Spatio-temporal fusion for Macro- and Micro-expression Spotting in Long Video Sequences

Li-Wei Zhang[1,2], Jingting Li[1], Su-Jing Wang[1,3,*], Xian-Hua Duan[2], Wen-Jing Yan[4], Hai-Yong Xie[5,6] and Shu-Cheng Huang[2]

[1] Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

[2] School of Computer, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China

[3] Department of Psychology, University of the Chinese Academy of Sciences, Beijing, China

[4] Institute of Psychology and Behavior Sciences, Wenzhou University, Wenzhou, Zhejiang, China

[5] School of Cyber Science, University of Science and Technology of China, HeFei, Anhui, China

[6] National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC), Beijing, China

***Abstract*— In this paper, we aim to construct a spotting framework automatically. It is still a great challenge to spot micro-expression(ME) intervals accurately due to short duration, low intensity, and shaking. Under the uncontrolled condition, the transformation is caused by head shaking. In order to remove the global movement caused by head shaking, we propose a simple yet effective method to disentangle local movement vector from the global optical flow field by the estimation of mean optical flow in the nose region. After preprocessing, we extract the completed specific pattern(SP) of ME in each region of interest(ROI). The pattern consists of two subpatterns: magnitude and angle. However, influenced by frame rate and different intensities of micro- and macro-expressions, we propose to use a multi-scale filter to improve the ability to spot both micro- and macro-expressions. The spotting result performed on CAS(ME)^2 and SAMM shows that our proposed method outperforms the basline method.**

I. INTRODUCTION

Facial expressions carry a lot of communication information, even more than verbal and limb behavior. Consider that micro-expression is difficult to conceal, micro-expression analysis provides the cue for revealing the people's intentions and physical state inside.

Research on micro-expressions requires FACS [2] to encode the micro-expression video frame by frame. However, the training of FACS encoding consumes too much time, and encoders generally need to receive 100 hours of training to be professional. The encoding process is also time-consuming, and it takes 2 hours to encode a 1-minute video on average [12]. Therefore, the automatic spotting system with high accuracy is very valuable.

Micro-expression research has gained a lot of attention over the past few years and is considered as a challenging task. The difficulty of micro-expression detection is that it is hard to define accurately. Although facial coding system FACS has been designed; the intensity of the change in the action unit corresponding to each person's expression is very different. In addition, the duration of the appearance of micro expressions is extremely short. Shen defines the micro-expression with a boundary of 500ms [15]. Duration below 500ms is classified as micro-expression.

In summary, the contributions of our paper are three-fold. First, we address the head motion problem in a simple way. We select the nose region as a standard global vector that contains only head motion. The local optical flow field is obtained by performing the operator of difference between superposition of optical field and standard global movement vector.

Second, we propose a Spatio-temporal feature fusion matrix which describes spatial and temporal information by row and column relationship. A specific pattern related to magnitude and angle is extracted from the matrix. We denoted it as SP-pattern, which contains all the information from a micro-expression interval. We can obtain onset, apex, and offset according to the SP-pattern.

Third, we use a multi-scale filter to remove high frequency noise and preserve crests of different intensities. In order to achieve good performance on both macro-expression and micro-expression, we comprehensively analyze information at different scales.

We organize the paper as follows. We present related work on micro-expression in section II. We describe our methodolgy in Section III. We present spotting experiments in Section IV. We draw our conclusions in Section V.

II. RELATED WORKS

Research on micro-expression relies on high-quality databases. Since 2011, nine databases of micro-expression have been developed(USF-HD [16], Polikovsky's dataset [13], York DDT [19], MEVIEW [3], SMIC [7], CASME [23], CASME2 [22], SAMM [24], CAS(ME)^2 [14]). The first two databases only contain posed micro-expressions and the others contain spontaneous micro-expressions. York DDT contains some unrelated facial movements. The three databases are not publicly available. All image sequences of SMIC, CASME, CASME II, SAMM and CAS(ME)^2 are

This paper is supported in part by grants from the National Natural Science Foundation of China (U19B2032, 61772511), in part by the Director Fund of National Engineering Laboratory for Public Security Risk Perception and Control by Big Data (18112403) and Natural Science Foundation of Zhejiang Province (No. LQ16C090002).

*Corresponding E-mail: wangsujing@psych.ac.cn.

photographed in a restricted laboratory environment. The six databases are publicly available for free.

