TSW-FD: A Novel Temporal and Spatial Domain Weight Analysis of Feature Difference for Micro-Expression Spotting

Zhihao Zhang\textsuperscript{1,2}, Fan Mo\textsuperscript{1,2}, Ke Zhao\textsuperscript{1,2}, Tong Chen\textsuperscript{3,4}, and Xiaolan Fu\textsuperscript{1,2,*}

\textsuperscript{1}State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China
\textsuperscript{2}Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China
\textsuperscript{3}Chongqing Key Laboratory of Non-linear Circuit and Intelligent Information Processing, Southwest University, Chongqing, 400715, China
\textsuperscript{4}Chongqing Key Laboratory of Artificial Intelligence and Service Robot Control Technology, Chongqing, 400715, China

*Email: fuxl@psych.ac.cn

Abstract. The micro-expression spotting has recently attracted increasing attention from psychology and computer vision community, since embraced in the second facial Micro-Expression Grand Challenge (MEGC 2019). Different from the original feature difference (FD) analysis, in this paper, we proposed a novel temporal and spatial domain weight analysis of feature difference (TSW-FD) to achieve micro-expression spotting. The experimental results showed that TSW-FD improved 17.86\% and 24.21\% in F1-Score comparing to the FD in CASME II and SMIC-E-HS.

1. Introduction

Micro-expression is briefly presented when people intentionally or unintentionally hide their real emotions \cite{1,2}. Micro-expressions are more likely to occur in high-stake situations \cite{1}, in which individuals present more micro-expressions when cheating \cite{3,4}. Therefore, micro-expressions were considered as a potential clue for lie recognition in medical diagnosis \cite{5}, criminal interrogation \cite{6} and security system \cite{7}.

The recognition of micro-expression is difficult by naked eyes because the duration of micro-expressions usually last only 1/25 to 1/2s \cite{8}, and the intensity is low and only relates to some areas of facial action unit (AU) \cite{9}. With the development of machine learning and video capture technology, micro-expressions were as a research hotspot. Micro-expression spotting means locating the temporal interval in raw videos, an essential pre-processing step in micro-expression analysis \cite{10}.

In the pioneer works about micro-expression spotting, Polikovsky et al. \cite{11,12} divided the video into multiple ROI based on facial AU and used 3D-gradient histogram as a descriptor to distinguish different stages of micro-expression. Shreve et al. \cite{13,14} used optical strain to detect macro and micro-expressions in videos. However, these efforts only used posed data. The posed micro-expression has obvious boundaries in each stage of micro-expression, and it is contrary to the nature of spontaneous micro-expression.

Moilanen et al. \cite{15} proposed FD analysis to detect micro-expression, which was the first work to spot micro-expressions in the spontaneous micro-expression database. However, only the features of
the current frame and the average feature of head frame and tail frame in sliding window were used for calculating the difference distance. Han et al. [16] proposed a collaborative feature difference to spot micro-expressions. Fisher linear discrimination was utilized to compute weights of each ROI, so that bigger weights were assigned to blocks with higher correlation of micro-expressions. The result significantly improved in CASME II, SMIC-E-HS, SMIC-E-NIR and SMIC-E-VIS. But they just used the spatial information, assigning weights to each ROI. Subsequently, Li et al. [17] utilized the time and space characteristics of micro-expressions and proposed an automatic quadratic ROI selection to detect micro-expressions. However, the strategy in temporal and spatial domain is too rough, that is, regional features were either retained or eliminated.

Due to the excellent performance of deep learning in face recognition and micro-expression recognition [18,19]. Convolutional network and recurrent network have been applied to solve the micro-expression spotting [20,21]. In our previous work [20], a convolutional network was trained with the ability to distinguish micro-expressions. The feature processing based on sliding window performs better than LBP and optical strain, however, the training process required sufficient samples and time consuming.

To solve the problems above, in this paper, we propose a novel temporal and spatial domain weight analysis of feature difference (TSW-FD) for micro-expression spotting. First, we extracted the LBP histogram for each frame in videos. Second, we computed the feature difference distance of each frame in a window interval. Then, a more refined temporal and spatial domain weight analysis of feature difference was applied, which fully consider the time and space characteristics of micro-expressions. Finally, a threshold technology was used to detect micro-expressions in videos.

The paper was organized as follows. We will describe the proposed method in Part 2. Experiments and analysis in Part 3 and conclusions in Part 4.

2. Method

The proposed method consists of three main parts, namely, pre-processing, the feature difference and temporal and spatial domain weight, and micro-expression spotting. Among them, the pre-processing includes automatic face key landmark detection, face alignment and face region segmentation; in feature difference and temporal and spatial domain weight, it mainly extracted LBP histogram and analysed the weight of feature difference in time and space domain. Threshold technology was utilized to determine the micro-expression interval in videos.

