	MaE	ME	Overall	MaE	ME	Overall
Pan [17]	-	-	0.0813	-	-	0.0595
Ours*	0.1504	0.0421	0.1017	0.0704	0.0075	0.0509
Ours**	**0.1543**	**0.0442**	**0.1050**	**0.0874**	0.0075	**0.0630**

in Micro-Expression Grand Challenge (MEGC) III [13, 10] 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.

As other methods uses different post-processing steps, we decided to use two different evaluation methods. The first method is Automated IoU Method and the second method is Multi-Scale Filter used by Zhang et al. [29].

5.1. Automated IoU Method

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, results shown in Fig. 4. 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).

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.

5.2. Comparison with the state of the art

Zhang et al. [29] and He et al. [10] are conventional approaches. These methods use post-processing steps to enhance ME spotting rate. Hence, it is not fair to compare our method directly. Instead, we use Zhang et al.'s post-processing steps (also named Multi-Scale Filter [29]) and the results are shown in Table 6. This method uses Savitzky-Golay filter for noise removal and signal merging as described in Zhang et al.'s paper. We obtained a notable improvement in ME and MaE spotting, particularly in CAS(ME)^2. By implementing these post-processing steps, our method outperforms in SAMM-LV and CAS(ME)^2 in MaE spotting and overall performance. Although we obtained better results using this evaluation, we noticed that this method requires selection of hyperparameters (e.g., window size and order of Savitzky-Golay filter, the upper limit of interval distance to merge etc).

Table 6. F1-score of ME and MaE spotting using Multi-Scale Filter (manual post-processing steps used by Zhang et al. [29]), where Ours* represents our proposed method with k-th frame skip and Ours** represents our proposed method with manual frame skip. This post-processing steps involves signal smoothing using Savitzky-Golay filter and signal merging when intervals are close to each other.

Method	SAMM-LV			CAS(ME)^2		
	MaE	ME	Overall	MaE	ME	Overall
He [10]	0.0629	0.0364	0.0445	0.1196	0.0082	0.0376
Zhang [29]	0.0725	**0.1331**	0.0999	0.2131	0.0547	0.1403
Ours*	0.1569	0.0512	**0.1083**	0.1880	0.0583	0.1449
Ours**	**0.1595**	0.0466	0.1084	**0.2145**	**0.0714**	**0.1675**

For the purpose of comparison with benchmark algorithms, we implemented these method. However, we will not recommend these post-processing steps as each hyperparameter can be customised to improve the results, which might result in overfitting. As stated in Zhang et al.'s paper: "the results are terrible" before the post-processing steps. In contrast, our proposed method is already competitive before these post-processing steps, as shown in Table 5. Overall, our method performed the best on SAMM-LV without post-processing steps, with an F1-score of 0.1050. With post-processing steps, our method achieved the best F1-Score of 0.1675 on CAS(ME)^2. It is noted that our method achieved the best result in MaE spotting on both datasets.

5.3. Visualisation using Grad-CAM

We visualise the activation of our network using Gradient-weighted Class Activation Mapping (Grad-CAM) [19]. This provides interpretable visualisation on the face region that the network is focusing on when spotting ME/MaE. We select the deepest interpretable layer, which is the last dense layer, and visualise its activation on SAMM-LV participants.

In Figure 5, we observe that the heatmaps closely resemble the Facial Action Units (AUs) of facial expression. The reliable AUs of happiness are associated with AU6 (Cheek Raiser) and AU12 (Lip Corner Puller). Figure 5 (a) illustrates the heatmap of happiness, where it shows high activation around eye corner (AU6) and the mouth region is

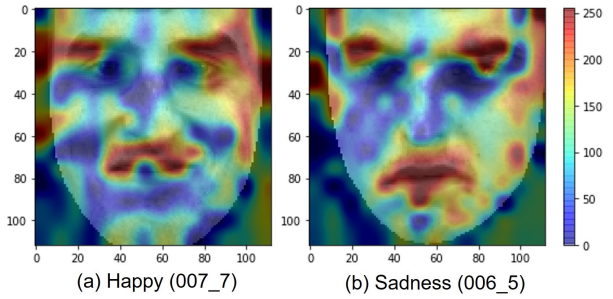

Figure 5. Grad-CAM visualisation on SAMM-LV participants. In (a), AU6 (Cheek Raiser) and AU12 (Lip Corner Puller) is detected; In (b), AU4 (Brow Lowerer) and AU15 (Lip Corner Depressor) is detected. AU12 and AU15 are distinctly distinguished: in (a), the mouth region heatmap is directed towards the upper part of the face; in (b), the heatmap at the mouth region forms a huge inverted curve extending towards the bottom of the face.

directed to the upper part of the face (AU12). On the other hand, the reliable AUs of sadness are AU4 (Brow Lowerer) and AU15 (Lip Corner Depressor). In Figure 5 (b), the heatmap shows activation on both the brows and the eyes region that indicates AU4, and the mouth region forms a huge inverted curve extending until the bottom of the face which resembles AU15.

