Signal Processing: Image Communication 110 (2023) 116875

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Spot-then-Recognize: A Micro-Expression Analysis Network for Seamless Evaluation of Long Videos

Gen-Bing Liong [a], John See [b,*], Chee-Seng Chan [a,*]

[a] CISiP, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[b] School of Mathematical and Computing Sciences, Heriot-Watt University Malaysia, Putrajaya, Malaysia

ARTICLE INFO

Keywords:
Facial Micro-Expression
Spotting
Recognition
Optical flow features
Shallow neural networks

ABSTRACT

Facial Micro-Expressions (MEs) reveal a person's hidden emotions in high stake situations within a fraction of a second and at a low intensity. The broad range of potential real-world applications that can be applied has drawn considerable attention from researchers in recent years. However, both spotting and recognition tasks are often treated separately. In this paper, we present Micro-Expression Analysis Network (MEAN), a shallow multi-stream multi-output network architecture comprising of task-specific (spotting and recognition) networks that is designed to effectively learn a meaningful representation from both ME class labels and location-wise pseudo-labels. Notably, this is the first known work that addresses ME analysis on long videos using a deep learning approach, whereby ME spotting and recognition are performed sequentially in a two-step procedure: first spotting the ME intervals using the spotting network, and proceeding to predict their emotion classes using the recognition network. We report extensive benchmark results on the ME analysis task on both short video datasets (CASME II, SMIC-E-HS, SMIC-E-VIS, and SMIC-E-NIR), and long video datasets (CAS(ME)^2 and SAMMLV); the latter in particular demonstrates the capability of the proposed approach under unconstrained settings. Besides the standard measures, we promote the usage of fairer metrics in evaluating the performance of a complete ME analysis system. We also provide visual explanations of where the network is "looking" and showcasing the effectiveness of inductive transfer applied during network training. An analysis is performed on the in-the-wild dataset (MEVIEW) to open up further research into real-world scenarios.

1. Introduction

Facial expression conveys a person's emotional state in various contexts, but are we able to perceive and interpret when it occurs? Generally, there are two types of facial expression, *i.e.* macro-expression (MaE) and micro-expression (ME). Previous researches have focused on the MaE and achieved an impressive result due to its characteristics of lasting longer duration and noticeable facial muscle movement. In contrast, ME occurs when people intend to suppress their emotions, which has drawn lesser attention compared to MaE [1]. The subtleness and rapid changes of ME typically last between 1/25 to 1/5 s, consequently is more difficult to be identified in real-time without good expertise [2]. Previous studies show that a person without proper training can only spot-then-recognize the ME slightly better than chance [3]. In a high-stake situation, an involuntary ME is elicited to reveal a person's genuine emotion. In 1969, [4] reported the occurrence of ME during psychotherapy when they found that the patient is hiding the true feeling of committing suicide to deceive the psychiatrist. Furthermore, analyzing ME is able to detect lie with a vast range of real-world applications, including police interrogation, government military, psychotherapy, and business negotiation. Fundamentally, the process of annotating temporal location and emotion class of ME is still relied on trained psychologist experts until now, which is time-consuming and costly. Conflicts that occasionally occur when the experts have different opinions would make the labeling process difficult and less reliable [5]. Therefore, it is essential to develop a well-established ME analysis system to ease the tedious annotating work.

In this paper, ME *analysis* is defined as a task consisting of two main steps that are seamlessly linked: *spotting* and *recognition*; the former localizes when the ME occurs, while the latter determines the emotion class of which the ME exhibits [6,7]. Spotting ME remains a challenging task because of the short-lived occurrences as aforementioned. Notably, three phases are involved in the expression, namely onset, apex, and offset. Theoretically, onset is the phase where the facial muscles start to contract and the visual appearance becomes apparent; apex phase refers to the instant where the facial expression reaches the peak intensity; offset is the phase where facial muscles are relaxed. Furthermore, the ME that is intertwined with MaE in long videos has brought additional difficulties for spotting tasks. On the other hand, there are six basic emotion classes, *i.e.* happy, surprise, anger, sadness, fear, and disgust.

* Corresponding authors.
E-mail addresses: genbing67@gmail.com (G.-B. Liong), j.see@hw.ac.uk (J. See), cs.chan@um.edu.my (C.-S. Chan).

https://doi.org/10.1016/j.image.2022.116875
Received 16 April 2022; Received in revised form 6 August 2022; Accepted 4 October 2022
Available online 18 October 2022
0923-5965/

As introduced by the Facial Action Coding System (FACS) [8], the Action Units (AUs) are the relevant individual components elicited by the emotion class of the expression. Most of the research works treat the spotting and recognition as separate tasks, motivated by the Micro-Expression Grand Challenge (MEGC) [9,10]. To the best of our knowledge, the current attempts in the ME analysis domain are limited [6,11]. In particular, they employed traditional algorithms to spot the ME on the short videos that consist of only one ME and without the involvement of MaE, which is inappropriate to be adapted on the long videos datasets. Likewise, the features extracted to tackle the spotting and recognition tasks are completely different, which require extra efforts to be implemented. Considering that both tasks are strongly correlated, therefore, we propose an unified approach to solve the tasks in a single network.

In the preliminary work of [12], a shallow network that infuses the optical flow features and a pseudo-labeling technique is proposed to spot the ME. The present work extends the initial work and presents a comprehensive model to seamlessly spot and recognize the ME in both short and long videos.[1] In this paper, the main contribution is three-fold:

1. We propose a novel shallow multi-stream multi-output network named MEAN (or Micro-Expression Analysis Network) for spotting the ME intervals from a given video and proceeding to predict its corresponding emotion class.
2. We reinforce the use of evaluation metric AP@[.5:.95] to provide a fairer measurement for the ME spotting task. For benchmarking purposes, we propose a metric, namely Spot-Then-Recognize Score (STRS), which combines both spotting and analysis results to gauge the performance of a complete ME analysis system.
3. We report detailed ME analysis experimental results of the proposed approach on publicly available short and long video datasets. Extensive qualitative and quantitative analysis are also provided, highlighting the benefits of the proposed approach.

The rest of the paper is organized as follows: Section 2 reviews the related work in literature; Section 3 introduces the proposed methodology for ME analysis; Section 4 explains the datasets, evaluation metrics, and experiments settings, to measure the performance of the proposed algorithm. Section 5 presents the results of ME spotting and analysis on short and long videos with further ablation studies and discussions; Section 6 draws the conclusion of this paper and highlights some possible research directions. Our source code is publicly available.[2]

2. Literature review

While several terms have been presented under the umbrella of the ME field, the commonly used terminology is clarified herein. The definition of the long videos refers to the videos that embed both MaE and ME with a longer average duration (longer than 10 s [1,13]), whereas short videos signify the videos that contain only a single ME with a shorter average duration (mostly lesser than 10 s [5,14,15]). Moreover, ME analysis is the task that performs the ME spotting then recognition, which is the main focus of this paper.

2.1. Pre-processing

Pre-processing is the first step in the general pipeline of an ME analysis system to reduce the computational complexity and to improve performance. Facial landmark detection is the prior yet important step in pre-processing to locate the landmark points from the face. In the early works of ME, a manual selection of 12 facial landmark points is proposed by [16], however, the process is time-consuming and inaccurate when a non-expert attempts to select the landmarks. Later, numerous techniques are introduced to automatically select the facial landmark points, Active Shape Model (ASM) [17], Constrained Local Model (CLM) [18], Dlib [19], Discriminative Response Maps Fitting (DRMF) [20], Face++ [21].

Face masking and Region of Interest (ROI) selection attempt to address the challenges caused by the enormous amount of false ME detected when the unwanted facial movement is present. In the work of [22], they introduced a masking technique at the mouth and eye regions when performing ME spotting. Similarly, [11] adopted the eye masking technique due to significant movement of the eye blinking. On the flip side, several works [11,23–25] have proven that ROI selection at certain regions, such as "eye", "eyebrow", and "mouth" can improve the spotting and recognition performance.

2.2. ME spotting

ME spotting is categorized into two major problems, which are the temporal spotting to localize the intervals of the ME from onset to offset frame, and the apex spotting to detect only the apex frame. For temporal spotting, [26] proposed the appearance-based feature difference analysis using Local Binary Pattern with Chi-Square (LBP-χ^2) distance to locate the peak frame. The spotted peak frame is considered a true positive if it falls in a certain ME interval. However, this manner of evaluation seemed less precise and highly dependent on the interval size. Also, [27] proposed the Local Temporal Pattern with Machine Learning (LTP-ML) to classify the ME and non-ME frames using Support Vector Machine (SVM). Besides, the motion-based approach introduced by [22] utilized the derivative of optical flow, known as optical strain to obtain the facial deformation information. Main Directional Maximal Difference (MDMD) [28] provided the baseline result for long videos spotting in MEGC 2020. They encode the main direction during the process when the MaE and ME occur. With the booming of deep learning in this era, [25] adopted the Histogram of Oriented Optical Flow (HOOF) features into Recurrent Neural Network (RNN) to learn the relevant micro-movements in a sequence. The Micro-Expression Spotting Network (MESNet) [7] consists of three modules for spotting multi-scale ME intervals, revealing the potential of using CNN-based methods. More recently, [12] introduced the pseudo-labeling technique with Shallow Optical Flow Three-stream CNN (SOFTNet) that incorporates the optical flow features to predict the likelihood of a frame in the expression intervals.