2.1. Pre-processing

We implemented pre-processing before the formal spotting for eliminating the interference caused by the head movement and image noise.

Firstly, a frontal face with neutral expression was selected as the template. Using the detector DRMF [22], the 66 key landmarks of template and first frame of raw video were detected automatically. The two inner eye corners and nasal spine are just used to calculate the affine transformation for face alignment [23]. As shown in Figure 1, (a) is the image in raw video, and (b) is the aligned image.

Secondly, according to the key landmarks of aligned face, the facial region was cropped by the horizontal distance between the two inner eye corners and the vertical distance from the nasal spine to the inner eye corner, in order to be suitable for faces of different sizes and shapes, as shown in (c).

Finally, we segmented the cropped face into $6 \times 6$ block [15] and normalized each block to $40 \times 40$, as shown in (d). According to the inner eye corner and the nasal spine, it ensures the contents in corresponding block still.
2.2. Feature Difference and Temporal and Spatial Domain Weight

As FD analysis, we used the Chi-square distance of LBP histogram to compute the difference distance.

For each current frame (CF) $f_i$ in raw video, it has a correspond window interval $v_i$ with fixed length $N$ (odd). The first and last frame of the $v_i$ were named tail frame (TH) and head frame (HF), that is, TH is the before $k$-th frame of CF and HF is the after $k$-th frame of CF, where $k=(N-1)/2$. We computed the difference distance of the $m$-th block in each frame located in window interval $v_i$, as:

$$F^m_{ij} = \frac{1}{2} \left( \chi(f^m_j, f^m_{i-k}) + \chi(f^m_j, f^m_{i+k}) \right)$$  \hspace{1cm} (1)

where $\chi(f^m_j, f^m_{i-k})$ represent the Chi-square distance of $f^m_{i-k}$ and $f^m_j$ in $m$-th block and $\chi(f^m_j, f^m_{i+k})$ is similarity, $j \in [i-k, i+k]$ and $m \in [1, 36]$. The Chi-square distance was computed as the equation:

$$\chi(p, q) = \sum \frac{(p_n - q_n)^2}{p_n + q_n}$$  \hspace{1cm} (2)

where $p$ and $q$ are the normalization histogram with same number of bins. Finally, we obtain a histogram difference vector with dimension $N \times 1$ of interval $v_i$.

In addition, we extend $k$ frames at the beginning and the end of videos because TF and HF would exceed the video boundaries when CF located in the first $k$ frames from the beginning and the end of a video.

In order to get temporal and spatial weight matrix, firstly, we randomly selected some window intervals with the annotated apex frame as the centre from micro-expression databases as trainset. As above, the difference distance vector of these intervals were computed. Therefore, the temporal weight matrix of $m$-th block represented as:

$$u^{(m)} = \frac{1}{\#X^{(m)}} \sum_i x^m_i$$  \hspace{1cm} (3)

where $x^m_i$ represent the $m$-th block difference distance vector of a window interval from trainset $X^{(m)}$, $\#X^{(m)}$ is the number of trainset.

For distance vector $u^{(m)}$ of the $m$-th block with dimension $N \times 1$, the element in the vector accumulated, as computed:

$$U^{(m)} = \sum u^{(m)}$$  \hspace{1cm} (4)

As the sum bigger, the block is more significant for micro-expression. Therefore, we obtained the normalized spatial weight matrix as equation (5):

$$V^{(m)} = \frac{U^{(m)} - U^{\text{min}}}{U^{\text{max}} - U^{\text{min}}}$$  \hspace{1cm} (5)

where $U^{\text{min}}$ and $U^{\text{max}}$ is the minimum and maximum of all blocks.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{The main steps in pre-processing raw face images.}
\end{figure}
For each CF corresponding a window interval in remaining video of database, the distance vector was firstly multiplied with $u^{(m)}$, then the result is multiplied with $V^{(m)}$ to obtain the final feature vector of videos.

2.3. Peak Detection
After the representation vector $F$ was calculated, in order to eliminate noise interference, the final features vector were expressed as $C$,

$$C_i = F_i - \frac{1}{2}(F_{i-k} + F_{i+k})$$

(6)

When $C_i$ equal to zero if it is less than zero. Finally, using the threshold of the feature vector $C$, the micro-expressions were detected. Different from previous work, the threshold was computed as:

$$T = p(C_{max} - C_{min})$$

(7)

where $C_{max}$ is the maximum of the final feature vector, $C_{min}$ is the minimum of non-zero and $p$ is the variable from 0.01 to 1, accumulated at an interval of 0.01.