The analysis of micro-expression has been studied for decades and various methods have been proposed. In order to improve the performance on the task, different features have been designed to represent ME. When ME occurs, facial local deformation can be captured by dynamic textures(DT). DT is an extension of texture that describes the local pattern at each pixel in the temporal domain. Local binary pattern(LBP) [6] is a popular feature descriptor used for the extraction of local texture. LBP eliminates the problem of lighting changes to a certain extent. VLBP [25] is an extension of the LBP operator which combines the motion and appearance together. The disadvantage of VLBP is high computational complexity. Zhao proposed the LBP-TOP method [26] to analyse DT in the spatial-temporal domain. LBPs from three orthogonal planes are concatenated to form a single descriptor. Optical flow features are widely used to capture the motion of objects. We estimate the magnitude and orientation of pixels between two frames by converting the optical flow vector to polar coordinates. Chaudhry et al. [1] propose Histogram of Oriented Optical Flow (HOOF) features to estimate the motion of all orientations. To analyse the main movement, Liu et al. [11] proposed the Main Directional Mean Optical Flow (MDMO) for micro-expression recognition. Optical strain [9] is an extension of optical flow that is capable of quantifying subtle changes on faces. Several works start to focus on ME spotting [21] and apex frame spotting [10], but still not much so far.

Recently, the development of ME research is facilitated by the process of deep learning [8] [5] [17] [18]. However, it is data starved to make full use of these methods. The models do not have enough capability to learn subtle ME representation.

Our proposed method spot automatically by directly extracting magnitude pattern and angle pattern. The entire spotting process is training free. In addition, a multi-scale filter is used to improve the performance of spotting both micro- and macro-expression.

III. METHODOLOGY

Our entire framework is composed of three sequential stages. They are pre-processing, feature extraction and spotting process that will be described below.

A. Pre-processing

Since the complexity for the computation of dense optical flow is very high. We detected the face of the first frame in every video and got the coordinate of upper-left point, the height of face, and the width of face. We can use these parameters to crop the rest of the faces in the video. In order to reduce the impact of head motion, we only focus on some ROIs related to the action unit. This process consists of two steps. The first is detecting 68 landmarks on the cropped face for each image by using the tool Dlib [4]. We select 12 regions based on landmarks that might be activated when micro-expressions occur. In this way, the shape of each ROI is rectangular, and the facial landmarks determine the center positions of these ROIs. We expect to utilize rich local information to improve the ME spotting robustness. The second is getting the coordinate of each pixel in the region, which can be used to get the optical flow vector at each pixel location. Fig 3 gives the result of pre-processing on the cropped face.

TABLE I
CORRESPONDENCE BETWEEN ROIS AND AUS

Facial region	ROI	AU
Eyebrows	1, 2, 3, 4, 5, 6	1, 2, 4, 5, 6, 7, 9
Nose	7, 12	11
Mouth	8, 9, 10, 11	10 12 14 15 17 20 23 24

We divide four parts on the face. The first part contains two eyebrows, composed of six ROIs. The second part contains one nose composed of two ROIs. The third part contains a mouth composed of four ROIs. In order to remove the global movement caused by the head moving, we calculate mean optical flow in the region of the nose part, which can be used to estimate head movements. The corresponding relationships are listed in Table I.

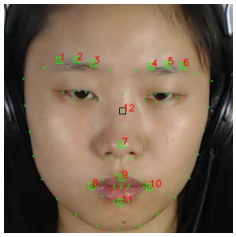


Fig. 1. 68 facial lanedmarks are detected and 12 ROIs are chosen.

B. Feature extraction

This section describes the process of extracting spatial features based on optical flow. Since directly calculating optical flow will be disturbed by the global movement, we propose a simple but effective method to remove the global movement and keep the local movement motivated by micro- and macro-expression.

1) Optical flow field: The optical flow field is a two-dimensional vector field, which reflects the changing trend of gray levels at each pixel in the image. It is usually used to estimate the displacement of each pixel between two frames in a video. We denote the flow vector as $[u_x, u_y]$. u_x and u_y represent horizontal and vertical displacements, respectively. By using multi-scale pyramid technology, optical flow algorithms can capture different changes of motion in video clips. Thus, we extract the main optical flow of dominant orientation at each ROI in consequence frames over time, which are the patterns of motion of facial regions.

We assume that a video has N frames, N-1 optical flow fields can be obtained. The optical flow field of each ROI R_i^r between the frame f_i and f_{i+1} is denoted as $\mathbf{M}_i^r$, where $i = 1, 2, ..., N-1$ is the index of frames and $r = 1, 2, ..., 12$

is the index of ROI. $\mathbf{M}_i^r \in \mathbb{R}^{w \times h \times 2}$ is a matrix of optical flow vectors, where w and h are the length and width of the R_i^r. Each optical flow vector of $\mathbf{M}_i^r$ at the location $p \in R_i^r$ is denoted as $\mathbf{u}_i^r(p)$.

