Abstract

Micro-expression usually occurs at high-stakes situations and may provide useful information in the field of behavioral psychology for better interpretation and analysis. Unfortunately, it is technically challenging to detect and recognize micro-expressions due to its brief duration and the subtle facial distortions. Apex frame, which is the instant indicating the most expressive emotional state in a video, is effective to classify the emotion in that particular frame. In this work, we present a novel method to spot the apex frame of a spontaneous micro-expression video sequence. A binary search approach is employed to locate the index of the frame in which the peak facial changes occur. Features from specific facial regions are extracted to better represent and describe the expression details. The defined facial regions are selected based on the action unit and landmark coordinates of the subject, in which case these processes are automated. We consider three distinct feature descriptors to evaluate the reliability of the proposed approach. Improvements of at least 20% are achieved when compared to the baselines.

1. Introduction

Micro-expression is a very brief and quick facial emotion triggered unintentionally, and it reflects our true feelings. It has two main characteristics: micro (lesser duration) and subtle (lower intensity) [4]. Similar to macro-expression, micro-expression can be categorized into six universal emotion states, including happy, fear, sad, surprise, anger and disgust. Micro-expressions last between one-fifth to one-twenty-fifth of a second, and they usually occur in several parts on the face where most people do not realize. It is not straightforward to recognize the genuine emotion shown on one’s face [12]. Thus, recognizing micro-expressions is beneficial in our daily life as we can interpret if someone is trying to conceal his/her feeling or trying to lie to you.

The purpose of this spotting task is to understand the pattern of the changes on the facial movements for behavioral and psychological research. It is essential to label the frames automatically because manual coding may be imprecise labeling and different coders may lead to inconsistency. Yan et al. [17] spot the apex frame in the CASME II [18] micro-expression database. The apex occurs when the change in facial muscle reaches the peak or the highest intensity of the facial motion. They utilized two distinct feature extraction methods: Local Binary Pattern [11] and Constraint Local Models [3], to encode the facial texture. The reported average frame distance between the groundtruth and the proposed method was not absolute mean and the experiment conducted only considered 50 samples out of 247 samples in CASME II. To the best of our knowledge, this is the only prior work for spotting apex frame performed on micro-expression database and we establish the baseline by duplicating their experiments.

This paper proposes a novel approach on automatic apex frame spotting. We employ a binary search strategy to search for the apex frame. Only features within certain facial regions are extracted to describe the texture changes. The rest of this paper is organized as follows: Section 2 gives a brief review of related literature, followed by Section 3 that describes the proposed method in detail. Next, Section 4 summarizes the experimental results while conclusion is drawn in Section 5.

2. Related Work

2.1. Landmark Detection and Facial Sub-region Extraction

FACS [5] expresses the movements of human facial muscles from the appearance. FACS can describe subtle and
ambiguous expressions, thus it is more adequate for describing the tiny instant changes on the face. Furthermore, an AU represents the facial muscle movements in a specific facial region in single direction. Note that AUs can occur either singly or in combination. [16] proved that there were improvements on the facial expression recognition rate with prior knowledge of AU information.

Many research papers demonstrated that extracting the features from certain facial regions improves the performance on facial expression recognition when compared to considering the entire face. The main reason is that those regions-of-interest (RoIs) contributes more facial changes information towards differentiation of the expressions and eliminates the parts that do not correspond to the desired facial movements. Zhong et al. [19] proposed a two-stage multi-task sparse learning framework to discover the common and specific patches based on the location of the facial motions and their AUs. Happy and Routray [7] introduced an automated salient facial patches selection method, in which the sub-regions selected depend upon the locations of facial landmarks detected using DRMF. However, there is no commonly agreed standard for specific combination of the facial patches on achieving better accuracy on facial expression analysis.

To extract the features of the RoIs at particular locations, it is essential to register and track the landmark coordinates as the pre-processing stage. CLM is proven to be more robust, effective and precise in terms of tracking performance when compared to the existing landmark detector [2]. Asthana et al. [1] proposed a facial landmark detector, namely, Discriminative Response Map Fitting (DRMF) that outperformed the state-of-the-art techniques [14].

3. Proposed Algorithm

The flowchart of the proposed approach is illustrated in Figure 1. The RoIs are extracted according to the landmark coordinates. Then a peak detector is used to search for the apex frame based on the features extracted using three different types of feature descriptors, namely, CLM, LBP and OS. The detail of each step is elaborated in the following subsections.

3.1. Landmark Detection and RoI extraction

Yan et al. [17] reported that the two most expressive facial parts are located in the eyebrow and mouth areas. On the other hand, Ringeval et al. [13] analyzed the facial features by splitting the landmarks into three groups, namely, “left eye + eyebrow”, “right eye + eyebrow” and “mouth”. Table 3.1 shows the statistics of the occurrence of the face regions based on the AUs on CASME II database. We discovered that among all the facial regions in the video samples, “eye+eyebrow” and “mouth” areas are most expressive as they appeared the most. In other words, these regions contribute the majority and meaningful micro-expression information and hence we treat these regions as our RoIs. The advantages to extract the features from the RoIs rather than the whole face are: (1) removal of regions that are less relevant or irrelevant to the expressions; (2) reduction of execution time due to smaller input image.