On the contrary, the apex frame, which contains the highest intensity of an expression, is spotted rather than an interval. [23] employed the divide-and-conquer strategy to locate the apex frame that has the highest optical strain magnitudes. Subsequently, [29] proposed detecting the maximum frequency amplitude changes to represent the apex frame of the expression. In the work of [30], they devised a Spotting Micro-Expression Convolutional Network (SMEConvNet) to extract the features with a sliding window technique to locate the apex frame. However, most of the works that cater to apex spotting have experimented on short videos only.

2.3. ME recognition

ME recognition is the second step of an ME analysis system, a classification task to assign the emotion label for the expression. LBP on Three Orthogonal Planes (LBP-TOP) has been widely used [14,31–33] in the appearance-based approach with the ability to extract spatio-temporal information in the facial regions. It is known as the baseline method for most of the short videos datasets. Several LBP variants [34,35] are then introduced to improve the discriminative power and computational complexity.

Motion-based approach implemented by Main Directional Mean Optical-flow (MDMO) [36] considers the spatial location and local

[1] *Long videos* are untrimmed sequences that could contain multiple MEs, or a mix of MEs and MaEs and other head rotations, eye blinks, etc.

[2] https://github.com/genbing99/MEAN_Spot-then-recognize.

statistic optical flow features based on ROI, reported to be insensitive to illumination variations. By using only apex frame, [37] proposed Bi-weighted oriented optical flow (Bi-WOOF) as a feature descriptor to learn the motion information. With the increasing data samples in ME, not surprisingly, the deep learning approach has obtained state-of-the-art performance in ME recognition in recent years. [38] proposed Dual Temporal Scale Convolutional Neural Network (DTSCNN), a two-stream network to deal with ME videos that has multiple frame rates. To minimize the computational cost and chances of overfitting, shallow networks are proposed in the works of [39,40] to learn the motion information. Several works [39,41–43] have been accepted in the MEGC2019 [9] to spur the ME recognition results. Most recently, [44] introduced the Identity-aware and Capsule-Enhanced Generative Adversarial Network (ICE-GAN) to augment the ME samples for network training.

2.4. ME analysis

To the best of our knowledge, there are only two known works in current literature that had attempted the "ME analysis" task, which is a sequential integration of the spotting and recognition tasks. Both these works relied on handcrafted feature descriptors. Li et al. [6] experimented with LBP and HOOF descriptors for the spotting task, while LBP, HOG, HIGO features were extracted for the recognition task. Meanwhile, Liong et al. [11] employed optical strain and Bi-WOOF features for spotting and recognition tasks respectively. In essence, there is no guarantee that a good spotting technique would carry through to the recognition task and vice versa. In addition, we observe that both works used different pre-determined feature descriptors for both spotting and recognition, which incurs additional decision-making and increases the computational cost. Moreover, their methods have only been tested on the short video datasets, which are far less challenging than samples in the long videos datasets.

3. Proposed approach

We present a novel shallow multi-stream multi-output MEAN architecture, to continuously spot and recognize the ME intervals seamlessly. Specifically, the MEAN architecture comprises of two task-specific networks, namely spotting and recognition. The spotting network outputs a regression score to spot the ME intervals, while the recognition network outputs the predicted emotion label. This section describes the techniques involved in the proposed approach: pre-processing and optical flow feature extraction, Inductive Transfer Learning (ITL) strategy, the MEAN architecture, and finally post-processing for evaluation. Theoretical and mathematical justifications are provided accordingly.

3.1. Pre-processing and optical flow features

We follow the pre-processing and optical flow features extraction in the work of [12], as illustrated in Fig. 1. We first apply the Dlib toolkit [19] to detect the frontal face and 68 facial landmark points from the first frame of each video, then the face is cropped to the size of 128 × 128 pixels for all frames in the video.

Feature extraction using optical flow features are in fashion in ME works [12,22,39,40] owing to its robustness in estimating the micro-movements between two frames. TV-L1 algorithm [45] is adopted to obtain the horizontal u and vertical v optical flow components due to its robustness and ability to preserve flow discontinuities. The optical flow vector:

$$OF = \{(u_{xy}, v_{xy}) | x = 1, \dots, X; y = 1, \dots, Y\} \tag{1}$$

is computed using two frames, *i.e.* the current frame F_i and the kth frame from i, F_{i+k}, where they have a similar height X, width Y; the value of k will be clarified in the following sub-sections.

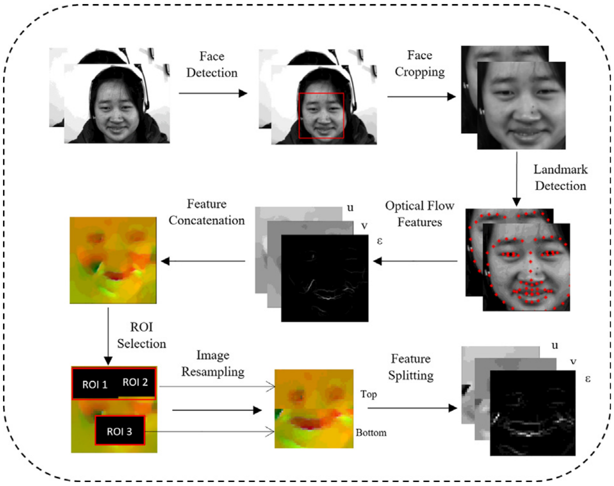


Fig. 1. Pre-processing steps of the proposed approach with optical flow feature extraction.

Subsequently, the optical strain is approximated from the optical flow components u and v according to infinitesimal strain theory. The optical strain, which captures the facial deformation information that is essential to measure the micro-movement, is computed as:

$$\epsilon = \begin{bmatrix} \epsilon_{xx} = \frac{\delta u}{\delta x} & \epsilon_{xy} = \frac{1}{2}(\frac{\delta u}{\delta y} + \frac{\delta v}{\delta x}) \\ \epsilon_{yx} = \frac{1}{2}(\frac{\delta v}{\delta x} + \frac{\delta u}{\delta y}) & \epsilon_{yy} = \frac{\delta v}{\delta y} \end{bmatrix} \tag{2}$$

whereby the diagonal components $(\epsilon_{xx}, \epsilon_{yy})$ represent normal strain components while $(\epsilon_{xy}, \epsilon_{yx})$ represent shear strain components. By taking the sum of squares, the optical strain magnitude ϵ, can be expressed as:

$$|\epsilon| = \sqrt{\frac{\partial u^2}{\partial x} + \frac{\partial v^2}{\partial y} + \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y})^2} \tag{3}$$

In this work, we refer to *optical flow features* as the concatenation of these three components (u, v, ϵ). We then perform ROI selection by adding extra 12 margin pixels for these three regions: (1) left eye and left eyebrow; (2) right eye and right eyebrow; (3) mouth, with the evidence that they contain relevant information; this follows the practice in [23]. To reduce the computational cost and retain the information, we perform image resampling, where the area that is bounded by regions (1) and (2) are resized into 21 × 21 pixels each, whereas region (3) is resized into 21 × 42 pixels; all three pieces are assembled to form an image of size 42 × 42 pixels. This procedure is applied to all channels of the optical flow features before they are split for the subsequent learning process.

3.2. ITL strategy

Inductive transfer learning (ITL) is a category of transfer learning approach where the source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$ are the same, while the source task $\mathcal{T}_S$ and target task $\mathcal{T}_T$ are different but related, as described in [46,47]. Since the data domain remains in ME, where $\mathcal{D}_S = \mathcal{D}_T$, this setting suits our learning objective to improve the performance of $\mathcal{T}_T$, *i.e.* recognition task with the knowledge learned from $\mathcal{T}_S$, *i.e.* spotting task. In all likelihood, we observe that the training sample size of the spotting task should be larger than that of the recognition task — this is because the spotting task takes each frame in the video as input while the recognition task uses only the ME frames. Thus, we hypothesize that transferring the knowledge by ITL can preserve meaningful features learned from the spotting task, hence easing the insufficient data issue while learning from the recognition task.

