CFD: A COLLABORATIVE FEATURE DIFFERENCE METHOD FOR SPONTANEOUS MICRO-EXPRESSION SPOTTING

Yiheng Han[†] *Bingjun Li*[†] *Yu-Kun Lai*[⋆] *Yong-Jin Liu*[†]

[†] Tsinghua University, China
[⋆] Cardiff University, UK

ABSTRACT

Micro-expression (ME) is a special type of human expression which can reveal the real emotion that people want to conceal. Spontaneous ME (SME) spotting is to identify the subsequences containing SMEs from a long facial video. The study of SME spotting has a significant importance, but is also very challenging due to the fact that in real-world scenarios, SMEs may occur along with normal facial expressions and other prominent motions such as head movements. In this paper, we improve a state-of-the-art SME spotting method called *feature difference analysis* (FD) in the following two aspects. First, FD relies on a partitioning of facial area into uniform regions of interest (ROIs) and computing features of a selected sequence. We propose a novel evaluation method by utilizing the Fisher linear discriminant to assign a weight for each ROI, leading to more semantically meaningful ROIs. Second, FD only considers two features (LBP and HOOF) independently. We introduce a state-of-the-art MDMO feature into FD and propose a simple yet efficient collaborative strategy to work with two complementary features, i.e., LBP characterizing texture information and MDMO characterizing motion information. We call our improved FD method *collaborative feature difference* (CFD). Experimental results on two well-established SME datasets SMIC-E and CASME II show that CFD significantly improves the performance of the original FD.

Index Terms— Micro-expression, spotting features, feature collaboration, weighted ROIs

1. INTRODUCTION

Facial expression is an important manner for human communication and information transfer. It can be notably classified into two types: macro- and micro-expressions. Macro-expressions are the normal expressions that we see everyday. In contrast, micro-expressions (MEs) are short, subtle facial movements that are not easily perceived by ordinary people, but they indicate the real emotion that people cannot suppress or fake.

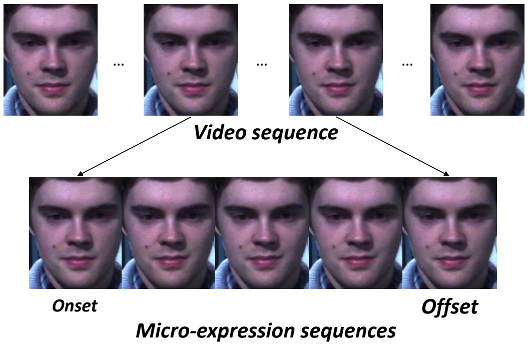


Fig. 1. Micro-expression spotting: tagging possible micro-expression subsequences from a long video sequence.

Research on micro-expression started from 1966 when Haggard et al. [1] proposed the concept of ME. In 1969, Ekman et al. [2] reported the finding of ME in an interview video, which was later considered as the critical clue for the patient's lie. Since then, people have realized the potential of ME in many applications, including education, mental illness diagnosis, airport security etc.

ME analysis is valuable, but also challenging. A ME usually lasts for very short time, mostly less than 0.5 seconds [3]. This means that MEs only exist in a few frames in normal camera videos. Moreover, ME movement has low strength and often shows in only part of the facial regions. These two significant characteristics — short duration and subtle movement — make ME analysis difficult.

As an essential and necessary preprocessing step for ME analysis, ME spotting is to identify subsequences of frames from a given long video that contain MEs (Figure 1). Most existing spotting methods (e.g., [4, 5]) are based on *posed* MEs. It is well known that spontaneous MEs are much more difficult for spotting than the posed MEs, because ME is considered involuntary and difficult to disguise; see a detailed discussion in [6]. So far only a few works address sponta-

Y. Han and B. Li are joint first authors. Y.J. Liu is the corresponding author. This work was supported by the Natural Science Foundation of China (U1736220, 61661130156), the National Key Research and Development Plan (2016YFB1001200), BNRist and Royal Society-Newton Advanced Fellowship (NA150431).

neous ME spotting; see [6] for a summary. In recent years, benefiting from the progress of technical advances in acquisition hardware, well-established spontaneous ME datasets [7, 8, 9] were released, providing essential data for research.