The landmark detector, i.e., DRMF, is employed to detect the sixty six facial landmarks, shown in Fig. 2(a). Then, the bounding boxes of the RoIs are determined according to the neighbouring landmark points. All three RoIs (i.e., “left eye + left eyebrow”, “right eye + right eyebrow” and “mouth”) are bounded in multiple rectangular boxes. The size of the rectangular boxes depends on the four borders (i.e., top, bottom, left and right) of the corresponding landmark points, as illustrated in Fig. 2(b). We added ten pixels margin in all four directions in the boxes to encode more local expression details and to overcome the imprecise landmark annotation problem.

3.2. Binary Search in Different Feature Extractors

Despite spotting the apex frame by determining the maximum peak (the conventional method) [17] in the video sequence, a binary search methodology is introduced to auto-
<table>
<thead>
<tr>
<th>Face regions</th>
<th>AU(s) in the region</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye + eyebrow</td>
<td>1, 2, 4, 7</td>
<td>200</td>
</tr>
<tr>
<td>Mouth</td>
<td>10, 12, 14, 15, 25</td>
<td>94</td>
</tr>
<tr>
<td>Chin</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>Nose</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>Cheek</td>
<td>6</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 1. Frequency of the face regions based on the action units.

Figure 2. Cropping out three RoIs: (a) 66 landmark points annotated by DRMF; (b) The rectangular boxes are set based on the coordinate of 12 landmark points of the four borders.

Figure 3(a) shows the density plot for the changes of each landmark coordinate along the video stream. Landmark 1 to 17 (contour of the face) are eliminated from the analysis. It can be seen that landmark 56 (indicates mouth corner) has the highest coordinate changes. To further illustrate, Figure 3(b) shows the coordinate changes of landmark 56 with annotated ground-truth and spotted apex frame. It shows that using the proposed binary search method can better estimate the apex frame with smaller frame differences.

### 3.2.1 Constraint Local Model

To minimize the noise caused by the participants’ head movements, the coordinate of each facial landmark is subtracted by the nostril (which indicates the center of the face) coordinate. This step is essential especially in employing geometric feature descriptor because little face movements that are irrelevant to the expressions may result in inaccurate feature extraction and classification. The corresponding landmark coordinates in each frame are subtracted by the coordinates in the first frame (assumed to be the neutral face) to denote the change for each landmark. The changes in difference are compared among three RoIs, and only the features of RoI with the highest changes in difference is extracted for apex frame investigation. Figure 3(c) shows the rate of difference at mouth region along the video sequence. From the figure, it demonstrates that the binary search method outperforms the conventional method.

### 3.2.2 Local Binary Pattern

LBP histograms for each RoI in each frame are calculated. Then, the apex frame is obtained by computing the correlation between the first frame and the rest of the frames. The correlation coefficient is defined by:

\[
  d = \frac{\sum_{i=1}^{nBins} h_{1i} \times h_{2i}}{\sqrt{\sum_{i=1}^{nBins} h_{1i}^2} \times \sqrt{\sum_{i=1}^{nBins} h_{2i}^2}},
\]

where \( h_1 \) is the gray-scale histogram of the first frame, and \( h_2 \) is the current frame. Here, \( 1-d \) indicates the rate of difference of the LBP features between two frames. The changes in difference are compared among three RoIs, and only the features of RoI with the highest changes in difference is extracted for apex frame investigation. Figure 3(c) shows the rate of difference at mouth region along the video sequence. From the figure, it demonstrates that the binary search method outperforms the conventional method.

### 3.2.3 Optical Strain

The optical strain magnitudes for each image pixel is first computed. Next, spatial pooling is performed by sum-

---

Algorithm 1 Binary Search

1. \( l \leftarrow \) split level
2. \( S \leftarrow \) set of candidate peaks, \( p_i \)
3. Initialize \( l = 0, S_{c} \epsilon p_i \)
4. **repeat**
   - Split half \( S_{c} \) to \( S_0, S_1 \)
   - \( S_{c} \leftarrow \max(|S_0, S_1|) \)
   - \( l \leftarrow l + 1 \)
5. **until** \( S_i = 1 \)
ming up all the strain magnitudes within each RoI in each frame to form a resultant strain value. Similar to the LBP approach, only the RoI with the highest strain magnitude among three RoIs will be considered for apex frame examination. Figure 3(d) illustrates the spotted apex frame using the optical strain method. The spotted frame using binary search strategy is very close to the ground-truth frame, which is only 1 frame distance.