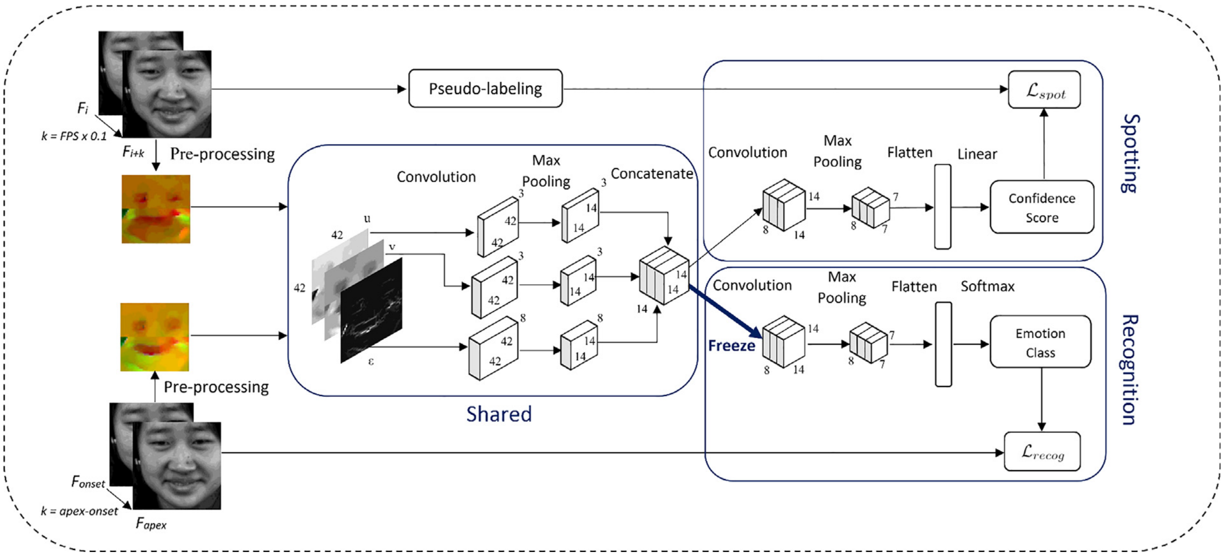


Fig. 2. Overview of our proposed MEAN architecture with three-stream inputs (u, v, ϵ) and two outputs (confidence score for spotting and emotion class for recognition). A two-step learning paradigm is applied where we first train the spotting network together with the shared layers, before training again on the recognition network with frozen shared layers.

3.3. MEAN architecture

The proposed Micro-Expression Analysis Network (MEAN) is a shallow architecture with three-stream inputs and two outputs, designed for ME spotting and recognition in a seamless manner (shown in Fig. 2). Precisely, MEAN architecture consists of two task-specific networks, in which the spotting network has a regression (linear) output, and the recognition network has a softmax output. Following the main architecture in [12], which incorporates three optical flow features (u, v, ϵ) as input, we revise the architecture to cover our interest in ME analysis. Specifically, the modifications are as follows: (1) The filters for the three-stream convolutional layer apply 3, 3, 8 instead of 3, 5, 8 respectively, with the consideration that horizontal and vertical components convey equally important features. (2) The final fully-connected layer is replaced with a "flatten layer" (which reshapes all feature maps into a single dimension) to reduce the number of parameters and network complexity. (3) A series of layers for the recognition network, *i.e.* convolutional layer, max-pooling layer, and flatten layer (followed by a softmax activation) are added after the shared portion of the network to predict the emotion class.

Formally, the MEAN model $\mathcal{M}$ takes three inputs (u, v, ϵ) of the ith frame, and predicts two outputs $(\hat{s}, \hat{c})$, where $\hat{s}$ refers to the predicted spotting confidence score and $\hat{c}$ refers to the predicted emotion class. Succinctly:

$$\hat{s}_i, \hat{c}_i = \mathcal{M}(u_i, v_i, \epsilon_i) \text{ for } i = 1, \ldots, F_{end} \tag{4}$$

where F_{end} refers to the last frame in the entire video. For clarity in representing the network layers in the architecture, we use the following notations: `conv`(number of filters) for the convolutional layer, `pool`(stride) for the max-pooling layer, `concat`(list of layers) for the concatenation layer, `flat` for the flatten layer, `linear` for the linear activation function, and `softmax` for the softmax activation function.

Suppose that each stream of the network takes an optical flow component input of fixed size (X, Y), the `conv` layer performs convolutional operations to form a feature map: $f_{conv}(z) = conv(\mathbb{R}^{X \times Y})$. Subsequently, the output feature map is passed through a ReLu activation function: $f_{ReLu}(z) = max(0, z)$. The `pool` layer is then used to minimize the computational complexity and highlight the most distinguished features in the feature map: $f_{pool}(z) = maxpool(z)$. We denote the layers up to the concatenation as the shared portion of MEAN, *i.e.* `shared` = `conv1a(3)` - `pool1a(3,3)` - `conv1b(3)` - `pool1b(3,3)` - `conv1c(8)` - `pool1c(3,3)` - `concat(pool1a, pool1b, pool1c)`. Subsequently, the output of `shared` (three concatenated streams) are passed on to the task-specific networks. We train the network using a two-step ITL learning paradigm by training the spotting network followed by the recognition network.

Spotting network: To train the spotting network, we construct the architecture using the following structure: `shared` - `conv2(8)` - `pool2(2,2)` - `flat` - `linear`. The training data is obtained from the optical flow features based on Eq. (1), using two frames.

Since all of the datasets contain onset and offset indices for each ME sample, an Intersection over Union (IoU)-based pseudo-labeling technique is employed to determine the confidence score s, similar to the work of [12]. To elaborate, we label each frame F_j using the sliding window W_j with length k' corresponding to the interval $[F_j, F_j + k' - 1]$, where $k' = (N + 1)/2$ is half of the average length of the ME, specific to each dataset. N denotes the average length of the ME in each dataset. The pseudo-labeling function, g applies the Heaviside step function:

$$g(IoU) = \begin{cases} 0, & \text{if } IoU \leq 0 \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where the IoU is computed as:

$$IoU = \frac{W \bigcap \mathcal{E}}{W \bigcup \mathcal{E}} \tag{6}$$

where $\mathcal{E} = [F_{onset}, F_{offset}]$

Finally, we obtain the pseudo-label set $S = \{s_i \text{ for } i = 1, \ldots, F_{end} - k'\}$ for subsequent loss computation.

After the `flat` layer, we obtain the predicted spotting confidence score $\hat{s}$ (which is a continuous value between 0 and 1) from activations z_i using a `linear` function:

$$\hat{s}(z_i) = f_{lin}(z_i) = \sum_i w_i z_i + b \tag{7}$$

Then, the spotting loss, $\mathcal{L}_{spot}$ is computed using conventional mean squared error (MSE) loss:

$$\mathcal{L}_{spot} = \frac{1}{n} \sum_{i=1}^{n} (s_i - \hat{s}(z_i)) \tag{8}$$

Recognition network: We apply ITL to transfer what has been learned from the spotting network to the recognition network by freezing the layers in `shared`, as depicted in Fig. 2. The recognition network is trained with the following architecture: `shared` − `conv2(8)` − `pool2(2,2)` − `flat` − `softmax`.

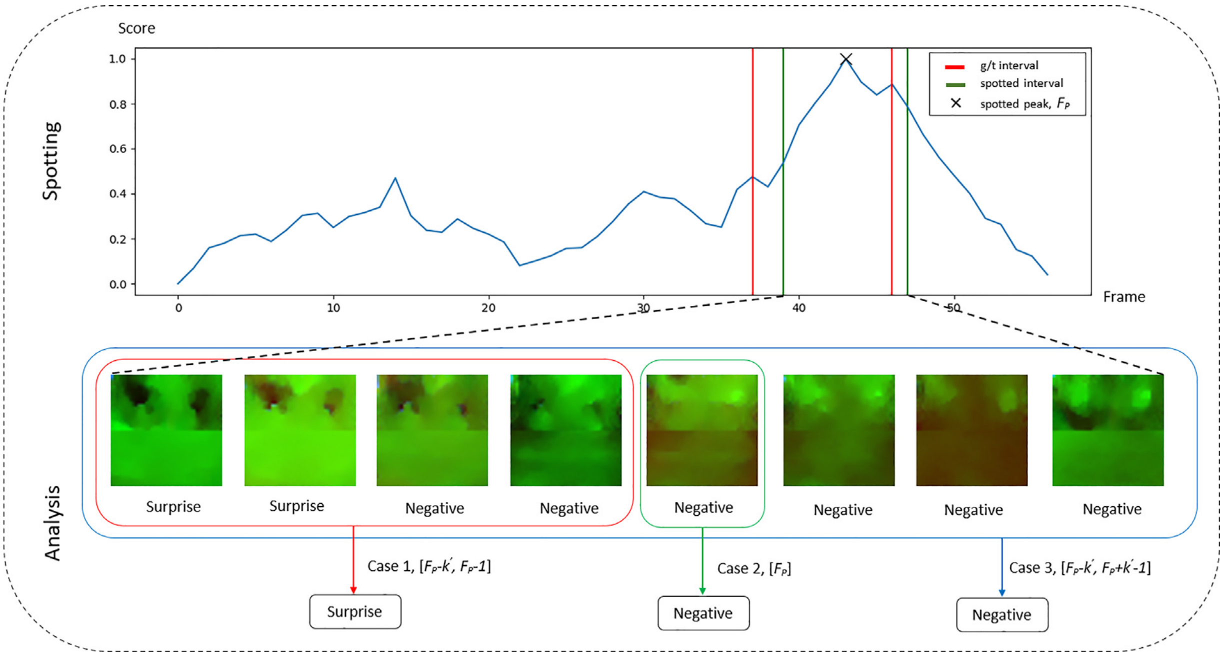


Fig. 3. Post-processing steps for ME spotting and analysis demonstrating the three different spotted intervals: (Case 1) onset to apex, $[F_P - k',\ F_P - 1]$; (Case 2) apex only, $[F_P]$; (Case 3) onset to offset, $[F_P - k',\ F_P + k' - 1]$.

