CAN MICRO-EXPRESSION BE RECOGNIZED BASED ON SINGLE APEX FRAME?

*Yante Li, Xiaohua Huang, Guoying Zhao**

Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{yante.li, xiaohua.huang, guoying.zhao}@oulu.fi

ABSTRACT

Micro-expressions are rapid and subtle facial movements such that they are difficult to detect and recognize. Most of recent works have attempted to recognize micro-expression by using the spatial and dynamic information from the video clip. Physiological studies have demonstrated that the apex frame can convey the most emotion expressed in facial expression. It may be reasonable to use apex frame for improving micro-expression recognition. However, it is wonder how much apex frames contribute to micro-expression recognition. In this paper, we primarily focus on resolving the contribution-level by using apex frame for micro-expression recognition. Firstly, we propose a new method to detect the apex frame in frequency domain, as it is found that apex frame has very correlated relationship with the amplitude change in frequency domain. Secondly, we propose to use deep convolutional neural network (DCNN) on apex frame to recognize micro-expression. Intensive experimental results on CASME II database shows that our method has achieved considerably improvement compared with the state-of-the-art methods in micro-expression recognition. These results also demonstrate that apex frame can express the major emotion in micro-expression.

Index Terms— Micro-expression, Apex frame spotting, 3D FFT, Deep Convolutional Neural Network

1. INTRODUCTION

Micro-expression (**ME**) is brief and involuntary facial movements that reveal the emotions in high-stake situation that people is trying to conceal his/her true feelings [1]. ME analysis has become an effective way to detect deception, and has several potential applications in fields such as security. Different from macro expressions, physiological studies [2] have shown ME occurs in short duration and with subtle change. All of the above characteristics make ME difficult to detect and recognize. Generally, two main tasks have been conducted in ME analysis. The first one is spotting that is to identify the occurrence of ME, and the other is recognition that aims to classify the ME into specific emotion categories [3, 4, 5]. This paper primarily focus on the second one.

The asterisk means Corresponding author.

Ekman declared that 'snapshot taken at an point when the expression is at its apex can easily convey the emotion message' [6]. In other words, the apex frame contributes major information for facial-expression recognition. Recently, Liong et al. [7] extract the feature from the onset, apex and offset frames for improving the recognition performance. Their results suggested that onset, apex and offset frames can provide sufficient information to classification. Furthermore, Liong et al. [8] firstly use optical flow based on regions of interest to spot the apex frame from a long-term video, and then extract bi-weighted orientation optical flow feature on this spotted apex frame for ME recognition. Motivated by [7, 8], we open two discussions in this paper including 'how much is the contribution of the apex frame to recognize ME?' and 'Could it achieve better results compared with the methods employing ME sequence?', and aim to deeply study these two questions. Specifically, we will go ahead both ME spotting and recognition tasks.

Currently, spontaneous ME spotting methods are mostly based on optical flow and feature contrast [9, 10, 11], respectively. However, these methods on optical flow have relatively slow in computation due to complicated feature operation. Feature contrast approaches miss ME dynamic information, as they just selected the first and last frames in the experientially designed interval. In our paper, we propose a new method to spot the apex frame in frequency domain. As the low intensity and short duration characteristics of ME makes ME non-distinctive occurred in the view of spatial, it is very tough to detect the apex frame in spatial domain. Instead, we attempt to make alternative way for spotting apex frame. Recently, Porter et al. [12] detected video cut in frequency domain by using Fast Fourier Transform (**FFT**), where the peak emotion happens in the high frequency. As we know, FFT can decompose a signal into different frequencies. Amongst those, high frequency can describe quick change of signal, for example, the edges in the image. Furthermore, our experimental experience shows that fast dynamic change of ME corresponds to a frequency hopping in the frequency domain. It may be possibly to analyze the signal change of ME in the frequency domain. Therefore, we transform ME to signal representation in frequency domain by using 3D Fast Fourier Transform (**3D FFT**), and further spot the apex frame according to the highest frequency amplitudes. Compared with the