3. Experiment

3.1. Materials
CASME II [24] and SMIC-E [25] are two popular micro-expression databases for spotting tasks. In CASME II, there are five types of micro-expressions, including happiness, disgust, surprise, repression and other. The annotation provided the onset frame, apex frame, offset frame and emotion information. SMIC-E contains three sub-datasets, namely SMIC-E-HS, SMIC-E-NIR and SMIC-E-VIS. The three sub-databases contain three categories of emotions, positive, negative and surprised, but only the onset frame and offset frames of micro-expressions annotated. There are just 71 micro-expression videos in SMIC-E-NIR and SMIC-E-VIS, therefore, we did not test the proposed method on these two databases. Details of the CASME II and SMIC-E-HS shown in Table 1.

| Databases      | SMIC-E-HS | CASME II |
|----------------|-----------|----------|
| Subjects       | 16        | 26       |
| Samples        | 157       | 247      |
| Frame Rate     | 100       | 200      |
| Resolution     | 690×480   | 680×480  |
| FACS           | No        | Yes      |
| Emotion        | 3         | 5        |
| Frame Annotation | Onset Offset | Onset Apex Offset |

3.2. Metric
In order to evaluate our method, the metrics are same as [26]. The TP, FP and FN of test-set firstly calculated under a threshold. If the detected interval $W_{\text{spotted}}$ satisfies:

$$\frac{W_{\text{spotted}} \cap W_{\text{groundtruth}}}{W_{\text{spotted}} \cup W_{\text{groundtruth}}} \geq 0.5$$

(8)

The $W_{\text{spotted}}$ is categorized as TP; otherwise, it is regarded as FP. The $W_{\text{groundtruth}}$ represent the ground truth of micro-expressions. When micro-expressions are not detected at certain threshold successfully, it is counted as FN. Finally, the TP, FP and FN were computed under different threshold, and the Precision, Recall and F1-Score were calculated for the overall evaluation.
Precision = \frac{TP}{TP + FP} \quad (9)

Recall = \frac{TP}{TP + FN} \quad (10)

F1-Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)

3.3. Results and Analysis
For each database, we used about half of videos to compute the temporal and spatial weight vector and used the other half for testing the method. When calculating the weight vector in CASME II, the apex frame as the middle of the window interval which contain the before and after k frames of apex frame. In SMIC-E-HS, we use the middle frame of onset frame and the offset frame as apex frame. The N equal to 65 and 33, that is, k is 32 and 16 for CASME II and SMIC-E-HS [23]. Besides the TSW-FD, we also tested the original FD. The experiment results were showed in Table 2 and Table 3.

In Table 2, the F1-score, Precision and Recall of TSW-FD are 0.3247, 0.2347 and 0.5266 in CASME II, respectively. Compared to FD, the TSW-FD improved F1-Score 17.86%. In Table 3, the F1-Score, Precision and Recall of TSW-FD are 0.0908, 0.0452 and 0.4986 in SMIC-E-HS. Compared with FD, the TSW-FD improved F1-Score 24.21%. The result indicated the weight analysis in temporal and spatial domains significantly improve the performance of micro-expressions spotting by taking advantage of the characteristics of micro-expressions. In addition, comparing the Table 2 and Table 3, we easily found both FD and TSW-FD achieved better performance in CASME II than SMIC-E-HS, which CASME II may have higher frame rate and smaller interference motion.

| Table 2. The experiment results in CASME II. |
|--------------------------------------------|
|                | TSW-FD | FD  |
| F1-Score       | 0.3247 | 0.2755 |
| Precision      | 0.2347 | 0.1917 |
| Recall         | 0.5266 | 0.4892 |

| Table 3. The experiment results in SMIC-E-HS. |
|--------------------------------------------|
|                | TSW-FD | FD  |
| F1-Score       | 0.0908 | 0.0731 |
| Precision      | 0.0452 | 0.0398 |
| Recall         | 0.4986 | 0.4471 |

4. Conclusion
Micro-expressions appeared when people try to hide their true emotions and is treated as an important clue for lie recognition. However, there is still no appropriate representation at temporal and spatial level for micro-expressions spotting. In this paper, we used weight analysis of feature difference in temporal and spatial domain. The F1-Score, Precision and Recall is 0.3247, 0.2347 and 0.5266 in CASME II. The F1-Score, Precision and Recall is 0.0908, 0.0452 and 0.4986 in SMIC-E-HS. The experiment results showed the validity of weight analysis of feature difference.

Acknowledgements
This research was supported in part by the National Key Research and Development Project (2018AAA0100205), the National Natural Science Foundation of China (61621136008), and the Major Project of National Social Science Foundation (19ZDA363).