From the `flat` layer, the predicted emotion probabilities for each sample are computed using the `softmax` function:

$$p(z_{ij}) = \frac{exp^{z_{ij}}}{\sum_{c=1}^{C} exp^{z_{ic}}} \quad \text{for } j = 1, \dots, C \tag{9}$$

where z_{ij} represents the predicted score for class j of the ith sample and C is the total number of classes. Subsequently, we optimize the recognition loss, $\mathcal{L}_{recog}$ using categorical cross-entropy loss which is defined as:

$$\mathcal{L}_{recog} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} c_{ij} log(p(z_{ij})) \tag{10}$$

where c_{ij} is the ground-truth emotion label for class j of the ith sample.

3.4. Post-processing

After the prediction of the entire video, we apply post-processing to obtain the spotted intervals and the corresponding predicted emotion. The process of ME spotting and analysis is illustrated in Fig. 3, where we experimented with three cases of spotted intervals: (case 1) onset to apex, $[F_P - k', F_P - 1]$; (case 2) apex only, $[F_P]$; (case 3) onset to offset, $[F_P - k',\ F_P + k' - 1]$. Note that F_P refers to the spotted peak frame.

ME spotting: The s_i from the output of $\mathcal{M}$ (i.e. Eq. (7)) is further aggregated using simple average smoothing function as follows:

$$\hat{s}_i = \frac{1}{2k' + 1} \sum_{j=i-k'}^{i+k'} \hat{s}_j \quad \text{for } i = F_1 + k', \dots, F_{end} - k' \tag{11}$$

where the predicted scores within the interval $[F_i - k', F_i + k']$ are averaged to avoid the false spotting of a sudden spike from the predicted scores. In addition, the aggregated score for each frame is interpreted as the likelihood of being in an interval of ME. Note that, the boundary frames, *i.e.* $F_1, \dots, F_1 + k' - 1$ and $F_{end} - k' + 1, \dots, F_{end}$, are discarded to prevent invalid spotted intervals in the subsequent steps.

We then use the simple peak detection technique in [48] to spot all the local maximas. Note that we set the threshold level to T and minimum peak distance to k'. Based on the thresholding technique suggested by [26], the threshold T is computed as:

$$T = \hat{S}_{mean} + p \times (\hat{S}_{max} - \hat{S}_{mean}) \tag{12}$$

where $\hat{S}_{mean}$ and $\hat{S}_{max}$ indicate the average and maximum score respectively in each video. We then extend the spotted peak frames F_P with k' frames on each side to obtain the interval $\hat{\mathcal{E}}_s = [F_P - k',\ F_P + k' - 1]$, which is the output of the spotting task.

ME analysis: To complete the ME analysis task, the emotion class $\hat{c}_i$ for each ith frame in the *spotted interval* $\hat{\mathcal{E}}_c = [F_P - k',\ F_P - 1]$ (case 1), is predicted. We then determine the predicted class of the sequence by its mode class, which takes the emotion class that has the highest number of occurrences in the sequence. In particular, we find that this method works well to avoid bias from a solitary frame (case 2), or the effect of noises from the entire window (case 3). We provide further details in Section 5.4 ablation studies.

4. Experiment

4.1. Dataset

To evaluate the proposed approach, we selected a total of six datasets: four short video datasets *i.e.* CASME II [5], SMIC [14] subsets (SMIC-E-HS, SMIC-E-VIS, SMIC-E-NIR), and two long video datasets *i.e.* CAS(ME)^2 [1], and SAMM Long Videos (SAMMLV) [13]; the summary is shown in Table 1 where the summary of annotation types, the emotion classes (number of samples) and more details of each dataset are provided. We note that this is the first known work that performs *ME analysis* on long videos datasets. Brief details on these datasets are as follows:

CASME II: The Chinese Academy of Sciences Micro-Expression (CASME II) is a well-known dataset comprising of 247 ME samples from 26 participants, which is the largest of the short video datasets. A Point Gray GRAS-03K2C camera is used to record the video at 200 FPS.

SMIC: The Spontaneous Micro-expression (SMIC) dataset contains three subsets, namely SMIC-E-HS, SMIC-E-VIS, and SMIC-E-NIR. In brief, SMIC-E-HS is captured using a high speed (HS) camera, PixeLINK PL-B774U at 100 FPS. It consists of 16 subjects with 157 ME samples. To improve the diversity of the samples, a visual camera (VIS) and near-infrared (NIR) are also used to record the ME of the selected subjects. Both these subsets were captured at 25 FPS, which is much lower than that of the HS subset.

Table 1
Summary of datasets used for ME analysis. Hap: Happiness, Dis: Disgust, Sur: Surprise, Rep: Repression, Oth: Others, Neg: Negative, Pos: Positive.

Dataset	CASME II	SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR	CAS(ME)^2	SAMMLV
Subjects	26	16	8	8	9	28
Videos	244	157	71	71	22	70
Samples	244	157	71	71	57	159
FPS	200	100	25	25	30	200
Resolution	640 × 480	640 × 480	640 × 480	640 × 480	640 × 480	2040 × 1088
Video Type	Short	Short	Short	Short	Long	Long
Emotion Class	Hap (32) Dis (61) Sur (25) Rep (27) Oth (99)	Neg (66) Pos (51) Sur (40)	Neg (24) Pos (28) Sur (19)	Neg (23) Pos (28) Sur (20)	Neg (21) Pos (8) Sur (9)	Neg (92) Pos (26) Sur (15)
Annotation	Onset, Apex, Offset	Onset, Offset	Onset, Offset	Onset, Offset	Onset, Apex, Offset	Onset, Apex, Offset
Parameter k'	34	17	4	4	6	37

CAS(ME)^2: The Chinese Academy of Sciences Macro-Expression and Micro-Expressions (CAS(ME)^2) dataset contains two parts, where part A has 87 long videos and part B has 300 MaE and 57 ME samples captured from a total of 22 subjects. The videos are filmed using a Logitech Pro C920 camera at 30 FPS.

SAMMLV: The Spontaneous Action and Micro-Movements Long Videos (SAMMLV) dataset is extended from the original SAMM dataset [15] to include videos of longer duration. It consists of 147 long videos with 343 MaE and 159 ME samples from 32 subjects. Videos were recorded using a Basler Ace acA2000-340 km at 200 FPS.

4.2. Performance metrics

While several performance metrics have been proposed in the current literature, we select widely used metrics to ensure fair comparisons and benchmarking while also introducing prospective metrics that are meaningful for ME spotting and analysis.

ME spotting: The Apex Spotting Rate (ASR) was introduced in the work of [11] to calculate the success rate of spotting the apex frame within the ground-truth onset and offset interval. The ASR is computed as:

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N}\delta, \qquad \text{where } \delta = \begin{cases} 1, & \text{if } f^{*} \in (f_{i,\text{onset}}, f_{i,\text{offset}}) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where N is the total number of videos in the dataset. However, this metric is only applicable in the short videos datasets where there exists only one ME in each video.

Thereafter, the MEGC2020 adopted the F1-score metric proposed by [49], where each ground-truth interval, $W_{groundTruth}$ in the video can be compared against the spotted interval, $W_{spotted}$. The spotted interval is determined as a True Positive (TP) if it satisfies the condition:

$$\frac{W_{spotted} \cap W_{groundTruth}}{W_{spotted} \cup W_{groundTruth}} \geq k_{IoU} \tag{14}$$

where k_{IoU} is typically set to 0.5. Otherwise, the spotted interval is considered as a False Positive (FP). In the entire dataset, suppose there are M ground-truth, N spotted intervals, and A spotted TPs, then False Positive, $FP = N - A$ and False Negative, $FN = M - A$. Thus, we can formulate the recall, precision, and F1-score metrics as follows:

$$\text{Recall} = \frac{A}{M}, \quad Precision = \frac{A}{N} \tag{15}$$

$$\text{F1-score}_s = \frac{2TP}{2TP + FP + FN} = \frac{2A}{M + N} \tag{16}$$

In addition, following [12], we propose the use of Average Precision (AP) over different IoU thresholds, ranging from 0.5 to 0.95 with a step size of 0.05, to measure the accuracy of spotted intervals over different tolerance levels. This measure is known as the AP@[.5:.95] metric which is made popular by the MS COCO object detection challenge [50]. Specifically, AP is computed using the definition of Area Under Curve (AUC) which is defined as:

$$\text{AP} = \sum_{n}^{N-1}(R_{n+1} - R_n)P_{interp}(R_{n+1}) \tag{17}$$

where the interpolated precision (P_{interp}) and recall (R) are taken for each predicted interval, and N is the total number of points after interpolation. Later, the AP@[.5:.95] is computed by taking the average of AP by varying IoU thresholds from 0.5 to 0.95, with increasing step size of 0.05. We encourage the use of this metric for spotting tasks to provide deeper insights into the capability of spotting based on different intervals.

ME analysis: The predicted emotion classes of the spotted intervals are obtained for the evaluation of ME analysis. The typical evaluation metric adopted is:

$$\text{Accuracy} = \frac{TP}{TP + FP} * 100\% \tag{18}$$

which measures the number of correctly predicted emotion classes over the total number of predictions.