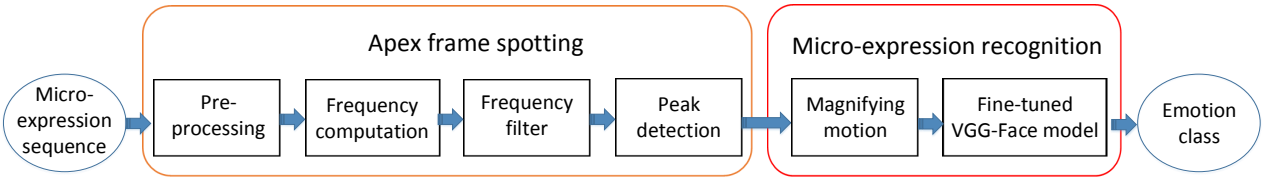


Fig. 1. Framework of the micro-expression recognition system.

commonly used methods, our method not only spots the exact apex frame, but also considers the change in temporal domain.

Considering deep learning has achieved considerable performance in facial expression recognition [13, 14], we propose to use deep Convolutional Neural Network (**DCNN**) based on the apex frame for ME recognition. As far as we know, there is few work exploiting deep learning for ME recognition. Patel et al. [15] firstly proposed transfer learning from objects and facial expressions based on DCNN models for avoiding to train deep model on small ME databases. Unfortunately, this method works worse than the state-of-the-art methods, such as spatio-temporal completed local quantized pattern (**STCLQP**) [4]. Peng et al. [16] presented a dual-template CNN model to recognize ME based on two optical flows. However, optical flow information over the whole video should be first extracted and fed into CNN. The poor performance of deep learning may be caused by two following reasons: (i) ME datasets are very small; (ii) ME is subtle. To address the above problems, we aim to fine-tune the VGG-Face model [17] based on the magnified apex frames.

The rest of the paper is organized as follows. We will describe our proposed method in apex frame spotting and ME recognition in Section 2. We will further present our experiments results in Section 3. We will give an conclusion in Section 4.

2. PROPOSED METHODS

This section will describe the proposed method. It primarily consists of spotting apex frames in ME video and feature extraction through deep learning architecture. The whole flowchart is illustrated in Figure 1.

2.1. Micro-expression apex frame spotting

According to [7, 8], the apex frame plays an important role in ME recognition. However, image pixels in spatial domain cannot provide sufficient discriminative information to spot apex frame due to subtle of ME. Instead, we spot the apex frame in frequency domain, as frequency represents the pixel change obviously. The flowchart of our spotting method is shown in Figure 2.

As the frequency of image is sensitive to illumination changes, we compute the texture map by local binary pattern to represent ME frames. During apex frame spotting

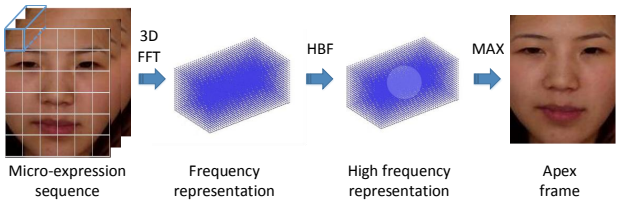


Fig. 2. The flowchart for the proposed ME spotting method. HBF represents high frequency band filter.

procedure, we focus on the changes between ME frames on temporal domain. Specifically, we divide the facial area to 6×6 blocks and compute frequency on all the blocks to obtain much information on pixel change in temporal domain. Furthermore, we compare the frequency of sequential video frames in a specified interval. We transform video blocks into frequency domain by using 3D FFT with a sliding temporal window. By masking a sliding window length with N in the current frame, we compute the frequency of frames located into the sliding window for it. For the i-th interval, we compute the frequency values for the interval on its 36 blocks by 3D FFT. The blocks are denoted as $\{b_{i1}, b_{i2}, \cdots, b_{i36}\}$. For the j-th block in the i-th interval, the frequency value is obtained as:

$$f_{b_{ij}}(x,y,z) = \int_{\frac{-N}{2}}^{\frac{N}{2}} \int_{\frac{-L_b}{2}}^{\frac{L_b}{2}} \int_{\frac{-W_b}{2}}^{\frac{W_b}{2}} F_{b_{ij}}(u,v,q) \times e^{j2\pi(us+vy+qz)} dv du dq, \quad (1)$$

where (x, y, z) represents the position in frequency domain, L_b represents the height of the j-th block b_{ij} in the i-th interval, the W_b represents the width of the j-th video interval block b_{ij} and $j = \{1, 2, \cdots, 36\}$.