5.4. Ablation Studies

Table 7 shows the ablation studies conducted using Automated IoU Method. With manual frame skip, our proposed method achieves the best result. Our proposed approach (with k-th frame skip) is not far behind, which shows that it is a viable method. We conduct our experiment without LCN using architecture with different depth, i.e. 3, 6, 10 and 20 layers. Even with a deeper network, without LCN, it performs worse than our proposed model, which indicates that LCN is a crucial component of our approach. We also replace GAP layer with Global Max Pooling. Although both average and max pooling, average pooling identifies all discriminative region more completely [30]. As ME is a subtle facial movement, detailed oriented average pooling is more suitable. This is further concluded in the performance of Global Max Pooling compared with our proposed network that uses GAP. Our proposed model performs worse without batch normalisation. We showed that each component of our network are essential and has a positive contribution to the overall performance.

We conducted ablation studies on weighted loss function (as described in Section 3.2). The weighted loss function is used to address the imbalance training dataset. The results are shown in Table 8. We only apply weighted loss in model trained on SAMM-LV as it shows no significant improvement in model trained on CAS(ME)^2. We weight ME more with respect to MaE by setting "M_ME" to 0.9 and "M_MaE" tp 0.1; and "W" is used to impose a harsher penalty when the network predicts ME/MaE as neutral wrongly. We can see that without weighted loss, the network performs worse. We also demonstrate fine-tuning of "W" and the setting of "W" as 3 achieves the best performance.

Table 7. Ablation studies on Automated IoU Method: F1-scores reported. Manual frame skip fine-tuning can produce slightly better results. With LCN removed, our network performance dropped. Even with deeper network (i.e. 20 convolution layers), it still under perform when compared to our proposed 3-layers deep network. The model performance dropped without batch normalisation and when GAP is replaced with Global Max Pooling.

	SAMM-LV			CAS(ME)^2		
	MaE	ME	Overall	MaE	ME	Overall
Without LCN						
- 3 Conv Layers	0.0297	0.0066	0.0198	0.0000	0.0000	0.0000
- 6 Conv Layers	0.1079	0.0275	0.0750	0.0041	0.0000	0.0030
- 10 Conv Layers	0.0825	0.0500	0.0518	0.0098	0.0000	0.0073
- 20 Conv Layers	0.0943	0.0160	0.0646	0.0000	0.0000	0.0000
Replace GAP with GlobalMaxPool	0.1311	0.0149	0.0795	0.0173	**0.0098**	0.0153
Without BatchNorm	0.1456	0.0252	0.0934	0.0510	0.0059	0.0359
Proposed	0.1504	0.0421	0.1017	0.0704	0.0075	0.0509
Manual Frame Skip	**0.1543**	**0.0442**	**0.1050**	**0.0874**	0.0075	**0.0630**

Table 8. Ablation studies on different weighing coefficients trained on SAMM-LV. For SAMM-LV, weighted loss function improves the detection rate. We did not report on CAS(ME)^2 dataset as we found that weighted loss function shows minimal effect in model performance.

	W	M_ME	M_MaE	MaE	ME	Overall
Proposed	3	0.9	0.1	**0.1504**	**0.0421**	**0.1017**
W/o Weighted Loss	1	1.0	1.0	0.1480	0.0238	0.0910
W/o Weighted M	3	1.0	1.0	0.1404	0.0099	0.0594
W/o Weighted W	1	0.9	0.1	0.1413	0.0268	0.0900
Vary coefficient W	6	0.9	0.1	0.1443	0.0213	0.0888
Vary coefficient W	10	0.9	0.1	0.1302	0.0240	0.0825
Vary coefficient W	0.5	0.9	0.1	0.1339	0.0262	0.0696

6. Discussion

Model comparison Our model is the state-of-the-art in SAMM-LV and competitive in CAS(ME)^2. Conventionally, optical flow methods have good performance but require extensive pre-processing and post-processing steps, which are computationally expensive. He et al. [10] and Zhang et al. [29] use image segmentation or ROI selection, followed by optical flow extraction and spatio-temporal fusion of each ROI. On the contrary, our method is an end-to-end solution with 3 layers of CNN.