4. Experiment

4.1. Dataset

To the best of our knowledge, micro-expression database is relatively few in the literature. The only database that is suitable for micro-expression apex frame spotting validation is CASME II [18] because each video sample contains the ground-truth apex frame that was manually annotated by two psychologists. The CASME II dataset contains 247 micro-expressions video samples elicited from 26 participants. Each clip only consists of one type of emotion. The videos were recorded using a Point Grey GRAS-03K2C camera with a high temporal resolution of 200 fps and a spatial resolution of $280 \times 340$ pixels. The ground-truths also include the labeling of emotion state (i.e., happiness, disgust, surprise, repression and tense), AUs and apex frame.

4.2. Performance Metric

There are two performance measurements to evaluate the method proposed, namely, mean absolute error (MAE) and Standard Error (SE), which can be written generically as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$  \hspace{1cm} (2)

and

$$\text{SE} = \frac{\sigma}{\sqrt{n}},$$  \hspace{1cm} (3)

where $n$ is the sample size, $e$ is the distance between the spotted apex frame and the ground-truth, $\sigma$ is the sample standard deviation. MAE is to calculate how close is the average estimated apex frames to the ground-truths, whereas SE is the standard deviation of the sample mean distribution.

4.3. Results and Discussion

We reported three methods to spot the apex frame, which are: (1) Baseline - the conventional method [17] that treats the maximum magnitude changes of the frame as the apex; (2) BS-whole face - the binary search strategy by extracting the features of the entire face image, and; (3) BS-RoIs - restricting the binary search strategy to the features found within three RoIs. Table 4.3 shows the results of apex frame spotting using three different feature descriptors (i.e., CLM, LBP and OS). The proposed method (for both whole face and RoIs areas) produces positive and promising performance when compared to the baselines. Explicitly, the features extracted within the defined RoIs contain significant and valuable micro-expression information and do not consider the regions that do not contribute to subtle movements. The average of frame difference between the detected and ground-truth apex is reduced by using the proposed BS-RoIs strategy. When compared to their respective baselines in CLM, LBP and OS, the percentage of reductions are 21.56%, 23.66% and 23.97%. LBP and OS feature descriptors achieved better results compared to CLM. CLM is computed according to the displacement of the landmark points across the video sequence. In contrary, LBP and OS are calculated based on the texture in a region. Thus, by considering the local regional expression information (instead of simply the coordinates of the landmark), the micro-expressions can be well encoded and represented.

Furthermore, the data of the predicted apex frame was analysed by an analysis of variance (ANOVA) with repeated measure design. This is to test whether the proposed method is statistical significant from the baseline method. It is observed that all the evaluated feature descriptors reveal significant difference (at $p < 0.05$) when compared with the baseline and binary search strategy.
Table 2. Results of apex frame spotting using the baseline and binary search method adopted in three different feature descriptors

<table>
<thead>
<tr>
<th>Methods</th>
<th>Baseline [17]</th>
<th>BS-whole face</th>
<th>BS-RoIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>SE</td>
<td>MAE</td>
</tr>
<tr>
<td>CLM</td>
<td>21.94</td>
<td>1.00</td>
<td>17.21</td>
</tr>
<tr>
<td>LBP</td>
<td>17.75</td>
<td>0.90</td>
<td>15.54</td>
</tr>
<tr>
<td>OS</td>
<td>18.98</td>
<td>0.95</td>
<td>16.57</td>
</tr>
</tbody>
</table>

Table 3. ANOVA repeated measure results for the baseline and binary search method adopted in three different feature descriptors to test for the significant difference (p < 0.05)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Whole face</th>
<th>RoIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-value</td>
<td>p-value</td>
</tr>
<tr>
<td>CLM</td>
<td>20.69</td>
<td>0</td>
</tr>
<tr>
<td>LBP</td>
<td>8.31</td>
<td>0.0043</td>
</tr>
<tr>
<td>OS</td>
<td>6.50</td>
<td>0.0114</td>
</tr>
</tbody>
</table>

5. Conclusion

In apex frame spotting of micro-expressions video, it is essential to label the frames automatically because manual coding may be imprecise labeling and different coders lead to inconsistency. This paper proposed a novel method to spot the apex frame by using the binary search strategy and restricting the region of interest to facial sub-region. The selection of the RoIs were based on the landmark coordinates of the face and it was completely automatic. Experiments on the CASME II micro-expression database demonstrated the effectiveness of the proposed method. Three distinct feature descriptors were adopted to further confirm the reliability of the method proposed. The overall result achieved an improvement of more than 20% when compared to the baseline. For future work, the spotted apex frames can be used for classifying the type of the micro-expressions to further validate the reliability of the proposed method.

6. Acknowledgment

This work was funded by Telekom Malaysia (TM) under the project 2beAware and by University Malaya Research Collaboration Grant (Title: Realistic Video-Based Assistive Living, Grant Number: CG009-2013) under the purview of University of Malaya Research.

References