Automatic Micro-expression Recognition from Long Video using a Single Spotted Apex

Sze-Teng Liong[1], John See[2], KokSheik Wong[1], Raphael Chung-Wei Phan[3]

[1] Faculty of Computer Science & Information Technology,
University of Malaya, Kuala Lumpur, Malaysia
szeteng1206@hotmail.com,koksheik@um.edu.my
[2] Faculty of Computing & Informatics,
Multimedia University, Cyberjaya, Malaysia
johnsee@mmu.edu.my
[3] Faculty of Engineering,
Multimedia University, Cyberjaya, Malaysia
raphael@mmu.edu.my

Abstract. Recently, micro-expression recognition has seen an increase of interest from psychological and computer vision communities. As micro-expressions are generated involuntarily on a person's face, and are usually a manifestation of repressed feelings of the person. Most existing works pay attention to either the detection or spotting of micro-expression frames or the categorization of type of micro-expression present in a short video shot. In this paper, we introduced a novel automatic approach to micro-expression recognition from long video that combines both spotting and recognition mechanisms. To achieve this, the apex frame, which provides the instant when the highest intensity of facial movement occurs, is first spotted from the entire video sequence. An automatic eye masking technique is also presented to improve the robustness of apex frame spotting. With the single apex, we describe the spotted micro-expression instant using a state-of-the-art feature extractor before proceeding to classification. This is the first known work that recognizes micro-expressions from a long video sequence without the knowledge of onset and offset frames, which are typically used to determine a cropped sub-sequence containing the micro-expression. We evaluated the spotting and recognition tasks on four spontaneous micro-expression databases comprising only of raw long videos – CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR. We obtained compelling results that show the effectiveness of the proposed approach, which outperform most methods that rely on human annotated sub-sequences.

1 Introduction

Micro-expression is a form of nonverbal communication that unconsciously reveals the true sentiments of a person. Micro-expressions are exhibited subtly and they typically occur very briefly, at a duration of about 1/5 to 1/25 of a second [1]. The intensity of the micro-expression is as small as twitching a tiny part of the facial muscles [2]. Thus, it is difficult to observe micro-expressions in real-time conversations due to the minuteness and the quickness of the motion. On the contrary, normal expressions, also known as macro-expressions, lasts between 3/4 of a second to 2 seconds, and can appear at multiple large areas of the facial regions [3], making them easier to be detected. Similar to macro-expressions, micro-expressions can be grouped into six basic expressions: happy, surprise, anger, sad, disgust and fear [4]. Various research groups, especially in the behavioral and computing fields, are interested to analyze micro-expressions mainly due to its usefulness in uncovering the emotional state of a person who is attempting to conceal it [5]. Hence, micro-expression recognition is useful in a wide range of applications, including clinical diagnosis, national security, and interrogation.

Micro-expression is a dynamic facial action which evolves in the following sequence of states: neutral-onset-apex-offset-neutral [6]. Starting from a neutral state, an onset frame indicates the beginning of a micro-expression where the facial muscles begin to undergo contraction, while offset frame is the end of the expression where the intensity of the muscles is reduced to zero. Apex frame is the instant when the micro-expression reaches its climax (the most intense movement). The apex is not necessarily located at the middle between the onset and offset frames, but it can be situated at any frame in the onset-offset range. In our work, we define the video sub-sequence that is composed only of frames from onset to offset as "*short video*". On the other hand, "*long video*" refers to the raw video sequence which may include the frames with micro-expressions and irrelevant motion that are present before the onset and after the offset. Fig. 1 illustrates the short and long video sequences with onset-apex-offset frame annotations. Notice how a micro-expression sequence (frames 30-120) in a long video can be easily shrouded by frames outside the onset-offset range that contain eye blinks and head rotations (such as in frames 15 and 150). In current literature, most works categorized micro-expressions using the pre-cropped short videos. For these cases, the locations of the onset and offset frames are required. These annotations can be obtained from the ground-truth, which are manually marked and verified by psychologists or "coders". Nonetheless, precision of ground-truth labeling is highly dependent on the judgment of the psychologists, who decide on the onset and offset locations by frame-by-frame observation [7, 8].

Apex frame contains vital information of a facial micro-expression as it is the best representative frame instant in the whole video sequence. A few recent works have been proposed for automatic apex frame spotting on the CASME II dataset, but all were tested on short videos with the luxury of onset and offset annotations. Yan et al. [9] demonstrated a pilot experiment by searching for the apex frame using two different feature extractors (i.e., Constraint Local Model (CLM) [10] and Local Binary Pattern (LBP) [11]) in 50 short video samples. The frame with the highest feature difference among the image frames is denoted as the apex. However, the CLM feature performed poorly as it was not able to annotate landmark points to a good degree of accuracy. Later on, Liong et al. [12] enhanced the work of [9] by employing Optical Strain (OS), a flow-based