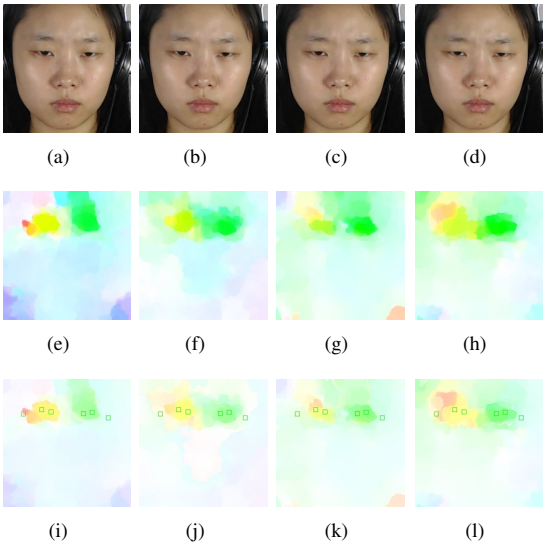


Fig. 2. An example of optical flow field for angry expression. Figure 2(a), 2(b), 2(c) and 2(d) are 4 consecutive frames in the video s15_0402 of CAS(ME)^2. Figure 2(e), 2(f), 2(g) and 2(h) illustrate the result of calculating optical flow directly. Figure 2(i), 2(j), 2(k) and 2(l) illustrate the result of calculating after removing the global movement.

2) Removing global movement: The proposed method aims to eliminate the influence of the global movement. We note that the nose region can be considered as a relatively rigid body compared to other regions. Hence, we estimate movement in the region of the nose. First, we calculate the sum of all vectors in ROI12 and take the unit vector. Then, we calculate the mean module of optical flow vectors in ROI12. Multiply the unit vector by the average module value to get motion estimation vector $\bar{\mathbf{v}}_i$:

$$\bar{\mathbf{v}}_i = \frac{\sum \mathbf{u}_i^r(p)}{||\sum \mathbf{u}_i^r(p)||_2} \cdot \frac{1}{|R_i^r|} \sum \mathbf{u}_i^r(p), \quad p \in R_i^{12} \tag{1}$$

where $||\cdot||_2$ denotes the L2-norm and $|\cdot|$ denotes the set cardinality. Then, each vector $\mathbf{u}_i^r(p)$ of the optical flow field $\mathbf{M}_i^r$ minus the motion estimation vector $\bar{\mathbf{v}}_i$:

$$\tilde{\mathbf{u}}_i^r(p) = \mathbf{u}_i^r(p) - \bar{\mathbf{v}}_i, \quad p \in R_i^r \tag{2}$$

where $\tilde{\mathbf{u}}_i^r(p)$ is the optical flow vector that removes the global movement. We can obtain the local optical flow field by formula (2). One example of local movement in the optical flow domain is illustrated in Fig 2.

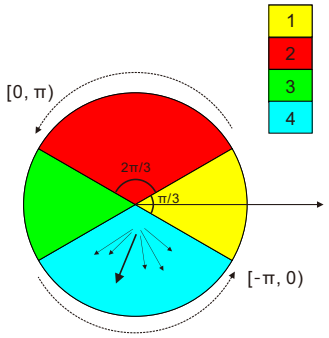


Fig. 3. The histogram of oriented optical flow with a bin number 4. Four color areas represent different angle range. The figure 3 illustrates the process of calculating main optical flow in dominant orientation

3) Spatial feature extraction: Different from MDMO, we divide 2π into four bins with two bins containing $\frac{2}{3}\pi$ angle range and two bins containing $\frac{1}{6}\pi$ angle range as illustrated in Fig 3. We assign each bin to a number to represent its orientation. For each pixel in the ROI, we first use its orientation angle to determine which angle bin it belongs to and then adds the corresponding magnitude number into the angle bin. The dominant orientation is the number of the angle bin with the largest magnitude sum. We select all the optical flow vectors belong to the dominant orientation and calculate the sum of all according to the law of vector addition and take the unit vector:

$$\bar{\mathbf{u}}_i^r = \frac{\sum \tilde{\mathbf{u}}_i^r(p)}{||\sum \tilde{\mathbf{u}}_i^r(p)||_2}, \quad \tilde{\mathbf{u}}_i^r(p) \in B_{max} \tag{3}$$

where B_{max} is is the set of optical flow vectors in the dominant orientation. Then, we calculate the average module of optical flow vectors belong to the dominant orientation. As follows:

$$\bar{m}_i^r = \frac{1}{|B_{max}|} \sum_{\tilde{\mathbf{u}}_i^r(p) \in B_{max}} |\tilde{\mathbf{u}}_i^r(p)| \tag{4}$$

similar to the section of removing global movement. Multiply the the unit vector by average module:

$$\boldsymbol{\gamma}_i^r = \bar{m}_i^r \cdot \bar{\mathbf{u}}_i^r \tag{5}$$

where $\boldsymbol{\gamma}_i^r$ is called the main vector of dominant orientation.

