OFF-ApexNet on micro-expression recognition system

Y.S. Gan [a], Sze-Teng Liong [b,*], Wei-Chuen Yau [c], Yen-Chang Huang [d], Lit-Ken Tan [e,f]

[a] *Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, ROC*
[b] *Department of Electronic Engineering, Feng Chia University, Taichung 40724, Taiwan, ROC*
[c] *School of Electrical and Computing Engineering, Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia*
[d] *School of Mathematics and Statistics, Xinyang Normal University, Henan, China*
[e] *Department of Computer Science, Faculty of Knowledge Engineering, Tokyo City University, 1-28-1, Tamazutsumi,Setagaya-ku, Tokyo, 158-8557, Japan*
[f] *Malaysia-Japan International Institute of Technology (MJIIT), University Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra (Jalan Semarak), 54100 Kuala Lumpur, Malaysia*

ARTICLE INFO

ABSTRACT

When a person attempts to conceal an emotion, the genuine emotion is manifest as a micro-expression. Exploration of automatic facial micro-expression recognition systems is relatively new in the computer vision domain. This is due to the difficulty in implementing optimal feature extraction methods to cope with the subtlety and brief motion characteristics of the expression. Most of the existing approaches extract the subtle facial movements based on hand-crafted features. In this paper, we address the micro-expression recognition task with a convolutional neural network (CNN) architecture, which well integrates the features extracted from each video. We introduce the Optical Flow Features from Apex frame Network (OFF-ApexNet). This is a new feature descriptor that combines the optical flow guided context with the CNN. Firstly, we obtain the location of the apex frame from each video sequence as it portrays the highest intensity of facial motion among all frames. Then, the optical flow information are attained from the apex frame and a reference frame (i.e., onset frame). Finally, the optical flow features are fed into a pre-designed CNN model for further feature enhancement as well as to carry out the expression classification. To evaluate the effectiveness of OFF-ApexNet method, comprehensive evaluations are conducted on three public spontaneous micro-expression datasets (i.e., SMIC, CASME II and SAMM). The promising recognition result suggests that the proposed method can optimally describe the significant micro-expression details. In particular, we report that, in a multi-database with leave-one-subject-out cross-validation (LOSOCV) experimental protocol, the recognition performance reaches 74.60% of recognition accuracy and F-measure of 71.04%. We also note that this is the first work that performs cross-dataset validation on three databases in this domain.

1. Introduction

Facial expression is one of the popular nonverbal communication types that plays an important role to reflect one's emotional state. Different combinations of facial muscular movement eventually represent specific type of emotions. According to the psychologists, people portray some particular emotions on the face in the same way, regardless the race or culture [1]. Furthermore, it was verified by [2] that there is no difference between the sighted and blind individuals on the configuration of the facial muscle movements to response to the emotional stimuli. In other words, facial expressions are universal. They can be commonly classified into six emotion classes: happiness, sadness, fear, anger, disgust and surprise.

Generally, facial expression is categorized into two types, namely, macro-expression and micro-expression. The formal expression typically lasts between three quarters of a second to two seconds, and the

muscle movements are possibly occurred simultaneously at multiple parts on the face. Therefore, macro-expressions are readily perceived by humans in real time conversations. Over the past few decades, the research in automated macro-expression recognition analysis has been an active topics. To date, plenty of the recognition systems developed achieved more than 95% of expression classification accuracy [3,4] and some of them even reached almost 100% perfect recognition performance [5–7]. However, it should be noted that macro-expression does not accurately implies one's emotion state as it can be easily faked. Hence, it is worth to investigate to deeper emotion states from the muscular movements.

Among several types of nonverbal communications, micro-expressions are discovered to be more likely to reveal one's true emotions. Micro-expressions often sustain within $0.04s$ to $0.2s$ [1] and

---

they may only present in a few small regions on the face. Besides, they are stimulated involuntary which means that people cannot control their appearance. Owing to its characteristic of potentially exposing a person's true emotions, it can be deployed in several applications such as national security, police interrogation, business negotiation, social interaction and clinical practice [8–12].

Thus far, the identification and annotation of micro-expressions are done manually by psychologists or trained experts. This may lead to reliability inconsistency as the labeling of the expression is solely dependent on the personal judgment. In addition, it is time and effort consuming as the annotators are required to inspect the tiny facial muscle changes in each frame transition. Therefore, it is essential to implement reliable computer-based micro-expression detection and classification systems to obtain trustable, accurate and precise ground-truths (i.e., emotion state, action unit, onset, apex and offset indices) of each video.

In general, a micro-expression recognition system involves three basic steps, include: (1) Image preprocessing — enhancement of image by preserving the significant features; (2) Feature extraction — identification of the important features from the image; (3) Expression classification — recognition of the emotion based on the features extracted. Fig. 1 illustrates the basic flowchart of the recognition process. Each step plays a vital role to obtain a promising recognition performance and they are all equivalently important because each of them is targeting unique strategies to address the desired features in different perspective. In the recent years, the automated micro-expression systems developed in the literature are increasing gradually. This might due to the lack of suitable databases for data training and testing purposes, and hence hindering further analysis study especially in performance assessment and investigation. To date, there are four spontaneous publicly-available micro-expressions databases (i.e., CASME [13], CASME II [14], SMIC [15] and SAMM [16]) that are popularly used for experimental evaluation. It should be noted that CASME II is an improved dataset version of CASME, which is established to address the inadequacies of poor illumination, low temporal and spatial resolution issues.

Recent works [17–19] have shown the feasibility of adopting deep learning (e.g., convolutional neural network (CNN)) in micro-expression recognition systems. The first work to adopt CNN architecture in the micro-expression recognition system is by [17]. They achieve 53.6% for SMIC in the SVM with leave-one-subject-out cross-validation (LOSOCV) protocol. However, the handcrafted method (i.e., STCLQP [20]) method can exhibit 64% recognition result. The main obstacles for deep feature learning approach are small scale of data and imbalance class distribution. To increase the data size to achieve better generalization in the neural network, [18] extends perform data augmentation to generate additional synthetic images from available datasets.

To the best of our knowledge, there has not been any attempt that performs three micro-expression databases cross evaluation for recognition task using CNN mechanism.

In this paper, a novel and robust feature extraction approach that can effectively represent the subtle facial muscle contractions for micro-expression recognition system is presented. Concretely, the contributions of this paper are listed as follows:

1. Adoption of only two frames (i.e., onset and apex) from each video to better represent the significant expression details and applying optical flow guided techniques to encode the motion flow features.
2. Proposal of a novel feature extractor that incorporates both the handcrafted (i.e., optical flow) and data-driven (i.e., CNN) features.
3. Implementation of a novel CNN architecture that is capable to highlight valuable input features and improve the emotion state prediction.

4. Comprehensive evaluation of the proposed approach on three recent spontaneous micro-expression databases is performed to validate its consistency and effectiveness.

The remainder of the paper is organized as follows. Section 2 discusses related works on the state-of-the-art apex frame spotting and feature extraction techniques. The proposal of the recognition system framework, theoretical derivations and the effective use of CNN are elaborated in Section 3. Overview of the databases used and the experimental settings are described in Section 4. Followed by Section 5 that reports the recognition performance, with discussion and analysis. Finally, conclusions are drawn in Section 6.

2. Related work

In the literature, most of the automated micro-expression studies focused on the first and second stages of the recognition system, i.e., image preprocessing and feature extraction. Some promising preprocessing techniques and feature extractors exploited in micro-expressions analysis systems will be discussed and elaborated in the following subsections.