Since there are multiple emotion classes according to the datasets, we compute the recall, precision, F1-score, UF1, and UAR metrics using:

$$\text{Recall} = \sum_{c=1}^{C}\frac{TP_c}{C \times (TP_c + FN_c)} \tag{19}$$

$$\text{Precision} = \sum_{c=1}^{C}\frac{TP_c}{C \times (TP_c + FP_c)} \tag{20}$$

$$\text{F1-score}_a = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{21}$$

$$UF1 = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{22}$$

$$UAR = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{n_c} \tag{23}$$

where C is the total number of classes and n_c is the number of samples of the cth class. These metrics are known to be effective in dealing with the imbalanced class distribution problem, which is obvious in these datasets.

Table 2
Performance comparison for ME spotting and analysis (using all spotted intervals) in terms of ASR and F1-score metrics respectively on the short video datasets. $\mathbb{S}$: Spotting, $\mathbb{A}$: Analysis.

Method	CASME II		SMIC-E-HS		SMIC-E-VIS		SMIC-E-NIR	
	$\mathbb{S}$	$\mathbb{A}$	$\mathbb{S}$	$\mathbb{A}$	$\mathbb{S}$	$\mathbb{A}$	$\mathbb{S}$	$\mathbb{A}$
Liong et al, 2016 [11]	0.823	0.59	0.3822	0.47	0.2817	0.53	0.2676	0.43
Zhang et al, 2018 [30]	0.828	N/A	–	N/A	–	N/A	–	N/A
MEAN (Ours)	**0.8443**	0.5723	**0.4904**	**0.4796**	**0.4789**	**0.5957**	**0.3239**	**0.449**

Overall: We review the metric Acc_{MESR} introduced in the work of [6], which multiplies the accuracy metric of spotting and recognition tasks. Due to the imbalanced class distribution, we propose a new metric, *i.e.* Spot-Then-Recognize Score or STRS in short, to evaluate the overall performance of the complete ME analysis system. It is simply defined as:

$$\text{STRS} = \text{F1-score}_s \times \text{F1-score}_a \tag{24}$$

where F1-score$_s$ and F1-score$_a$ are computed from the spotting and analysis tasks respectively as defined in Eqs. (16) and (21). Specifically, this metric measures the ability of an approach at spotting ME intervals and then classifying them into the correct emotions.

4.3. Experiment settings

For the spotting network in the proposed MEAN architecture, we select the value $k = FPS \times 0.1$ to compute the optical flow features between two frames, where FPS refers to frame per second. Typically, the upper limit duration of a micro-expression is 0.2 s, therefore we select the value k based on the number of frames within 0.1 s, which indicates half of the duration of the expression. Meanwhile, for the recognition network, we set the value $k = F_{apex} - F_{onset}$ for extracting optical flow features. Since the SMIC dataset does not provide the apex labels, we compute the apex frame $F_{apex} = \{max(F_{onset}, F_{onset+i})$ for $i = F_{onset} + 1, \dots, F_{offset}\}$, to represent the frame with maximum difference from onset to offset. For each ME sample, the F_{onset} and $F_{onset} + k$ frames are selected to compute the optical flow features. Note that, the values of k and k' are different, where k is used in the pre-processing phase while k' is used in the pseudo-labeling technique and post-processing phase.

The two-step learning process is required to train the proposed MEAN architecture. To train the spotting network with the shared network, we use the Adam optimizer with a learning rate of 5×10^{-4}, trained to a maximum of 200 epochs. To address the issue of dataset imbalance, we sample the non-ME and ME classes with a ratio of 8:1. Subsequently, we use the similar setting to train the recognition network. Considering that the training samples for recognition are fewer than that of the spotting task, we train the network up to a maximum of 800 epochs.

To validate the performance of the approaches without subject bias, we employ Leave-One-Subject-Out (LOSO) cross-validation protocol, where the dataset is split into the training set and testing set whilst leaving out samples belonging to one subject for testing while the remaining samples are used for training. This is repeated for a number of folds corresponding to the number of subjects, and the average performance is reported. As empirically tested in the work of [12], we set the parameter $p = 0.55$ for peak detection. Since short videos have only one ME sample in each video, we omitted the thresholding step during post-processing and consider only the highest peak frame for evaluation. On the contrary, the thresholding step is applied for the long videos.

All experiments are performed on a single NVIDIA GeForce GTX 1080 Ti GPU. Our proposed MEAN architecture is considered to be shallow with only 7608 parameters (3.5 million FLOPs). To determine the computational time in a single fold of LOSO cross-validation (CASME II dataset), we split the training set following a ratio of 80:20 for network training and validation. The first training stage (*i.e.*, spotting network) takes 109.91 s and 0.56 s for training and validation, respectively. Then, the second training stage (*i.e.*, recognition network) takes 7.71 s and 0.09 s for training and validation, respectively. The inference time of our approach (MEAN) takes only 0.0014 s per frame (~714 FPS) and this is highly promising for real-time applications.

Table 3
Performance comparison for ME analysis on SMIC-E-VIS dataset under different settings.

Method	Accuracy (%)	Settings
Liong et al, 2016 [11]	53.52	Evaluate all spotted intervals
MEAN (Ours)	**59.15**	
Li et al. 2017 [6]	56.67	Evaluate only spotted TP
MEAN (Ours)	**77.42**	

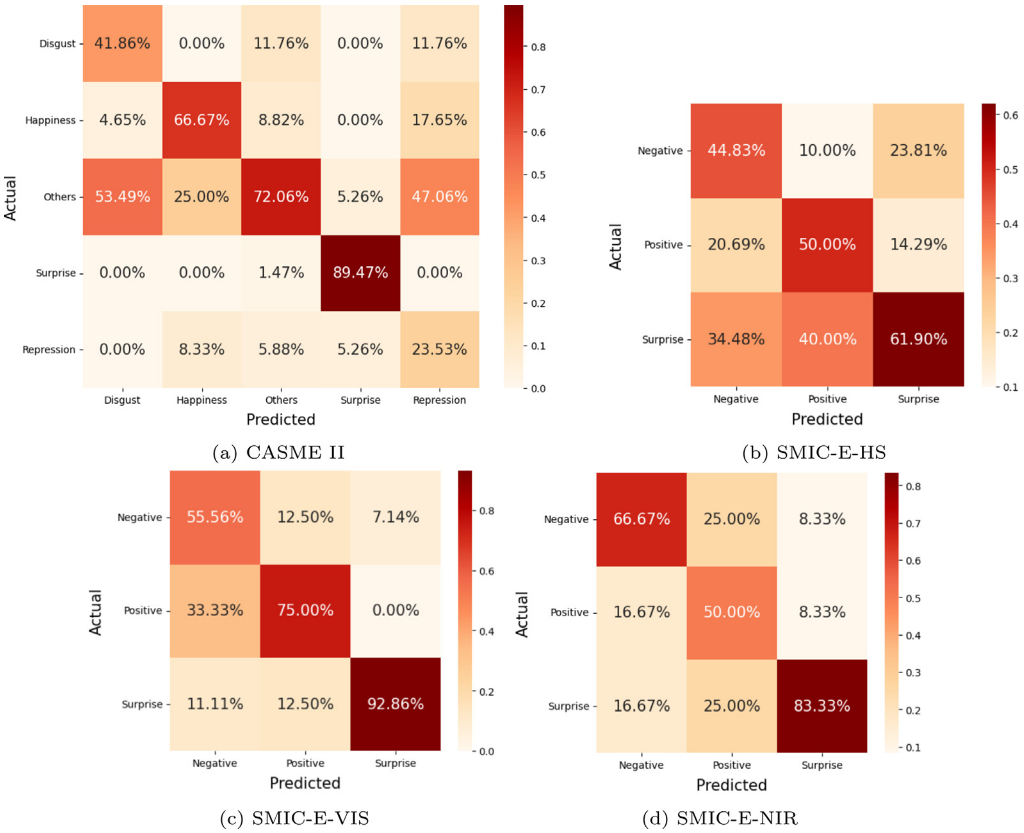
Table 4
Performance comparison for ME spotting on the long videos in terms of F1-score.

Method	CAS(ME)^2	SAMMLV
Yap et al, 2019 [13]	–	0.0508
He et al, 2020 [28]	0.0082	0.0364
Gan et al, 2020 [51]	0.0098	–
Zhang et al, 2020 [52]	0.0547	0.0725
Wang et al, 2021 [7]	0.0360	0.0880
Liong et al, 2021 [12]	0.1173	**0.1520**
MEAN (Ours)	**0.1214**	0.0949

5. Result and discussion

5.1. ME spotting and analysis on short videos

In current literature, the works that perform apex frame spotting on short video datasets are limited. In Table 2, we compare our proposed approach with existing works on four short videos datasets, namely CASME II, SMIC-E-HS, SMIC-E-VIS, and SMIC-E-NIR, where the spotting and analysis results are reported in terms of ASR and F1-score respectively. Note that the analysis result is obtained by evaluating the spotted intervals ($\hat{\mathcal{E}}_c$) for a fairer comparison against previous works such as [11]. We observe that the MEAN network outperforms other methods on the spotting task while achieving the best analysis task results on SMIC-E-HS, SMIC-E-VIS, and SMIC-E-NIR, while remaining comparable on the CASME II dataset. In addition, we also evaluate the proposed approach on the SMIC-E-VIS versus the work of [11] that considers all spotted intervals, and against the work of [6] that uses only the spotted TP that relies on the ground-truth labels (see Table 3). On both scenarios, we achieved superior performance in terms of accuracy metrics (for fair comparison against these methods). The confusion matrices of ME analysis for each short videos dataset are shown in Fig. 4 for more in-depth understanding of their class-wise performances. Generally, we notice that (1) the 'Others' class in CASME II is easily confusable with the 'Disgust' and 'Repression' classes and that future works should pay close attention towards distinguishing these classes better, and (2) the VIS and NIR subsets of SMIC were obviously very accurate at predicting 'Surprise' emotions as these modalities seemed to provide better discriminability.