We extract the feature of each ROI R_i^r by casting $\boldsymbol{\gamma}_i^r$ from the cartesian coordinates to the polar coordinates:

$$\boldsymbol{\eta}_i^r = (\rho_i^r, \theta_i^r) \tag{6}$$

where $\boldsymbol{\eta}_i^r$ represents the main feature of dominant orientation of r-th ROI in i-th optical flow field, m_i^r represents the corresponding magnitude and θ_i^r represents the corresponding angle.

In order to fuse spatial information, we connect the features extracted from 11 ROIs into a sequence:

$$L_i = (\boldsymbol{\eta}_i^1, \boldsymbol{\eta}_i^2, \boldsymbol{\eta}_i^3, ..., \boldsymbol{\eta}_i^{11}), \quad i = \{1, 2, ..., N-1\} \tag{7}$$

where L_i can be considered as spatial feature of the i-th frame in the video clip.

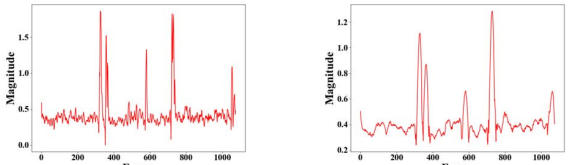
4) Fusion of Spatio-temporal information: After obtaining the spatial feature for every optical flow field. We arrange the spatial feature L_i in the order of the columns to obtain the feature fusion matrix $\mathbf{F}$:

$$\mathbf{F} = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_{N-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\eta}_1^1 & \boldsymbol{\eta}_1^2 & \cdots & \boldsymbol{\eta}_1^{11} \\ \boldsymbol{\eta}_2^1 & \boldsymbol{\eta}_2^2 & \cdots & \boldsymbol{\eta}_2^{11} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\eta}_{N-1}^1 & \boldsymbol{\eta}_{N-1}^2 & \cdots & \boldsymbol{\eta}_{N-1}^{11} \end{bmatrix} \tag{8}$$

The $\mathbf{F}$ describes the Spatio-temporal information of the video in a simple way. Each row of the matrix represents the movement of multiple ROIs on the face. Since micro- and macro-expression intervals in CAS(ME)^2 and SAMM databases are spontaneous. When micro- and macro-expression occurs, multiple AUs related to face regions are activated. Rows of the matrix reflect the momentum distribution in facial space. Each column of the matrix represents how the magnitude and angle of the corresponding ROI change over time. The expression process consists of three nodes, which are the onest, apex, and offset. The entire phrase contains rich temporal information. Columns reflect the distribution of a fixed region in the temporal domain. Therefore, Spatio-temporal feature provides more robustness than using a single dimension to improve the performance of spotting.

C. Spotting process

This section is aim to describe the process of automatic spotting system. After obtaining the feature fusion matrix $\mathbf{F}$ of the video. We track the trajectory of each column over time corresponding to the ROI. 11 trajectories can be obtained for one video and the trajectory length is equal to the number of frames of the video minus one. Next, 11 trajectories are directly passsed to the multi-scale filter stage. Finally, we extrat SP-pattern and merge all results of spotting at each ROI.



(a) (b)

Fig. 4. An example of filter result at ROI3 in the video s16_0102 of CAS(ME)^2. Fig 4(a) and Fig 4(b) have window sizes of 11 and 41, respectively.

1) Multi Scale Filter: Savitzky-golay filter [20] is a low-pass filter for noise in signal measurements. It's widely used in spectral and data stream analysis. We implemented the filter by using a polynomial form to fit a smoother curve and preserve the extreme value information as much as possible. We use a sliding window which moves from left to right until pass the entire curve. Assuming that the window size is 2s+1, the points in the window are denoted as:

$$B = \{x_1, x_2, ..., x_{2s+1}\} \tag{9}$$

next, k-th order polynomial is used to obtain the filter value:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{2s+1} \end{bmatrix} = \begin{bmatrix} x_1^0 & x_1^1 & \cdots & x_1^{k-1} \\ x_2^0 & x_2^1 & \cdots & x_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2s+1}^0 & x_{2s+1}^1 & \cdots & x_{2s+1}^{k-1} \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{bmatrix} \tag{10}$$

formula (10) can be written as:

$$\begin{aligned} \mathbf{Y} &= \sum_{j=0}^{k-1} a_j \cdot \begin{bmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_{2s+1}^j \end{bmatrix} \\ &= \sum_{j=0}^{k-1} a_j \cdot \boldsymbol{\beta}_j \end{aligned} \tag{11}$$

formula (10) can be simplified as follows:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{A} \tag{12}$$