A state-of-the-art spotting technique called *feature difference analysis* (FD) was proposed in [6], which is an extension of the first spontaneous ME spotting method [10]. Given a long facial video sequence, FD divides the facial region of each frame into equal-sized regions of interest (ROIs). Then, the local binary pattern (LBP) histogram feature [11] or the histogram of optical flow (HOOF) feature [12] in each ROI is extracted. For each frame, a subsequence centered at this frame with a fixed length is formed, and the feature difference of this subsequence is defined as the distance of this frame's feature and the average feature of the first and the last frames of this subsequence. Finally, after removing background noise, a threshold is used to tag the possible peak frames and subsequences.

FD achieves superior performance among existing spontaneous ME spotting methods including [13, 14]. However, it still has space for improvement. First, the contribution of each facial ROI is considered equally, which is unreasonable. Since MEs are often local, different regions may have different contributions. Second, FD only considers the LBP and HOOF features independently. A collaboration of different complementary features may achieve better performance. Based on these observations, in this paper, we propose an improved FD method called *collaborative feature difference* (CFD). The main contributions are:

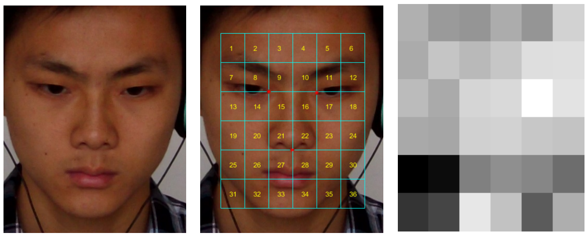
- We propose a novel evaluation method to compute the contribution of each ROI by utilizing the Fisher linear discriminant, such that the weighted ROIs are more semantically meaningful.
- We introduce the state-of-the-art MDMO feature [15] and propose a collaborative strategy using both the LBP and MDMO features: LBP characterizing texture information and MDMO characterizing motion information, such that they complement each other and the collaborative strategy improves the spotting performance.

The paper is organized as follows. We describe our method in Section 2. Experiments and discussions are presented in Section 3 and conclusion is in Section 4.

2. CFD METHOD

2.1. Weighted regions of interest

Given a long facial video, for ME spotting analysis, the following preprocessing steps are performed first, including facial region detection, face alignment and region of interest division. Discriminative response map fitting (DRMF) method [16] is used to detect the facial landmarks and determine the facial region. Then, similar to [6], three points (two inner eye corners and the nasal spine point) are used to align the face and divide the area into equal-sized regions of interest (ROIs), as illustrated in Figures 2(a) and 2(b).



(a) Facial Region (b) Uniform ROIs (c) Weights

Fig. 2. The division of ROIs and evaluated weights.

This division method is consistent with subtle and local movement of micro-expressions. However, due to the local nature of ME, it is unreasonable to assume that each ROI has an equal contribution. We instead evaluate a suitable weight for each ROI to make them more semantic.

Given a labeled spontaneous ME dataset[1], we consider the two-class classification problem $\{Yes, No\}$, where Yes means having MEs and No for not having MEs in a sample video clip. Denote by R_i the ith ROI, $i = 1, 2, \cdots, 36$. For each ROI R_i, we collect all the feature values[2] computed in it for all video clips with the Yes label, and denote this collection as a set $X_{yes}(i) = \{x^j_{yes}(i)\}$. Similarly we collect all feature values $X_{no}(i) = \{x^j_{no}(i)\}$ computed in R_i for all video clips with the No label.

We estimate a weight for each ROI R_i based on the following key observation. If the feature sets $X_{yes}(i)$ and $X_{no}(i)$ contributed by R_i are good indicators for ME spotting, the weight for R_i should be large and vice versa. We evaluate the goodness of the feature sets $X_{yes}(i)$ and $X_{no}(i)$ by adopting the Fisher linear discriminant. Let $u_{yes}(i)$ and $u_{no}(i)$ be the mean vectors of $X_{yes}(i)$ and $X_{no}(i)$ respectively:

$$u_{yes}(i) = \frac{1}{\#X_{yes}(i)} \sum_{x_{yes} \in X_{yes}(i)} x_{yes} \tag{1}$$

$$u_{no}(i) = \frac{1}{\#X_{no}(i)} \sum_{x_{no} \in X_{no}(i)} x_{no} \tag{2}$$

where $\#X$ is the cardinality of the set X. We further define the within-class variances for $X_{yes}(i)$ and $X_{no}(i)$ as

$$s^2_{yes}(i) = \sum_{x_{yes} \in X_{yes}(i)} (x_{yes} - u_{yes}(i))^2 \tag{3}$$

$$s^2_{no}(i) = \sum_{x_{no} \in X_{no}(i)} (x_{no} - u_{no}(i))^2 \tag{4}$$

[1] In our experiment, we use half of the CASME II dataset for evaluating the ROIs' weights, and use the other half and the SMIC-E dataset for testing the spotting performance.