(a) CASME II

(b) SMIC-E-HS

(c) SMIC-E-VIS

(d) SMIC-E-NIR

Fig. 4. Confusion matrices of ME analysis (using only spotted TP) for short video datasets: (a) CASME II; (b) SMIC-E-HS; (c) SMIC-E-VIS; (d) SMIC-E-NIR. Results shown in Accuracy (%).

5.2. ME spotting and analysis on long videos

Notably, this is the first work that reports ME analysis on long videos (CAS(ME)^2 and SAMMLV), which we intend to establish baseline results for future research. Table 4 compares the ME spotting results compiled from MEGC2020 and the recent works [7,12]. The proposed MEAN achieved good results on the spotting task, surpassing the MEGC2020 state-of-the-art approaches on CAS(ME)^2 while ranking second on SAMMLV. While this paper is the extension of our earlier work [12] which achieved the highest spotting result on SAMMLV, we now use a more justifiable k value to compute the optical flow features. Instead of selecting the best k empirically, we formulate the value based on the FPS of the collected videos. For the new 3-class analysis task where there were no prior works, we achieve an F1-score of 0.6667 and 0.5263 on CAS(ME)^2 and SAMMLV respectively. It is worth mentioning that the number of ME samples in the long videos is very few (38 for CAS(ME)^2 and 133 for SAMMLV on a 3-class basis) due to the enormous amount of neutral frames, unrelated head movements, and also the involvement of MaE samples. The lack of training data remains a challenging issue for long videos analysis and future work should attempt to address this. The confusion matrices of the ME analysis for CAS(ME)^2 and SAMMLV datasets are reported in Fig. 5. It is observed that the network performs poorly on the 'Surprise' emotion, which can be attributed to the limited samples of this class in the datasets.

5.3. Discussion on metrics

We show a more detailed view of the ME spotting and analysis results using our approach on short and long video datasets in Table 5. Importantly, we observe that a larger number of TPs and lesser FPs were spotted in short videos compared to long videos. This is likely due to these possible reasons: (1) short videos mainly consist of only a single ME while long videos could have multiple MEs intertwined with MaEs in each video; (2) short videos have a smaller number of neutral frames which significantly reduces the spotting difficulty; (3) different peak detection techniques were applied on both short and long videos, as elaborated in Section 4.3.

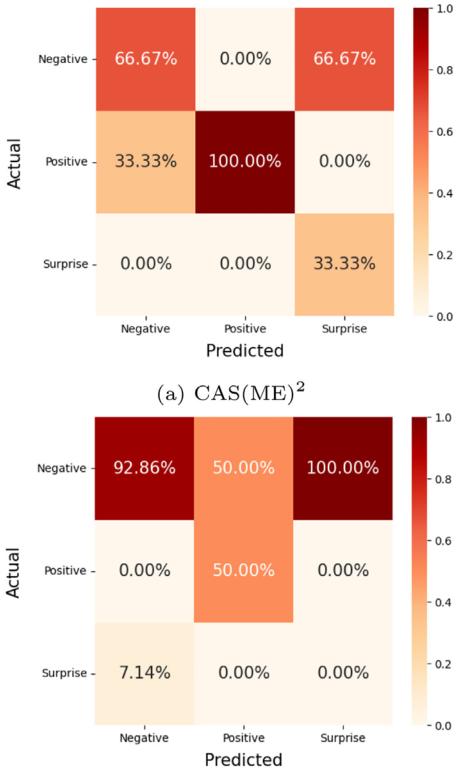
To bridge this gap, we promote the usage of the metric AP@[.5:.95] from object detection [50] for the spotting task. Specifically, it measures the quality of the spotted intervals by considering different overlapping ratios. This can help strike a balance between intervals that are sparsely intersecting and those that are overlap heavily, providing a fairer assessment of the spotting quality. Thus, the spotted intervals are expected to be as close as possible to the actual intervals in order to achieve a high AP@[.5:.95].

Further to the ME analysis, only the spotted TP intervals are taken into consideration for the second recognition step. This is because the spotted sequences that were not able to precisely locate the exact onset and offset frames may not yield the correct emotion type. Comparing between the spotted TP intervals and their corresponding ground-truth windows, our proposed approach uses a fixed k frame distance (note: k is half the average length of ME in the dataset) to generate the optical flow features (u, v, ϵ) of Eq. (1). Hence, the performance of ME analysis using features computed from these spotted intervals is highly dependent on the spotting accuracy.

Finally, the proposed STRS metric multiplies the F1-score for both spotting and analysis tasks. Generally, the short videos are expected to achieve better performance compared to long videos due to the higher spotting result obtained. It is observed that SMIC-E-NIR has a slightly better STRS than the SMIC-E-HS dataset, which was let down by poorer spotting accuracy. On the contrary, CASME II performed well in the spotting task, which in turn, contributed to a much higher STRS

Table 5
Detailed results for ME spotting and analysis on short and long videos. 𝕊: Spotting, 𝔸: Analysis.

Metrics	Short video								Long video			
	CASME II		SMIC-E-HS		SMIC-E-VIS		SMIC-E-NIR		CAS(ME)^2		SAMMLV	
	𝕊	𝔸	𝕊	𝔸	𝕊	𝔸	𝕊	𝔸	𝕊	𝔸	𝕊	𝔸
Total	244	159	157	60	71	31	71	22	57	7	159	19
TP	159	96	60	31	31	24	22	16	17	4	26	14
FP	85	63	97	29	40	7	49	6	206	3	363	5
FN	85	63	97	29	40	7	49	6	40	3	133	5
Precision	0.6516	0.5983	0.3822	0.5076	0.4366	0.7492	0.3099	0.6667	0.0762	0.6667	0.0668	0.5882
Recall	0.6516	0.5872	0.3822	0.5224	0.4366	0.7447	0.3099	0.6667	0.2982	0.6667	0.1635	0.4762
F1-score	0.6516	0.5927	0.3822	0.5149	0.4366	0.7470	0.3099	0.6667	0.1214	0.6667	0.0949	0.5263
AP@[.5:.95]	0.1180	N/A	0.0653	N/A	0.0708	N/A	0.0393	N/A	0.0152	N/A	0.0104	N/A
UF1	N/A	0.6755	N/A	0.5000	N/A	0.7425	N/A	0.6667	N/A	0.5794	N/A	0.5018
UAR	N/A	0.5872	N/A	0.5224	N/A	0.7447	N/A	0.6667	N/A	0.6667	N/A	0.4762
STRS	0.3862		0.1968		0.3261		0.2066		0.0809		0.0499	



(a) CAS(ME)^2

(b) SAMMLV

Fig. 5. Confusion matrices of ME analysis (using only spotted TP) for long videos datasets: (a) CAS(ME)^2; (b) SAMMLV datasets in terms of accuracy (%).

score. This demonstrates the feasibility of this metric in providing an overall measure of a complete ME analysis system; a high STRS can only be obtained when both tasks perform reasonably well. Broadly speaking, the STRS scores also intuitively reveals the *trainability* of each dataset — the CASME II being obviously more well-trained than the SMIC counterparts (likely due to its larger sample size or higher FPS), and SAMMLV being an obviously more complex dataset to work on compared to CAS(ME)^2.

5.4. Ablation studies

In this section, we report some ablation studies conducted on the proposed approach, as well as a qualitative analysis of the results.

Optical flow length k: The parameter k is used to determine the distance between two frames for the computation of optical flow features in Section 3.1. It is of paramount importance to capture the significant movement in an expression. Therefore, we tested the value of k empirically by varying the duration that multiples with the FPS of the dataset. In Table 6, it is observed that $k = FPS \times 0.10$ clearly obtained the best results on ME spotting, analysis and overall. As suggested by [53], the ME onset phase contains the fundamental features with a duration lower limit of about 0.065 s and an upper limit of about 0.260 s. Our k (0.1 s) falls within the limit range and is reasonable.

Table 6
Performance comparison of ME spotting (F1-score), analysis (F1-score), and overall (STRS) using different k values on the CASME II dataset.

k	Spotting	Analysis	Overall
$FPS \times 0.05$	0.6398	0.5739	0.3672
$FPS \times 0.10$	**0.6516**	**0.5927**	**0.3862**
$FPS \times 0.15$	0.5000	0.5169	0.2585

Table 7
Performance comparison of ME spotting (F1-score), analysis (F1-score), and overall (STRS) using different smoothing functions on the CASME II dataset.

