Feng Chia University
Outstanding Academic Paper by Students

Lightweight Apex-based Enhanced Network (LAENet)

for Micro-expression Recognition

Lightweight Apex-based Enhanced Network (LAENet)神經網路

進行微表情辨識

Author(s): Sung-En Lien, Yi-Chen Chiang

Class: 3rd year of Department of ISTM

Student ID: D0656129, D0759923

Course: Digital Signal Processing Application

Instructor: Sze-Teng Liong, Y.S. Gan

Department: International School of Technology and Management

Academic Year: Semester 1, 2020

Abstract

Micro-expression is an expression that reveals one's true feelings and can be potentially applied in various domains such as healthcare, safety interrogation, and business negotiation. The micro-expression recognition is thus far being judged manually by psychologists and trained experts, which consumes a lot of human effort and time. Recently, the development of the deep learning network has proven promising performance in many computer vision related tasks. Amongst, micro-expression recognition adopts the deep learning methodology to improve the feature learning capability and model generalization. This paper introduces a Lightweight Apex-based Enhanced Network (LAENet) that improves by extending one of the state-of-the-art, Shallow Triple Stream Three-dimensional CNN (STSTNet). Concretely, the network is first pre-trained with a macro-expression dataset to encounter the small data problem. The features extracted from the datasets are the optical flow guided features. Besides, an eye masking technique is introduced to reduce noise interference such as eye blinking and glasses reflection issues. The results obtained are accuracy of 79.19% and F1-score of 75.9%. Comprehensive experimentation had been conducted on the composite dataset that consists of CASME II, SMIC, and SAMM datasets. Moreover, a thorough recognition results comparison is provided by comparing it with recent methods. Detail qualitative and quantitative results are reported and discussed.

Keyword： micro-expressions, eye masking, apex, optical strain, recognition

中文摘要

微表情可表露人們隱藏的真實情緒，可運用在醫療照顧、安全審訊、商業協商上。由心理學家等相關專家辨識微表情，將耗費相當多的時間和精力，近期透過神經網路進行深度學習已證實其可靠性。這份研究報告將介紹神經網路 Lightweight Apex-based Enhanced Network (LAENet)，此為其前身 Shallow Triple Stream Three-dimensional CNN (STSTNet)的延伸研究。具體來說，神經網路會先經過較為明顯的表情資料庫訓練，這解決了微表情資料庫過小的問題。表情特徵是由 optical flow 取得。另外，運用 eye masking 來減少眨眼與眼鏡反光所造成的雜訊。最終的精確度為 79.19% F1-score 為 75.9%。全面性的實驗是運用在由 CASME II、SMIC 和 SAMM 所組成的複合資料庫。此外與近期其它研究結果進行比較，並報告與討論詳細結果。

關鍵字：微表情辨識

Table of Content

List of Figures

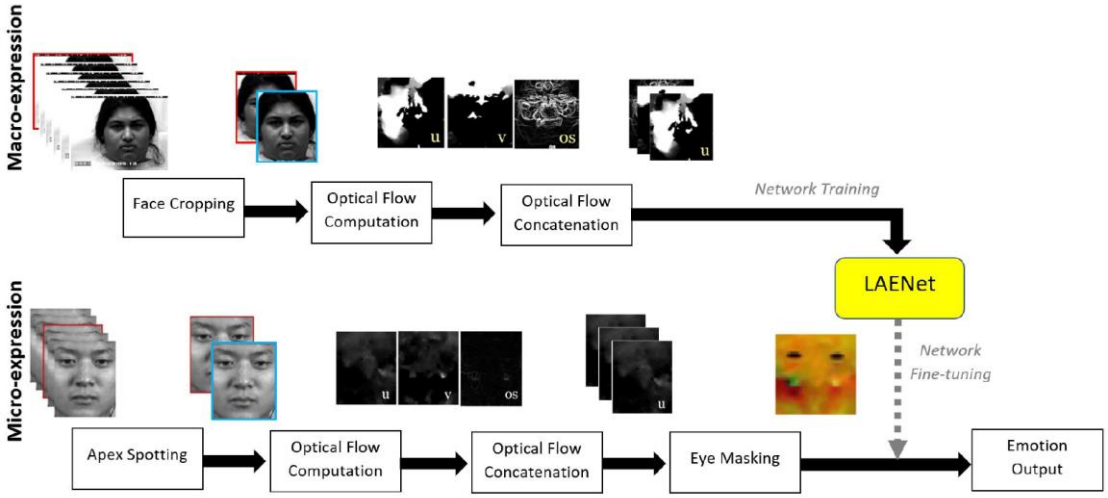

Figure 1: Flow diagram of proposed framework approach

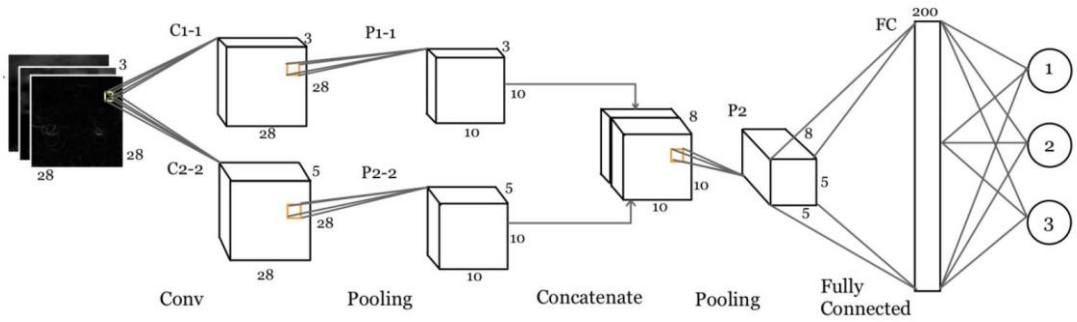

Figure 2: The proposed network architecture: LAENet, that is modified from STSTNet [5]

