A Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition System

Sze-Teng Liong[1] and Y.S. Gan[2],* and John See[3] and Huai-Qian Khor[4]
[1] Department of Electronic Engineering, Feng Chia University, Taichung 40724, Taiwan R.O.C.
[2] Department of Mathematics, Xiamen University Malaysia, Jalan Sunsuria,
Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia
[3,4]Faculty of Computing and Informatics Multimedia University, Persiaran Multimedia Cyberjaya, Malaysia

***Abstract*— In the recent year, the state-of-the-arts of facial micro-expression recognition task have been significantly advanced by the emergence of data-driven approaches based on deep learning. Due to the superb learning capacity of deep learning, it generates promising performance beyond the traditional handcrafted approaches. Recently, many researchers have focused on developing better networks by increasing its depth, as deep networks can effectively approximate certain function classes more efficiently than shallow ones. In this paper, we aim to design a shallow network to extract the high level features of the micro-expression details. Specifically, a two-layer neural network, namely Shallow Triple Stream Three-dimensional CNN (STSTNet) is proposed. The network is capable to learn the features from three optical flow features (i.e., optical strain, horizontal and vertical optical flow images) computed from the onset and apex frames from each video. Our experimental results demonstrate the viability of the proposed STSTNet, which exhibits the UAR recognition results of 76.05%, 70.13%, 86.86% and 68.10% in composite, SMIC, CASME II and SAMM databases, respectively.**

I. Introduction

Facial expression is a form of nonverbal communication by contracting the muscular patterns on the face in reflecting ones emotional state. Different muscular movements and patterns eventually reflect different types of emotions. However, the expressions portray on the face may not accurately implies ones emotion state as its can be faked easily. Among several types of nonverbal communications (i.e., facial expression, vocal intonation and body posture), micro-expression (ME) is discovered to be more likely to reveal ones deepest emotions [3]. Since the ME is stimulated involuntary, it allows the competent in exposing ones concealed genuine feelings without deliberately control. In contrast to the facial macro-expression, which normally lasts between 0.75s to 2s, micro-expression usually occurring with less duration (0.04s to 0.2s) and lower intensity [4].

In recent years, there has been an increasing interest in computer vision aspect in developing robust automated ME recognition systems. The state-of-the-art performance for ME recognition task are less than 70% [1], [12], even the methods are tested on the datasets that are elicited in one constrained laboratory condition. In contrast, the macro-expression (a.k.a. normal expression) recognition systems can exhibit almost 100% perfect recognition accuracy [9], [25]. All ME videos are captured using high frame rate cameras (i.e., $>100 fps$) and there exists a lot of redundant frames, it is essential to eliminate interference from unrelated facial information, in the meantime emphasizing important characteristics of ME. Temporal Interpolation Method (TIM) [34] is one of the techniques used in ME systems to address the problem of different video length [12], [29], [24]. It normalizes the length of all image sequence into a certain frame number, either through downsampling or upsampling. TIM has been adopted by original ME databases (i.e., CASME II and SMIC) in order to standardize the frame length before performing the feature extraction. Liong et al. [20], [15] proposed to identify only the single apex frame (i.e., the frame with highest emotion intensity). They demonstrate that it is sufficient to encode facial micro-expression features by utilizing the apex and onset (first frame of ME video) frames.

For feature extraction, many researchers propose the algorithms based on Local Binary Pattern (LBP) [22]. The LBP family includes, Local Binary Pattern on Three-Orthogonal Planes (LBP-TOP) [33], Local Binary Pattern with Six Intersection Points (LBP-SIP) [30], Local Binary Pattern with Mean Orthogonal Planes (LBP-MOP) [31] and Spatiotemporal Completed Local Quantization Patterns (STCLQP) [7]. LBP is a popular appearance-based feature extraction methods due to its characteristics of discrimination ability, compact texture representation and low computational complexity. Apart from LBP variants, there exists an optical flow family, that estimates the motion of objects from frame to frame, based on the brightness patterns in the frame and it is capable to capture the tiny facial muscle movements. For example, Optical Strain Feature (OSF) [16], Optical Strain Weight(OSW) [17], Fuzzy Histogram of Oriented Optical Flow (FHOOF) [6], Bi-Weighted Oriented Optical Flow (Bi-WOOF) [20] and Main Directional Mean optical flow (MDMO) [21].