Smoothing function	Spotting	Analysis	Overall
Simple average	**0.6516**	**0.5927**	**0.3862**
Gaussian ($\sigma = 1$)	0.2869	0.3874	0.1111
Gaussian ($\sigma = 3$)	0.2787	0.3557	0.0991
Exponential ($\tau = 1$)	0.2746	0.3521	0.0967
Exponential ($\tau = 3$)	0.2705	0.3314	0.0896

Spotting smoothing functions: The simple average smoothing function is implemented in the ME spotting post-processing phase (see Section 3.4). In particular, we also experimented with different smoothing functions to aggregate the predicted scores. Table 7 reports the performance comparison of selected functions: simple average, Gaussian with standard deviation 1 and 3, and exponential with decay rate 1 and 3 [48] on CASME II dataset. Interestingly, the simple average outperforms all tasks by a significant margin. This can be explained by the pseudo-labeling function (see Eq. (5)) that assigns an equal value of 1 to the ME frames regardless of the percentage of intersection, which corresponds to the effect of simple averaging computed based on equal weights.

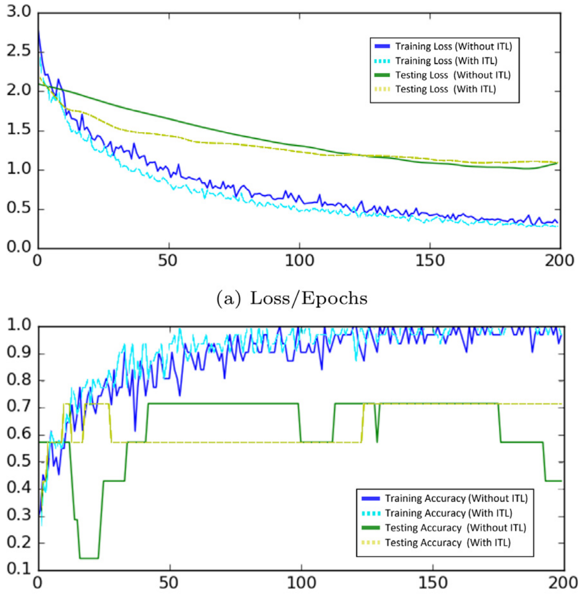
Spotted interval cases: An experiment is performed to determine the length of the spotted intervals for ME analysis during the post-processing phase. Table 8 compares the results of three different cases of spotted intervals for ME analysis to better justify our choice of using the interval (case 1) onset to apex, $[F_P - k', F_P - 1]$. A superior performance is achieved using case 1 interval, alluding to the fact that psychological studies [53] have shown that the "onset phase" (*i.e.* interval between the onset and apex) possesses the fundamental characteristics of MEs and hence, optical flow features derived from this interval are likely to contain the most relevant features for learning. Our experimental results verify this fact quite strongly.

Table 8
Performance comparison of ME analysis (using only spotted TP) between three different spotted interval cases as indicated in Section 3.4. Results shown in F1-score.

Intervals	CASME II	SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR	CAS(ME)^2	SAMMLV
Case 1	**0.5927**	**0.5149**	**0.7470**	**0.6667**	**0.6667**	**0.5263**
Case 2	0.2626	0.3273	0.502	0.2558	0.3704	0.2619
Case 3	0.4483	0.4455	0.6428	0.2927	0.5072	0.2759

Table 9
Performance comparison of ME analysis (using only spotted TP) between MEAN architecture with and without ITL. Results shown in F1-score.

Strategy	CASME II	SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR	CAS(ME)^2	SAMMLV
Without ITL	0.5423	0.4710	0.4251	0.4058	0.4938	0.2917
With ITL	**0.5723**	**0.4796**	**0.5958**	**0.4490**	**0.6667**	**0.5263**



(a) Loss/Epochs

(b) Accuracy/Epochs

Fig. 6. Loss and accuracy of the proposed MEAN with ITL and without ITL over 200 epochs. A single LOSOCV fold with Subject 3 (CASME II dataset) held-out is shown in this example.

Impact of ITL: We also conducted a control experiment (MEAN network without ITL strategy) on the ME analysis task without enforcing knowledge transfer after the spotting task. To demonstrate this, we implemented two separate networks for spotting and recognition tasks without the weight sharing in the first few layers. As shown in Table 9, we observe an improvement in performance by applying the ITL strategy in ME analysis. The pre-trained MEAN on the spotting task is able to improve the generalization ability of the recognition task that has lesser samples. Thus, it is evident that the proposed network allows the preservation of lower-level features learned earlier from the initial training stage. Theoretically, we believe that using the low-level features in the shared layers can address the issue of insufficient data samples, which is well-known for the ME recognition task.

Further, we also analyze the effectiveness of applying ITL in the proposed approach by plotting the loss and accuracy over the training iterations during feature learning. The learning curves of the MEAN with ITL and without ITL are shown in Fig. 6. Specifically, we notice that the convergence speed increases at the beginning of training for the network with ITL, where the training loss decreases and the training accuracy reaches the peak at a much faster pace. At the same time, the testing loss also converges earlier, obtaining a higher and more consistent testing accuracy compared to the network trained without ITL. In brief, we show that the ITL strategy can improve the performance and convergence speed through weight sharing in the first few layers of the network model.

Qualitative analysis: To further analyze what the proposed network is learning, we apply Gradient-weighted Class Activation Mapping (GradCam) [54] to visualize the class activation heatmaps. The gradients that flow into the concatenated layers of the spotting and recognition network in the MEAN architecture are computed to produce the coarse localization heatmaps. We further overlay the facial region with the generated heatmaps to better visualize the regions that highly activate the predicted emotion class. In Fig. 7, we show a few examples that are spotted and classified correctly by both spotting and recognition networks respectively. The salient parts are colored according to a color gradient (from blue, green, yellow, to red in increasing importance) to pinpoint areas of high activation. Concretely, we observe that the activated regions are closely related to the AUs of the specific predicted emotion; these locations intuitively show where the model is "looking" at. In Fig. 7, we make the following observations: (1) Negative (self-reported as anger) emotion is emphasized by an activated AU 4 (brow lowerer); (2) Positive (self-reported as happy) emotion shows strong activations at AU 6 (cheek raiser) and AU 12 (lip corner puller); (3) Surprise emotion clearly indicates a distinct contribution from AU 1 (inner brow raiser) and AU 2 (outer brow raiser). Additionally, we also note that the activated regions that are not related to the supposed emotion, such as the lip corner for 'Surprise' emotion, warrants further investigation.

In-the-wild Analysis: As proven in the main experiments, the proposed approach achieved groundbreaking results for ME analysis for samples captured in the constrained lab environment. In addition, we also evaluate the ME analysis task on samples from the unconstrained environment (*a.k.a* in-the-wild) which may contain occlusion, pose variation, and illumination changes. Thus far, there is only one dataset — the Micro-Expression VIdEos in-the-Wild (MEVIEW) [55] dataset with limited attempts to perform recognition [56]. Succinctly, MEVIEW consists of 21 ME samples (re-categorized into 9 negatives, 6 positives, and 6 surprises) from 14 subjects. The videos are recorded at 25 FPS with an image resolution of 1280 × 720. To properly align the facial viewpoints, we employed FacePoseNet [57] following the work of [56] for a fairer comparison. We used similar experimental settings (see Section 4.3) with the k' computed to be 6.

The performance comparison of our proposed approach with the recent prior work [56] for ME spotting and analysis on MEVIEW is shown in Table 10. Notably, our proposed approach outperforms their method on the ME spotting task with an ASR of 0.4286. This is ascribed to the capability of our deep learning-based method in utilizing the salient motion features in a meaningful way. However, we perform less well on the ME analysis in terms of F1-score, UF1, and UAR. This is because of the limited ME samples for training the recognition network. Thereby, we reckon that with the increasing amount of in-the-wild ME samples available in the near future, the ME analysis performance can be further improved.

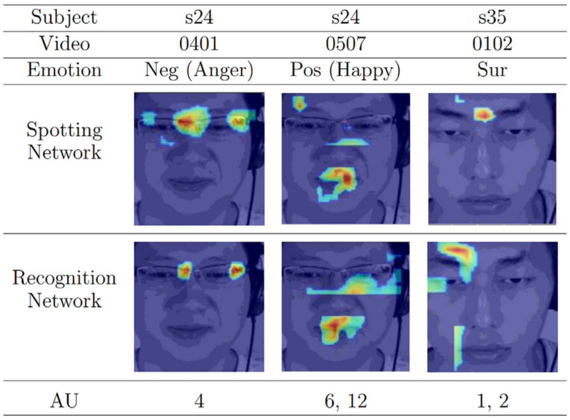


Fig. 7. Visualization of the activated facial regions correspond to the elicited emotion using GradCam. Neg: Negative, Pos: Positive, Sur: Surprise.

Table 10
Performance comparison for ME spotting and analysis (using all spotted intervals) on MEVIEW. 𝕊: Spotting, 𝔸: Analysis.