Figure 3: Sample image of annotated landmark points detected by DRMF

Figure 4: Example of before and after performing face cropping

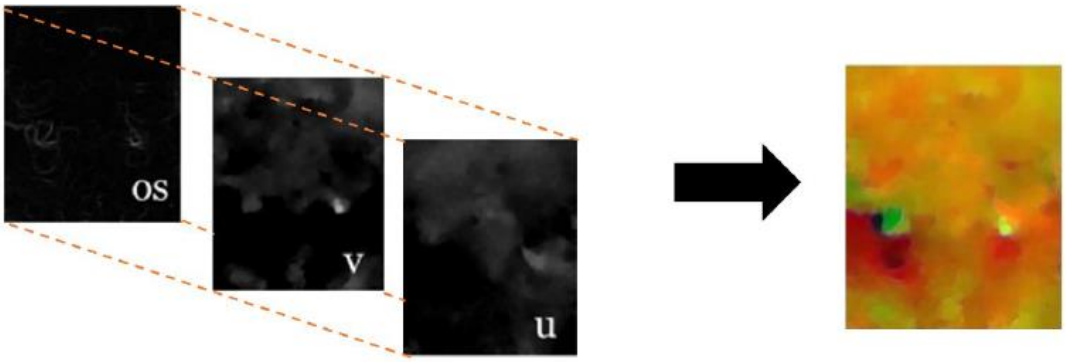

Figure 5: Concatenating the optical flow guided components of u, v, and ε to form a single color image

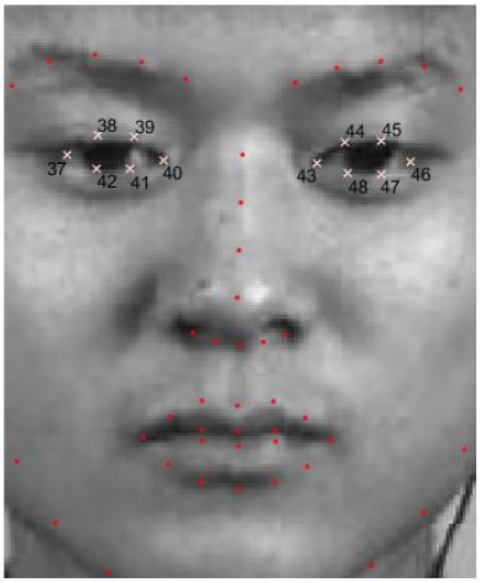

Figure 6: Landmarks annotated by using DRMF

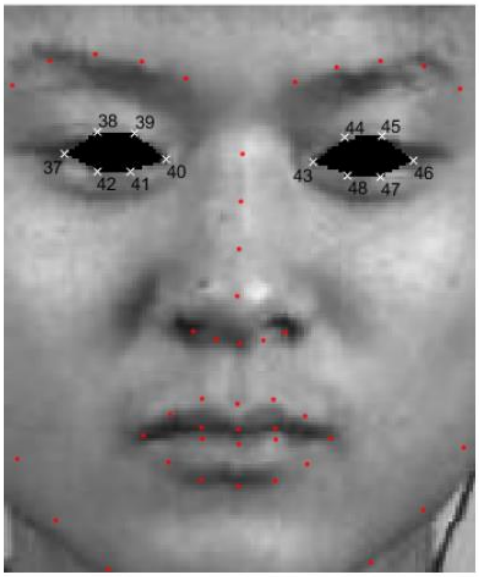

Figure 7: Sample of eye masking

List of Tables

Table 1: Network information of the proposed network and the state-of-the-art

Network	Depth	Parameter (Million)	Image Input Size
LAENet	**2**	**0.000843**	**28 x 28 x 3**
STSTNet	2	0.00167	28 x 28 x 3
OFF-ApexNet	5	2.77	28 x 28 x 2
AlexNet	8	61	227 x 227 x 3
SqueezeNet	18	1.24	227 x 227 x 3
GoogleNet	22	7	224 x 224 x 3
VGG16	16	138	224 x 224 x 3

Table 2: The configuration of LAENet includes convolutional (C)layers, pooling (P) layers, fully connected (FC) layer and outputs of max layer

Layer	Filter size	Kernel	Stride	Padding	Output size
C1-1	3 × 3 × 3	3	[1,1]	Same	28 × 28 × 3
C1-2	3 × 3 × 3	5	[1,1]	Same	28 × 28 × 5
P1-1	3 × 3	-	[3,3]	Same	10 × 10 × 3
P1-2	3 × 3	-	[3,3]	Same	10 × 10 × 5
P2	2 × 2	-	[2,2]	[0,0,0,0]	5 × 5 × 8
FC	-	-	-	-	200 × 1
Output	-	-	-	-	3 × 1

Table 3: Summary of the CASMEII, SMIC, SAMM, and CK+ datasets

Dataset		CASME II [15]	SMIC [23]	SAMM [32]	CK+ [37]
Expression Type		Micro-	Micro-	Micro-	Macro-
Subjects	Total number	24	16	28	118
	Age (mean/range)	28.1	22.03	33.24	18∼50
Camera frame rate		200	100	200	30
Expression Classes	Positive	-	51	-	-
	Negative	-	70	-	-
	Surprise	25	43	15	83
	Happiness	32	-	26	69
	Disgust	61	-	9	59
	Repression	27	-	-	-
	Anger	-	-	57	45
	Contempt	-	-	12	18
	Fear	-	-	8	25
	Sadness	-	-	6	28
	Total	145	164	133	327
Resolution (pixels)	Original	640 x 480	640 x 480	2040 x 1088	640x490 or 640x480
	Cropped	170 x 140	170 x 140	170 x 140	-
Frame Number	Average	70	34	73	18
	Maximum	126	58	101	71
	Minimum	24	11	30	6
Video duration (seconds)	Average	0.35	0.34	0.36	0.6
	Maximum	0.63	0.58	0.51	2.39
	Minimum	0.12	0.11	0.15	0.2
Ground-truth	Onset index	v	v	v	v
	Offset index	v	v	v	x
	Apex index	v	x	v	v
	Action unit	v	x	v	x

Table 4: Micro-expression recognition results when comparing to state-of-the-art

No.	Methods	Full				SMIC		CASME II		SAMM	
		Acc	F1-score	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
1	LBP-TOP [32], [23], [15]	-	-	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
2	Bi-WOOF [21]	0.6833	0.6304	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
3	OFF-ApexNet [41]	0.7460	0.7104	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
4	AlexNet [42]	0.7308	0.6959	0.6933	0.7154	0.6201	0.6373	0.7994	0.8312	0.6104	0.6642
5	SqueezeNet [43]	0.6380	0.5964	0.5930	0.6166	0.5381	0.5603	0.6894	0.7278	0.5039	0.5362
6	GoogLeNet [44]	0.6335	0.5698	0.5573	0.6049	0.5123	0.5511	0.5989	0.6414	0.5124	0.5992
7	VGG16 [45]	0.6833	0.6439	0.6425	0.6516	0.5800	0.5964	0.8166	0.8202	0.4870	0.4793
8	Part-based + Adversarial + EMR [46]	-	-	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152
9	CapsuleNet [36]	-	-	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
10	Dual-Inception [35]	-	-	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
11	STSTNet [47]	0.7692	0.7389	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
12	LAENet	**0.7919**	**0.7597**	**0.7568**	**0.7405**	**0.6620**	**0.6523**	**0.9101**	**0.9119**	**0.6814**	**0.6620**