The aim is to minimize $||\mathbf{Y} - \mathbf{B}||_2$. We formulate this as a minimum distance problem. Note that Y is a vector of linear space $L = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{k-1}\}$ and $||\mathbf{Y} - \mathbf{B}||_2$ is the distance

between the point Y and B. If $||\mathbf{Y}-\mathbf{B}||_2$ is the minimum, then B is orthogonal to the entire linear space L:

$$\begin{bmatrix} \boldsymbol{\beta}_0^T \\ \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_{k-1}^T \end{bmatrix} \cdot (\boldsymbol{Y}-\boldsymbol{B}) = 0 \tag{13}$$

let $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{k-1})^T = \mathbf{P}$:

$$\mathbf{P} \cdot (\mathbf{B} - \mathbf{XA}) = 0 \tag{14}$$

expand parentheses and move items:

$$\mathbf{P} \cdot \mathbf{B} = \mathbf{P} \cdot \mathbf{XA} \tag{15}$$

multiply the left side of the equal sign by $(\mathbf{PX})^{-1}$:

$$\mathbf{A} = (\mathbf{PX})^{-1} \cdot \mathbf{PB} \tag{16}$$

where $\mathbf{A}$ is a coefficient matrix for filtering. Smooth curve can be obtained according to the formula (12).

Consider that the unstable facial landmarks and uncontrollable shaking. Given a set of trajectories, we set the sliding window size from small to large. Windows with different scales can preserve the crests of different intensities. An example of filter results is illustrated in Fig 4.

2) Pattern Extraction: This part aims to capture the pattern of expression. Since micro- and macro-expressions are similar in patterns, we focus on describing the SP-pattern of micro-expressions. Fig 5 shows examples of SP-pattern in different conditions.

The pattern we want to capture based on two restrictions: angle and magnitude. The process of micro-expressions can be divided into two phases. The first phase is from onset to apex. The second is from apex to offset. The magnitude of optical flow is the displacement between the current frame and the next frame for each ROI. Acceleration can be observed at the beginning of each phase in Fig 5(a) and Fig 5(b). In some cases, two phases merge into one as illustrated Fig 5(c) and Fig 5(d). We still observe that the pattern with obvious positive acceleration step, apex and negative acceleration step. When one phase transitions to another, the domain orientation of movement is converted from positive to negative or from negative to positive. Since we use radians, the positive value of angle belong to the angle range from 0 to π and the negative value of angle belong to the angle range from $-\pi$ to 0. The two subpatterns are superimposed into a complete SP-pattern. Although different action units have complex motion characteristics, ME intervals fit the SP-pattern in most cases.

We qualify the process of pattern extraction. The process is consists of two steps. The first is detecting the maximum Max in the magnitude trajectory for each ROI. We select $b \cdot Max$ as a threshold. Points that exceed the threshold may be the vertices of the magnitude pattern. Then, we extract the crests whose apexes are these points, the apex of the movement is the value at the top of the crest. The second is to compare the number of frames with positive and negative angles in the crest range:

$$\delta_n^r = \sum_{f_i^r \in B_{crest} \cap \theta_i^r > 0} f_i^r - \sum_{f_i^r \in B_{crest} \cap \theta_i^r < 0} f_i^r \tag{17}$$

where B_{crest} is the set of frames belong to crests of magnitude trajectories. δ_n^r is the difference between frames with positive angle and frames with nagative angle in the n-th crest range of r-th ROI.

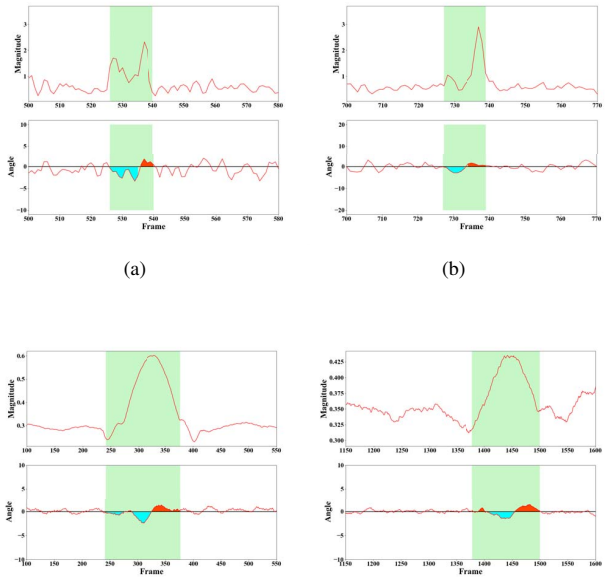
Due to the angle pattern is very sensitive to the deformation of motion, even a subtle movement will be detected. In order to improve the robustness of the method, we set a threshold t. If the absolute value of δ_n^r is less than the threshold t, we argue that the frames in the crest range conform to the angle pattern. They are the frames that belong to the ME interval. The complete SP-pattern is shown in Fig 5.