For the deep learning methods, one of the earliest works that adopts convolutional neural network (CNN) is carried out by [23]. However, their method do not outperform the conventional methods as the architecture designed is likely being overfitted due to limited sample size. On the other hand, [14] applies a VGG-Face model on the apex frame for each video sequence then fine-tunes the weights of the network with the small scale data (i.e., CASME II). The recognition accuracy reported is $\sim$63% in CASME II in

* Corresponding author: Email addresses:ysgan@xmu.edu.my (Y.S. Gan)

leave one-subject-out cross validation (LOSOCV) protocol, but the number of learnable parameters (weights and biases) in the network is very large (i.e., 138 million). On the other hand, Wang et al. [28] adopt CNN architecture and Long Short Term Memory (LSTM) to learn the spatial-temporal information for each image frame. The total number of learnable parameters in the CNN network is about 80 million. Prior to passing the image data into the designed architecture, a TIM technique is applied to each video sequence, in order to fix the frame length to either 32 or 64. Besides, a three-stream CNN network is proposed by Li et al. [11], where each stream takes in the grayscale frame, the horizontal and vertical optical flow field, respectively. The recognition results of the proposed architecture for CASME II database is performed as good as many recent methods (∼60%) [7], [20]. However, it did not show the effectiveness in SMIC, where the recognition accuracy is ∼55%.

To the best of our knowledge, [15] is the first work that perform cross-dataset validation on three distinct databases (i.e., CASME II, SMIC, SAMM). Succinctly, they suggest a three-step framework: 1) Apex frame acquisition from each video; 2) Computation for optical flow guided features (i.e., horizontal and vertical optical flow images) from the apex and onset frames; 3) Features fusion and enhancement using OFF-ApexNet architecture. Hence, motivated by [15], this paper aims to improve the recognition performance by simplifying the neural network but has sufficient capacity to learn the real structure of the ME details. The main contributions of this paper include:

1) Proposal of a relatively small and shallow neural network whilst remaining the effectiveness in generating rich and discriminative features representation.
2) Feature extraction from three optical flow information (i.e., optical strain, horizontal and vertical optical flow).
3) Re-implementation on several state-of-the-art methods and provide certain analysis based on the quantitative experimental observations.

II. Proposed Method

While many architectures proposed in the literature relied on increasing the number of neurons or increasing the number of layers to allow the network to learn more complex functions, this paper attempts to present a shallow neural network architecture that comprises two learnable layers. Similar to [15], the proposed micro-expression recognition system consists of three components, namely, apex frame spotting, optical flow features computation and features enhancement with CNN. The overview of the recognition system is illustrated in Fig. 1.

Firstly, the apex frame spotting stage is to identify the index frame that consists the highest intensity of ME in a video sequence. Since SMIC database does not provide the ground-truth apex frame, the D&C-RoIs [19] approach is directly employed to obtain the apex frame location. D&C-RoIs has been utilized by several ME works such as in [20], [15], [18], [5] as it facilitates to produce reasonably