2.1. Image preprocessing

The two properties of the micro-expressions are low intensity and often occur in specific facial regions. Therefore, some of the previous works aim to emphasize the facial muscle movements in some particular areas, instead of extracting the features from the entire face. By focusing to extract features from several small facial regions can omit the noticeable background noises captured by the camera (which are probably due to the flickering lights). In addition, considering the regions of interest (RoIs) is able to accelerate the feature extraction and classification processes as irrelevant data are eliminated. For instance, [21] encode the expression features from 16 RoIs based on the Facial Action Coding System (FACS) [22] which indicate the relation between the facial muscle changes and the emotion state. However, the shapes and sizes of the 16 RoIs are not flexible as they are heavily rely on the feature coordinates detected by the landmark detector. On the other hand, [23] proposed to reduce number of RoIs to three regions (i.e., "left eye + left eyebrow", "right eye + right eyebrow" and "mouth"). The selection of these three areas are identified according to the occurrence frequency of the muscle movements in the videos provided by CASME II and SMIC databases. Although the size and location of the 3 RoIs are not fixed, they are merely dependent on the position of the landmark coordinates. Unfortunately, the landmark-based approach might not be sufficiently accurate and the 3 regions selected are not always the optimal areas that can capture the perfect expression information. In addition, it is pointed by [24], that a fine-scale alignment is essential to be performed as the preprocessing step. This is because the subtle misalignment resulted from the conventional facial registration and alignment tools could cause degradation in the recognition performance.

Moreover, there are some works that minimize the information redundancy in micro-expressions by emphasizing only a portion of all frames of each video. For example, [25] select several important frames for extraction. This is intuitive as the images are captured using high frame rate cameras, there will be similar facial motion patterns appearing in consecutive frames. Therefore, they intent to identify and remove unfavorable redundant frames as the preprocessing step. Besides, this could boost the discrimination power of the feature vectors. On a similar note, another recent method proposed by [26] also describes the expression details from a reduced set of frames. Concretely, Temporal Interpolation Model (TIM) [27] is applied to normalize all the videos in SMIC dataset to 20 frames and CASME II to 30 frames. It should be noted that the average frame length for SMIC and CASME II are 33 and 67, respectively. Although shorten the video length improves the
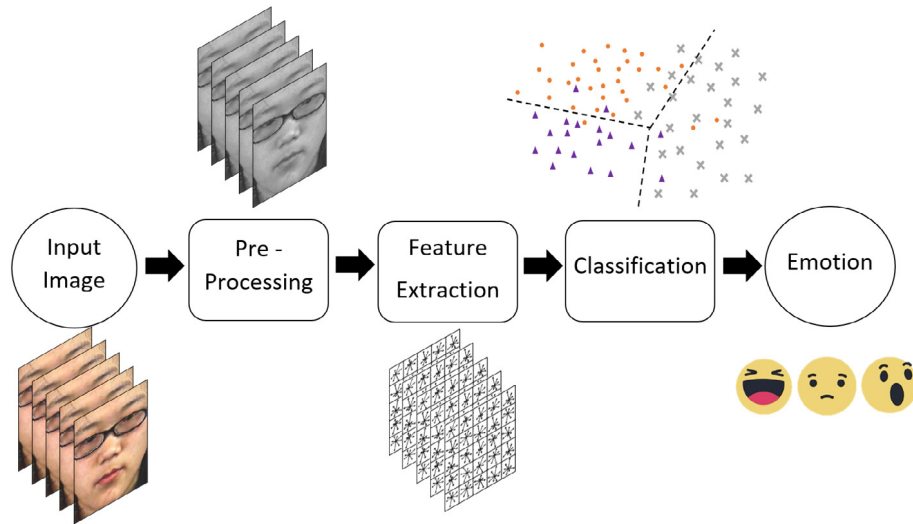
Fig. 1. Block diagram of a typical facial micro-expression recognition system.

efficiency and accuracy performance, an arbitrary decision has to be made about what frame length should be used.

Another remarkable preprocessing technique proposed in [28–30] well-represent the entire video by utilizing only the apex frame (and onset frame as reference frame). To be concise, there are generally three temporal segments in each micro-expression videos (i.e., onset, apex and offset). The onset is the instant that the facial muscles begins to contract and grow stronger. The apex frame indicate the most expressive facial action when it reaches the peak. The offset is the moment where the muscles are relaxing and the face returns to its neutral appearance. From the results reported in [28,29], it supports that, encoding the features from apex frame provides more valuable expression details than a series of frames. Furthermore, the apex-based approach is employed in the other work of [31], where they tested on other micro-expression databases comprising only of raw long videos and promising performance results are obtained.

2.2. Feature extraction

The primitive feature extraction method that evaluated on spontaneous micro-expression databases (i.e., CASME II and SMIC) is known as Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [32]. LBP-TOP was eventually designed to describe dynamic texture patterns. In brief, it is capable to capture the local spatio-temporal motion information (i.e., pixel, region and volume levels). Furthermore, it is robust against geometric variations caused by scaling, rotation, or translation. With great discriminative feature representation as well as its computational simplicity, LBP-TOP has been comprehensively studied and modified to accommodate in different applications. As a result, several LBP variants are proposed and some of them are examined in micro-expression analysis, such as Local Binary Patterns with Six Intersection Points (LBP-SIP) [33], Spatiotemporal Local Binary Pattern with Integral Projection (STLBP-IP) [34], Completed Local Quantization Pattern (STCLQP) [20], Spatiotemporal local Radon-based Binary Pattern (STRBP) [35], Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection (DiSTLBP-RIP) [36].

Apart from LBP-TOP, optical flow [37] is one of the most popular feature extractors, as it has been very successful in a variety of computer vision tasks, such as action recognition [38], face tracking [39], medical image reconstruction [40]. Succinctly, optical flow measures the apparent motion of the brightness patterns in a sequence of images in terms of velocity vector field. Owing to its robust feature representations with data from multiple domains, a number of researchers unleashed the potential of optical flow in micro-expression recognition

systems. For instance, [41] proposed to construct a RoI-based feature vector using optical flow to describe the local motion information and the spatial location. Thus, aside having compact feature representation (i.e., feature dimension of 72 per video), it is robust to translation, rotation and illumination changes. As an extension of optical flow, [42] derived a higher order accurate differential approximation, namely optical strain. Optical strain leads to better performance in determining the motion changes compared to optical flow, as it is capable to preserve relatively meaningful facial muscle movements [43,44].

Deep learning has emerged as a family of machine learning technique that operates such that the important information are iteratively extracting from data and transforming them into the final output features. Deep learning has significant impacts on a variety of application domains as it yields numerous state-of-the-art results, such as speech recognition [45], face recognition [46] and scene recognition [47]. However, deep learning has yet to have a widespread impact on micro-expression studies. Particularly, the first work that adopts Convolutional Neural Network (CNN) is established by [17] to evaluate the proposed algorithm in CASME II and SMIC databases with LOSOCV protocol during the data training and testing stages. However, the accuracy results obtained by their work do not outperform the conventional methods as the model is possibly being overfitted. Besides, [18] intent to increase the number of samples to the double of each dataset using data augmentation. They partitioned all the images into three sets, namely training, testing and validation, which consist the portion of 80%, 1% and 1%, respectively. On the other hand, a recent work by [19] directly feeds the CNN model with high level features (i.e., optical flow). Later, [48] performed a CASME II-SAMM cross database experimental validation using the following two approaches: Holdout-database Evaluation (HDE) and Composite Database Evaluation (CDE). HDE is like a two-fold cross validation, where it trains on CASME II and test on SAMM, vice versa. Whereas, CDE first combines the two databases into a single composite database, and then performs LOSOCV. The best overall recognition accuracy and F-measure obtained are 74.70% and 0.64, respectively, when evaluated on the composite database (i.e., CASME II + SAMM).