In the ME frames, most of the energy belongs to low frequency which is useless for apex frames spotting. In practice, we aim to find frame intervals containing more high frequency signal in the frequency domain. In our method, we employ high frequency band filter (HBF) to remove the lower frequency, such that we reduce the influence of unchanging pixels in video frames. The high frequency filter $H_{b_{ij}}$ is defined as Equation 2, where D_0 is the threshold.

$$H_{b_{ij}}(x,y,z) = \begin{cases} 1 & \text{if } \sqrt{x^2+y^2+z^2} \geq D_0 \\ 0 & \text{if } \sqrt{x^2+y^2+z^2} < D_0 \end{cases}. \quad (2)$$

We filter video blocks in frequency domain following Equation 3,

$$G_{b_{ij}}(x,y,z) = f_{b_{ij}}(x,y,z) \times H_{b_{ij}}(x,y,z). \quad (3)$$

Subsequently, we accumulate the frequency amplitude value $G_{b_{ij}(x,y,z)}$ of all 36 blocks in the i-th video interval by the following formulation,

$$A_i = \sum_{j=1}^{36} \sum_{x=1}^{N} \sum_{y=1}^{L_b} \sum_{z=1}^{W_b} G_{b_{ij}}(x,y,z), \quad (4)$$

where A_i is the frequency amplitude for the i-th interval. It represents the range of rapid facial movements in the i-th internal. In the same way, we can get all the video interval frequency information. The peak interval corresponding to maximum frequency amplitude of interval indicates the highest intensity frames of rapid facial movements. Therefore, we choose the middle frame of the interval as the apex frame.

2.2. Micro-expression recognition with deep learning

Deep learning has been shown efficient on facial expression recognition [14], we present to use deep learning to recognize ME. As ME has low intensity, it is hard to recognize ME. As previously discussed, apex frame can provide sufficient emotion information to ME, but the intensity level of apex frame is not distinguished due to subtle change, especially for deep model that describes high-layer abstract information. Therefore, we present to magnify the apex frame captured by our spotting method to resolve the problem caused by low intensity and improve the discrimination of ME. Recently, Eulerian magnification method [18] has been used to enlarge the difference between ME categories such that classifier can better distinguish them than by using original ME [10]. Motivated by [10], the subtle motion of apex frame is magnified by using the Eulerian magnification method. Bigger values of motion amplification level lead to larger scale of motion amplification, but also can cause bigger displacement and artifacts. In our method, we set the magnified ratio as 30.

Since ME datasets have small sample size, apex frames are not enough to train a good model. Based on high-speed camera, it is simple to obtain ME video that has at least 5 to 15 frames. According to this advantage, we select the apex frame and its two previous and latter frames for extending our training ME image dataset, such that the sample size for deep model has been increased by five times compared to original dataset.

Recently, the VGG-face descriptor [17] is proposed by University of Oxford based on VGG-Very-Deep-16 CNN architecture and achieves good performance on face recognition. Additionally, we can avoid overfitting through finetuning on VGG-face with our small datasets. Therefore, we fine-tune the VGG-FACE model with our extended dataset.

3. EXPERIMENTS

In this section, we make experiments on the apex frame spotting and ME recognition on CASME II [19] database and analyze their corresponding results. CASME II dataset consists of 247 ME elicited from 26 participants with high resolution (640×480 pixel) and recording rate (200 fps). It contains five kinds of expressions: happiness, surprise, disgust, repression and other.

3.1. Experiments on the apex frame spotting

In the experiments, we use all the preprocessing samples provided by [19], which only contains the face regions and are cropped from onset to offset in CASME II dataset. Figure 3 shows the frequency amplitude change in the ME clip. According to Figure 3, it is observed that apex frame appears when frequency amplitude achieves the maximum. It is consistent with the finding in [12].