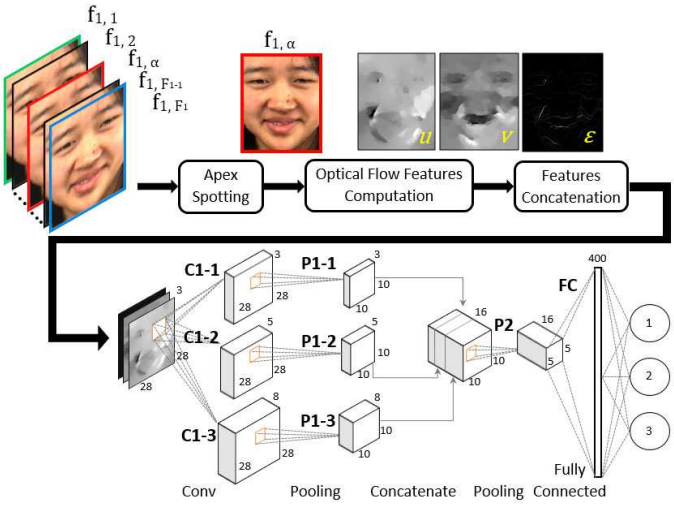

Fig. 1: Flow diagram of proposed STSTNet approach

good performance on ME recognition task. In brief, the D&C-RoIs analyzes the difference between local appearance-based features of sequential frames, whereby the features are computed using the LBP descriptor. Then, a Divide & Conquer strategy is utilized and search the frame in which the maximum facial muscle changes occurs.

For clarity, we define some notation frequently used in this paper. A ME video sequence is denoted as:

$$S = [s_1, s_2, ..., s_n], \tag{1}$$

where n is the number of video clips. The i-th of the video refers to:

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

where F_i is the total number of images in the i-th sequence. Note that, each video only contain one onset (begin) frame, one apex (maximum) frame and one offset (end) frame. The onset, apex and offset frames are indicated as $f_{i,1}$, $f_{i,\alpha}$ and f_{i,F_i}, respectively. The apex frame is expressed as:

$$f_{i,\alpha} \in f_{i,1}, \ldots, f_{i,Fi}. \tag{3}$$

Note that, $f_{i,\alpha}$ in SMIC is obtained after applying the D&C-RoIs approach, whereas for CASME II and SAMM are exploiting the ground-truths apex frame.

Next, several optical flow guided features are computed from the onset and apex frames. The optical flow map that computed from the two frames (i.e., onset and apex) can be formulated as:

$$O_i = \{(u(x, y), v(x, y)) | x = 1, 2, ..., X, y = 1, ..., Y\}, \tag{4}$$

where X and Y denote the width and height of the images, $f_{i,j}$, respectively. $u(x, y)$ and $v(x, y)$ represent the horizontal and vertical component of O_i, respectively. Optical strain is capable to approximate the deformation intensity and can be defined as:

$$\varepsilon = \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T], \tag{5}$$

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

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \varepsilon_{yx} = \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix}, \quad (6)$$

where the diagonal strain components, $(\varepsilon_{xx}, \varepsilon_{yy})$, are normal strain components and $(\varepsilon_{xy}, \varepsilon_{yx})$ are shear strain components. The optical strain magnitude for each pixel can then be computed by taking the sum of squares of the normal and shear strain components, such that:

$$|\varepsilon_{x,y}| = \sqrt{\frac{\partial u}{\partial x}^2 + \frac{\partial v}{\partial y}^2 + \frac{1}{2}(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial x})^2}. \quad (7)$$

To summarize, each video can be derived into the following three optical flow derived representations:

1) u - Horizontal component of the optical flow field O_i,
2) v - Vertical component of the optical flow field O_i,
3) ε - Optical strain

The final step is to further process the optical flow derived features with a shallow triple stream 3D-CNN. The three optical flow features are concatenated to form a single 3D image, then resize to 28×28×3. Then the image is pass through a one-layer three-stream that comprised of a convolutional layer (with different number of kernel) and a maximum pooling layer in each stream. This is because to deal with the small scale input data, we utilize three 3×3 different number of kernels number in order to avoid the underfit phenomena. The maximum pooling operation is used to highlight dominant features while eliminating redundancy. Next, the output are concatenated to form a resultant 3D block of features. Lastly, an average pooling is applied and a fully connected is performed before classifying the three ME emotion states. The average pooling operation treats every pixel equally important and it could retain more high level information.

The exact network configurations we used on STSTNet are shown in Table I. The initial learning rate is set to 0.00005 and a epoch value of 500.

III. Experiment

A. Database

There are three databases that are commonly used in the ME research specifically in the computer vision field, namely, SMIC, CASME II and SAMM. The detailed information of the three databases is shown in Table II. It is observed that the databases are not comprehensive and there is an imbalanced distribution of samples per emotion. For generalization, the video samples of three databases are fused together to perform the feature extraction and classification.