3) Merge Spotting Results: For each video, the fusion feature matrix $\mathbf{F}$ is obtained according to the formula (8). Columns of the matrix are denoted as $T = \{C_1, C_2, ..., C_{11}\}$. We serve E as the operator of pattern extraction, F as the operator of the filter. We first perform F on T and get the result $F(T) = \{F(C_1), F(C_2), ..., F(C_{11})\}$. Then, we extract patterns and get $I = E(F(T))$, where I are spotting intervals of expressions. We maintain a series $\Gamma = (0, 0, .., 0)$, which can be used to record if there are expressions by 0 and 1. The length of Γ is the frame number of the video. Before spotting, we initialize the value in Γ to 0. In order to get the final spotting result, we set the corresponding position of Γ to 1 according to the pattern of extraction I. Thus, each main local movement of ROI can be captured. Spotting results with the same indexes will be overwritten, and spotting results with different indexes will be merged. In addition, spotting intervals that are too close are also merged.

IV. RESULTS AND DISCUSSION

A. Databases

CAS(ME)^2 and SAMM are used to evaluate our spotting method. The database CAS(ME)^2 is divided into Part A and Part B: Part A contains 87 long videos with both micro- and macro-expressions, and Part B contains 300 macro-expression samples and 57 micro-expression samples. Emotions fall into four categories: positive, negative, surprised, and others. CAS(ME)^2 encode the onset, apex, and



(a) (b) (c) (d)

Fig. 5. Temporal patterns under different conditions. The X axis is the index of the frame number, the Y axis is magnitude and angle of main optical flow, respectively. Figure 5(a) illustrates the pattern at ROI1 in the video s23_0102 of CAS(ME)^2. Figure 5(b) illustrates the pattern at ROI2 in the video 23_0102 of CAS(ME)^2. Figure 5(c) illustrates the pattern at ROI8 in the video 26_1 of SAMM. Figure 5(d) illustrates the pattern at ROI7 in the video 18_3 of SAMM. They are similar in magnitude and angle patterns.

TABLE II
THE CHARACTERISTICS OF CAS(ME)^2 AND SAMM

Datasets	CAS(ME)^2	SAMM
Sample	57	159
Subject	22	32
FPS	30	200
Resolution	640x480	2040x1088
Face size	N/A	400x400
Emotion category	4	7

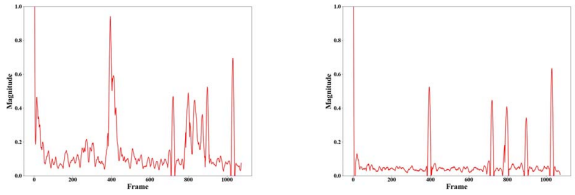
offset. Based on the combination of AU, emotion-evoked video emotion type, and self-reported emotion, the micro-expression labels are determined.

There are 159 samples in the SAMM database (the picture sequence contains spontaneous micro-expressions). A high-speed camera with a frame rate of 200 fps and a resolution of 2040x1088 records them. As the same as the CAS(ME)^2 database, these samples are recorded in a well-controlled laboratory environment and lighting conditions are strictly designed. Table II shows the characteristics of databases.

B. Metric

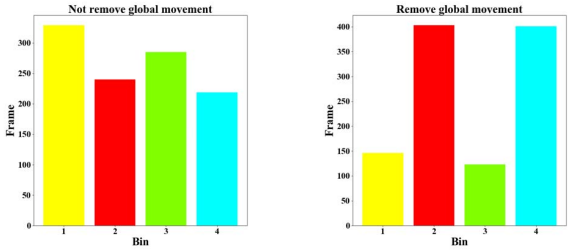
As evaluation metric, we adopted F1 score to evaluate the performance of the method $F_1 = \frac{2RP}{R+P}$, where R and P are denoted as recall and precision, respectively. One of the challenges of MEGC2020 is to spot both micro- and macro-expression.