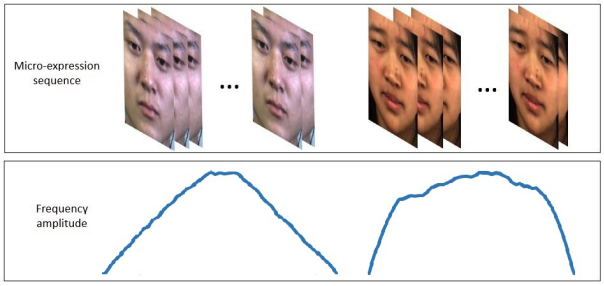


Fig. 3. Examples of frequency amplitude change

According to [8], we use the Mean Absolute Error (**MAE**) as a metric to determine the effectiveness of the apex frame spotting. Specifically, the MAE is defined as the average frame distance between the spotted and the ground-truth apex frames:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |e_i|, \quad (5)$$

where e_i is the frame distance between the spotted and the ground-truth apex frames of the i-th sample, and N is the number of ME samples in CASME II. For CASME II dataset, our method achieves 11.83 in term of MAE, when the sliding detection window size N is set to $N = 61$, and the HBF threshold D_0 is set to $D_0 = 31$.

Table 1 shows the results of apex frame spotting using different methods. According to Table 1, our method achieves the lowest MAE by decreasing 7.15 compared with the baseline algorithm [20]. As well, our method has very competitive result compared with automatic apex frame (**AAF**) proposed by Wong et al. [21]. These comparisons demonstrate that analyzing ME in frequency domain is helpful to apex frame spotting.

Table 1. Results of apex frame spotting on CASME II dataset

Methods	Baseline[20]	AAF[21]	**Ours**
MAE	18.98	14.43	**11.83**

3.2. Experiments on micro-expression recognition

In this section, we present the results of ME recognition on CASME II dataset. In the recognition experiments, the samples in CASME II are classified as: happiness (**H**), disgust (**D**), surprise (**S**), Repression (**R**) and others (**O**). The leave-one-subject-out cross validation protocol is used in our experiments, where we use the samples of one subject as test set and the rest as training set. The recognition accuracy is used as performance metric.

In our method, we magnified the ME with the ratio 30. During the fine-tuning VGG-Face model process, we set the drop-out rate as 0.8 for avoiding over-fitting, and clip the gradient to 30 for preventing gradient exploration.

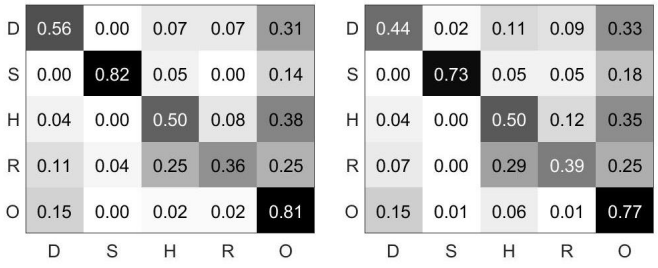
Firstly, we evaluate the influence of magnification technology based on Eulerian magnification method [18] to ME recognition based on apex frames. ME recognition based on apex frame achieves the recognition accuracy of 55.6%. But the performance is significantly improved by by the increased accuracy of 7.7%, when apex frames are magnified. Since magnification method strongly enlarges the ME texture information, DCNN explores more discriminative features from those for enhancing classification performance. According to these comparative results, it is believing that magnification can be very helpful to the improvement of recognition performance. As well, it is consistent to the finding [10], in which they exploited magnification method for ME recognition.

We further compare our recognition method (shown in the column **Methods A** of Table 2) with magnified ground-truth apex frame (**MagGA**), magnified spotted apex frame (**MagSA**) with the state-of-the-art methods based on ME sequence (shown in the column **Methods S** of Table 2) and the methods [22, 15, 4, 7]. Table 2 presents the comparative results of all algorithms.

From Table 2, both MagGA and MagSA achieve promising results compared with Method S [22, 15, 4]. It shows that the apex frame contributes the most emotion information to ME recognition in the video. Furthermore, MagGA considerably boosts the performance by increasing 16% in term of the recognition accuracy compared to [15]. Even though using spotted apex frames, MagSA achieves the promising improvement by increasing 13.34% in term of the recognition accuracy. Moreover, as the training process is offline, the average time cost for ME recognition is only 0.61 seconds. To the best of our knowledge, this is the first work to obtain promising results on raw ME frames with deep learning. We finally provide the confusion matrices of our methods shown in Figure 4. It is found that 'Others' and 'Surprise' classes have highest correct prediction. While the 'Repression' class is always classified to other classes. The main reason is due to the limited samples of 'Repression'.