Table 5: The confusion matrix of STSTNet on databases: Full, SMIC, CASME II and SAMM (measured by recognition rate %)

Full

	Neg	Pos	Sur
Neg	**87.60**	8.80	3.60
Pos	36.70	**56.88**	6.42
Sur	25.30	3.61	**71.08**

SMIC

	Neg	Pos	Sur
Neg	**77.14**	14.29	8.57
Pos	33.33	**58.82**	7.84
Sur	32.56	2.33	**65.12**

CASME II

	Neg	Pos	Sur
Neg	**94.32**	5.68	0
Pos	37.50	**59.38**	3.13
Sur	8	0	**92**

SAMM

	Neg	Pos	Sur
Neg	**89.13**	7.61	3.26
Pos	42.31	**50.00**	7.69
Sur	33.33	13.33	**53.33**

Table 6: The confusion matrix of LAENet on databases: Full, SMIC, CASME II and SAMM (measured by recognition rate %)

Full

	Neg	Pos	Sur
Neg	**89.60**	7.2	3.2
Pos	31.19	**61.47**	7.34
Sur	22.89	6.02	**71.08**

SMIC

	Neg	Pos	Sur
Neg	**77.14**	15.71	7.14
Pos	29.41	**62.75**	7.84
Sur	34.88	9.30	**55.81**

CASME II

	Neg	Pos	Sur
Neg	**95.45**	4.55	0
Pos	15.63	**78.13**	6.25
Sur	0	0	**100**

SAMM

	Neg	Pos	Sur
Neg	**93.48**	3.26	3.26
Pos	53.85	**38.46**	7.69
Sur	26.67	6.67	**66.67**

Table 7: Recognition accuracy with and without considering CK+dataset and adopting eye masking using LAENet architecture

		LAENet
without CK+	Without eye masking	74.43%
	With eye masking	74.66%
with CK+	Without eye masking	77.60%
	With eye masking	**79.19%**

1. Introduction

Facial expressions are ways for communication or to show one’ s emotional states without verbal form. However, facial expressions might not accurately indicate one’ s emotion from time to time since a person can intentionally try to suppress facial expressions that may reveal the emotion, they consider inappropriate [1].

The notation of micro-expressions was proposed by Haggard and Isaacs in 1966 [2] which applied on various field of research later.

The duration of the occurrence of a micro-expression is around (0.04s to 0.2s) and it usually lasts for a shorter moment compared with macro-expression which lasts between 0.75s to 2s [3].

Besides, micro-expression has lesser intensity and is less obvious.

Most importantly, the micro-expression can reveal one's genuine emotion that is trying to conceal. This finding was indicated by Ekman that micro-expression is a trail for finding lies and cheats [4].

It is interesting to highlight that suppress and mask over the original facial-expression might be accompanied by a micro-expression.

Therefore, due to the subtlety of facial motion, people often missed or misinterpreted one's actual feelings.

In the last several years, the interest in utilizing computer vision techniques in automatically recognizing micro-expression has been increased dramatically. Thus far, the approaches for micro-expression recognition can barely reach the accuracy of 75% [5], even evaluated on the datasets built in strictly fixed environment conditions. In contrast, the macro-expression recognition system [6, 7] can generate remarkable recognition accuracies ($>$90%) in many public released databases.

A basic micro-expression recognition system comprises three stages: image pre-processing, feature extraction, and emotion classification. However, most of the research groups focused on developing and enhancing the algorithm for the feature extraction stage.

For instance, the baseline or the conventional solution in obtaining the discriminant facial features is to utilize the Local Binary Pattern (LBP)-based [8] algorithms. For example, Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [9] that was widely used for dynamic texture recognition and can be utilized to describe the subtle spatio-temporal facial features. Owing to the advanced capability of the LBP-based family in extracting discriminant texture with low computational complexity and compact representation, many research groups modified the LBP method to improve

the recognition performance specifically in micro-expression applications. For instance, LBP-Six Intersection Points (LBP-SIP) [10] , Local Binary Pattern-Three Mean Orthogonal Planes (LBP-MOP) [10], Spatiotemporal Completed Local Quantized Patterns [11], Spatiotemporal Local Radon Binary Pattern (STRBP) [12] and Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection (DistSTLBP - RIP) [13].

Apart from the LBP-based features, another type of feature descriptor that has been widely used in encoding the facial motion features is known as the optical flow features. In brief, optical flow computes the muscle movement by approximating the change of the brightness intensities between consecutive frames. One of the earliest works that applied optical flow technique is [14]. They utilized the optical strain magnitude which is the derivative of the optical flow vectors and the recognition result achieved was ~5\% higher than the baseline method [15].

In the same year, Liong et al [16] considered the optical strain features on certain important facial regions, meanwhile eliminating the background noises that might generating unwanted information.

As a result, a recognition accuracy of 66.4% and is obtained when evaluated on CASME II and SMIC database, respectively.

Other recent methods that implementing optical flow components include the fuzzy histogram of optical flow orientation (FHOFO) feature [17], Main Directional Mean optical flow (MDMO) [18], Optical Strain Feature + Optical Strain Weight (OSF + OSW) [19], Fusion of Motion Boundary Histograms (FMBH) [20], and Bi-Weighted Oriented Optical Flow (Bi-WOOF) [21].

Instead of conducting extensive experiments to heuristically determining the appropriate values of the algorithm's parameters, a deep learning algorithm is capable to automatically learn the high-level features from data in an incremental manner.

The deep learning method in the micro-expression recognition system was first utilized by [22]. However, the overfitting phenomena occurred when evaluated on both the SMIC and CASME II databases.