C. Results



(a) (b)

Fig. 6. An example of trajectories of magnitude at ROI3 in the video s15_0102 of CAS(ME)^2. Fig 6(a) illustrates the trajectory without removing the global movement. Fig 6(b) illustrates the trajectory after removing the global movement.



(a) (b)

Fig. 7. The historam of frames categoried in four orientations in the video s16_0102. The X axis is the orientation bin, the Y axis is the frame number. Statistical regions contain ROI 1,2,3,4,5,6

1) Eliminating head shaking and preserving ME movement: As described in section III, we propose a method to disentangle ME motion from the entire shaking face area.

Fig 6 shows an example of the contribution of the method. We observe that the trajectory with superimposed global movement has other crests caused by head motion. They include much basic noise and cover up the local movement.

In order to evaluate the contribution quantitatively, we analyze the trend of the main optical flow in the four directions as illustrated in Fig 3. The analysis is performed on a long video, which brings many active AUs in eyebrow regions. Since the motion pattern of eyebrows is vertical in most

TABLE III

MEAN OF SCALE IN CAS(ME)^2 AND SAMM.

Database	CAS(ME)^2			SAMM		
Characteristic	micro	macro	FPS	micro	macro	FPS
Mean	19.3	52.8	30	128.7	401.0	200

cases, we mainly focus on ROI 1,2,3,4,5,6. For example, raised eyebrows(AU4) and frowned(AU1). Fig 7 presents the effect of eliminating head shaking. The frames belong to an upward direction(bin 2), and a downward direction(bin 4) are more than the situation which does not perform the method of eliminating head shaking.

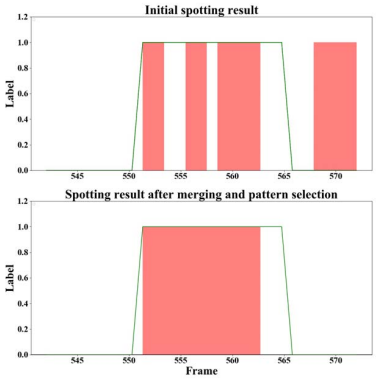


Fig. 8. An example of pattern selection and merge process in s16_0101 of CAS(ME)^2. The green curve represents the ground truth and the red area represents the result of spotting.

2) Pattern extraction and Merge process: Pattern extraction and merge process are beneficial to spotting result. Fig 8 illustrates an example of the contribution of the process. As shown in the first layer of the figure, three peaks are distributed in the ground-true interval and one peak outside the range. The result of the first layer fits the magnitude pattern described in section III. If we only extract the pattern according to magnitude, the result will be very terrible. However, the merge process and angle pattern in the second layer help a lot. The peak on the right is deleted because of the wrong angle pattern. The neighboring peaks with short distances are merged into a complete interval.

The restriction of the angle pattern reduces the quantity of TPs. It also reduces a large number of FPs at the same time. From the perspective of reducing FPR, the advantages outweigh the disadvantages.

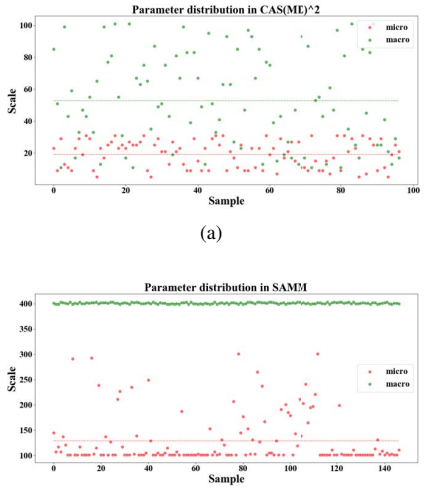


Fig. 9. Analysis of the distribution of scale. The X axis is the sample index, the Y axis is the scale of sample. The red points represent the mico-expression and the green points represent macro-expression.

TABLE IV

FINAL SPOTTING RESULT OF CAS(ME)^2 AND SAMM

Database	CAS(ME)^2			SAMM		
Expression	micro	macro	overall	micro	macro	overall
Total	57	300	357	159	343	502
TP	26	119	145	41	27	68
FP	867	698	1565	416	375	791
FN	31	181	212	118	316	434
Precision	0.0291	0.1457	0.0848	0.0897	0.0672	0.0792
Recall	0.4561	0.3967	0.4062	0.2579	0.0787	0.1355
F1-score	**0.0547**	**0.2131**	**0.1403**	**0.1331**	**0.0725**	**0.0999**

3) Analysis of multi-scale filter: The filter process plays a role in spotting task. One main parameter for filtering is the scale of sliding window W_s. Hence, we analyze the impact of W_s on spotting performance. In order to avoid losing the original intensity information, we set the scale of the window from small to large to preserve information from fine to coarse. In this way, we adjust the sliding window to achieve good performance. According to the definition of micro- and macro-expression duration, the crest width of macro-expression is larger than half of frame rate. Table III shows the average scale of each database. We discover that the average scale of micro- expression is far smaller than that of macro-expression. The frame rate and scale parameter closely related in Table III. Larger frame rate means smaller

deformation between adjacent frames, especially for the optical flow field. Therefore, large W_s is suitable to extract SP-pattern in the video with large frame rate.