B. Performance Metric

As performance metrics for quantitative evaluation for our imbalance data, we use accuracy, F1-score, Unweighted F1-score (UF1) and Unweighted Average Recall (UAR), individually for each class. Mathematically,

$$\text{Accuracy} := \frac{\sum_{i=1}^{M}\sum_{j=1}^{k} TP_i^j}{\sum_{i=1}^{M}\sum_{j=1}^{k} TP_i^k + \sum_{i=1}^{M}\sum_{ij=1}^{k} FP_i^k} \quad (8)$$

$$\text{F1-score} := \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

such that

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

and

$$\text{Precision} := \sum_{i=1}^{M} \frac{\sum_{j=1}^{k} TP_i^j}{M \times \sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FP_i^j} \quad (11)$$

$$\text{UF1} := 2 \times \frac{\sum_{i=1}^{M} \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}}{M} \quad (12)$$

such that

$$\text{Precision}_i := \frac{\sum_{j=1}^{k} TP_i^j}{\sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FP_i^j} \quad (13)$$

and

$$\text{Recall}_i := \frac{\sum_{j=1}^{k} TP_i^j}{\sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FN_i^j} \quad (14)$$

$$\text{UAR} := \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{j=1}^{k} TP_i^j}{M \times \sum_{j=1}^{k} TP_i^j + \sum_{j=1}^{k} FN_i^j} \quad (15)$$

where M is the number of classes; TP, FN and FP are the true positive, false negative and false positive, respectively.

All the results presented are evaluating on a three-class recognition task and leave-one-subject-out cross validation (LOSOCV) protocol is utilized in the classifier. Hence, the M we used is 3 and k is 68 (68 subjects).

IV. Results and Discussion

The benchmark method (i.e., LBP-TOP [33]), state-of-the-art methods (i.e., Bi-WOOF [20] and OFF-ApexNet [15]) and some popular deep learning methods (i.e., AlexNet [10], Squeezenet [8], GoogLeNet [27], VGG16 [26]) are reproduced to compare with the proposed method, as tabulated Table III. *Full* refers to the combination of the three databases. It is observed that, the proposed STSTNet outperforms other methods in all the scenarios, except the case of OFF-ApexNet in CASME II. Overall, the STSTNet approach produced an average improvement (compared to LBP-TOP) of approximately 15%, 48%, 14%, 26% for composite, SMIC, CASME II and SAMM databases, respectively.

From the confusion matrix in Table IV, it can be seen that the proposed method is good in distinguishing the negative emotion. This might due to there are more than half of the video (250/ 442) belongs to negative sample. In addition, a high frame rate camera was used in CASME II to record the subtle motion. Therefore, it is expected that the exact apex frame is being captured and thus leads to more accurate

optical flow computation that better describe the motion changes. For SMIC database, it exhibits lower recognition performance compared to CASME II, because the apex spotting technique (i.e., D&C-RoIs) is not really robust, which has a mean absolute error of $\sim$13 frames [19]. Besides, the images may include some background noises such as the shadows, highlights, illumination, flickering lights due to the database elicitation setup. Note that, SMIC images are recorded using a relatively low frame rate camera ($100fps$). As for SAMM database, Table IV shows that it can perform very well for negative ($\sim$90%) and about 50% for the positive and surprise emotions. This may because SAMM has severe imbalance data issue, whereby the surprise and positive video samples occupy 10% and 20% of the entire database.

Table V summarizes the properties of all the neural networks mentioned in Table III, include: 1) Depth - the largest number of sequential convolutional or fully connected layers in an end-to-end neural network; 2) Learnable parameters - the weights and biases of the network; 3) Image input size - the size of the input image to the network; 4) Execution time - The training and testing time for a single fold in LOSOCV protocol. For example, to test the first subject (i.e., *sub01*) in CASME II, there are 3 testing samples and 439 training samples. The STSTNet has the least network depth (i.e., 2), learnable parameters size (i.e., 1670 of weights and biases) and computational time (i.e., 5.7s to train the model and examine on the test data).