Nevertheless, the proposed approach performed better when compared to the baseline method [23]. Besides, Li et al [24] demonstrated 3D flow-based CNN (3D-FCNN) takes in three different types of data (i.e. grayscale, the vertical and horizontal component of optical flow) as the input. Before the feed in the data into the network, a temporal interpolation model (TIM) technique is employed to standardize the frame length of each video in the dataset. On the other hand, [25] proposed the Transfer Long-term

Convolutional Neural Network (TLCNN) method which extracting the features from the frames of the micro-expression video clip by using Deep CNN and provide them to Long Short Term Memory (LSTM) which can learn the temporal sequence information of micro-expression. However, this method involves processing an enormous number of learnable parameters.

An alternative approach to tackle a different frame length of videos is to consider the motion features of one most expressive frame among the video. Thus, to identify the location of the apex frame (the frame with the highest intensity of the facial changes), [26] proposed to spot the apex frame using a D\&C-RoIs method. In short, D\&C-RoIs computes the LBP features of certain regions of interest (RoIs) such as the eye, eyebrow, and mouth areas. Then, the Divide and Conquer strategy is utilized on the proportion of the change differences to acquire the apex frame. This apex spotting method is useful especially on those datasets that do not have ground-truth apex information like SMIC database [23].

Besides, the apex spotted frames had been used in several works as an extension to the emotion recognition framework. For instance, [21] utilizes the optical flow guided components as the features that are computed from the apex and onset frames for each video. In addition, Gan et al [27] introduced

OFF-ApexNet which contains takes in the vertical and horizontal optical flow computed from the onset and apex frames as the individual input to the shallow network architecture.

Other related works that adopting the apex frame features instead of considering all the frames in the video are [28-31].

To date, the micro-expression databases that are publicly available for the algorithm evaluation are still limited.

Concisely, the databases include SMIC [23], CASME II [15], SAMM [32], CAS(ME)^2 [33]. Note that these databases are elicited under a constrained laboratory experiment setup and the participants are required to portray a poker face during the video recording. This is to avoid the acquisition of external factors that may contribute to the feature noises, such as the background and head movement. Moreover, the common characteristics among the databases are the emotions collected are all spontaneous and were recorded using a high-speed camera (i.e., >100fps). The emotion type contained in the databases is slightly different, whereby the number of classes possessed in the SMIC, CASME II, SAMM, and CAS(ME)^2 are 3, 5, 8, and 4, respectively. Since the emotions in the databases can be further categorized into three basic classes of emotion: positive, negative, and surprise, some

works [5, 34-36] attempted to fuse the databases to form a composite database. Thence, more videos with different natures are involved in the algorithm evaluation and it increases the algorithm's generality. As a result, the overall recognition result obtained for the composite database is promising (i.e., ~80\% [34]) with manageably finite of data.

This work tends to extend the method of [5] by inheriting its strengths such as high computational speed and easy implementation. Concretely, [5] designed a shallow network architecture to extract the features from the optical flow guided features that are computed from two frames (i.e., the onset and apex frames) in each video. This work introduces region-based masking on part of the face to avoid the interference of eye blinking motions. Besides, the network architecture is modified to a more compact model such as lesser learnable parameters are involved to further increase the computational speed. Moreover, motivated by the idea presented in [34], the model is pre-trained using a macro-expression dataset that poses more obvious muscle facial movements. In short, the contributions of this paper are listed as follows:

1. Application of eye masking to eliminate the eye blinking action which is not considered as an emotion.

2. Adoption of the macro-expression dataset to utilize the feature knowledge for recognition improvement in micro-expression data.
3. Improvement of state-of-the-art network architecture by reducing the number of learnable parameters, namely Lightweight Apex-based Enhanced Network (LAENet).

2. Proposed Method

The framework proposed in this work follows the main flow of [5]. The main difference of this work and [5] has been highlighted in the previous section. The flow of this work is primarily divided into two stages: 1) LAENet pre-training using macro-expression videos - To acquire the dominant motion features by learning the emotion characteristics from more obvious and noticeable muscle movements. 2) LAENet fine-tuning using micro-expression videos - To adopt the enriched features by leveraging the model on similar facial movement regions. The illustration of the proposed framework is illustrated in Figure 1 and the example of a visualization for the LAENet architecture is shown in Figure 2. Following subsections elaborates the implementation of the proposed framework in more detail.

2.1. LAENet pre-training using macro-expression videos

Referring to the work presented in [34] whereby a transfer learning technique is adopted to encode the significant domain knowledge extracted from the macro-expression videos to realize the micro-expression recognition task. Specifically, the macro-expression dataset considered in the experiment is CK+ [37]. The reason employing CK+ in this experiment is because it contains a relatively complete ground-truths such as the information of emotion label, onset index, apex index, action unit and the number of videos is sufficiently large (i.e., 327 samples). A series of image processing techniques are applied to these videos before passing the data into the LAENet. The processes involved include face detection, face cropping, optical flow computation, and feature concatenation. The details for each step are described as follows.

2.1.1. Face Detection and Face Cropping

To determine the face boundary, a face detection method is utilized, namely, Discriminative Response Map Fitting (DRMF) [38]. This detector emphasizes on a generic face fitting scenario and can be promptly used for delineating the 66 landmark points on the face with multiple angles. The example of the landmark points annotated on one of the CK+ images is shown in Figure 3. The most top, bottom, left, and right detected landmark points are

enlarged by 25, 0, 5, 5 pixels, respectively. Then the rectangle boundary with the added margin pixels are cropped out to form the face image, as illustrated in Figure 4.

2.1.2. Macro-expression Optical Flow Feature Computation

Optical flow is a technique that is capable to quantify the motion of objects in a video stream. Note that three key assumptions should be considered, which are: (1) Brightness of the same pixel should remain the same throughout each frame; (2) The object's movement is limited; (3) The object moves like their neighbors. Inspired by [21] that points out the apex frame in a video is adequate and sufficient to represent the entire video. Therefore, rather than processing all the images in the video which may involve relatively high computational time and redundancy, this work considers the apex frame for feature extraction.