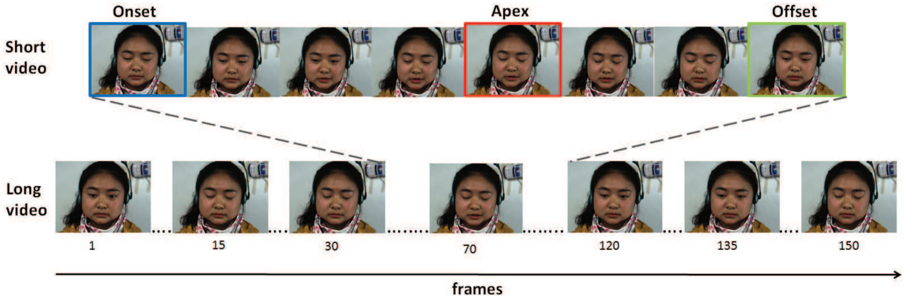


Fig. 1. An example of a long and short video with annotated ground-truth labels indicating the onset, apex and offset frames.

feature which was motivated by the work of Shreve et al. [3]. The authors also discovered that instead of considering the whole face for feature representation, the features from three regions-of-interest (i.e., "left eye + left eyebrow", "right eye + right eyebrow" and "mouth") provided more salient features with respect to micro-expressions. On a separate direction, several other works [3] attempted to detect a micro-expression sequence from a long video by detecting the frames that make up the sequence. "Peak frames" that show a hike in intensity indicate likely onset and offset frames. However, the success of these methods depends greatly on the choice of threshold parameters used to determine peak frames.

A recent work by Liong et al. [13] proved that the utilization of information from the apex frame alone is sufficient for micro-expression recognition. They validated their method on short videos from the CASME II and SMIC micro-expression databases. Features were obtained from the apex frame and the first frame of the short video (or onset frame) using the Bi-Weighted Oriented Optical Flow (Bi-WOOF) [13] feature extractor. However, the authors followed the assumption that the onset and offset frames are already annotated, and hence, constraints the finding of the apex to that particular range. This is unrealistic considering that long videos may contain many irrelevant motions that can be falsely spotted as micro-expressions.

As far as we are aware, there is only a single attempt [14] to work on long micro-expression videos in the literature, to realize a seamless automatic recognition system. They utilized two kinds of features, LBP and Histogram of Oriented Optical Flow (HOOF) [15], to characterize the frames in the sequence. A chi-square dissimilarity metric is used to compute the feature difference between each frame and a reference frame. Then, all spotted frames are thresholded and cross-checked against a pre-defined frame interval to determine the spotting accuracy. The spotting threshold was chosen (at true positive rate of 74.86%) to obtain the spotted micro-expression sequences which are fed to the recognition component. Although evaluation on the SMIC-E-VIS showed promising intent of such a scheme, the reliance on the annotated onset and offset frames, and use of a tunable threshold parameter warrants the need for manual intervention.

Evaluation of the micro-expression system on long videos is particularly challenging, primarily because of the presence of unwanted facial movements. These motions correspond to falsely detected micro-expressions, which may appear before the actual onset frame and after the offset frame. One common irrelevant facial movement that is unavoidable during the elicitation of micro-expression database is the eye blinking motion. Shreve et al. [3] suggested to remove the eye regions because eye blinking can adversely affect optical flow estimation, causing false detection of the micro- and macro-expressions. In their work, the boundaries of the eye regions were automatically marked using a landmark annotator, unlike the work of [16] which was done manually.

In this paper, we present a novel approach that can automatically recognize the type of facial micro-expression given a long video without ground-truth annotations of the onset, apex and offset frames. A complete micro-expression system which combines both apex frame spotting and micro-expression recognition components that is capable of operating on long videos is introduced. In the apex frame spotting component, pre-processing is first performed to automatically mask the eye regions to prevent ambiguous eye behaviors (i.e., eye blinking). After which, optical strain magnitudes are computed and sum-aggregated for twelve facial blocks, and a max operator pinpoints the apex frame. In the recognition component, we employ the Bi-Weighted Oriented Optical Flow (Bi-WOOF) feature using only the spotted apex frame and a neutral reference frame (we take the first frame as the most neutral expression) to describe the video sequence. We validate the reliability of the system and the effectiveness of the proposed methods in four spontaneous micro-expression databases: CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR. To the best of our knowledge, these are the only databases which contained long videos.

The structure of the paper is organized as follows: Section 2 explains the algorithms of the proposed micro-expression system in detail, Section 3 describes the databases, performance metrics and settings used in the experiments, Section 4 presents the results for both apex frame spotting and micro-expression recognition, with further analysis. Finally, Section 5 concludes the paper.

2 Proposed Approach

The micro-expression system proposed includes two main stages: apex frame spotting and micro-expression recognition. Firstly, the apex frame in a long video sequence is identified by applying Optical Strain feature extractor after eye masking and regions of interest selection techniques. The spotted apex frame is then fed into the micro-expression recognition stage, which is made up of a feature extractor and a classifier. The framework of the proposed algorithm is illustrated in Fig. 2, with detail of each stage elaborated as follows.