There are a few other works [49–51] perform knowledge transfer learning from the existing data (i.e., macro-expression and speech features) to the micro-expression field, in order to improve the recognition rate. This is to leverage the useful information that has been massively trained in the respective domains. On the other hand, a tensor-based method [52] is proposed by to describe the geometry-preserving discriminative features.

3. Proposed algorithm

The impressive recognition performance presented in the earlier work by [28] has brought the significance of the apex frame into a sharp focus, especially in the feature extraction stage. With rich motion patterns obtained from the apex frame (with onset as the reference frame), it is possible to select the features with minimal redundancy. As a result, the facial regions containing relevant details of the expression can be easily noticed and encoded. It should be noted that [28] did not employ CNN to automatically identify and learn relevant spatio-temporal context information. Nevertheless, the highest recognition accuracy achieved was promising (about 60% in CASME II and SMIC databases). Thus, the method proposed in this paper tend to integrate their frameworks by introducing a mechanism that is capable to better recognize the subtle micro-expression.

Particularly, the proposed method is targeted to emphasize on the preprocessing and feature extraction stages. In brief, it incorporates the following three steps:

1. Apex frame acquisition: to spot the apex frame location from each video.
2. Optical flow features elicitation: to estimate the horizontal and vertical optical flow from the apex and onset frames.
3. Feature enhancement with CNN: to enrich the optical flow features that can automatically identify and learn relevant spatio-temporal context information in a hierarchical way.

A conceptual framework in this paper is illustrated in Fig. 2. The detailed procedures for each step are described in the following sub-sections.

3.1. Apex frame acquisition

There are three micro-expression databases exploited in the experiment, namely, CASME II [14], SMIC [15] and SAMM [16]. The location of the ground-truth apex frame has been provided in CASME II and SAMM, which are annotated by at least 2 trained experts. Since the apex frame index in SMIC is absence, an automatic apex spotting system has to be applied to approximate the location of apex frame. It has been demonstrated that the apex spotting mechanism, D&C-RoIs [53], is capable to exhibit reasonable good recognition performance [28,31]. Succinctly, the D&C-RoIs method first computes the LBP features from three facial sub-regions (i.e., "left eye+eyebrow", "right eye+eyebrow" and "mouth") of each image. Then, a correlation coefficient principle is employed to acquire the changes in difference of the LBP features between the onset frame to the rest of the frames. Finally, a Divide & Conquer strategy is utilized on the rate of the feature difference to search for the apex frame, whereby it indicates the frame index of the local maximum.

For clarity, the notations used in this paper are defined and explained in the following sections. A micro-expression video clip is expressed as:

$$S = [s_1, s_2, \ldots, s_n], \qquad (1)$$

where n is the number of video clips. The ith of the sample video clip is molded to:

$$s_i = \{f_{i,j} | i = 1, \ldots, n; j = 1, \ldots, F_i\}, \qquad (2)$$

where F_i is the total number of image frames in the ith sequence. There will be one apex frame in each video sequence and it can be located at any frame index between the onset (first frame) and offset (last frame). The onset, apex and offset frames are denoted as $f_{i,1}$, $f_{i,\alpha}$ and f_{i,F_i}, respectively. The apex frame can be denoted as:

$$f_{i,\alpha} \in f_{i,1}, \ldots, f_{i,Fi} \qquad (3)$$

Thus, $f_{i,\alpha}$ is predicted after adopting the D&C-RoIs approach.

3.2. Optical flow features elicitation

In this process, a higher level with reduced dimension features are produced in this stage. Consequently, the optical flow features are obtained prior to passing the raw onset and apex images to the CNN architecture. Optical flow is able to indicate the apparent facial motion changes between frames. It is an approximation of the image patterns based on the local derivatives between two images. Specifically, it aims to generate a two-dimensional vector field, i.e., motion field, that represents the velocities and directions of each pixel. In order to attain the dynamical movement of the desired optimal expression (i.e., $p_{i,\alpha}$), the intensity difference between the onset (i.e., $f_{i,1}$) and apex (i.e., $f_{i,\alpha}$) is estimated.

To estimate the optical flow, it is generally assumed that:

• The apparent brightness of the moving objects remains unchanged between the source and target frames. Thus the noises generated by a large variety of imaging variables such as the shadows, highlights, illumination and surface translucency phenomena are entirely neglected.
• The movement between two consecutive frames are small as the motion changes gradually over time.
• Image flow field is continuous and differentiable in both the space and time domains.
• The scene is static, the objects in the scene are rigid, and the changes of the objects' shape are ignored.

Suppose that the intensity of the reference frame that locates at tth of a video sequence is defined as $I_t(x, y)$. The intensity of the next consecutive frame, $(t + 1)$-th is denoted as $I_{t+1}(x + \delta x, y + \delta y)$. According to the brightness constancy constraint, the intensity of the two adjacent frames is achieved as:

$$I_t(x, y) = I_{t+1}(x + \delta x, y + \delta y), \qquad (4)$$

where $\delta x = u^t \delta t$ and $\delta y = v^t \delta t$. Explicitly, $u^t(x, y)$ and $v^t(x, y)$ refer to the horizontal and vertical of the optical flow field, respectively. By adopting Taylor series expansion on (4), it becomes an expanded form:

$$I_{t+1}(x + \delta x, y + \delta y) \approx I_t(x, y) + \delta x \frac{\partial I}{\partial x} + \delta y \frac{\partial I}{\partial y} + \delta t \frac{\partial I}{\partial t} \qquad (5)$$

We then combines (4) and (5), the optical flow equation can be succinctly formulated as follows:

$$I_t(x, y) = I_t(x, y) + \delta x \frac{\partial I}{\partial x} + \delta y \frac{\partial I}{\partial y} + \delta t \frac{\partial I}{\partial t},$$
$$0 = \delta x \frac{\partial I}{\partial x} + \delta y \frac{\partial I}{\partial y} + \delta t \frac{\partial I}{\partial t} \qquad (6)$$

By dividing both sides of the equations by δt:

$$0 = \frac{\delta x}{\delta t} \frac{\partial I}{\partial x} + \frac{\delta y}{\delta t} \frac{\partial I}{\partial y} + \frac{\delta t}{\delta t} \frac{\partial I}{\partial t},$$
$$0 = u^t(x, y) \frac{\partial I}{\partial x} + v^t(x, y) \frac{\partial I}{\partial y} + \frac{\partial I}{\partial t} \qquad (7)$$

For a sufficiently small interval time between the onset and apex frames (i.e., less than 0.2 s), it is assumed that the brightness of the surface patches remains constant. Hence, the optimal expression flow feature $p_{i,\alpha}$, can be obtained as:

$$I_{t=1}(x, y) = I_{t+\alpha}(x + u^t(x, y)\delta t, y + v^t(x, y)\delta t) \qquad (8)$$

Finally, the optical flow map that computed from the two frames (i.e., onset and apex) is formed to represent the entire video:

$$O_i = \{(u(x, y), v(x, y)) | x = 1, 2, \ldots, X, y = 1, \ldots, Y\}, \qquad (9)$$

where X and Y denote the width and height of the images, $f_{i,j}$, respectively.