The optical flow guided features can be obtained from these two images, viz, onset and apex frames. The onset frame is served as the reference frame and is assumed to have a neutral expression in this case. Mathematically, the constraint equation of the optical flow is formulated as:

$$\nabla I \cdot \vec{p} + I_t = 0, \quad (1)$$

where $I_{(x,y,t)}$ is the change of temporal image brightness in intensity value

with respect to point (x, y) and time t. $\nabla \mathrm{I} = (I_x, I_y)$ is the gradient vector that interpret the intensity of the image at (x, y) and I_t is the temporal gradient of the intensity functions. Consider the pixel location (x, y) as the point of interest is initially located, it is differentiated with respect to time to represent the horizontal and vertical movement in a change of time. Thus, the horizontal (p) and vertical (q) components of the optical flow can be expressed as follows:

$$\vec{p} = [p = \frac{dx}{dt}, q = \frac{dy}{dt}]^T, \quad (2)$$

For each video, the optical flow field that are approximated from the onset and apex frames can be molded to:

$$O_i = \{(u_{x,y}, v_{x,y}) | x = 1,2,\dots,X, y = 1,2,\dots,Y\} \quad (3)$$

where x ranges from 1 to X, indicating the width of the image, whereas y ranges from 1 to Y denoted as the height of the frame. u and v represents the horizontal and vertical components of O_i respectively.

The optical flow components can be further extended by applying a derivative function. Hence, the component known as optical strain is obtained and the typical infinitesimal strain ε is denoted as:

$$\varepsilon = \frac{1}{2}[\nabla\vec{u} + (\nabla\vec{v})^T] \quad (4)$$

where $\vec{u} = [u, v]^T$ is the displacement vector. It can also be written as:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \\ \varepsilon_{yx} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \quad (5)$$

where $(\varepsilon_{xx}, \varepsilon_{yy})$ are normal strain components and $(\varepsilon_{xy}, \varepsilon_{yx})$ are the shear strain components. A normal strain measures the change in length in either x or y direction, whereas shear strain measures the change in both the angular directions. The resultant optical strain magnitude of each pixel can be obtained by applying the square root of the sum of squares of the strain components, formulated as b

$$|\varepsilon_{x,y}| = \sqrt{\frac{\partial u^2}{\partial x} + \frac{\partial v^2}{\partial y} + \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2} \quad (6)$$

The optical flow method utilized in the experiment is TV-L1, owing to its advantages of robustness in approximating dense and subtle motion. Besides, it allows some discontinuity and tolerance for the outliers. This provides the advantage of finer noise processing while preserving flow discontinuous. In short, the optical flow technique yields these three desired motion features: u, v, and ε.

2.1.3. Macro-expression Feature Concatenation

The three optical flow guided components are equally as important as the other. They are assumed to carry the same weights independently and hence a

simple concatenation along the third dimension is performed. The sample image shown in Figure 5 clearly visualizes the resultant concatenated feature map, whereby the red, green, and blue color components are represented by the u, v, and ε features, respectively. The figure displays the majority of the pixels are in yellow indicating the image has an overall movement in both the horizontal and vertical directions (i.e., the combination of red and green is yellow). Besides, a noticeable red portion of pixels appears around the mouth corner area. This suggests that an action unit 12 is detected (a lip corner puller action), and thus the participant most likely showing a happiness emotional expression.

2.1.4. Network Training using LAENet

The optical flow motion details obtained from the macro-expression videos are served as the additional features to supplement the micro-expression recognition. Thus, the macro-expression features are first trained using a newly designed shallow network architecture. Concretely, the convolutional neural network contains two layers and is known as the Lightweight Apex-based Enhanced Network (LAENet).

Inspired by STSTNet [5] that requires a significantly less learnable parameter, LAENet is comparatively simpler and quicker. Particularly, the main

difference between LAENet and STSTNet is the removal of one of the network streams, which is from the original three streams to two streams. Thus, this greatly affects the computational complexity as the number of the learnable parameter (i.e., weights and biases) has been reduced by half, viz, from 1670 to 843. A detailed comparison regarding the key information of the state-of-the-art neural networks such as the network depth, number of parameters required, and the input size is shown in Table 1. It can be seen that compared to the popular network like SqueezeNet, LAENet is ~1500 times smaller.

The network diagram of LAENet is illustrated in Figure 2. Concisely, the input layer of the network is having the size of 28 x 28 x 3. The following layer contains two-stream convolutions with a kernel size of 3 and 5, respectively. Since the input size is relatively small, the filter size is set to 3 x 3 x 3. Then, a rectified linear unit (ReLU) action function and batch normalization layers are applied to avoid the vanishing gradients problem and accelerate the convergence of stochastic gradient descent as it selectively activates some of the neurons. Next, the activations undergo a max-pooling operation with the filter size 3 x 3. This is to downsample the image and also to prevent the over-fitting phenomenon by providing an abstracted form of the

feature representation. After that, a drop out regularization with the dropout rate of 0.5 is applied to improve the network generalization by decorrelating the weights. The activations from both the two streams are concatenated along the third dimension to form an activation of 10 x 10 x 8. Then an average pooling operation is performed to further reduce the size of features. It is noticed that an average pooling is utilized instead of a max-pooling operation. This is because an average pooling works better in preserving local features. Thus, all the meaningful and important features are reserved after passing through the process. Since this is a three-class classification task which is to categorize the emotions of `Positive', `Negative', and `Surprise', a fully connected layer with three neurons is applied as the final layer of the LAENet. To summarize the configuration of LAENet, a summary of the layers' size and the parameter is provided in Table 2.

2.2. LAENet fine-training using micro-expression videos

The following subsections focus on discussing the processes in extracting the features from micro-expression videos. Following the suggestions in MEGC 2019 [39], the databases considered in this experiment are SMIC, CASME II, and SAMM. Specifically, the processes involved are: 1) apex frame spotting - to identify the most expressive frame in each video; 2) optical

flow computation - to approximate the location and strength of the subtle muscular contraction on the face; 3) optical flow guided features concatenation - to summarize the approximated motions into a single compact image representation; 4) eye masking - to remove the unwanted motions such as eyes blinking; 5) LAENet fine-tuning - to train the LAENet by tweaking the parameters in order to adapt the unnoticeable motion characteristics. Subsections below discuss each process in great detail, justify and explain the techniques involved, and elaborate upon each step.

2.2.1. Apex Frame Spotting

Similar to the steps in macro-expression feature extraction, the apex frame is first identified to estimate the optical flow features. However, an extra image processing step is required to perform on micro-expression, viz, automatic apex frame spotting. This is because the SMIC database does not provide ground truth apex frame. Therefore, since the proposed method in this paper extends the work in [5], whereby they employed D\&C-RoIs [26] approach to acquire the apex frame. In brief, the automatic apex frame spotting contains five steps, viz: 1) For each image, the facial landmarks are annotated using the DRMF method; 2) Certain regions of interest that are useful for micro-expression are identified; 3) The LBP features of those

regions are computed; 4) For each video, a correlation coefficient formula is applied to calculate the magnitude difference of LBP features between those regions of the first frame and the subsequent frames. 5) The higher the magnitude difference, the most likely it is the apex frame. However, to avoid false detection due to the presence of noises, a peak detector that adopted the divide-and-conquer strategy is employed.

