A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition

Yong-Jin Liu, *Member, IEEE*, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, *Member, IEEE*, Guoying Zhao, *Senior Member, IEEE*, and Xiaolan Fu, *Member, IEEE*

Abstract—Micro-expressions are brief facial movements characterized by short duration, involuntariness and low intensity. Recognition of spontaneous facial micro-expressions is a great challenge. In this paper, we propose a simple yet effective Main Directional Mean Optical-flow (MDMO) feature for micro-expression recognition. We apply a robust optical flow method on micro-expression video clips and partition the facial area into regions of interest (ROIs) based partially on action units. The MDMO is a ROI-based, normalized statistic feature that considers both local statistic motion information and its spatial location. One of the significant characteristics of MDMO is that its feature dimension is small. The length of a MDMO feature vector is $36 \times 2 = 72$, where 36 is the number of ROIs. Furthermore, to reduce the influence of noise due to head movements, we propose an optical-flow-driven method to align all frames of a micro-expression video clip. Finally, a SVM classifier with the proposed MDMO feature is adopted for micro-expression recognition. Experimental results on three spontaneous micro-expression databases, namely SMIC, CASME and CASME II, show that the MDMO can achieve better performance than two state-of-the-art baseline features, i.e., LBP-TOP and HOOF.

Index Terms—Micro-expression, optical flow, recognition, feature

1 Introduction

Facial expressions can provide rich information in social life. Full facial expressions typically last for 0.5-4 seconds [1] and can thus be readily recognized by humans. However, psychological studies have shown that a person may conceal but occasionally leak their genuine emotions [2]. Micro-expressions, once named micro-momentary facial expressions that might be unknown to or uncontrollable for humans, were first discovered in 1966 [3]. Three years later, Ekman [4] used the term *micro-expression* when he analyzed an interview video of patients who tried to commit suicide. The work in [4] presented evidence that micro-expressions can reveal concealed emotions; this evidence has recently drawn extensive attention from psychologists.

Micro-expression recognition has a wide range of applications in diverse fields, including clinical diagnosis and national security. However, micro-expressions are fleeting and easily neglected by the naked eye. Ekman [5] developed a Micro-Expression Training Tool (METT) in 2003. In 2009, Frank et al. [6] performed a real-life micro-expression test and found that when recognizing at real-time speed, trained participants (with the help of the METT) only had an accuracy of less than 50 percent, not to mention the ordinary people without training. In [7], a Point Grey GRAS-03K2C camera was used to collect micro-expression data at high temporal resolution: the sampling rate was 200 fps and the resolution was 640×480. Two expert coders examined these video data frame-by-frame without any time restrictions to spot and code micro-expressions. Furthermore, these two coders discussed and arbitrated the disagreements. Thanks to recent well-developed databases [7], [8], [9], the demand for computer vision techniques to improve the performance of micro-expression recognition is increasing.

Full-expression recognition has been widely studied in the computer vision field [10]. However, micro-expressions have the following characteristics that make micro-expression recognition quite different from full-expression recognition. First, micro-expressions are rapid facial movements, typically occurring in less than 0.5 second [11]. Second, the intensity of these fleeting micro-expressions is also very low in terms of facial muscles' movement. Third, no complete micro-expressions involving both the upper and lower halves of the face simultaneously were observed in [12]; in other words, micro-expressions typically involve a fragment of the facial region. Therefore, previous work that were suitable for full-expression recognition may not work well for micro-expressions. Automatic micro-expression recognition algorithms have recently received attention [7], [8], [9], [13], but there is still considerable room for improvement in recognition accuracy.

2 Related Work

The flourishing of full-expression recognition methods [10] depends largely on the well-established facial expression databases, such as CK+, MUG, MMI, JAFFE and Multi-PIE.

- *Y. J. Liu and J. K. Zhang are with the Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, P. R. China. E-mail: liuyongjin@tsinghua.edu.cn, zhangjinkai999@gmail.com.*
- *S. J. Wang and X. L. Fu are with the State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101. E-mail: {wangsujing, fuxl}@psych.ac.cn.*
- *W. J. Yan is with the College of Teacher Education, Wenzhou University, Wenzhou, China, 325035. E-mail: yanwj@wzu.edu.cn.*
- *G. Y. Zhao is with the Center for Machine Vision Research, Infotech Oulu and Department of Electrical and Information Engineering, University of Oulu, PO Box 4500, FI-90014, Finland. E-mail: gyzhao@ee.oulu.fi.*

Manuscript received 13 Mar. 2015; revised 19 Sept. 2015; accepted 29 Sept. 2015. Date of publication 30 Sept. 2015; date of current version 5 Dec. 2016.
Recommended for acceptance by M. Valstar.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2015.2485205

However, there are few well-established micro-expression databases due to the difficulty of eliciting micro-expressions. To the best knowledge of the authors, there are only five publicly-available micro-expression databases (USF-HD [14], Polikovsky's database [15], SMIC [8], CASME [9], [16] and CASME II [7]), and only three of them (SMIC, CASME and CASME II) are spontaneous.