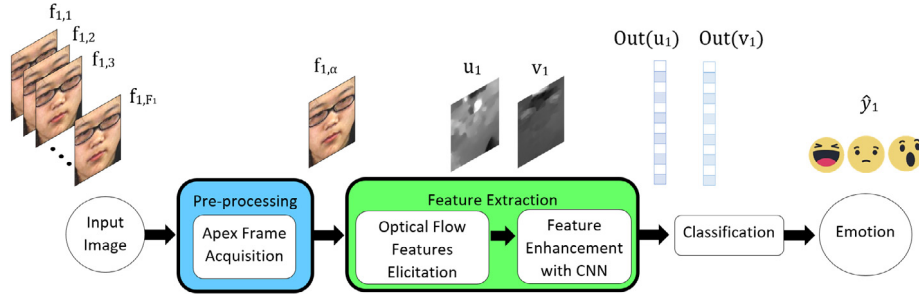In short, each video sequence, s_i is summarized into the following two optical flow derived representations:

Fig. 2. Overview of the proposed micro-expression recognition system. It consists of three main steps, namely apex frame acquisition, optical flow features elicitation and feature enhancement with CNN.

1. $u(x, y)$ — Horizontal component of the optical flow field O_i
2. $v(x, y)$ — Vertical component of the optical flow field O_i

The optical flow technique utilized in the experiments later is TV-L1 [54] method. This is because it is better in preserving the flow discontinuities and is more robust compared to the classical optical flow method (i.e., Black and Anandan [55]) [42].

3.3. Feature enhancement with convolutional neural network

The optical flow features contain the spatio-temporal expression details. They are then fed into a CNN architecture which is expected to further improve the feature information by reconstructing and refining the selection of more significant motion details. CNN is one of the deep artificial neural networks that has been widely used in analyzing visual imagery [56–58]. It consists of several layers, such as the input layer, convolutional layer, pooling layer, fully connected layer and output layer. CNN has also been recently exploited in micro-expression recognition mechanisms. For example, [19] designed a 3D-CNN architecture to effectively learn the high-level features (i.e., optical-flow data). However, in contrast to [19], the optical flow representations obtained from the previous stage (i.e., Section 3.2) are having two dimensional maps (i.e., $X \times Y$). Therefore, a new 2D-CNN architecture is proposed to perform the feature learning task.

Fig. 3 illustrates the conceptual visualization of our proposed OFF-ApexNet (Optical Flow Features from Apex frame Network) architecture. The horizontal and vertical components of the optical flow are used as the input data of the CNN. Two independently trained CNN models (i.e., to train u and v separately) will be merged to form a resultant feature vector at the fully connected layers. The basic overview of the duty of each layer is described and explained as follows.

First, for the input layer, all the input data are normalized to a fix size (i.e., $\aleph \times \aleph$), whereby the input data in this case is the optical flow based components, such that:

$$u = \frac{\delta x(t)}{\delta t}, \tag{10}$$

and

$$v = \frac{\delta y(t)}{\delta t}, \tag{11}$$

where u and v refers to the horizontal and vertical components of optical flow, respectively. The normalized data is then multiplied with a convolution kernel to form a feature map in the following convolutional layer. Concretely, each e_{ij} pixel in the feature map is calculated by:

$$e_{ij}^l = \{f^l(x_{ij}^l + b^l) | i = 1, 2, \dots, \aleph, j = 1, \dots, \aleph\}, \tag{12}$$

where $x_{ij}^l = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab}^l y_{(i+a)(j+b)}^{l-1}$,

x_{ij}^l is the pixel-value vector of the units in the local receptive field corresponding to e_{ij} pixel at layer l, whereas f^l denotes the ReLu activation function at layer l. w^l and b^l are the coefficient vector and

bias respectively at layer l, determined by the feature map. Thus for an input x, the ReLu function can be indicated as:

$$f(x) = max(0, x) \tag{13}$$

The input optical flow features (i.e., u and v) are now transformed into feature maps (i.e., e) representation. The size of generated feature map is rely on the number of convolution kernels. Conventional kernel sizes chosen in the past research are 3×3, 5×5 and 7×7.

The subsequent layer is the pooling layer. It is used as a subsampling operator to progressively reduce the spatial size of the feature map representation. As a result, it can effectively minimize the computational complexity of the CNN architecture. The $k\text{-}th$ unit in the feature map in the pooling layer can be achieved by:

$$Pool_k = f(down(C) * W + b), \tag{14}$$

where W and b are the coefficient and bias, respectively. $down(\cdot)$ is a subsampling function, which can be expressed as:

$$down(C) = max\{C_{s,l} | s \in Z^+, l \in Z^+ \le m\}, \tag{15}$$

where $C_{s,l}$ refers to the pixel value of C in the feature map e. m denotes the sampling size.

It is observed that each layer (i.e., convolutional layers and pooling layers) in the CNN architecture deliberately learn and convert the optical flow features to higher level features in other subsequence layers. After passing through all the convolution network layers (which may consists of several convolution layers and pooling layers), the final feature representation (denoted as $Out(\tau)$) comprises significant expression information, where τ is the optical flow based features of input images (i.e., u and v).

Since the total number of videos used in the experiments is relatively few (i.e., 441 from three datasets), the proposed CNN architecture is composed of only four layers (i.e., two convolution layers and two pooling layers). These layers are responsible to generate meaningful features from the input data, where the final output $Out(\tau)$ can be concisely expressed as follows:

$$Out(u) = f^4(down(f^3((f^2(down(f^1(u * W^1 + b^1)) \\ * W^2 + b^2)) * W^3 + b^3)) * W^4 + b^4) \tag{16}$$

and

$$Out(v) = f^4(down(f^3((f^2(down(f^1(v * W^1 + b^1)) \\ * W^2 + b^2)) * W^3 + b^3)) * W^4 + b^4) \tag{17}$$

The high-level reasoning features (i.e., Out_u and Out_v) derived from the input data are then flattened and merged before passing to the following fully connected layer. In general, the fully connected layers transforms the features to a set of desired number of classes from the analysis of frequencies based on the importance of features. There are three emotion classes in the experiments, namely positive, negative, and surprise. Note that similar to the convolutional layer, a ReLu activation function is applied to all of the output after the fully connected layer.
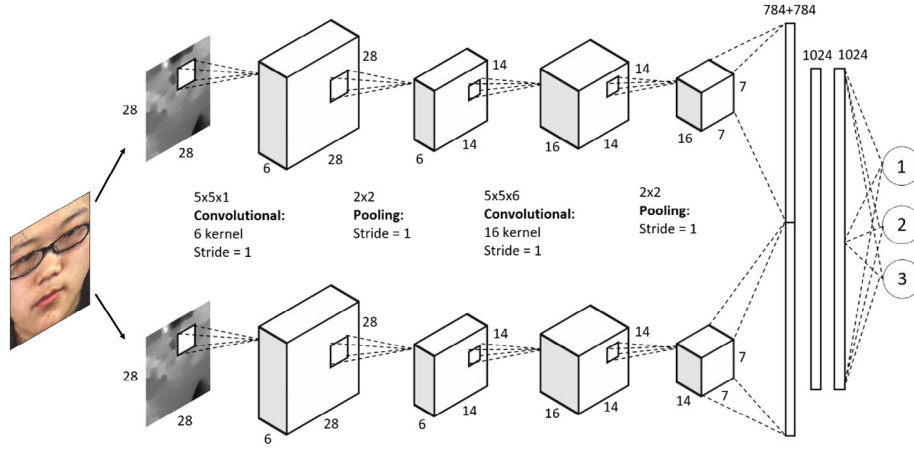
Fig. 3. Desired CNN architecture that takes optical flow derived components (upper stream is u and bottom stream is v) as input data.