2.2.2. Micro-expression Optical Flow Feature Computation

The methods involved in the optical flow computation for videos in micro-expressions are similar to that of macro-expression, which had been described in Subsection 2.1.2. It is important to emphasize that although the same optical flow technique is utilized (i.e., TV-L1), the feature maps obtained for both the micro- and macro-expressions are greatly different from each other. Both the feature maps can be observed from Figure 1, where the top stream is processing the macro-expression videos and the bottom stream is for the micro-expression videos. It can be noticed that the [u, v, ε] for macro- is remarkably obvious compared to that of micro-.

2.2.3. Micro-expression Feature Concatenation

This feature concatenation step is the same as Subsection 2.2.3 where the [u, v, ε] features obtained from the previous step are integrated by

undergoing a simple concatenation operation along the third dimension. As such, each video can be represented using a single color-like image. Since the magnitude for all the desired expression features approximated are relatively smaller, the resultant output image might contain meaningless features caused by noises.

2.2.4. Eye Masking

Before feeding the concatenation features into the LAENet, an eye masking step is carried out to avoid the noises captured such as the eye blinking or glasses reflection that potentially contribute misleading information. Thus, the eyes are covered and motions of eyes are not considered during the feature extractions. To achieve this goal, the eye regions are identified using the landmark points annotated by DRMF. Concisely, the right and left eyes are enclosed by the landmark points of 37-42 and 43-48, respectively. Hence, two polygons can be formed by connecting the neighboring landmark points as mentioned. Figure 6 depicts the landmark points detected using DRMF and Figure 7 shows the example of the face with two binary masks on the eye regions. Note that, since the positions of the eyes are different for each participant, the masked regions differ across the dataset.

2.2.5. Network Fine-tuning using LAENet

To recognize the subtle micro-expression, the LAENet architecture is leveraged by first training the network using CK+. Thus, the model that has learned to identify the face locations of the emotion appeared is useful in adapting the pre-trained features to the micro-expressions dataset that is suffering from the limitation of data. Each feature map is resized to 28 x 28 x 3 to fit the requirement of the input layer in LAENet. Compared to training the micro-expressions video from scratch, this network fine-tuning strategy reduces training time, prevents over-fitting on a small dataset, and most importantly helps in recognition performance improvement.

3. Experiment

This section elaborates the datasets involved in the experiments, such as SMIC [23], CASME II [15], SAMM [32], and CK+ [37]. Besides, several performance metrics are introduced to verify the effectiveness of the proposed method. The metrics include Accuracy F1-score, UF1, and UAR.

3.1. Databases

Since database for micro-expression is relatively scarce, the experiment fuses SMIC [23], CASME II [15], and SAMM [32], to form a larger database.

It should be noted that all these databases were recorded in a controlled environment. For instance, SAMM and CASME II require a high-quality professional studio lighting setup. The details of the databases are summarized as tabulated in Table 3. It can be seen that all the databases suffer from class imbalanced distribution issues. There is a total of 10 emotion states exist in the databases. Amongst, the only common emotion for the databases is `Surprise'. To accomplish the database fusion, the emotions are re-categorized into three basic emotions, viz, `Positive', `Negative', and `Surprise'. Specifically, the `Positive' emotion includes the `Positive' itself and `Happiness', whereas `Negative' emotion includes 'Negative' itself, `Disgust', `Repression', `Anger', `Contempt', `Fear', and `Sadness'. The following subsections provide details on each database.

3.1.1. SMIC

The original SMIC dataset comprises three subsets: SMIC-HS, SMIC-VIS, and SMIC-NIR. Basically, the subsets are different by camera types where HS, VIS, and NIR refer to high-speed, normal visual, and near-infrared, respectively. All three cameras were placed parallel and records simultaneously. For this experiment, only SMIC-HS is considered since it was recorded at a higher frame rate (i.e., 100fps). The average video duration is

0.34 seconds. There are two types of image resolutions provided, viz, the original size 640 x 480 and the face cropped size 170 x 140. The total number of micro-expression clips contained in the SMIC-HS is 164 that is made up of 16 participants (i.e., six females and ten males). Concisely, the races include Caucasians Asians, with the mean age of 28.1 years. The ground-truth information presence is the onset and offset indices.

3.1.2. CASME II

CASME II dataset surpasses its predecessor CASME [40] dataset in both the frame rate and resolution. It provides the frame rate of 200 fps and a spatial resolution of 170 x 140 for cropped faces. This is a relatively complete dataset that contains almost all the essential ground-truth information like action unit, onset, apex, and offset indices. However, for those videos that are lacking some of the ground-truth details are eliminated from the experiment conducted herein. Besides, for the purpose of datasets merging, the emotions left are `Surprise', `Happiness', `Disgust', and `Repression'. After the emotion re-categorization, the resultant videos are `Positive' (33 samples), `Negative' (88 samples), and `Surprise' (25 samples). Note that the `Negative' emotion refers to `Disgust' and `Repression'.

3.1.3. SAMM

SAMM dataset comprises micro-movements captured at 200 fps with the initial spatial resolution of 2040 x 1088. Among the three micro-expression datasets discussed, SAMM has the highest resolution and contains the most number of participants with larger age standard deviation. In terms of ethnicity, there are 17 White British, three Chinese, two Arab, two Malay, and one each of African, Afro-Caribbean, Black British, White British / Arab, Indian, Nepalese, Pakistani, and Spanish. Nevertheless, the number contains the same number of males and females. Similar to CASME II, the emotions are re-categorized into the three basic emotions, where the `Negative' emotion comprises videos of `Disgust', `Anger', `Contempt', `Fear', and `Sadness'. The number of resultant videos for each emotion are `Positive' (26 samples), `Negative' (92 samples), and `Surprise' (15 samples).