Zhang et al. [29] is the only method that did poorly in spotting MaE compared to other categories in SAMM-LV. Commonly, MaE (regular facial expressions) is easier to detect when compared to ME. As SAMM-LV is a dataset with high frame-rate of 200 fps, Zhang et al.'s optical flow on consecutive frames approach is unable to capture the long range dependency of MaE, which explains their relative poor results on MaE of SAMM-LV. Zhang et

al. [29] is also the only method that shows better performance in CAS(ME)^2 than SAMM-LV. This may imply that Zhang et al. is heavily biased towards CAS(ME)^2. Another possible reason is the post-processing method may be more suitable in CAS(ME)^2. Our method using Zhang et al.'s post-processing has also shown notable improvement in CAS(ME)^2. The merging process in Zhang et al's post-processing is questionable and can be a potential source of overfitting the results. For example, 3 false positives can be merged into 1 true positive, which greatly improves the results (as shown in [29]). Moreover, this method cannot improve with additional data, whereas ours is expected to improve [8]. We provide an important contribution, justifying collection of further data.

To date, Pan et al. [17] is the only deep learning approach for spotting MaE and ME in long videos, evaluated using IoU method of MEGC III. Comparing with this method, our model with manual frame skipping has better performance in both datasets. We also produce a complete report on all three spotting categories. Our method is able to spot ME, MaE and co-occurrence of both types of facial expression, which are the features absent in [17].

***k*-th frame skip** We investigate the effectiveness of k-th frame used by manually vary the frame skips. By varying the frame skips by taking k-th frame as initial reference, the results show only slight improvement. This indicates that k-th frame method remains a good measurement for frame skip.

Automated vs manual method Both Automated IoU Method (automated method) and Multi-Scale Filter (manual method) show similar performance on model trained on SAMM-LV. This shows that Automated IoU evaluation works well on SAMM-LV with only a minimal performance increment via manual method. However, it is not the case for model trained on CAS(ME)^2. The disparity of the performance on CAS(ME)^2 in both method might be a result of different noise removal method used (Butterworth filter and Savitzky-Golay filter). The automated Butterworth filter is not adaptive enough in handling different noises, whereas using Savitzky-Golay, we can decide a suitable window size and order of filter for each respective noise. Despite higher performance detected in Savitzky-Golay, in real-world applications, automation is preferred as it is not realistic to fine-tune hyperparameters when we make prediction. With further refinement, our proposed Automated IoU Method has potential for real-world applications.

7. 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. We demonstrated our 3-layer network with LCN outperforms deep network with 20 convolutional layers. Further improvements include embedding facial landmark detection into the algorithm and simplifying the spotting algorithm to allocate more computational resources for real-time ME analysis.

References

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

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

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

[4] 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*, 9(1):116–129, Jan 2018.

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

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

[7] Paul Ekman and Gavin Yamey. 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*, 12:140–142, 2004.

[8] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

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

[10] Ying He, Su-Jing Wang, Jingting Li, and Moi Hoon Yap. Spotting macro-and micro-expression intervals in long video sequences. *arXiv preprint arXiv:1912.11985*, 2019.

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 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. What is the best multi-stage architecture

for object recognition? In *2009 IEEE 12th international conference on computer vision*, pages 2146–2153. IEEE, 2009.

[13] J Li, S Wang, Moi Hoon Yap, John See, Xiaopeng Hong, and Xiaobai Li. Megc2020-the third facial micro-expression grand challenge. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 234–237.

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

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

[16] Antti Moilanen, Guoying Zhao, and Matti Pietikainen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1722–1727, Aug 2014.

[17] Hang Pan, Lun Xie, and Zhiliang Wang. 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)*, pages 343–347, 2020.

[18] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. Cas (me)ˆ 2: A database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 2017.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

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

[21] Bo Sun, Siming Cao, Jun He, and Lejun Yu. Two-stream attention-aware network for spontaneous micro-expression movement spotting. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pages 702–705. IEEE, 2019.

[22] Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, and Guoying Zhao. Dense prediction for micro-expression spotting based on deep sequence model. *Electronic Imaging*, 2019(8):401–1, 2019.

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

[24] Michiel Verburg and Vlado Menkovski. 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)*, pages 1–6. IEEE, 2019.

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

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

[27] Chuin Hong Yap, Connah Kendrick, and Moi Hoon Yap. 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)*, pages 194–199, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

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

[29] L-w Zhang, Jingting Li, S Wang, X Duan, W Yan, H Xie, and S Huang. 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)*, pages 245–252, 2020.

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.