Next, the transformed features from the fully connected layer is passed into the output layer. The amount of neurons in the output layer is associated with the number of classes to be classified, which is three in this case. The output probabilities of each class are computed using an activation function, which will result to a sum of one. However, practically the output given by the former layers do not guarantee that the total sum of the probabilities over all classes equals to one. To resolve this issue, a softmax regression is utilized as the activation function. Specifically, the probability of classifying into class c is given by:

$$\hat{y} = p\left(y = c | x_j\right) = \frac{e^{x_j}}{\Sigma_{n=1}^{N} e^{x_n}}, 1 \leq c \leq C, \qquad (18)$$

where y is the ground-truth value of input x_j, C is the number of the classes. The loss function can be defined as follows:

$$L(y, \hat{y}) = -\Sigma_{i=1}^{N} l(y_i) \log(\hat{y}_i), \qquad (19)$$

where $l\{\cdot\}$ is eigenfunction. When $l\{\cdot\}$ is true, the loss function will return a number of one. The gradient of error can be calculated using (19). Then sum of errors from multiple inputs is anticipated to be minimized by updating the weights of networks using a stochastic Adam gradient descent. This particular type of gradient descent is known as an optimization algorithm. It aims to search for the weights and coefficient in the neural network by performing backpropagation, so that the actual output to be closer the target output. Thereby, decreases the error of each output neuron and the network as a whole.

4. Experiment

4.1. Database

There are a total of three micro-expression databases involve in the experiment, namely SMIC [15], CASME II [14] and SAMM [16]. This is to avoid the issue of overfitting, which will happen when the gap between training and testing errors is large. Since the number of video of each single database is considered small (i.e., ≈ 150), it will fit the training dataset very well but underperform on new datasets. Besides, more training data can improve the data generalization capability. As such, by considering all the three datasets as a whole, it could lead to constructing a good predictive model. Thus, better in recognizing the new (i.e., unseen) faces with different imaging conditions and environments.

Note that, all the image data from these databases are captured under constrained laboratory condition and have undergone face registration and alignment before releasing to the recorded videos to the public. Particularly, facial alignment is carried out in order to standardize all the faces into a uniform size and shape. Besides, it is

Table 1
Detailed information of the SMIC, CASME II and SAMM databases used in the experiment.

		SMIC	CASME II	SAMM
Participants		16	24	28
Frame rate (fps)		100	200	200
Cropped resolution (pixels)			170×140	
Avg. frame number		34	68	74
Avg. video duration (s)		0.34	0.34	0.37
Expression	Negative	70	88	91
	Positive	51	32	26
	Surprise	43	25	15
	Total	164	145	132
Ground-truth (index)	Onset	Yes	Yes	Yes
	Offset	Yes	Yes	Yes
	Apex	No	Yes	Yes
Number of coder		2	2	3
Inter-coder reliability		N/A	0.846	0.82

also to ensure that the data extracted later are capable of integration. Succinctly, face alignment is a process of detecting the transforming a set of landmark coordinates to map the face to the model face. Specifically, both the SMIC [15] and CASME II [14] utilized Active Shape Model (ASM) [59] to allocate the 68 facial landmark points then Local Weighted Mean (LWM) [60] is employed to transform the faces based on the model face. For SAMM, the faces are first registered with a Face++ automatic facial point detector [61], then dlib [62] is adopted as the face alignment tool.

An overview of the micro-expression datasets information that used in the experiment is shown in Table 1. More details are elaborated as follows.

4.1.1. SMIC

The Spontaneous Micro-expression (SMIC) dataset comprises 16 subjects with 164 video clip. The camera used to capture the video was PixeLINK PL-B774U with a temporal resolution of 100 fps. The cropped images have an average spatial resolution of 170×140 pixels, and each video consists of 34 frames (viz., $0.34s$). The ground-truths are labeled by two annotators, which include the emotion state, the action unit, the onset, offset frame indices. However, the apex frame information of each video is not provided. The videos include three classes: positive (51 videos), negative (70 videos) and surprise (43 videos). A three-class baseline recognition accuracy is reported as 48.78% by employing LBP-TOP as the feature descriptor and SVM with LOSOCV protocol.

4.1.2. CASME II

The Chinese Academy of Sciences Micro-Expression (CASME II) consists of 255 videos, elicited from 26 participants. The videos are recorded using Point Gray GRAS-03K2C camera which has a frame rate

Table 2
OFF-ApexNet configuration for two convolution layers, two pooling layers, two fully connected layers and an output layer.

Layer	Filter size	Kernel size	Stride	Padding	Output size
Conv 1	$5 \times 5 \times 1$	6	[1,1,1,1]	Same	$28 \times 28 \times 6$
Pool 1	2×2	–	[1,2,2,1]	Same	$14 \times 14 \times 6$
Conv 2	$5 \times 5 \times 6$	16	[1,1,1,1]	Same	$14 \times 14 \times 16$
Pool 2	2×2	–	[1,2,2,1]	Same	$7 \times 7 \times 16$
FC 1	–	–	–	–	1024×1
FC 2	–	–	–	–	1024×1
Output	–	–	–	–	3×1

Table 3
Overall micro-expression recognition accuracy and F-measure evaluated on SMIC, CASME II and SAMM databases using the proposed method, OFF-ApexNet.

Epoch	Accuracy (%)	F-measure
1000	72.56	.6905
2000	73.47	.7027
3000	**74.60**	**.7104**
4000	72.79	.6918
5000	73.70	.6998

of 200 fps. The average video length is 0.34 s, equivalent to 68 frames. Each video's emotion label is annotated by two coders, where the reliability is 0.846. All the images are cropped to 170×140 pixels. The ground-truth information provided by the database include the emotion state, the action unit, the onset, apex and offset frame indices. The videos are grouped into seven categories: others (99 videos), disgust (63 videos), happiness (32 videos), repression (27 videos), surprise (25 videos), sadness (7 videos) and fear (2 videos). A 5-class recognition baseline result of 63.41% is reported which the feature extractor utilized was LBP-TOP and the classifier was Support Vector Machine (SVM) with Leave-One-Video-Out Cross-Validation (LOVOCV) protocol. To perform cross database evaluation in the experiment later, some of the videos are recategorized based on the emotion state. This is to cope with the database (i.e., SMIC) that has few expressions. As a result, three main emotion classes are standardized: positive, negative and surprise. Negative class include repression and disgust expressions; happiness is regarded as positive class, while the videos with others expression are not considered in the experiment.

4.1.3. SAMM

The Spontaneous Actions and Micro-Movements (SAMM) dataset contains 159 spontaneous videos, elicited from 32 participants. The videos are recorded using Basler Ace acA2000-340km camera with a temporal resolution of 200 fps. The average number of frames of the micro-expression video sequences is 74 frames (viz., 0.37 s). This dataset provides the cropped face video sequence with a spatial resolution of 400×400 pixels. In an attempt to standardize the image resolution so that it is equivalently behaved as the other two databases, all the images are resized to 170×140 pixels. Each video is assigned with its emotion label, action unit, frame indices of apex, onset and offset. The reliability of the marked labels by 3 coders is 0.82. This database composes of eight classes of expressions: anger (57 videos), happiness (26 videos), other (26 videos), surprise (15 videos), contempt (12 videos), disgust (9 videos), fear (8 videos) and sadness (6 videos). A recognition accuracy of 80.06% is achieved with LBP-TOP as the feature extractor and Random Forest as the classifier with LOSOCV protocol. For the experiment purpose, the videos are reclassified, such that it consists of three main classes: negative (i.e., anger, contempt, disgust, fear and sadness), positive (happiness) and surprise. Note that videos with other expression are neglected.

4.2. Experiment settings

In the OFF-ApexNet, the input features (i.e., u and v) are resized into $[\aleph \times \aleph] = [28 \times 28]$. After that, they are processed by the convolutional, pooling, fully connected layers and finally the output layer. The parameter setting for each layer is tabulated in Table 2. To reduce the overfitting phenomena, a dropout regularization operation is applied after the two fully connected layers. A ratio of 0.5 is set, so that it keeps 50% of the original output. The initial learning rate is set to 0.0001 and a set of epochs values (i.e., 1000, 2000, 3000, 4000 and 5000) are examined.