2.1 Apex Frame Spotting

In the apex frame spotting task, some of the frames in the long videos might contain irrelevant micro-expression movements, such as eye blinking action, which

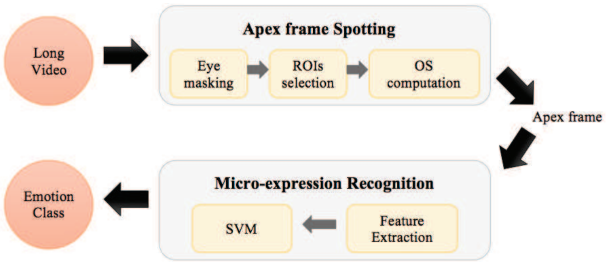


Fig. 2. Flow diagram of the proposed micro-expression system for long videos.

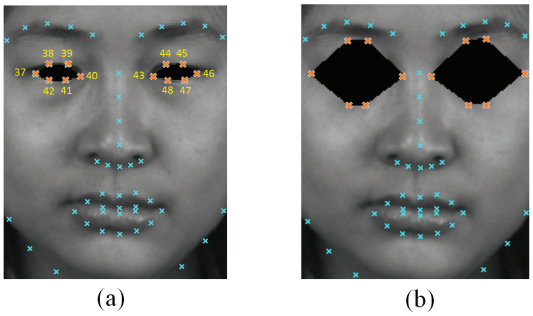
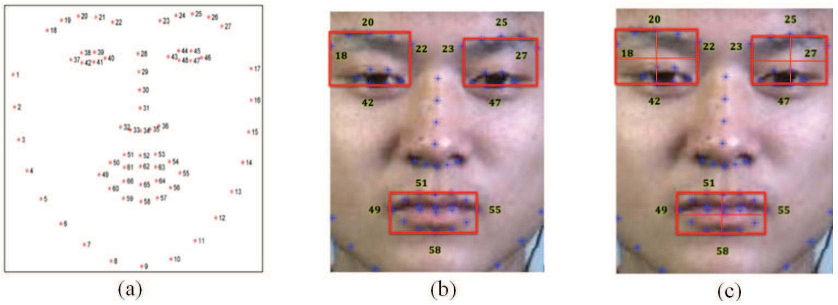


Fig. 3. The eye masking process: (a) There are six landmark coordinates which marked the boundaries of the left (landmark points 37, 38, 39, 40, 41 and 42) and right (landmark points 43, 44, 45, 46, 47 and 48) eye regions; (b) The eye regions are removed after adding some pixel margins.

can possibly lead to erroneous apex frame spotting; and further to that, misclassification of micro-expressions. Therefore, we present a novel eye masking approach to address the eye blinking issue. Besides that, instead of using the whole face for feature extraction, we select a number of facial regions that contain meaningful micro-expression details, particularly at the eye and mouth regions [9, 12]. The optical strain magnitudes are then computed for each region of interest (ROI); the frame with the highest sum of optical strain magnitudes (from any region) is chosen as the apex. We note that the eye masking and ROIs selection steps are fully automated and completely rely on the facial locations of the first frame of each video, that are marked by the landmark detector.

Eye Masking Eye blinking is a natural motion of rapid opening and closing of the eyelids, and cannot be considered a micro-expression. Since the micro-expression databases are typically recorded at a high frame rate, the blinking action is clearly visible when displaying the video frame-by-frame; hence, it is significantly more intense compared to micro-expressions. Thus, it is a nagging issue that exists in some of the long video sequences. We overcome this issue by masking the left and right eye parts to reduce the false spotting of the apex frame.



(a) (b) (c)

Fig. 4. Illustration of extraction of the three RoIs: (a) Sixty six landmark coordinates labeled by DRMF; (b) The four edges (i.e., top, bottom, left and right) are determined based on the landmark point locations; (c) Each ROI is partitioned into four blocks with the same size.

To ensure this is done automatically, the eye regions are removed based on the location of landmark points annotated by a robust landmark detector, Discriminative Response Map Fitting (DRMF) [17]. DRMF has shown to outperform state-of-the-art landmark detection methods [18, 19], with lower computational time and real-time capabilities. The process of eliminating the eye regions is shown in Fig. 3. Landmark coordinates 37 to 42 indicate the boundary of the right eye region, while landmark coordinates 43 to 48 are the boundary points of the left eye region. To overcome potential inaccurate landmark annotation, a fifteen pixel margin is added to expand the eye boundaries.

ROI Selection Subsequently, the choice of region to perform accurate spotting is crucial. In [12], features from two main facial regions that contribute important micro-expression information, i.e. “eye and eyebrow” and “mouth” regions, were considered rather than the whole face region. The cropping of these region-of-interests (ROIs) are done in a completely automatic way: (1) using the facial landmark points annotated earlier, the three ROIs are identified using rectangular bounding boxes determined based on the landmark locations; (2) the ROI bounding boxes are widened by a margin of ten pixels on all four edges to compensate for potentially imprecise landmark annotation; (3) each ROI is equally divided into four blocks to encode more local appearance features. Thus, there is a total of twelve facial region blocks in a frame. Fig. 4 illustrates the steps involved in the ROIs selection.