References

[1] Ekman, P., and Friesen, W. V. 1969 Nonverbal leakage and clues to deception. Psychiatry, 32(1), 88-106
[2] Ekman, P. 2009. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (revised edition) (WW Norton & Company)
[3] Ten Brinke, L., Porter, S., & Baker, A. 2012 Darwin the detective: Observable facial muscle contractions reveal emotional high-stakes lies. Evolution and Human Behavior, 33(4), 411-6
[4] Vrij, A. and Ganis, G. 2014 Theories in Deception and Lie Detection. In Credibility Assessment (USA: Academic Press) pp 301-374
[5] Russell, T. A., Chu, E., & Phillips, M. L. 2006 A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro expression training tool British Journal Of Clinical Psychology 45(4), 579-83
[6] Ekman, P. 2009 Lie catching and micro-expressions. The Philosophy of Deception 1(2), 5
[7] Weinberger, S. 2010 Intent to deceive? Can the science of deception detection help to catch terrorists? Sharon Weinberger takes a close look at the evidence for it. Nature 465(7297) 412-6
[8] Yan, W. J., Wu, Q., Liang, J., Chen, Y. H., and Fu, X. 2013 How fast are the leaked facial expressions: The duration of micro-expressions? Journal of Nonverbal Behavior, 37(4), 217-30
[9] Porter, S. and Ten Brinke, L. 2008 Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. Psychological Science, 19(5), 508-14
[10] Oh, Y. H., See, J., Le Ngo, A. C., Phan, R. C. W., and Baskaran, V. M. 2018 A survey of automatic facial micro-expression analysis: databases, methods, and challenges. Frontiers in Psychology, 9, 1128
[11] Polikovsky, S., Kameda, Y., and Ohta, Y. 2009 Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. Proc. Adv. Crime. Detect. and. Prevent 2010 1-6
[12] Polikovsky, S., Kameda, Y., and Ohta, Y. 2013 Facial micro-expression detection in high-speed video based on facial action coding system (FACS). IEICE Transactions on Information and Systems 96(1) 81-92
[13] Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D., and Sarkar, S. 2009 Towards macro-and micro-expression spotting in video using strain patterns. In 2009 Workshop on Applications of Computer Vision (WACV) IEEE pp 1-6
[14] Shreve, M., Godavarthy, S., Goldgof, D., and Sarkar, S. 2011 Macro-and micro-expression spotting in long videos using spatio-temporal strain. In Face and Gesture 2011 IEEE pp 51-56
[15] Moilanen, A., Zhao, G., and Pietikäinen, M. 2014 Spotting rapid facial movements from videos using appearance-based feature difference analysis. In 2014 22nd International Conference on Pattern Recognition IEEE pp 1722-1727
[16] Han, Y., Li, B., Lai, Y. K., and Liu, Y. J. 2018 CFD: a collaborative feature difference method for spontaneous micro-expression spotting. In 2018 25th IEEE International Conference on Image Processing (ICIP) pp 1942-1946
[17] Li, B., Zhang, Z., Cao, R., Zhang, J., and Chen, T. 2020 Automatic double region of interest selection for spotting micro-expression from long videos. In 2020 International Conference on Computer Engineering and Application (ICCEA) IEEE pp. 903-906
[18] Wen, S., Liu, W., Yang, Y., Huang, T., and Zeng, Z. 2018 Generating realistic videos from keyframes with concatenated GANs. IEEE Transactions on Circuits and Systems for Video Technology 29(8) 2337-48
[19] Peng, M., Wang, C., Chen, T., Liu, G., and Fu, X. 2017 Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in Psychology, 8 1745
[20] Zhang, Z., Chen, T., Meng, H., Liu, G., & Fu, X. 2018 SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos. IEEE Access 6 71143-51
[21] Verburg, M. and Menkovski, V. 2019 Micro-expression detection in long videos using optical flow and recurrent neural networks. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) IEEE pp 1-6
[22] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. 2013 Robust discriminative response map fitting with constrained local models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 3444-3451
[23] Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., and Pietikäinen, M. 2017 Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Transactions on Affective Computing, 9(4) 563-77
[24] Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., Chen, Y. H., and Fu, X. 2014 CASME II: An improved spontaneous micro-expression database and the baseline evaluation PloS One 9(1), e86041
[25] Li, X., Pfister, T., Huang, X., Zhao, G., and Pietikäinen, M. 2013 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 pp 1-6
[26] See, J., Yap, M. H., Li, J., Hong, X., and Wang, S. J. 2019 MEGC 2019—the second facial micro-expressions grand challenge. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) IEEE pp 1-5