V. Conclusion

This paper presents a shallow triple stream three-dimensional CNN (STSTNet) to extract the optical flow guided images. A compact and discriminative feature representation is constructed from three optical flow images (i.e., horizontal optical flow, vertical optical flow and optical strain map). Overall, the proposed STSTNet approach demonstrated promising recognition results on three spontaneous micro-expression databases. STSTNet is capable to yield good results of 76.05%, 70.13%, 86.86% and 68.10% for UAR in composite, SMIC, CASME II and SAMM databases, respectively. For the future work, the apex spotting technique can be improved to extract more accurate and meaningful motion data. Furthermore, other variation of optical flow guided features (such as magnitude and orientation) can be considered to include as the input for the neural network.

VI. Acknowledgments

This work was funded by Ministry of Science and Technology (MOST) (Title: Micro-expression Analysis and Evaluation based on Convolutional Neural Network, Grant Number: MOST 107-2218-E-035-006-).

TABLE I: STSTNet configuration for the convolution (*C*) layers, pooling (*P*) layers, fully connected (*FC*) layer and output layer

Layer	Filter size	Kernel	Stride	Padding	Output size
C1-1	$3 \times 3 \times 3$	3	[1,1]	Same	$28 \times 28 \times 3$
C1-2	$3 \times 3 \times 3$	5	[1,1]	Same	$28 \times 28 \times 5$
C1-3	$3 \times 3 \times 3$	8	[1,1]	Same	$28 \times 28 \times 8$
P1-1	3×3	-	[3,3]	Same	$10 \times 10 \times 3$
P1-2	3×3	-	[3,3]	Same	$10 \times 10 \times 5$
P1-3	3×3	-	[3,3]	Same	$10 \times 10 \times 8$
P2	2×2	-	[2,2]	[0,0,0,0]	$5 \times 5 \times 16$
FC	-	-	-	-	400×1
Output	-	-	-	-	3×1

TABLE II: Detailed information of three micro-expression databases

		CASME II	SMIC	SAMM
Participants		24	16	28
Samples		145	164	133
Frame rate (*fps*)		200	100	200
Cropped image resolution (pixels)		170×140	170×140	170×140
Frame number	Average	70	34	73
	Maximum	126	58	101
	Minimum	24	11	30
Video duration (*s*)	Average	0.35	0.34	0.36
	Maximum	0.63	0.58	0.51
	Minimum	0.12	0.11	0.15
Expression	Negative	88	70	92
	Positive	32	51	26
	Surprise	25	43	15
Ground-truth provided	Onset index	✓	✓	✓
	Offset index	✓	✓	✓
	Apex index	✓	✗	✓

TABLE III: Comparison of micro-expression recognition performance in terms of Accuracy (*Acc*), F1-score, Unweighted F1-score (*UF1*) and Unweighted Average Recall (*UAR*) on the composite (*Full*), CASME II, SMIC and SAMM databases

No.	Methods	Full				SMIC		CASME II		SAMM	
		Acc	F1-score	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
1	LBP-TOP [2], [13], [32]	-	-	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
2	Bi-WOOF [20]	0.6833	0.6304	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
3	AlexNet [10]	0.7308	0.6959	0.6933	0.7154	0.6201	0.6373	0.7994	0.8312	0.6104	0.6642
4	SqueezeNet [8]	0.6380	0.5964	0.5930	0.6166	0.5381	0.5603	0.6894	0.7278	0.5039	0.5362
5	GoogLeNet [27]	0.6335	0.5698	0.5573	0.6049	0.5123	0.5511	0.5989	0.6414	0.5124	0.5992
6	VGG16 [26]	0.6833	0.6439	0.6425	0.6516	0.5800	0.5964	0.8166	0.8202	0.4870	0.4793
7	OFF-ApexNet [15]	0.7460	0.7104	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
8	**STSTNet**	**0.7692**	**0.7389**	**0.7353**	**0.7605**	**0.6801**	**0.7013**	**0.8382**	**0.8686**	**0.6588**	**0.6810**

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

(a) 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**

(b) 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**

(c) CASME II

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

(d) 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 V: Properties of the neural networks