Next, in the softmax classification layer, a cross-database micro-expression recognition will be performed, which means the videos from the three databases are combined in the experiment. Therefore, the

total number of video involved in the experiment is 441, which is made up from SMIC (164 videos), CASME II (145 videos) and SAMM (132 videos). There are three main emotion classes: negative, positive and surprise. Specifically, a LOSOCV protocol is employed to examine the robustness of the proposed framework. The principle of LOSOCV protocol is to iteratively leave out the videos of a single subject or participant as the testing set, while the rest of the videos will be served as training set. This procedure is repeated for k times, where k is the number of participants in the experiment. Finally, the sum of all the true positive samples (correctly predicted labels) from the resultant confusion matrix is divided by the total number of sample (i.e., 441) to indicate the final recognition accuracy. It should be reminded that, the video of the same subject will not be appearing in both the training and testing sets simultaneously. Thus, it is considered as a person-independent approach.

To deal with the imbalance class distribution (i.e., 249 negative videos, 109 positive videos and 83 surprise videos), an alternative recognition performance measurement is exploited, namely F-measure. Concretely, F-measure is defined as:

$$\text{F-measure} := 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{20}$$

for

$$\text{Recall} := \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{21}$$

and

$$\text{Precision} := \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{22}$$

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

5. Results and discussion

5.1. Recognition performance

To the best of our knowledge, this is the first attempt that evaluates the feature extractor on three micro-expression databases. Table 3 reports the micro-expression recognition performance in both accuracy and F-measure of OFF-ApexNet method with various epoch size. Concisely, all the three databases (i.e., SMIC, CASME II and SAMM) are merged and treated as a single database. Therefore, a LOSOCV classification will be applied for $k = 68$ times, which the 16 times are from SMIC, 24 times from CASME II and 28 times from SAMM. From Table 3, it is noticed that OFF-ApexNet approach achieves the highest accuracy of 74.60% and F-measure of 71.04% when the epoch value is set to 3000.

On the other hand, Table 4 shows the comparison of the micro-expression recognition performances of the proposed method (i.e., OFF-ApexNet) with other state-of-the-art feature extraction methods when evaluated on SMIC, CASME II and SAMM databases individually. Particularly, the previous research works (i.e., methods #1 to #11) focus to conduct the training and testing videos on a single database separately. For methods #1 to #11, the number of the expression to be predicted are based on the suggested expression category from the original papers [14–16]. Some of the number of videos for certain expressions are quite few (i.e., less than 10 samples), thus those videos are neglected in the experiments. Concisely, there are a total of three expressions

Table 4

Comparison of micro-expression recognition performance in terms of *Acc* (Accuracy (%)) and *F-mea* (F-measure) on the SMIC, CASME II and SAMM databases for the state-of-the-art feature extraction methods, and the proposed method.

	Methods	SMIC		CASME II		SAMM	
		Acc	F-mea	Acc	F-mea	Acc	F-mea
		3 classes		5 classes		5 classes	
1	LBP [14–16]	45.73	.4600	39.68	.3589	35.56	.1768
2	OSF [63]	31.98	.4461	–	–	–	–
3	OSW [44]	53.05	.5431	41.70	.3820	–	–
4	LBP-SIP [33]	54.88	.5502	43.32	.3976	–	–
5	MRW [64]	34.15	.3451	46.15	.4307	–	–
6	STLBP-IP [34]	57.93	.5829	59.51	.5679	–	–
7	FDM [24]	54.88	.5380	41.96	.2972	–	–
8	Sparse sampling [25]	58.00	.6000	49.00	.5100	–	–
9	STCLQP [20]	64.02	.6381	58.39	.5836	–	–
10	MDMO [41]	–	–	44.25	.4416	–	–
11	Bi-WOOF [28]	61.59	.6110	57.89	.6125	–	–
		3 classes					
12	LBP-TOP	38.41	.3875	60.00	.5222	59.09	.3640
13	Bi-WOOF [28]	61.59	.6110	80.69	.7902	58.33	.3970
14	**OFF-ApexNet**	**67.68**	**.6709**	**88.28**	**.8697**	**68.18**	**.5423**

Table 5

Emotion ratio distribution of the three databases.

	SMIC	CASME II	SAMM
Negative	4	6	7
Positive	3	2	2
Surprise	3	2	1

Table 6

Confusion matrices of OFF-ApexNet for the recognition task on all the databases.

	Negative	Positive	Surprise
Negative	**.84**	.11	.05
Positive	.35	**.58**	.07
Surprise	.20	.11	**.69**

Table 7

Confusion matrices of OFF-ApexNet for the recognition task on SMIC database.

	Negative	Positive	Surprise
Negative	**.76**	.17	.07
Positive	.25	**.65**	.10
Surprise	.28	.14	**.58**

Table 8

Confusion matrices of OFF-ApexNet for the recognition task on CASME II database.

	Negative	Positive	Surprise
Negative	**.93**	.07	0
Positive	.31	**.66**	.03
Surprise	0	0	**1**

Table 9

Confusion matrices of OFF-ApexNet for the recognition task on SAMM database.

	Negative	Positive	Surprise
Negative	**.81**	.10	.09
Positive	.58	**.35**	.08
Surprise	.33	.20	**.47**

(i.e., positive, negative and surprise) in SMIC, five expressions (i.e., disgust, happiness, repression, surprise and others) in CASME II and five expressions (i.e., anger, happiness, contempt, surprise and other) in SAMM.

In Table 4, method #1 (i.e., LBP-TOP) is commonly known as the baseline approach in this automated micro-expression recognition domain, the recognition results reported are obtained by reproducing the experiments for each database. Since the SAMM database is released very recently, methods #2 to #11 did not examine the methods on this database. It can be seen that method #11 (i.e., Bi-WOOF) outperformed the feature descriptors #1 to #10. As such, Bi-WOOF approach is adopted to compare with the proposed method OFF-ApexNet later.

To establish a fair comparison on the effectiveness of the proposed method, two state-of-the-art approaches (i.e., LBP-TOP and Bi-WOOF approach) are selected and the experimental configurations are set to similar across the comparing methods. More precisely, the videos from the three databases (i.e., SMIC, CASME II and SAMM) are recategorized into exclusively three expressions (i.e., positive, negative and surprise). This video relabeling is similar to the work by [65], where they evaluated the cross-database micro-expression for CASME II and SMIC. They also have a total of three expression categories.

As a result, the recognition performance is presented as methods #12 and #13. Particularly, for the proposed OFF-ApexNet approach (i.e., #14), the feature extraction process follows the procedure as described in Section 3. Firstly, the OFF-ApexNet model is trained by three databases as a whole using a LOSOCV strategy, then tested on each database separately. It can be seen that, among all the methods shown in Table 4, OFF-ApexNet method achieves the best recognition results across all the three databases.

5.2. Analysis and discussion

In Table 4, it can be seen that the accuracy result in SMIC database is the lowest among the three databases, when utilizing OFF-ApexNet. It might because of the apex frames of each video are spotted using an automatic apex spotting system, instead of utilizing the ground-truths. Referring to [53], the average of frame difference between the detected and ground-truth apex is 13 frames. Thus, extracting the features from imprecise apex frame could affect the classification performance. For SAMM database, the F-measure is only 0.5423. This is due to the imbalance emotion class distribution where the ratio distribution is summarized in Table 5. SAMM database has the most severe imbalance data issue, whereby there are only 10% surprise videos and 20%

positive videos. It is also observed that the although SMIC is having balanced data distribution, the recognition performance (i.e., accuracy and F-measure) exhibited is lower than CASME II. This is possibly due to the prominent expressive frames in SMIC database are not being captured by the camera as it has a much lower frame rate (i.e. 100 fps), compared to CASME II (200 fps). In a consequence, it fails to spot the precise apex frame in such circumstances.