In the USF-HD database [14], 100 acted facial micro-expressions were collected, and strain patterns were used for feature description. However, acted facial expressions have been found to differ significantly from natural facial expressions that occur in daily life [17]. In this paper, we study micro-expression recognition using three spontaneous micro-expression databases (SMIC, CASME, CASME II).

The SMIC database [13] contains 77 spontaneous micro-expressions recorded from six subjects in two categories (negative or positive), which was further extended in [8] to include 164 micro-expression video clips elicited from 16 participants. The CASME database was developed in [9], [16] and contains 195 spontaneous micro-expressions recorded from 20 subjects. Seven categories (happiness, sadness, disgust, surprise, fear, repression, and tenseness) were annotated for these 195 spontaneous micro-expressions based on (1) action unit coding, (2) the main emotion of the video episode, and (3) participants' self-reports. Because there are few and insufficient samples for sadness, fear and repression, only four categories, namely, *positive* (happiness), *negative* (sadness, disgust, fear), *surprise* and the *others* (repression, tenseness) were used to train and evaluate the algorithm. CASME II was developed in [7] and has improved video quality and sample size compared with CASME:

- More details of facial muscle movement: compared to 100 fps in SMIC and 60 fps in CASME, the recording rate in CASME 2 is 200 fps;
- A larger face size in video clips: compared to 190×230 pixels of facial regions in SMIC and 150×190 pixels in CASME, the facial region in CASME II is approximately 280×340 pixels.

Micro-expressions in CASME II were elicited in a well-controlled laboratory environment, and proper illumination was used to remove light flickering. Among approximately 3,000 facial movements in 26 subjects, 247 micro-expressions were selected.

The recognition method developed in this paper is based on the optical flow field in micro-expression video clips. Optical flow has been widely studied in computer vision for more than three decades, and has also been successfully applied in full facial expression recognition [18]. Recently, the accuracy of optical flow estimation has improved significantly. The reader is referred to [19] for a comprehensive survey of state-of-the-art optical flow methods.

In this paper, we propose a simple yet effective feature, called the *Main Directional Mean Optical-flow* (MDMO) feature, for spontaneous micro-expression recognition. We apply a robust optical flow computation method [20], [21] on a textural part of images and postprocess it using an affine transformation such that the resulting optical flow field is insensitive to lighting conditions and head movements. Then we detect and partition the facial area into regions of interest (ROIs). Based on estimated optical flow fields, the MDMO feature is then designed to be a ROI-based, normalized statistic feature considering both local statistic motion information and its spatial location. We show that the MDMO feature can be used to efficiently recognize micro-expressions by working with a SVM classifier. We evaluate our proposed method on three spontaneous micro-expression databases (SMIC, CASME, and CASME II) and compare MDMO with two state-of-the-art baseline features (HOOF [22] and LBP-TOP [13], [23]). The experimental results show that the MDMO achieves better performance than HOOF and LBP-TOP.

3 Proposed Method

To build an effective feature for micro-expression recognition, we first detect and divide the face region into ROIs in video clips by utilizing a set of facial feature points obtained from an instance, called discriminative response map fitting (DRMF) [24], of the constrained local models [25], [26] (Section 3.1). Then, we compute the optical flow field in each frame of a micro-expression video clip (Section 3.2). We propose an alignment method in the optical flow field domain to reduce the influence of noise induced by head movements (Section 3.3). Based on the ROIs and aligned facial regions, we propose an novel MDMO feature (Section 3.4). Finally, we use the MDMO feature to train a SVM classifier for micro-expression recognition (Section 3.5). Experimental results with comparison to two baseline features (HOOF and LBP-TOP) on three spontaneous micro-expression databases (SMIC, CASME, CASME II) are presented in Section 4.

3.1 ROI Partitioning in Facial Regions

We use the DRMF method [24] to robustly detect a set of facial feature points in the facial region of the first frame in each micro-expression video clip. We briefly summarize this method below.

First, the Viola-Jones face detector [27] is used to locate the facial region in each frame. Second, a set of initial feature points is computed by extracting response patches followed by a low-dimensional projection. Third, DRMF iteratively disturbs these initial feature points by correlating with the target image a generated feature template that controls the shape and appearance variation learned from a training set. The statistical models of shape and appearance patterns of variability, which are presented in [28], [29], are applied in the DRMF method in a robust and accurate manner to locate 68 feature points in the facial region.

In our application, we use 66 feature points obtained from the DRMF method, i.e., two feature points identifying inner lip corners were not used in our micro-expression recognition. One example is shown in Fig. 1 (left). All 66 feature facial points in a frame are denoted as $FP = \{fp_1, fp_2, \ldots, fp_{66}\}$. The portion of the facial region in different frames often varies in practice. Our method uses the detected set FP of feature points to normalize the facial region in each frame. See Fig. 1 (right) for an example.