[2] The extracted features are presented in Section 2.2.

The weight for the ROI R_i is then defined as

$$w_i = \frac{(u_{yes}(i) - u_{no}(i))^2}{s_{yes}^2(i) + s_{no}^2(i)} \tag{5}$$

The higher the weight w_i is, the more significant the ROI R_i is. Figure 2 illustrates the evaluated weights of all ROIs.

2.2. Feature extraction

In ME analysis, appearance-based texture features and optical-flow-based motion features are the two main classes of ME features. In our study, we select LBP [11] and MDMO [15] as representative features from these two classes.

Local binary patterns (LBP) and its variants are widely used in many computer vision applications (e.g., [17]) including ME spotting [6]. LBP labels the pixels of an image region by thresholding the neighborhood of each pixel and generating binary numbers. LBP is a powerful feature for texture classification due to its discriminative power and computational simplicity. We implement the same LBP feature in [6] which extracts a normalized 59-dimensional LBP feature for each ROI.

LBP is a good texture descriptor. However, micro-expression is a facial movement. Naturally, we can use optical-flow based motion features as a complement to LBP for characterizing the motion information. Main directional mean optical-flow (MDMO) is a state-of-the-art ME feature [15]. MDMO selects the main direction of the optical flow histogram of each ROI, represented by the magnitude and orientation. MDMO can maximize the subtle movement of micro-expression and has been demonstrated to achieve significant performance on ME recognition. For implementation detail of MDMO, please refer to [15].

Due to their complementary nature, we extract both the LBP and MDMO features, and propose a collaborative strategy to leverage their advantages (see Section 2.4).

2.3. Computation of feature difference contrast

The original FD method [6] only considers the LBP and HOOF features independently. In our work, we adapt the FD method by incorporating both LBP and MDMO features as follows.

For the ith frame of a given facial video clip, the spotted sequence v_i is formed with a fixed length N, which starts with the $(i-k)$th frame and ends with the $(i+k)$th frame, where $k = \lfloor (N-1)/2 \rfloor$ denotes the half length of the spotted sequence. Thus, the first and last k frames of the raw video clip are not used for generation of subsequences. The feature difference between the ith frame and the starting/ending frames can be used to describe the changes of features within v_i. Intuitively, with higher feature difference value, v_i is more likely to contain micro-expressions.

The feature difference F_j^i of the jth ROI in v_i is computed as:

$$F_j^i = \frac{1}{2} w_j \left(Dis(f_j^i, f_j^{i-k}) + Dis(f_j^i, f_j^{i+k}) \right), \tag{6}$$

where f_j^i denotes the jth ROI feature vector of the ith frame which includes 59-dimensional histogram of LBP and 2-dimensional MDMO feature (ρ_j^i, θ_j^i). $Dis(f_1, f_2)$ denotes the distance function between the two feature vectors, where Chi-squared (χ^2) distance is used for LBP and Euclidean distance is used for MDMO.

Compared to the original FD method [6], we use two feature differences in F_j^i (see Eq.(6)):

- Instead of using the average feature vector of f_j^{i-k} and f_j^{i+k} to compute the distance, which may lose useful difference information, we compute the two feature differences separately which are then averaged afterwards. Our experimental results show that our adaption better preserves feature differences among these frames.
- We incorporate the weights evaluated in Section 2.1 to make F_j^i more discriminative.

After obtaining the feature difference of each ROI in v_i, we follow [6] to apply the post-processing steps. To maximize the effect of micro-expression, $F_1^i, F_2^i, \cdots, F_{36}^i$ are sorted in descending order as $F_{j_1}^i, F_{j_2}^i, \cdots, F_{j_{36}}^i$. Then the average value of the top M ROIs with larger feature differences is computed. After relatively local peaks and background noise removal, the final feature difference contrast C^i of v_i is expressed as:

$$F^i = \frac{1}{M} \sum_{m=1}^{M} F_{j_m}^i \tag{7}$$

$$C^i = F^i - \frac{1}{2}(F^{i-k} + F^{i+k}) \tag{8}$$

where $M = 12$ (i.e., $\frac{N}{3}$) is used. Through the computation of feature difference, the ME characteristic of each spotted video is obtained for the downstream peak detection (see Section 2.4).