Optical Strain Computation Shreve et al. [3] employed optical strain magnitudes for macro- and micro-expression spotting. Inspired by this work, we adapted the idea to better characterize micro motions, by obtaining optical strain magnitudes based on a reference frame. Optical strain is the extension of optical flow, and is more effective than the latter in identifying the subtle deformable facial muscle within a time interval [20]. In this work, we use the TV-L1 optical flow method [21], which is able to preserve flow discontinuities and arguably more robust compared to the classic Black and Anandan [22] optical flow method employed in [3].

Given a micro-expression video, $s_i = \{f_{i,j} | i = 1, \dots, S; j = 1, \dots, F_i\}$ of length F_i, we first compute the optical flow vector $\boldsymbol{p} = [p = \frac{u}{\Delta t}, q = \frac{v}{\Delta t}]$ between each frame in the video sequence (except the first frame) and the reference frame (first frame chosen as it contains the most neutral expression). Optical strain can be described by a two-dimensional displacement vector, $\boldsymbol{u} = [u, v]$ and its magnitude for each pixel can be calculated by taking the sum of squares of the normal and shear strain components, expressed as follows:

$$\begin{aligned} |\varepsilon| &= \sqrt{{\varepsilon_{xx}}^2 + {\varepsilon_{yy}}^2 + {\varepsilon_{xy}}^2 + {\varepsilon_{yx}}^2} \\ &= \sqrt{\frac{\partial u}{\partial x}^2 + \frac{\partial v}{\partial y}^2 + \frac{1}{2}(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial x})^2} \end{aligned} \tag{1}$$

A more detailed discourse on optical strain can be found in [3, 20].

As mentioned earlier, there are twelve facial blocks in each frame. The optical strain magnitudes are calculated for each of these regions after applying eye masking. The optical strain magnitudes in each block b are summed up and the frame with the highest block value (or sum of magnitudes) is designated as the spotted apex frame f^* for the sequence:

$$f^* = \arg\max_j \left\{ \sum_{j,b} |\varepsilon_{j,b}| \right\} \qquad \text{for } j = [1, F_i - 1], b = [1, 12] \tag{2}$$

2.2 Micro-expression Recognition

To describe the features of the spotted apex frame, we employ the Bi-Weighted Oriented Optical Flow (Bi-WOOF) feature extractor in [13]. This is the only work that makes use of information from a single apex frame for representing a micro-expression video. For each video sequence, the orientation θ and magnitude ρ of the flow vector $\boldsymbol{p} = [p, q]$ computed from the spotted apex frame and a neutral reference frame (we choose the first frame) are calculated as follows:

$$\theta_{x,y} = tan^{-1} \frac{q_{x,y}}{p_{x,y}} \tag{3}$$

$$\rho_{x,y} = \sqrt{{p_{x,y}}^2 + {q_{x,y}}^2} \tag{4}$$

We utilize the first frame of each video instead of the onset frame used in [13] since we do not rely on any ground-truth annotations when processing long video sequences. Then, the orientation values are locally and globally weighted by the magnitude and optical strain values respectively to form the Bi-WOOF features. The flow magnitudes are used to weight the flow orientations at a finer bin level (local). The spatially-pooled optical strain magnitudes provide coarser (global) weighting to $N \times N$ equally-partitioned blocks in the frame.

3 Experiment

To assess the performance of the proposed approach to micro-expression spotting and recognition, experiments were performed on four publicly accessible spontaneous micro-expression datasets: CASME II-RAW [7], SMIC-E-HS [8], SMIC-E-VIS [8] and SMIC-E-NIR [8]. Note that all these datasets have been recorded under well-controlled laboratory conditions because an unconstrained environment poses a great challenge to the elicitation of emotions and further machine processing. For standardized experimentation, we first align the faces with DRMF [17], detect them using a standard face detector [23], and then resize the cropped faces to 170 × 140 pixels.

3.1 Datasets

CASME II-RAW This database comprises of 246 micro-expressions from 26 subjects with a mean age of 22.03 years. The videos were collected using Point Grey GRAS-03K2C camera at a resolution of 640 × 480 pixels and a frame rate of $200fps$. The average frame length is 244 frames ($\sim$1.22s), with the longest being 1,024 frames ($\sim$5.12s) and the shortest being 51 frames ($\sim$0.26s). There are five main categories of micro-expressions, in the following distribution: 25 surprise videos, 27 repression videos, 32 happiness videos, 63 disgust videos and 99 other videos. The micro-expressions are elicited from the subjects by showing them some video clips and asking them to keep a poker face when watching the videos. The emotion types are labeled based on the action units decided by two coders, participants' self report and the content of the video shown. A reliability score of the action unit labeling is reported at 0.846. The ground-truths provided include the onset, apex, offset frame indices and the emotion class.