Method	𝕊	𝔸		
	ASR	F1-score	UF1	UAR
Gan et al, 2022 [56]	0.3333	**0.6758**	**0.6575**	**0.6852**
MEAN (Ours)	**0.4286**	0.5190	0.4592	0.4620

6. Conclusion

In summary, this paper proposes a novel shallow multi-stream multi-output MEAN architecture to perform a "spot-then-recognize" task on ME sequences. This work presents a fresh attempt at performing ME analysis, where both spotting and recognition tasks are integrated seamlessly using only a single feature extraction step brought upon by shared layers of the neural network. The network is trained in two stages, namely spotting and recognition modules, where the knowledge of the former is induced to the latter to achieve better performance. We achieved promising results on both ME spotting and analysis for short videos datasets (CASME II, SMIC-E-HS, SMIC-E-VIS, SMIC-E-NIR), while also providing baseline results for ME analysis on two long video datasets (CAS(ME)^2 and SAMMLV). While we encourage the ME community to adopt AP@[.5:.95], which is a fairer metric for measuring window overlaps, we also report our results on a newly introduced STRS metric, which measures the performance of a complete ME analysis system. Finally, we also provide additional ablation studies that justify various choices in our methodology, offer further insights into the AUs that correspond to the predicted emotions, and tested our approach in challenging in-the-wild scenarios.

Future Directions. There are some limitations in the proposed approach that are potential future directions. Firstly, the ME interval is obtained by extending the spotted peak frame in each video in the post-processing phase. This is not a guaranteed assumption since MEs could naturally have different durations. Hence, future work may be dedicated towards estimating their respective intervals based on the spotting confidence scores. Secondly, the evaluation performed on each dataset could do better with more data samples. To deal with this issue, these datasets could be merged to learn more robust models for composite dataset evaluation. Apart from that, the network can be trained on a larger dataset [58] to improve the generalization ability. Lastly, further explorations can be taken to include the depth information [59] to facilitate the ME analysis.

CRediT authorship contribution statement

Gen-Bing Liong: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft. **John See:** Conception and design of study, Acquisition of data, Writing – original draft, Writing – review & editing. **Chee-Seng Chan:** Conception and design of study, Analysis and/or interpretation of data, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code in the manuscript.

Acknowledgments

All authors approved the version of the manuscript to be published.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, X. Fu, CAS(ME)^2: a database for spontaneous macro-expression and micro-expression spotting and recognition, IEEE Trans. Affect. Comput. 9 (4) (2017) 424–436.

[2] P. Ekman, Telling Lies: Clues To Deceit in the Marketplace, Politics, and Marriage (Revised Edition), WW Norton & Company, 2009.

[3] P. Ekman, MicroExpression Training Tool (METT), University of California, San Francisco, 2002.

[4] P. Ekman, W.V. Friesen, Nonverbal leakage and clues to deception, Psychiatry 32 (1) (1969) 88–106.

[5] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, PLoS One 9 (1) (2014) e86041.

[6] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietikäinen, Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods, IEEE Trans. Affect. Comput. 9 (4) (2017) 563–577.

[7] S.-J. Wang, Y. He, J. Li, X. Fu, MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos, IEEE Trans. Image Process. 30 (2021) 3956–3969.

[8] E. Friesen, P. Ekman, Facial action coding system: a technique for the measurement of facial movement, Palo Alto 3 (2) (1978) 5.

[9] J. See, M.H. Yap, J. Li, X. Hong, S.-J. Wang, Megc 2019–the second facial micro-expressions grand challenge, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[10] L. Jingting, S.-J. Wang, M.H. Yap, J. See, X. Hong, X. Li, MEGC2020-the third facial micro-expression grand challenge, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 777–780.

[11] S.-T. Liong, J. See, K. Wong, R.C.-W. Phan, Automatic micro-expression recognition from long video using a single spotted apex, in: Asian Conference on Computer Vision, Springer, 2016, pp. 345–360.

[12] G.-B. Liong, J. See, L.-K. Wong, Shallow optical flow three-stream cnn for macro-and micro-expression spotting from long videos, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2643–2647.

[13] C.H. Yap, C. Kendrick, M.H. 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), IEEE, 2020, pp. 771–776.

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

[15] A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE Trans. Affect. Comput. 9 (1) (2016) 116–129.

[16] S. Polikovsky, Y. Kameda, Y. Ohta, Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor, 2009.

[17] S. Milborrow, F. Nicolls, Active shape models with SIFT descriptors and MARS, in: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Vol. 2, IEEE, 2014, pp. 380–387.

[18] D. Cristinacce, T.F. Cootes, et al., Feature detection and tracking with constrained local models, in: Bmvc, Vol. 1, Citeseer, 2006, p. 3.

[19] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[20] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3444–3451.

[21] I. Megvii, Face++ research toolkit, 2013.

[22] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, S. Sarkar, Automatic expression spotting in videos, Image Vis. Comput. 32 (8) (2014) 476–486.

[23] S.-T. Liong, J. See, K. Wong, A.C. Le Ngo, Y.-H. Oh, R. Phan, Automatic apex frame spotting in micro-expression database, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 665–669.

[24] J. Li, C. Soladie, R. Seguier, Ltp-ml: Micro-expression detection by recognition of local temporal pattern of facial movements, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 634–641.

[25] M. Verburg, V. 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), IEEE, 2019, pp. 1–6.

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

[27] J. Li, C. Soladie, R. Seguier, Local temporal pattern and data augmentation for micro-expression spotting, IEEE Trans. Affect. Comput. (2020).

[28] Y. He, S.-J. Wang, J. Li, M.H. Yap, 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, 2020, pp. 742–748.

[29] Y. Li, X. Huang, G. Zhao, Can micro-expression be recognized based on single apex frame? in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3094–3098.

[30] Z. Zhang, T. Chen, H. Meng, G. Liu, X. Fu, SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos, IEEE Access 6 (2018) 71143–71151.

[31] Y. Guo, Y. Tian, X. Gao, X. Zhang, Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method, in: 2014 International Joint Conference on Neural Networks (IJCNN), IEEE, 2014, pp. 3473–3479.

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

[33] Y. Wang, J. See, Y.-H. Oh, R.C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, X. Li, Effective recognition of facial micro-expressions with video motion magnification, Multimedia Tools Appl. 76 (20) (2017) 21665–21690.

[34] Y. Wang, J. See, R.C.-W. Phan, Y.-H. Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: Asian Conference on Computer Vision, Springer, 2014, pp. 525–537.

[35] Y. Wang, J. See, R.C.-W. Phan, Y.-H. Oh, Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition, PLoS One 10 (5) (2015) e0124674.

[36] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A main directional mean optical flow feature for spontaneous micro-expression recognition, IEEE Trans. Affect. Comput. 7 (4) (2015) 299–310.

[37] S.-T. Liong, J. See, K. Wong, R.C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Process., Image Commun. 62 (2018) 82–92.

[38] M. Peng, C. Wang, T. Chen, G. Liu, X. Fu, Dual temporal scale convolutional neural network for micro-expression recognition, Front. Psychol. 8 (2017) 1745.

[39] S.-T. Liong, Y.S. Gan, J. See, H.-Q. Khor, Y.-C. Huang, Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[40] Y.S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, OFF-ApexNet on micro-expression recognition system, Signal Process., Image Commun. 74 (2019) 129–139.

[41] Y. Liu, H. Du, L. Zheng, T. Gedeon, A neural micro-expression recognizer, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–4.

[42] L. Zhou, Q. Mao, L. Xue, Dual-inception network for cross-database micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[43] N. Van Quang, J. Chun, T. Tokuyama, CapsuleNet for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–7.

[44] J. Yu, C. Zhang, Y. Song, W. Cai, ICE-GAN: Identity-aware and Capsule-Enhanced GAN for Micro-Expression Recognition and Synthesis, 2020.

[45] J.S. Pérez, E. Meinhardt-Llopis, G. Facciolo, TV-L1 optical flow estimation, Image Process. Line 2013 (2013) 137–150.

[46] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[47] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (1) (2020) 43–76.

[48] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in python, Nature Methods 17 (2020) 261–272, http://dx.doi.org/10.1038/s41592-019-0686-2.

[49] J. Li, C. Soladie, R. Seguier, S.-J. Wang, M.H. Yap, Spotting micro-expressions on long videos sequences, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[51] Y. Gan, S. Liong, D. Zheng, S. Li, C. Bin, Optical strain based macro-and micro-expression sequence spotting in long video, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2020.

[52] L.-W. Zhang, J. Li, S.-J. Wang, X.-H. Duan, W.-J. Yan, H.-Y. Xie, S.-C. Huang, Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences, in: 15th IEEE FG, 2020, pp. 245–252.

[53] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, X. Fu, How fast are the leaked facial expressions: The duration of micro-expressions, J. Nonverbal Behav. 37 (4) (2013) 217–230.

[54] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[55] P. Husák, J. Cech, J. Matas, Spotting facial micro-expressions “in the wild”, in: 22nd Computer Vision Winter Workshop (Retz), 2017, pp. 1–9.

[56] Y. Gan, J. See, H.-Q. Khor, K.-H. Liu, S.-T. Liong, Needle in a haystack: Spotting and recognising micro-expressions “in the wild”, Neurocomputing (2022).

[57] F.-J. Chang, A. Tuan Tran, T. Hassner, I. Masi, R. Nevatia, G. Medioni, Faceposenet: Making a case for landmark-free face alignment, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1599–1608.

[58] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[59] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, IEEE Trans. Pattern Anal. Mach. Intell. (2022).