3.1.4. CK+

Different from the three datasets mentioned above, CK+ contains videos of macro-expression. As observed in Table~\ref{table:database}, the average video duration of this dataset is 0.6 seconds, which is about doubled of the micro-expressions. However, the average frame number in this dataset is only 18 frames. This is because the camera frame rate is 30fps, about one-seventh compared to SAMM and CASME II datasets. There is a total of 327 videos

that are made up of 7 types of emotions, viz, `Surprise' (83 samples), `Happiness' (69 samples), `Disgust' (59 samples), `Anger' (45 samples), `Contempt' (18 samples), `Fear' (25 samples), and `Sadness' (28 samples). During the emotion elicitation, the participants are asked to portray specific basic emotional expressions. Thus, CK+ is a type of posed expression database and the muscle movements are larger.

3.2. Performance Metric

Apart from the common performance metric *Accuracy*, other types of metric adopted in the method evaluation are *F1-score*, Unweighted F1-score (*UF1*) and Unweighted Average Recall (*UAR*). This is mainly to tackle the multi-class classification problem especially in a class imbalanced database. The equations for the metrics are defined as follows:

$$Accuracy := \frac{\sum_{i=1}^{M}\sum_{j=1}^{k} TP_i^j}{\sum_{i=1}^{M}\sum_{j=1}^{k} TP_i^k + \sum_{i=1}^{M}\sum_{j=1}^{k} FP_i^k} \quad (7)$$

$$F1-score := \frac{2 \times Percision \times Recall}{Precision + Recall} \quad (8)$$

such that

$$Recall := \sum_{i=1}^{M} \frac{\sum_{j=1}^{k} TP_i^j}{M \times \sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FN_i^j} \quad (9)$$

$$Precision := \sum_{i=1}^{M} \frac{\sum_{j=1}^{k} TP_i^j}{M \times \sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FP_i^j} \quad (10)$$

$$UF1 := 2 \times \frac{\sum_{i=1}^{M} \frac{Precision_i \times Recall_i}{Precision_i + Recall_i}}{M} \quad (11)$$

such that

$$Precision_i := \frac{\sum_{j=1}^{k} TP_i^j}{\sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FP_i^j} \quad (12)$$

$$Recall_i := \frac{\sum_{j=1}^{k} TP_i^j}{\sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FN_i^j} \quad (13)$$

$$UAR := \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{j=1}^{k} TP_i^j}{M \times \sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FN_i^j} \quad (14)$$

where M is the number of classes; TP, FN, and FP are true positive, false negative, and false positive, respectively. Following the conventional fashion to evaluate in the micro-expression recognition system, the experiment employs a leave-one-subject-out cross-validation (LOSOCV) protocol for the three-class classification.

4. Result

The recognition performance of the proposed LAENet is compared to several existing methods that had been examined on the composite database that consists of 442 videos from three databases (i.e., SMIC, CASME II, and SAMM). The results comparison tabulated in Table 4.The methods can be categorized into four main groups the baseline methods (#1 LBP-TOP [15, 23,

32]), apex-based method (#2 Bi-WOOF [21] and #3 OFF-ApexNet [27]), popular networks (#4 AlexNet [42], #5 SqueezeNet [43], #6 GoogLeNet [44], and #7 VGG16 [45]), methods participated in MEGC 2019 [39] (#8 Part-based + Adversarial + EMR [46], #9 CapsuleNet [36], #10 Dual-Inception Network [30], and #11 STSTNet [5]). Note that the results for methods #1-#7 are reproduced by following the steps described in the corresponding papers.

In overall, LAENet outperforms all methods except method #8. Succinctly, method #8 adopts the CK+ database to increase the size of the data sample. Then, adversarial training and motion magnification is performed to improve the visualization of muscle movement intensity. Besides, it is important to highlight that method #8 won the first prize in the recognition challenge in MEGC 2019. Concisely, the accuracy, F1-score, UF1, and UAR attained for LAENet are 79.19%, 73.89%, 73.53%, and 76.05%.

To further analyze this recognition task, the confusion matrices for both the STSTNet and LAENet methods are provided in Table 5 and Table 6, respectively. The common trend in both the confusion matrices is SMIC database produces a relatively lower recognition score compared to CASME II and SAMM. This may be due to the inaccurate apex frame of SMIC that is

estimated using the D&C-RoIs strategy. In addition to that, the camera used to capture the SMIC database is has the lowest frame rate, (i.e., 100fps). Therefore, the desired apex frame might not be captured and collected. Besides, there is still room for improvement when classifying the surprise emotion in the SMIC database, as it dropped about 10$\%$ when comparing LAENet to STSTNet methods. Nevertheless, LAENet maintains its classification capability in negative class and outperforms by ~4% for positive class.

In contrast, CASME II exhibits the best result, as it is capable to achieve 100% in surprise emotion for LAENet. It seems that applying the CK+ dataset, which has a decent amount of surprise expression, could make up the imbalanced database to some degree. As for the SAMM dataset, an increase in results is seen in recognizing the negative and surprise emotions. Despite the increase, there is an 11.54% drop for positive expressions.

On the other hand, it is observed that among the three emotions, negative emotion generates the highest result across all the databases.
This is because of the class imbalance problem. As discussed in Section 3, the emotion distribution of the negative/surprise/positive is 70/51/43, 88/25/33, and 92/15/26 for the SMIC, CASME II, and SAMM databases, respectively.

Thus, this unequal distribution of classes in the training dataset causes the network tends to be more biased towards the majority class.

To inspect the impact of adopting the CK+ database and the eye masking technique, the recognition performance is summarized in Table 7. The table shows that with the eye masking technique, the result does not differ much (i.e., <0.3%) when CK+ is not involved in the LAENet network training progress. However, the result difference becomes noticeable when adopting CK+ in the experiment. Concretely, the accuracy increases by 1.6% when the eye masking technique is applied, which is from 77.6% to 79.19%. As a whole, the adoption of CK+ and eye masking improves the result by ~5\% when using the same experimental setup and fixed network configuration.

In summary, this experiment demonstrates the effectiveness of the proposed method by integrating several image processing techniques and the positive impact in enhancing the network architecture to extract meaningful features. Besides, this paper has discussed the eye blinking interference and how it influences the recognition performance. Our future endeavors will be concentrated on empirical verification as well as validation of the framework when adopted on other spontaneous macro-expression datasets. Moreover, the implementation of a real-time recognition system is expected to be analyzed

in the near future. Most importantly, the development of the system in real-world applications is desired that contains a variety of different backgrounds.