SMIC-E-HS It consists of 157 micro-expression clips from 16 subjects (mean age of 28.1 years). The videos were recorded using PixeLINK PL-B774U camera with a temporal resolution of $100fps$ and spatial resolution of 640 × 480 pixels. The average frame length is 590 frames ($\sim$5.9s); the longest is 1200 frames ($\sim$12s) while the shortest is 120 frames ($\sim$1.2s). There are three micro-expression classes: negative (66 videos), positive (51 videos) and surprise (40 videos). The micro-expression categories are determined by two coders based on the participants' self-report data. The ground-truths provided are onset, offset frame indices, and the emotion label. No apex frame indices are given.

SMIC-E-VIS This dataset is made up of 71 micro-expression videos from 8 subjects. The videos were recorded using a standard visual camera with a frame rate of $25fps$ at 640 × 480 pixels. The video clips have an average length of 150 frames ($\sim$6s); the longest is 300 frames ($\sim$12s) and the shortest is 30 frames ($\sim$1.2s). It consists of three micro-expression classes: negative (24 videos), positive (28 videos) and surprise (19 videos). The procedure and the ground-truth labels provided are the same as that in SMIC-E-HS.

SMIC-E-NIR It is a collection of 71 micro-expression video sequence obtained from 8 subjects. The videos were recorded with a near-infrared camera at a resolution of 640 × 480 pixels at $25fps$. The average frame length is 150 frames ($\sim 6s$); the longest is 300 frames ($\sim 12s$) and the shortest is 30 frames ($\sim 1.2s$). The micro-expression videos are categorized into three classes: negative (23 videos), positive (28 videos) and surprise (20 videos). The process of the ground-truth acquisition is the same as that in SMIC-E-HS.

3.2 Performance Metrics

Spotting Task The effectiveness of apex frame spotting can be determined using the Mean Absolute Error (MAE), which was also used in [9, 12]. MAE indicates the average frame distance between the ground-truth and the spotted apex frame, and it can be computed by the following equation:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |e_i| \tag{5}$$

where N is the total number of video sequence in the database and e is the distance (in frames) between the ground-truth apex and the spotted apex. However, among the four databases used in the experiments, only CASME II-RAW provided ground-truth apex frame indices. Thus, we propose to evaluate the performance of apex frame spotting using another measurement, Apex Spotting Rate (ASR), which calculates the success rate in spotting apex frames within the onset and offset range given a long video. An apex frame is scored 1 if it is located between the onset and offset frames, and 0 otherwise:

$$\begin{gathered} \text{ASR} = \frac{1}{N}\sum_{i=1}^{N} \delta \\ \text{where } \delta = \begin{cases} 1, & \text{if } f^* \in (f_{i,onset}, f_{i,offset}) \\ 0, & \text{otherwise} \end{cases} \end{gathered} \tag{6}$$

Recognition Task The classifier adopted in all experiments reported in this paper is the Support Vector Machine (SVM) with linear kernel. To ensure subject independence in the classification process, a leave-one-subject-out (LOSO) cross validation is employed. In LOSO, for each k-fold (where k is the total number of the subjects), the video samples of one subject are held out as testing set, while the remaining video samples form the training set. Following the experiment settings in [13], the block size for the Bi-WOOF feature extractor is set to 8 × 8 for CASME II-RAW, and 5 × 5 for the SMIC databases. Since the video samples in different micro-expression classes are distributed unequally [24], we measure the recognition accuracy using the F1-score which conveys the balance by averaging the precision (exactness) and recall (completeness):

$$\text{F1-score} := 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Table 1. Performance of apex frame spotting with and without eye masking on CASME II-RAW measured by MAE. Average frame number per video is 244.

Feature Extractor	W/o eye mask	With eye masked	Improvement
LBP	51.86 frames	55.26 frames	-6.56%
OS	42.77 frames	**27.21 frames**	**36.38%**

$$\text{Precision} := \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} := \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

where TP, FN and FP are the true positive, false negative and false positive.

4 Results and Discussion

4.1 Results

Performance of Apex Frame Spotting The MAE result for apex frame spotting task on the CASME II-RAW dataset shown in Table 1 compares the technique with and without applying eye masking on two types of feature extractors - LBP and OS. The LBP feature was utilized in [14] to spot the micro-expression frames while we propose the use of OS in this work. The lower the MAE (in frames), the closer the spotted apex frame is to the ground-truth apex frame, implying more accurate spotting. The spotting performance for OS outperforms the LBP. This result also emphasizes the importance of using eye masking (a more detailed look into the impact of this step can be found in Fig. 5). Eye masking improves the spotting accuracy with OS features by 36.38%.