2.4. Collaborative peak detection

In the original FD method [6], the peak in the raw video is detected by the following threshold value T:

$$T = C_{mean} + p(C_{max} - C_{mean}) \tag{9}$$

where C_{max} and C_{mean} are the max and the mean values of C in the whole video, $0 < p < 1$ is a parameter to determine the threshold level. If $C^i \geq T$, then v_i is considered as a micro-expression sequence.

In our approach, we propose a simple yet effective collaborative strategy which combines LBP and MDMO features to

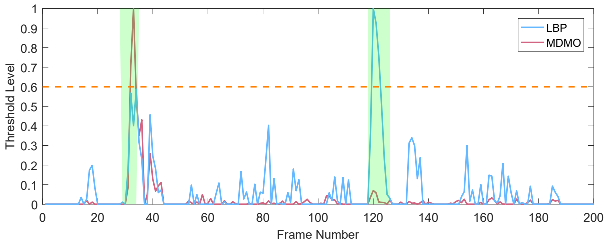


Fig. 3. There are two true micro-expression clips (green shades) in the video sequence. With a high threshold level, LBP and MDMO can each spot one of them, while the collaborative strategy can spot both.

leverage both their advantages. Denote the feature difference contrast computed by each feature as C^i_{LBP} and C^i_{MDMO}, and the threshold as T_{LBP} and T_{MDMO} with the same threshold level p. The collaborative rule is expressed as:

$$label_i = \begin{cases} 1, & C^i_{LBP} >= T_{LBP} \mid\mid C^i_{MDMO} >= T_{MDMO} \\ 0, & otherwise \end{cases} \tag{10}$$

where $label_i$ indicates whether v_i is a micro-expression sequence or not. Figure 3 illustrates the mechanism of the collaborative strategy.

3. EXPERIMENTS

3.1. Experimental settings

We implement the proposed CFD method using MATLAB and the source code is available[3]. For a fair comparison between the state-of-the-art FD method [6] and our proposed CFD method, we follow [6] to use the same datasets and the same evaluation metric, which are briefly summarized below.

The SMIC dataset [9] was extended to include more frames before and after the ME span. The three new subsets of extended SMIC are denoted as SMICE-HS, SMICE-VIS and SMICE-NIR. The original CASEME II dataset [8] already provided long video clips that include extra frames before and after ME span, and then can be directly used. We randomly select half of the CASEME II dataset for determining the weights of ROIs and test the performance of FD and CFD on the other half and SMIC dataset.

After peak detection, if the spotted peak is located within the frame range $[\text{onset} - (N-1)/4, \text{offset} + (N-1)/4]$ of a labeled ME clip, it is considered as one true positive ME; otherwise, it is a false positive ME. In [6], $N = 9, 9, 33, 65$ for SMICE-VIS, -NIR, -HS and CASME II, respectively. The true positive rate (TPR) is the percentage of frames of correctly spotted MEs, divided by the total number of ground truth ME frames in the dataset. The false positive rate (FPR) is the percentage of incorrectly spotted frames, divided by the total number of non-ME frames in the dataset. The performance of spontaneous ME spotting is evaluated by the ROC curves with TPR as the y axis and FPR as the x axis.

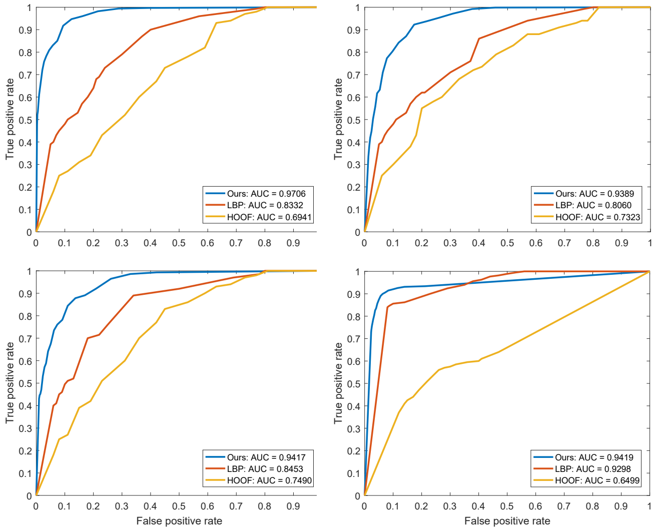


Fig. 4. The ROC curves of FD (evaluated on LBP and HOOF independently) and CFD (ours) on the datasets of SMIC-E-HS, SMIC-E-VIS, SMIC-E-NIR and CASEME II.