Table 2. Micro-expression recognition accuracy on CASME II dataset, where the bold text is our proposed method and the corresponding result. 'Method S' and 'Method A' represent the methods using ME sequence and using the apex frame, respectively.

Methods S	Accuracy	Methods A	Accuracy
Baseline[22]	38.00 %	Less [7]	59.67 %
Selective[15]	47.30 %	**MagGA**	**63.30 %**
STCLQP[4]	59.00 %	**MagSA**	**60.64 %**



(a) MagSA (b) MagGA

Fig. 4. Confusion matrices on CASME II dataset

4. CONCLUSION

In micro-expression research, the apex frames are very important, as they convey emotion easily. In this paper, we study the contribution of apex frame to micro-expression recognition. We proposed a new method to locate the apex frame accurately through analyzing micro-expression in frequency domain and then recognize the micro-expression with deep model only based on the magnified apex frame. The results show that our method is effective compared with the state-of-the-art methods and the only apex frame can work well for micro-expression recognition.

5. ACKNOWLEDGEMENT

This work is supported by Academy of Finland, Tekes Fidipro program (Grant No.1849/31/2015), Business Finland project (Grant No.3116/31/2017), Infotech Oulu and the National Natural Science Foundation of China (Grants No.61772419), Jorma Ollila Grant of Nokia Foundation, Central Fund of Finnish Cultural Foundation, Apurahoja tekoalytutkimukseen of Kaute Foundation, NVIDIA GPU Grant Program. The authors wish to acknowledge CSC - IT Center for Science, Finland, for computational resources. Thanks for the help of Henglin Shi.

6. REFERENCES

[1] P. Ekman and W.V. Friesen, "Constants across cultures in the face and emotion," *Personal. Soc. Psychol.*, vol. 17, no. 2, 1971.

[2] P. Ekman, "Lie catching and microexpressions," *Phil. Decept.*, pp. 118–133, 2009.

[3] S. Wang, W. Yan, X. Li, and G. Zhao, "Micro-expression recognition using dynamic textures on tensor independent color space," in *International Conference on Pattern Recognition*, 2014, pp. 4678–4683.

[4] X. Huang, G. Zhao, X. Hong, and W. Zheng, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, no. PA, pp. 564–578, 2016.

[5] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[6] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384, 1993.

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

[8] S. T. Liong, J. See, K. Wong, and R. C. Phan, "Automatic micro-expression recognition from long video using a single spotted apex," *Lecture Notes in Computer Science*, vol. 10, no. 177, 2016.

[9] S. Matthew, G. Sridhar, G. Dmitry, and S. Sudeep, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, 2011, pp. 51–56.

[10] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, , no. 99, pp. 1–1, 2015.

[11] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao, "Spontaneous micro-expression spotting via geometric deformation modeling," *Computer Vision and Image Understanding*, vol. 147, pp. 87–97, 2016.

[12] S. V. Porter, M. Mirmehdi, and B. T. Thomas, "Video cut detection using frequency domain correlation," in *International Conference on Pattern Recognition*, 2000, p. 3413.

[13] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *ACM on International Conference on Multimodal Interaction*, 2015, pp. 435–442.

[14] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE International Conference on Computer Vision*, 2016, pp. 2983–2991.

[15] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *International Conference on Pattern Recognition*, 2017, pp. 2258–2263.

[16] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition:," *Frontiers in Psychology*, vol. 8, pp. 1745, 2017.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, pp. 41.1–41.12.

[18] H. Wu, E. Shih, E. Shih, J. Guttag, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *Acm Transactions on Graphics*, vol. 31, no. 4, pp. 65, 2012.

[19] W. Yan, X. Li, S. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu, "Casme ii: an improved spontaneous micro-expression database and the baseline evaluation," *Plos One*, vol. 9, no. 1, pp. e86041, 2014.

[20] W. J. Yan, S. J. Wang, Y. H. Chen, G. Zhao, and X. Fu, "Quantifying micro-expressions with constraint local model and local binary pattern," *European Conference on Computer Vision Workshops*, pp. 296–305, 2014.

[21] S. T. Liong, J. See, K. Wong, and N. Le, "Automatic apex frame spotting in micro-expression database," *Asian Conference on Patten Recognition (ACPR)*, pp. 665–669, 2015.

[22] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.