On the other hand, Table 2 shows the apex spotting accuracy measured in terms of ASR. With eye masking, we observe important improvements of 20%, 41.68%, 20.02% and 31.58% on the CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR databases respectively. From these results, we show that the elimination of eye regions (but not up to the extent of eyebrows) from consideration is able to increase the precision of searching for the apex frame. It is worth mentioning that the overall performance on the SMIC databases is still quite low (even with eye masking). We discuss this in detail in Section 4.2.

Performance of Micro-expression Recognition At this point of time, there are no papers in literature that reported the F1-score recognition performance on micro-expression long videos (i.e., CASME II-RAW, SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR). For reference purpose, we provide the state-of-the-art methods that worked on the short videos (i.e., CASME II, SMIC-HS, SMIC-VIS and SMIC-NIR), shown in Table 3. Whereas, the results for long videos are

Table 2. Performance of apex frame spotting with and without eye masking on the CASME and SMIC databases measured by ASR.

Databases	W/o eye mask	With eye masked	Improvement
CASME II-RAW	0.6584	**0.8230**	**20.00%**
SMIC-E-HS	0.2229	**0.3822**	**41.68%**
SMIC-E-VIS	0.2253	**0.2817**	**20.02%**
SMIC-E-NIR	0.1831	**0.2676**	**31.58%**

Table 3. State-of-the-art recognition performance using F1-score in short videos databases.

#	Methods	CASME II	SMIC-HS	SMIC-VIS	SMIC-NIR
1	Baselines [7, 8]	.39	.39	.39	.40
2	Le et al. [24]	.33	.47	-	-
3	Le et al. [25]	.51	-	-	-
4	Le et al. [26]	.51	.60	-	-
5	Wang et al. [27]	.40	.55	-	-
6	Liong et al. [28]	-	.45	-	-
7	Liong et al. [20]	.38	.54	-	-
8	Oh et al. [29]	.43	.35	-	-
9	Huang et al. [30]	.57	.58	-	-
10	Xu et al. [31]	.30	.54	**.60**	**.60**
11	Liong et al. [13]	**.61**	**.62**	.58	.58

tabulated in Table 4. For the spotting task in Table 4, our methods employed the OS feature. Method #1 randomly spots the apex frame in the video sequence; method #2 spots the apex frame without masking the eye regions; method #3 spots the apex frame after applying eye masking. We can see that our best approach (#3) generates the best performance in all the long video databases.

We also compared the proposed method (method #3, with eye masking) with the work of Li et al. [14], which is the only other work that implemented a micro-expression spotting and recognition system for long videos (Table 5). This could only be done using the Accuracy measure, and only on one database, SMIC-E-VIS. It is observed that our proposed method is comparable to that of Li et al. [14] but with several advantages. Our method does not rely on the ground-truth onset and offset labels, and has the computational benefit of only needing to find the apex frame for recognition purpose.

4.2 Discussion

Although we evaluated the proposed method (#3) in long videos, we obtained a superior result on CASME II-RAW in Table 4, compared to all other methods (#1 to #10) in Table 3 that were conducted on CASME II (short videos) despite not having onset, offset and apex frame labels. The performance of our

Table 4. Comparison of recognition performance using F1-score in long videos databases.

#	Methods	CASME II-RAW	SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR
1	Spotting (random) + recognition	.36	.37	.33	.28
2	Spotting (w/o eye mask) + recognition	.46	.36	.44	.38
3	Spotting (with eye mask) + recognition	**.59**	**.47**	**.53**	**.43**

Table 5. Comparison of recognition accuracy (%) and methodology between the state-of-the-art method and the proposed method on the SMIC-E-VIS database.

Methods	Recognition (%)	Remarks
Li et al. [14]	56.67	Onset and offset frames used. Only correctly spotted sequences used.
Proposed method	**53.52**	No onset, offset, apex labels required. All spotted apices are used.

proposed method on SMIC-E-HS, SMIC-E-VIS and SMIC-E-NIR databases are slightly mixed, compared to the results of SMIC-HS, SMIC-VIS and SMIC-NIR databases respectively, in Table 3. This may be due to the fact that most of the frames in these databases do not contain micro-expression-like motions. We show in Table 6 the average percentage of frames of the long videos that consists of micro-expressions. We note that only approximately 6% of the frames consists of micro-expressions in the three SMIC databases. In other words, 94% of the frames contain either neutral faces, macro-expressions or other forms of irrelevant motions such as head rotations and eyeball movements. In addition, SMIC has a much lower frame rate i.e. $100fps$, $25fps$ and $25fps$ in HS, VIS and NIR datasets respectively, compared to CASME II ($200fps$). This points to the possibility of macro-expression and irrelevant movements become more prominent while micro-expressions may occur only in a few frames. Hence, attempting to spot the apex frame in these circumstances is an arduous task. In future, better techniques can be designed to differentiate between these different states of emotions (macro and micro).