3.2. Results

We evaluate the FD method (which only considers using LBP and HOOF independently [6]) and the proposed CFD method. Their ROC curves are shown in Figure 4. The values of area under the ROC curve (AUC) are summarized in Table 1, showing that (1) in CFD, the combination of ROIs' weights and collaborative strategy achieves the best performance, i.e., better than LBP+ROIs' weights and MDMO+ROIs' weights, and (2) CFD is significantly better than FD.

Method	SMICE-HS	SMICE-VIS	SMICE-NIR	CASME II
FD(HOOF)	69.41%	74.90%	73.23%	64.99%
FD(LBP)	83.32%	84.53%	80.60%	92.98%
LBP+W	95.85%	92.15%	92.57%	92.81%
MDMO+W	94.37%	92.67%	90.68%	93.56%
CFD	**97.06%**	**94.17%**	**93.89%**	**94.19%**

Table 1. AUC values of FD and CFD methods with different combinations (W is for ROIs' weights) on four datasets.

4. CONCLUSION

In this paper, we propose an efficient CFD method for spontaneous ME spotting by leveraging the ROIs' weights and a feature collaborative strategy. Experimental results on four datasets SMICE-HS, SMICE-VIS, SMICE-NIR and CASME II show that CFD achieves a significantly better performance than the state-of-the-art FD method.

[3] http://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm

5. REFERENCES

[1] Ernest A Haggard and Kenneth S Isaacs, “Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy,” in *Methods of Research in Psychotherapy*, pp. 154–165. Springer, 1966.

[2] Paul Ekman and Wallace V Friesen, “Nonverbal leakage and clues to deception,” *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[3] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu, “How fast are the leaked facial expressions: The duration of micro-expressions,” *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[4] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta, “Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor,” in *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*. IET, 2009, pp. 1–6.

[5] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar, “Macro-and micro-expression spotting in long videos using spatio-temporal strain,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 51–56.

[6] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen, “Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods,” *IEEE Transactions on Affective Computing*, DOI:10.1109/TAFFC.2017.2667642, 2018.

[7] Wen-Jing Yan, Su-Jing Wang, Yong-Jin Liu, Qi Wu, and Xiaolan Fu, “For micro-expression recognition: Database and suggestions,” *Neurocomputing*, vol. 136, pp. 82–87, 2014.

[8] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu, “CASME II: An improved spontaneous micro-expression database and the baseline evaluation,” *PloS one*, vol. 9, no. 1, pp. e86041, 2014.

[9] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen, “A spontaneous micro-expression database: Inducement, collection and baseline,” in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.

[10] Antti Moilanen, Guoying Zhao, and Matti Pietikäinen, “Spotting rapid facial movements from videos using appearance-based feature difference analysis,” in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 1722–1727.

[11] Timo Ojala, Matti Pietikäinen, and David Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[12] Ce Liu, *Beyond pixels: exploring new representations and applications for motion analysis*, Ph.D. Dissertation, Massachusetts Institute of Technology, 2009.

[13] Zhaoqiang Xia, Xiaoyi Feng, Jinye Peng, Xianlin Peng, and Guoying Zhao, “Spontaneous micro-expression spotting via geometric deformation modeling,” *Computer Vision and Image Understanding*, vol. 147, pp. 87–94, 2016.

[14] Su-Jing Wang, Shuhang Wu, Xingsheng Qian, Jingxiu Li, and Xiaolan Fu, “A main directional maximal difference analysis for spotting facial movements from long-term videos,” *Neurocomputing*, vol. 230, pp. 382–389, 2017.

[15] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu, “A main directional mean optical flow feature for spontaneous micro-expression recognition,” *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2016.

[16] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic, “Robust discriminative response map fitting with constrained local models,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3444–3451.

[17] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.