Fig 9 illustrates the distribution of the parameter in two databases. We can see that the parameters of SAMM distribute more tightly around the average line compared with the parameters of CAS(ME)^2. This is mainly because a high frame rate will cause the width of the SP-pattern to increase. When filtering with a large window, more noise caused by random motion can be filtered out, making the SP-pattern caused by micro-expressions less prone to distortion. Table IV shows the performance of the final spotting.

V. CONCLUSIONS AND FUTURE WORKS

The propose of the method was to spot both micro- and macro-expressions automatically. To obtain decoupled local movement, we proposed a simple yet effective method in the optical flow domain to eliminate head shaking. The decoupled local movement is discriminative for spotting. After preprocessing, we extract features based on the optical flow field and fuse temporal and spatial information according to ROIs on the face. When MEs occur, they fit the same pattern. The SP-pattern contains two subpattern, magnitude pattern and angle pattern. We extract pattern by using the Spatio-temporal fusion matrix proposed. In addition, we propose to use a multi-scale filter to improve the spotting performance.

The performance of the proposed method is sensitive to the change of the ROI number and area. In future work, more facial areas connected to ME should be considered to enhance the robustness of SP-pattern.

REFERENCES

[1] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1932–1939. IEEE, 2009.

[2] P. Ekman and W. Friesen. Facial action coding system (facs): A technique for the measurement of facial action, palo alto, ca: Consulting. 1978.

[3] P. Husák, J. Cech, and J. Matas. Spotting facial micro-expressions "in the wild". In *22nd Computer Vision Winter Workshop (Retz)*, 2017.

[4] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[5] J. Li, Y. Wang, J. See, and W. Liu. Micro-expression recognition based on 3d flow convolutional neural network. *Pattern Analysis and Applications*, 22(4):1331–1339, 2019.

[6] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE transactions on affective computing*, 9(4):563–577, 2017.

[7] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, pages 1–6. IEEE, 2013.

[8] X. Li, J. Yu, and S. Zhan. Spontaneous facial micro-expression detection based on deep learning. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pages 1130–1134. IEEE, 2016.

[9] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Processing: Image Communication*, 47:170–182, 2016.

[10] 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 *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 665–669. IEEE, 2015.

[11] 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, 2015.

[12] M. Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3505–3513, 2009.

[13] S. Polikovsky, Y. Kameda, and Y. Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. 2009.

[14] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu. CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 9(4):424–436, 2017.

[15] X.-b. Shen, Q. Wu, and X.-l. Fu. Effects of the duration of expressions on the recognition of microexpressions. *Journal of Zhejiang University Science B*, 13(3):221–230, 2012.

[16] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar. Macro-and micro-expression spotting in long videos using spatio-temporal strain. In *Face and Gesture 2011*, pages 51–56. IEEE, 2011.

[17] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao. Recognizing spontaneous micro-expression using a three-stream convolutional neural network. *IEEE Access*, 7:184537–184551, 2019.

[18] T.-K. Tran, Q.-N. Vo, X. Hong, and G. Zhao. Dense prediction for micro-expression spotting based on deep sequence model. *Electronic Imaging*, 2019(8):401–1, 2019.

[19] G. Warren, E. Schertler, and P. Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69, 2009.

[20] H. J. Wayt and T. R. Khan. Integrated savitzky-golay filter from inverse taylor series approach. In *2007 15th International Conference on Digital Signal Processing*, pages 375–378. IEEE, 2007.

[21] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao. Spontaneous micro-expression spotting via geometric deformation modeling. *Computer Vision and Image Understanding*, 147:87–94, 2016.

[22] W.-J. Yan, X. Li, S.-J. 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*, 9(1):e86041, 2014.

[23] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu. Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–7. IEEE, 2013.

[24] C. H. Yap, C. Kendrick, and M. H. Yap. Samm long videos: A spontaneous facial micro-and macro-expressions dataset. *arXiv preprint arXiv:1911.01519*, 2019.

[25] 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(6):915–928, 2007.

[26] 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*, (6):915–928, 2007.