For method #1 in Table 4, the apex frame is spotted randomly, as a control method. As expected, the recognition performance is the poorest among all evaluated methods in all databases. This indicates the importance of obtaining the apex frame correctly. Both the spotting and recognition results (in Table 2 and Table 4) prove that the eye masking technique enhances the micro-expression system by removing noises from eye blinking, resulting in more meaningful features. Fig. 5 demonstrates the difference in the selection of spotted apex, with and without applying eye masking. Without the eye masking technique (the

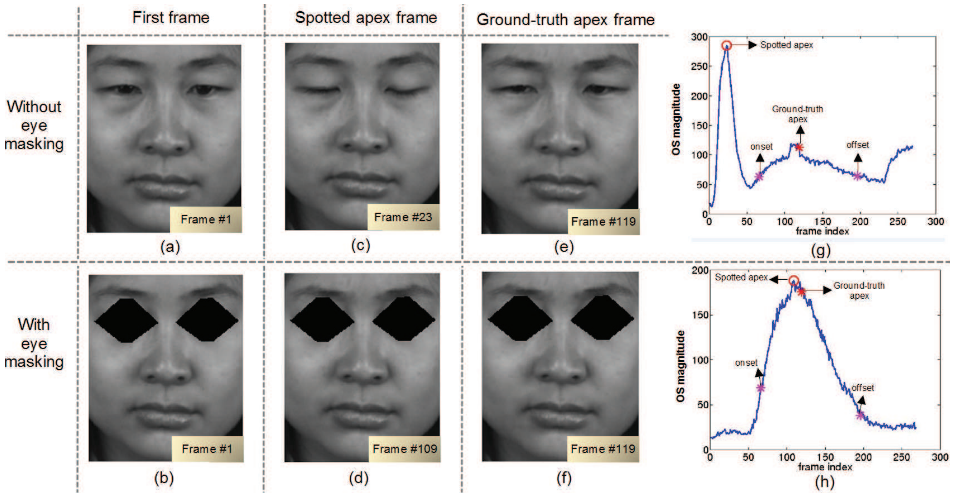


Fig. 5. Top row without eye masking, bottom row with eye masking: (a-b) The first frame in the video; (c-d) The spotted apex frame; (e-f) The ground-truth apex frame; (g-h) Plots of optical strain magnitudes across the video sequence. Relevant labeled frames are marked.

upper row in Fig. 5), the detected apex frame (frame 23) contains an eye closing motion, a falsely spotted micro-expression. This occurred because the facial movement is relatively more intense among all the frames in the video. On the contrary, the spotted apex frame with eye masking (frame 109) is much closer to the ground-truth apex frame (frame 119).

In Table 5, Li et al. [14] tested their full micro-expression system on the SMIC-E-VIS. Although their reported recognition performance is slightly better than that of our method, there are several glaring differences. Firstly, they utilized the ground-truth onset and offset frame labels to form a frame interval which was used to determine the spotted micro-expression sequence. Secondly, the incorrectly spotted micro-expression sequences are not considered for recognition. The authors pointed out that the reported performance was achieved using only the correctly spotted ME sequences (TPR=74.86%) [14]. Hence, we believe that our proposed approach eliminates the need for human intervention; it does not make use of any hand-labeled ground-truth frames (i.e., onset, apex and offset). It also mimics a fully automatic and realistic system which considers the likelihood of a less-than-desirable spotted apex. For a closer inspection into how each class performed, we provide confusion matrices of the recognition task for CASME-II-RAW and SMIC-E-HS databases, tabulated in Table 7 and 8.

5 Conclusion

A novel fully automatic micro-expression recognition system, which combines both apex frame spotting and micro-expression recognition, is proposed in this

Table 6. Average number of frames in the short and long videos of the CASME II and three SMIC databases.

Databases	Short Video	Long Video	Frames with micro-expression
CASME II	67 frames	244 frames	∼27%
SMIC-HS	33 frames	590 frames	∼6%
SMIC-VIS	9 frames	150 frames	∼6%
SMIC-NIR	9 frames	150 frames	∼6%

Table 7. Confusion matrices for recognition task on the CASME-II-RAW database using our proposed method, measured by accuracy rate (%).

	disgust	happiness	tense	surprise	repression
disgust	**42.62**	8.20	44.26	1.64	3.28
happiness	6.25	**56.25**	25.00	6.25	6.25
others	23.47	9.18	**61.22**	1.02	5.10
surprise	12.00	4.00	16.00	**68.00**	0
repression	0	7.41	33.33	0	**59.26**

Table 8. Confusion matrices for recognition task on the SMIC-E-HS database using our proposed method, measured by accuracy rate (%).

	negative	positive	surprise
negative	**53.85**	32.31	13.85
positive	39.22	**49.02**	11.76
surprise	53.66	12.20	**34.15**

paper. We present the first known work that classifies the emotion class given a long micro-expression video without the knowledge of ground-truth onset and offset frames. In the spotting task, a major problem which exists in long videos is the presence of eye blinking motion, which can easily be mistaken as a micro-expression. We remove the eye regions by applying an automatic eye masking technique which depends entirely on detected landmark coordinates. Optical strain magnitudes of all frames are computed with the aid of eye masking and region selection in order to determine the apex frame. In the recognition task, we apply the recent Bi-WOOF feature extractor to capture discriminative features from each video, using only the spotted apex frame and a neutral reference frame (the first frame). Using a SVM classifier with linear kernel, we evaluated the proposed approaches on four spontaneous micro-expression databases that contain long videos. The proposed recognition approach outperforms all existing state-of-the-art methods on the CASME II database, achieving a promising F1-score of 59%.