No.	Network	Depth	Parameter (Million)	Image Input Size	Execution Time (s)
1	**STSTNet**	**2**	**0.00167**	**28 $\times$ 28 $\times$ 3**	**5.7366**
2	OFF-ApexNet [15]	5	2.77	28 $\times$ 28 $\times$ 2	5.5632
3	AlexNet [10]	8	61	227 $\times$ 227 $\times$ 3	12.9007
4	SqueezeNet [8]	18	1.24	227 $\times$ 227 $\times$ 3	14.3704
5	GoogLeNet [27]	22	7	224 $\times$ 224 $\times$ 3	29.3022
6	VGG16 [26]	16	138	224 $\times$ 224 $\times$ 3	95.4436

REFERENCES

[1] B. Allaert, I. M. Bilasco, and C. Djeraba. Consistent optical flow maps for full and micro facial expression recognition. In *VISIGRAPP (5: VISAPP)*, pages 235–242, 2017.

[2] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, 2016.

[3] P. Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009.

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

[5] Y. Gan and 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)*, pages 168–172. IEEE, 2018.

[6] S. L. Happy and A. Routray. Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing*, 99, 2017.

[7] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 175:564–578, 2016.

[8] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[9] G. U. Kharat and S. V. Dudul. Emotion recognition from facial expression using neural networks. In *Human-Computer Systems Interaction*, pages 207–219. Springer, 2009.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] J. Li, Y. Wang, J. n. See, and W. Liu. Micro-expression recognition based on 3d flow convolutional neural network. *Pattern Analysis and Applications*, pages 1–9, 2018.

[12] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 9(4):563–577, 2018.

[13] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition*, pages 1–6, April 2013.

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

[15] S.-T. Liong, Y. Gan, W.-C. Yau, Y.-C. Huang, and T. L. Ken. Off-apexnet on micro-expression recognition system. *arXiv preprint arXiv:1805.08699*, 2018.

[16] S. T. Liong, R. C.-W. Phan, J. See, Y. H. Oh, and K. Wong. Optical strain based recognition of subtle emotions. In *International Symposium on Intelligent Signal Processing and Communication Systems*, pages 180–184, December 2014.

[17] S. T. Liong, J. See, R. C. W. Phan, A. C. Le Ngo, Y. H. Oh, and K. Wong. Subtle expression recognition using optical strain weighted features. In *Asian Conference on Computer Vision*, pages 644–657. Springer, 2014.

[18] S. T. Liong, J. See, and K. Wong. Micro-expression recognition using apex frame with phase information. In *APSIPA ASC 2017*, December 2017.

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

[20] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018.

[21] Y. J. Liu, J. K. Zhang, W. J. Yan, S. J. Wang, G. Zhao, and X. Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 7(4):299–310, 2016.

[22] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29:51–59, January 1996.

[23] D. Patel, X. Hong, and G. Zhao. Selective deep features for micro-expression recognition. In *23rd International Conference on Pattern Recognition (ICPR)*, December 2016.

[24] T. Pfister, X. Li, G. Zhao, and M. Pietikinen. Recognising spontaneous facial micro-expressions. In *International Conference on Computer Vision*, November 2011.

[25] A. R. Rivera, J. R. Castillo, and O. O. Chae. Local directional number pattern for face analysis: Face and expression recognition. *IEEE transactions on image processing*, 22(5):1740–1752, 2013.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[28] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing*, 2018.

[29] S. J. Wang, W. J. Yan, G. Zhao, X. Fu, and C. G. Zhou. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. *European Conference on Computer Vision - ECCV 2014 Workshops*, pages 325–338, March 2015.

[30] Y. Wang, J. See, R. C. W. Phan, and Y. H. Oh. LBP with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Computer Vision–ACCV*, pages 525–537, 2014.

[31] Y. Wang, J. See, R. C. W. Phan, and Y. H. Oh. Efficient spatiotemporal local binary patterns for spontaneous facial micro expression recognition. *PLoS ONE*, May 2015.

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

[33] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:915–928, April 2007.

[34] Z. Zhou, G. Zhao, Y. Guo, and M. Pietikainen. An image-based visual speech animation system. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(10):1420–1432, 2012.