5. Conclusion

In short, this paper presents an end-to-end system in recognizing the micro-expression by utilizing the optical flow guided features. The proposed method extends the state-of-the-art network namely STSTNet by introducing LAENet network architecture. As a result, the recognition performance yields an accuracy 79.91% and F1-score of 75.97% when evaluated on three spontaneous micro-expression datasets (SMIC, CASME, and SAMM). These results outperform STSTNet by ~2%.

The promising result is mainly contributed to the enrichment of the network that had been previously trained by the CK+ dataset that has more obvious expression motions. Besides, the newly proposed network has been modified into a smaller network with lower computational complexity by removing one of the streams. In addition, it is also important to emphasize the utilization of the apex frame to represent the entire video supplements in the feature extraction stage.

6. Acknowledgement

This work was funded by Ministry of Science and Technology (MOST)

(Grant Number: 109-2221-E-035-065-MY2, 108-2218-E-009-054-MY2,

MOST 108-2218-E-035-007-, and 108-2218-E-227-002-.).

References

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

[2] E. A. Haggard, K. S. Isaacs, Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy, in: Methods of research in psychotherapy, Springer, 1966, pp. 154 – 165.

[3] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion., Journal of personality and social psychology 17 (2) (1971) 124.

[4] P. Ekman, M. O' Sullivan, Who can catch a liar?, American psychologist 46 (9) (1991) 913.

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

[6] G. U. Kharat, S. V. Dudul, Emotion recognition from facial ex-pression using neural networks, in: Human-Computer Systems Interaction, Springer, 2009, pp. 207 – 219.

[7] A. R. Rivera, J. R. Castillo, O. O. Chae, Local directional number pattern for face analysis: Face and expression recognition, IEEE transactions on image processing 22 (5) (2012) 1740 – 1752.

[8] T. Ojala, M. Pietik □ainen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern recognition 29 (1) (1996) 51 – 59.

[9] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence29 (6) (2007) 915 – 928.

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

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

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

[13] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietik □ainen, Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition, IEEE Transactions on Affective Computing 10 (1)(2017) 32 – 47.

[14] S.-T. Liong, R. C.-W. Phan, J. See, Y.-H. Oh, K. Wong, Optical strain based recognition of subtle emotions, in: 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), IEEE, 2014, pp. 180 – 184.
[15] W.-J. Yan, S.-J. Wang, G. Zhao, X. Li, Y.-J. Liu, Y.-H. Chen,X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, PLoS ONE 9 (2014)e86041.doi:10.1371/journal.pone.0086041.
[16] 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.
[17] S. Happy, A. Routray, Fuzzy histogram of optical flow orientations for micro-expression recognition, IEEE Transactions on Affective Computing.
[18] 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 Transactions on Affective Computing 7 (4) (2015) 299 – 310.
[19] 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 Processing: Image Communication 47 (2016) 170 – 182.
[20] H. Lu, K. Kpalma, J. Ronsin, Motion descriptors for micro-expression recognition, Signal Processing: Image Communication 67 (2018) 108 – 117.
[21] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Processing: Image Communication 62 (2018) 82 – 92.
[22] D. Patel, X. Hong, G. Zhao, Selective deep features for micro-expression recognition, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2258 – 2263.
[23] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietik □ainen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),IEEE, 2013, pp. 1 – 6.
[24] J. Li, Y. Wang, J. See, W. Liu, Micro-expression recognition based on 3d flow convolutional neural network, Pattern Analysisand Applications 22 (4) (2019) 1331 – 1339.
[25] 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.
[26] S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, R. Phan, Automatic apex frame spotting in micro-expression database, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 665 – 669.

[27] Y. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, Off-apexnet on micro-expression recognition system, Signal Processing: Image Communication 74 (2019) 129 – 139.
[28] 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.
[29] 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.
[30] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, W. Lin, Dual-stream shallow networks for facial micro-expression recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 36 – 40.
[31] S.-T. Liong, Y. Gan, D. Zheng, S.-M. Li, H.-X. Xu, H.-Z. Zhang,R.-K. Lyu, K.-H. Liu, Evaluation of the spatio-temporal features and gan for micro-expression recognition system, Journal of Signal Processing Systems (2020) 1 – 21.
[32] A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE Transactions on Affective Computing 9 (1) (2016) 116 – 129.
[33] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, X. Fu, Cas (me)ˆ2: A database for spontaneous macro-expression and micro-expression spotting and recognition, IEEE Transactions on Af-fective Computing 9 (4) (2017) 424 – 436.
[34] Y. Liu, H. Du, L. Zheng, T. Gedeon, A neural micro-expression recognizer, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1 – 4.
[35] Q. M. L. Zhou, L. Xue, Dual-inception network for cross-database micro-expression recognition, Automatic Face Ges-ture Recognition.
[36] J. C. N. V. Quang, T. Tokuyama, Capsulenet for micro-expression recognition, Automatic Face Gesture Recognition.
[37] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010, pp. 94 – 101.
[38] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3444 – 3451.
[39] J. See, M. H. Yap, J. Li, X. Hong, S.-J. Wang, Megc 2019 – the second facial micro-expressions grand challenge, in: 2019 14thIEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1 – 5.

[40] 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: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition(FG), IEEE, 2013, pp. 1 – 7.
[41] S.-T. Liong, Y. Gan, W.-C. Yau, Y.-C. Huang, T. L. Ken, Off-apexnet on micro-expression recognition system, arXiv preprintarXiv:1805.08699.
[42] I. S. A. Krizhevsky, G. E. Hinton., Imagenet classification with deep convolutional neural networks., Advances in NIPS (2012)1097 – 1105.
[43] M. W. M. K. A. W. J. D. F. N. Iandola, S. Han, K. Keutzer., Squeezenet: Alexnet-level accuracy with 50x fewer parameters andi 0.5 mb model size., arXiv.
[44] Y. J. P. S. S. R. D. A. D. E. V. V. C. Szegedy, W. Liu, A. Ra-binovich., Going deeper with convolutions., Proc. of the IEEECVPR (2015) 1 – 9.
[45] K. Simonyan, A. Zisserman., Very deep convolutional networks for large-scale image recognition., arXiv.
[46] L. Z. Y. Liu, H. Du, T. Gedeon, A neural micro-expression recognizer, Automatic Face Gesture Recognition.
[47] J. S. H.-Q. K. S.-T. Liong, Y. Gan, Y.-C. Huang, Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition, Automatic Face Gesture Recognition.