References

1. Porter, S., ten Brinke, L.: Reading between the lies identifying concealed and falsified emotions in universal facial expressions. Psychological Science **19(5)** (2008) 508–514
2. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Journal for the Study of Interpersonal Processes **32** (1969) 88–106
3. Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D., Sarkar, S.: Towards macro- and micro-expression spotting in video using strain patterns. In: Applications of Computer Vision (WACV). (2009) 1–6
4. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology **17** (1971) 124
5. Ekman, P.: Lie catching and microexpressions. The philosophy of deception (2009) 118–133
6. Valstar, M.F., Pantic, M.: Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. HumanComputer Interaction (2007) 118–127
7. Yan, W.J., Wang, S.J., Zhao, G., Li, X., Liu, Y.J., Chen, Y.H., Fu, X.: CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLoS ONE **9** (2014) e86041
8. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikainen, M.: A spontaneous micro-expression database: Inducement, collection and baseline. In: Automatic Face and Gesture Recognition. (2013) 1–6
9. Yan, W.J., Wang, S.J., Chen, Y.H., Zhao, G., Fu, X.: Quantifying micro-expressions with constraint local model and local binary pattern. In: Computer Vision-ECCV workshop. (2014) 296–305
10. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition **41(10)** (2008) 3054–3067
11. Ojala, T., Pietikinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29(1)** (1996) 51–59
12. Liong, S.T., See, J., Wong, K., Le Ngo, A.C., Oh, Y.H., Phan, R.C.W.: Automatic apex frame spotting in micro-expression database. In: Asian Conference on Pattern Recognition (ACPR). (2015)
13. Liong, S.T., See, J., Wong, K., Phan, R.C.W.: Less is more: Micro-expression recognition from video using apex frame. arXiv (2016)
14. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikinen, M.: Reading hidden emotions: Spontaneous micro-expression spotting and recognition. arXiv preprint arXiv:1511.00423 (2015)
15. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European conference on computer vision. (2006) 428–441
16. Moilanen, A., Zhao, G., Pietikainen, M.: Spotting rapid facial movements from videos using appearance-based feature difference analysis. In: ICPR. (2014) 1722–1727
17. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Computer Vision and Pattern Recognition. (2013) 3444–3451
18. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision **91(2)** (2011) 200–215

19. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition. (2012) 2879–2886
20. Liong, S.T., See, J., Phan, R.C.W., Le Ngo, A.C., Oh, Y.H., Wong, K.: Subtle expression recognition using optical strain weighted features. In: Asian Conference on Computer Vision Workshops on Computer Vision for Affective Computing. (2014) 644–657
21. Prez, J.S., Meinhardt-Llopis, E., Facciolo, G.: Tv-l1 optical flow estimation. Image Processing On Line (2013) 137–150
22. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer vision and image understanding **63(1)** (1996) 75–104
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE CVPR. Volume 1. (2001) I–511
24. Le Ngo, A.C., Phan, R.C.W., See, J.: Spontaneous subtle expression recognition: Imbalanced databases & solutions. In: Asian Conference on Computer Vision. (2014) 33–48
25. Le Ngo, A.C., Liong, S.T., See, J., Phan, R.C.W.: Are subtle expressions too sparse to recognize? In: Digital Signal Processing (DSP). (2015) 1246–1250
26. Le Ngo, A.C., See, J., Phan, R.C.W.: Sparsity in dynamics of spontaneous subtle emotions: Analysis & application. Transactions on Affective Computing (2016)
27. Wang, Y., See, J., Phan, R.C.W., Oh, Y.H.: Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In: Computer Vision–ACCV, IEEE (2015) 525–537
28. Liong, S.T., Phan, R.C.W., See, J., Oh, Y.H., Wong, K.: Optical strain based recognition of subtle emotions. In: International Symposium on Intelligent Signal Processing and Communication Systems. (2014) 180–184
29. Oh, Y.H., Le Ngo, A.C., See, J., Liong, S.T., Phan, R.C.W., Ling, H.C.: Monogenic riesz wavelet representation for micro-expression recognition. In: Digital Signal Processing, IEEE (2015) 1237–1241
30. Huang, X., Wang, S.J., Zhao, G., Pietikainen, M.: Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In: International Conference on Computer Vision Workshops. (2015) 1–9
31. Xu, F., Zhang, J., Wang, J.: Microexpression identification and categorization using a facial dynamics map. Transactions on Affective Computing (2016)