To further analyze the three-class recognition performance, confusion matrices are computed and shown in Tables 6 to 9. Generally, confusion matrix is a typical measurement to illustrate the classification rate for each expression. The confusion matrix in Table 6 indicates the overall performance, which means all the three databases are treated as a single database for training and testing purposes. The other three confusion matrices (i.e., in Tables 7 to 9) are tested on each database independently. It can be seen that the negative emotion can always exhibit the highest prediction rate compared to positive and surprise. The main reason is that, the negative emotion is the dominant class across the three databases (refer to Table 5).

On the other hand, instead of utilizing both the horizontal and vertical optical flow component as the input data for OFF-ApexNet approach, the performance results for the individual flow component are also evaluated. A comparison of the choice of input features is tabulated in Table 10, with a variation of the epoch values. Concretely, u is simply taking account only the horizontal optical flow features, while v considers the vertical optical flow features. $u + v$ refers to the proposed OFF-ApexNet method which fuses u and v motion information as the input data. It is observed that OFF-ApexNet approach exhibits consistent high performance results compared to both the u and v methods.

Table 11 provides a closer look into the effects of varying the number of convolution and pooling layers in the proposed approach. It

Table 10
Comparison of the micro-expression recognition accuracy and F-measure when the input data to the network are: u, the horizontal optical flow features; v, the vertical optical flow features and $u + v$, both horizontal and vertical optical flow.

Epoch	u		v		u + v	
	Accuracy (%)	F-measure	Accuracy (%)	F-measure	Accuracy (%)	F-measure
1000	67.35	0.6224	66.89	0.6199	72.56	0.6905
2000	68.03	0.6307	67.57	0.6253	73.47	0.7027
3000	66.21	0.6134	65.31	0.6047	**74.60**	**0.7104**
4000	67.35	0.6287	65.99	0.6111	72.79	0.6918
5000	66.67	0.6220	66.21	0.6150	73.70	0.6998

Table 11
Recognition results when varying the convolution and pooling layer amount in the neural network.

Layer	Accuracy (%)	F-measure
2	**74.60**	**.7104**
3	71.88	.6945
4	70.75	.6742
5	73.70	.7099

is observed that, the recognition accuracies and F-measure used in the proposed neural network architecture (i.e., 2 layers) are outperformed the in all the cases. The results conclude that adding more layers to a neural network, making it possible to represent more complex functions, but more prone to potentially overfitting the features. The input image fed into the neural network considered small (i.e., 441 images in total), experimental results suggest that the optimum layer number to be selected for our application is 2.

There are some recent works adopt CNN architecture for micro-expression recognition. For example, Li et al. [66] employ VGG-Face model on the apex frames, Wang et al. [67] utilizes both LSTM and CNN after applying TIM64, Li et al. [68] design a three-stream flow CNN to extract the TIM10 image data. The OFF-Apexnet approach is a shallow CNN, which only learns a total of 1.3 million of parameters (i.e., weights and biases), whereas the parameters to be learned in [66], [67] and [68] are 138 million, 80 million and 3 million, respectively.

Furthermore, comparisons against the state-of-the-art methods show that the use of the apex frame with optical flow and CNN feature extraction can provide more valuable information. It has been pointed out by [65] that the class imbalance problem is always one of the factors to affect the difficulty degree of cross-database micro-expression recognition task. Nevertheless, our proposed method successfully encountered this issue and achieving overall satisfactory and promising results across the three micro-expression databases.

6. Conclusion

In a nutshell, a novel feature extraction approach, Optical Flow Features from Apex frame Network (OFF-ApexNet) is introduced to recognize the micro-expressions. As its name implies, it combines both the handcrafted features (i.e., optical flow derived components) and the fully data-driven architecture (i.e., convolutional neural network). First, the horizontal and vertical optical flow features are computed from onset and apex frames. Then, the features are proceed to feed into a neural network to further highlight significant expression information. The utilization of both the handcrafted and data-driven features is capable to achieve promising performance results on three recent state-of-the-art databases, namely SMIC, CASME II and SAMM. As a result, a highest three-class classification accuracy of 74.60% was achieved with its F-measure of 0.71, when considering the three databases as a whole.

While most works in literature tend to work on new feature extraction techniques, we believe that the pre-processing and classification steps are equally important. Our work presents an additional pre-processing strategy (i.e., identification of the apex frame) to complement state-of-the-art feature extractor (i.e., optical flow and CNN).

From the recognition performance, it explicitly affirms that our hypotheses and intuition hold true: Extracting the optical flow guided features solely on apex frame with neural network architecture enhancement approach improves the performance of all databases in recognition task. In addition, this is the first attempt that performs three recent and publicly micro-expression databases cross evaluation for recognition task using CNN mechanism. It is also worth to highlight that the proposed approach is simple and effective, as it accomplished the two important factors in designing all the automated system: recognition accuracy and computational complexity. Therefore, we believe that the reported findings will be of interest to the readers and is beneficial to the development of micro-expression related research domain.

The contributions of this work point to some avenues for further research. For instance, rather than utilizing optical flow feature, other feature extractors (i.e., LBP, HOG, SIFT, etc.) can be applied to better represent the motion details. As a result, valuable input data will be passed to the convolutional neural network architecture for feature enrichment and selection, thereby improve the classification performance. Besides, attention can be devoted to handling the issues of imbalance data in these databases so that the methods proposed can lead to consistent good recognition results across all the expressions.

Conflict of interest:

- Xiamen University Malaysia
- Feng Chia University
- Xinyang Normal University
- Tokyo City University

References

[1] P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion., J. Personality Soc. Psychol. 17 (2) (1971) 124.
[2] D. Matsumoto, B. Willingham, Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals., J. Personality Soc. Psychol. 96 (1) (2009) 1.
[3] A.T. Lopes, E. de Aguiar, A.F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognit. 61 (2017) 610–628.
[4] Z. Wang, Q. Ruan, G. An, Facial expression recognition using sparse local fisher discriminant analysis, Neurocomputing 174 (2016) 756–766.
[5] G.U. Kharat, S.V. Dudul, Emotion recognition from facial expression using neural networks, in: Human-Comput. Syst. Interact., Springer, 2009, pp. 207–219.
[6] H. Ali, M. Hariharan, S. Yaacob, A.H. Adom, Facial emotion recognition using empirical mode decomposition, Expert Syst. Appl. 42 (3) (2015) 1261–1277.
[7] A.R. Rivera, J.R. Castillo, O.O. Chae, Local directional number pattern for face analysis: face and expression recognition, IEEE Transact. Image Process. 22 (5) (2013) 1740–1752.
[8] P. Seidenstat, F.X. Splane, Protecting Airline Passengers in the Age of Terrorism, ABC-CLIO, 2009.

[9] M. OSullivan, M.G. Frank, C.M. Hurley, J. Tiwana, Police lie detection accuracy: the effect of lie scenario, Law Hum. Behav. 33 (6) (2009) 530.

[10] D. Matsumoto, H.S. Hwang, Evidence for training the ability to read microexpressions of emotion, Motivation and Emotion 35 (2) (2011) 181–191.

[11] J.H. Turner, The Eevolution of Emotions: The Nonverbal Basis of Human Social Organization, Lawrence Erlbaum Associates, Publishers: Mahwah, New Jersey, 1997.

[12] M. Frank, M. Herbasz, K. Sinuk, A. Keller, C. Nolan, I see how you feel: training laypeople and professionals to recognize fleeting emotions, in: The Annual Meeting of the International Communication Association. Sheraton New York, New York City, 2009.

[13] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, X. Fu, Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces, in: Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on, IEEE, 2013, pp. 1–7.

[14] 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.

[15] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: inducement, collection and baseline, in: Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on, IEEE, 2013, pp. 1–6.

[16] 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) (2018) 116–129.

[17] D. Patel, X. Hong, G. Zhao, Selective deep features for micro-expression recognition, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 2258–2263.

[18] M.A. Takalkar, M. Xu, Image based facial micro-expression recognition using deep learning on small datasets, in: Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on, IEEE, 2017, pp. 1–7.

[19] 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.

[20] X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikäinen, Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns, Neurocomputing 175 (2016) 564–578.

[21] S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, C.-G. Zhou, Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features, in: Workshop at the European Conference on Computer Vision, Springer, 2014, pp. 325–338.

[22] P. Ekman, W.V. Friesen, Facial Action Coding System: Investigatoris Guide, Consulting Psychologists Press, 1978.

[23] S.-T. Liong, J. See, R.C.-W. Phan, K. Wong, S.-W. Tan, Hybrid facial regions extraction for micro-expression recognition system, J. Signal Process. Syst. 90 (4) (2018) 601–617.

[24] F. Xu, J. Zhang, J.Z. Wang, Microexpression identification and categorization using a facial dynamics map, IEEE Trans. Affect. Comput. 8 (2) (2017) 254–267.

[25] A.C. Le Ngo, J. See, R.C.W. Phan, Sparsity in dynamics of spontaneous subtle emotions: analysis and application, IEEE Trans. Affect. Comput. 8 (3) (2017) 396–411.

[26] J. He, J.-F. Hu, X. Lu, W.-S. Zheng, Multi-task mid-level feature learning for micro-expression recognition, Pattern Recognit. 66 (2017) 44–52.

[27] Z. Zhou, G. Zhao, Y. Guo, M. Pietikainen, An image-based visual speech animation system, IEEE Trans. Circuits Syst. Video Technol. 22 (10) (2012) 1420–1432.

[28] 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.

[29] S.-T. Liong, J. See, K. Wong, Micro-expression recognition using apex frame with phase information, in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, IEEE, 2017, pp. 534–537.

[30] Y. Gan, S.-T. Liong, Bi-directional vectors from apex in cnn for micro-expression recognition, in: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), IEEE, 2018, pp. 168–172.

[31] 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.

[32] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[33] 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.

[34] X. Huang, S.-J. Wang, G. Zhao, M. Piteikainen, Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection, in: Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 1–9.

[35] X. Huang, G. Zhao, Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern, in: the Frontiers and Advances in Data Science (FADS), 2017 International Conference on, IEEE, 2017, pp. 159–164.

[36] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikainen, Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition, IEEE Trans. Affect. Comput. (2018) 1–1.

[37] J.J. Gibson, The Perception of the Visual World, Houghton Mifflin, 1950.

[38] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1932–1939.

[39] D. Decarlo, D. Metaxas, Optical flow constraints on deformable models with applications to face tracking, Int. J. Comput. Vis. 38 (2) (2000) 99–127.

[40] N. Weng, Y.-H. Yang, R. Pierson, Three-dimensional surface reconstruction using optical flow for medical imaging, IEEE Trans. Med. Imag. 16 (5) (1997) 630–641.

[41] 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) (2016) 299–310.

[42] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, S. Sarkar, Towards macro- and micro-expression spotting in video using strain patterns, in: Applications of Computer Vision (WACV), 2009 Workshop on, IEEE, 2009, pp. 1–6.

[43] S.-T. Liong, J. See, R.C.-W. Phan, Y.-H. Oh, A.C. Le Ngo, K. Wong, S.-W. Tan, Spontaneous subtle expression detection and recognition based on facial strain, Signal Process., Image Commun. 47 (2016) 170–182.

[44] S.-T. Liong, J. See, R.C.-W. Phan, A.C. Le Ngo, Y.-H. Oh, K. Wong, Subtle expression recognition using optical strain weighted features, in: Asian Conference on Computer Vision, Springer, 2014, pp. 644–657.

[45] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., Deep speech 2: end-to-end speech recognition in english and mandarin, in: International Conference on Machine Learning, 2016, pp. 173–182.

[46] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: Face recognition with very deep neural networks, arXiv preprint arXiv:1502.00873, 2015.

[47] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.

[48] M. Peng, Z. Wu, Z. Zhang, T. Chen, From macro to micro expression recognition: deep learning on small datasets using transfer learning, in: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, IEEE, 2018, pp. 657–661.

[49] X. Zhu, X. Ben, S. Liu, R. Yan, W. Meng, Coupled source domain targetized with updating tag vectors for micro-expression recognition, Multimedia Tools Appl. 77 (3) (2018) 3105–3124.

[50] X. Jia, X. Ben, H. Yuan, K. Kpalma, W. Meng, Macro-to-micro transformation model for micro-expression recognition, J. Comput. Sci. 25 (2018) 289–297.

[51] X. Ben, X. Jia, R. Yan, X. Zhang, W. Meng, Learning effective binary descriptors for micro-expression recognition transferred by macro-information, Pattern Recognit. Lett. 107 (2018) 50–58.

[52] X. Ben, P. Zhang, R. Yan, M. Yang, G. Ge, Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation, Neural Comput. Appl. 27 (8) (2016) 2629–2646.

[53] 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: Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on, IEEE, 2015, pp. 665–669.

[54] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime tv-l 1 optical flow, in: Joint Pattern Recognition Symposium, Springer, 2007, pp. 214–223.

[55] M.J. Black, P. Anandan, The robust estimation of multiple motions: parametric and piecewise-smooth flow fields, Comput. Vision Image Understand. 63 (1) (1996) 75–104.

[56] Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, J. Yuan, Multi-stream CNN: learning representations based on human-related regions for action recognition, Pattern Recognit. 79 (2018) 32–43.

[57] D. Bai, C. Wang, B. Zhang, X. Yi, X. Yang, Sequence searching with cnn features for robust and fast visual place recognition, Comput. Graph. 70 (2018) 270–280.

[58] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, S.W. Baik, Action recognition in video sequences using deep bi-directional lstm with cnn features, IEEE Access 6 (2018) 1155–1166.

[59] B. Van Ginneken, A.F. Frangi, J.J. Staal, B.M. ter Haar Romeny, M.A. Viergever, Active shape model segmentation with optimal features, IEEE Trans. Med. Imag. 21 (8) (2002) 924–933.

[60] A. Goshtasby, Image registration by local approximation methods, Image Vis. Comput. 6 (4) (1988) 255–261.

[61] M.T. Inc., Face++ research toolkit, URL https://www.faceplusplus.com.cn/face-detection, Graphics.

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

[63] S.-T. Liong, R.C.-W. Phan, J. See, Y.-H. Oh, K. Wong, Optical strain based recognition of subtle emotions, in: Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on, IEEE, 2014, pp. 180–184.

[64] Y.-H. Oh, A.C. Le Ngo, J. See, S.-T. Liong, R.C.-W. Phan, H.-C. Ling, Monogenic riesz wavelet representation for micro-expression recognition, in: Digital Signal Processing (DSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 1237–1241.

[65] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, G. Zhao, Domain regeneration for cross-database micro-expression recognition, IEEE Trans. Image Process. 27 (5) (2018) 2484–2498.

[66] 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.

[67] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, X. Fu, Micro-expression recognition with small sample size by transferring long-term convolutional neural network, Neurocomputing 312 (2018) 251–262.

[68] J. Li, Y. Wang, J. See, W. Liu, Micro-expression recognition based on 3d flow convolutional neural network, Pattern Anal. Appl. (2018) 1–9.