We further partition the normalized facial region into 36 ROIs (see Fig. 2, left). The locations of these ROIs are

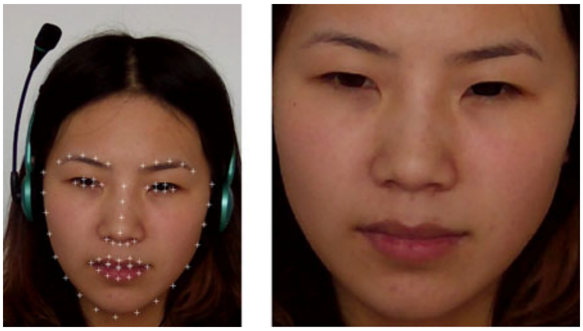


Fig. 1. Left: detection of 66 feature points in a facial region using the DRMF method. Right: the normalized facial area.

uniquely determined by the 66 feature points. For example, as illustrated in Fig. 2 (right), the position of the vertex shared by ROIs 21, 22, 25, and 26 is the average of positions of two feature points, f_4 and f_{33}. The specification rules of all vertices in the 36 ROIs are provided in Supplemental Material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2015.2485205. The partitioning of these 36 ROIs is partially based on the facial action coding system [30], e.g., the eyebrow is divided into outer and inner parts. Two guidelines are applied:

- ROI partitioning should not be too coarse; otherwise, many AUs will locate at similar or overlapping portions of the face. On some portions, such as the mouth, more partitions are provided to better discriminate different AUs.
- ROI partitioning should not be too dense. It is generally sufficient for each ROI to correspond to at least one AU; for example, ROI one is only related to AU2 and does not need to be subdivided further.

The correspondence between our ROI partitioning and AUs is summarized in Table 1.

In the following sections, we use grey-scale images of micro-expressions for a clear representation.

3.2 Computation of Optical Flow Fields

Optical flow infers the motion of objects by detecting the changing intensity of pixels between two image frames over time. In a video clip, a pixel at location (x, y, t) with intensity $I(x, y, t)$ will have moved by Δx, Δy and Δt between the two frames. According to the brightness constancy constraint, we have

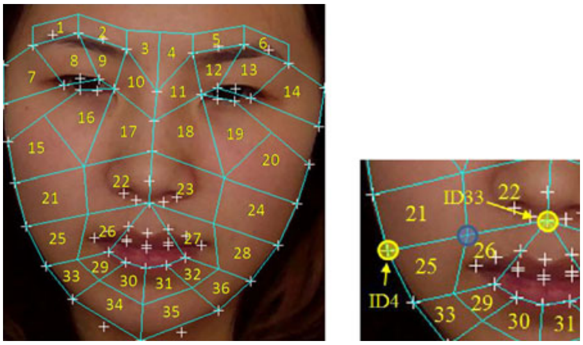


Fig. 2. Left: the partitioning of 36 regions-of-interest (ROIs), which are uniquely determined by the 66 feature points as shown in Fig. 1. Right: the position of the vertex shared by ROIs 21, 22, 25 and 26 is the average of positions of two feature points with IDs 4 and 33.

TABLE 1
Correspondence between ROI Numbers and the Numbers of Action Units (AUs)

ROI No	AU No	ROI No	AU No
1	AU2	19	AU6, AU7
2	AU1, AU4	20	AU6
3	AU1, AU4	21	AU6
4	AU1, AU4	22	AU10, AU11, AU12
5	AU1, AU4	23	AU10, AU11, AU12
6	AU2	24	AU6
7	AU6	25	AU12, AU15
8	AU2, AU5	26	AU10, AU12, AU13, AU14 AU15, AU18, AU23, AU24
9	AU1, AU4	27	AU10, AU12, AU13, AU14 AU15, AU18, AU23, AU24
10	AU9	28	AU12, AU15
11	AU9	29	AU16, AU18, AU20 AU23, AU24
12	AU1, AU4	30	AU18, AU23, AU24
13	AU2, AU5	31	AU18, AU23, AU24
14	AU6	32	AU16, AU18, AU20 AU23, AU24
15	AU6	33	AU16, AU20
16	AU6, AU7	34	AU17
17	AU9	35	AU17
18	AU9	36	AU16, AU20

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{1}$$

Assuming that the movement is small, the image constraint at $I(x, y, t)$ can be developed with a Taylor series to obtain:

$$\begin{aligned} & I(x + \Delta x, y + \Delta y, t + \Delta t) \\ & = I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \tau \end{aligned} \tag{2}$$

where τ is a higher-order infinitesimal. From these equations it follows that:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{3}$$

and

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t} = 0 \tag{4}$$

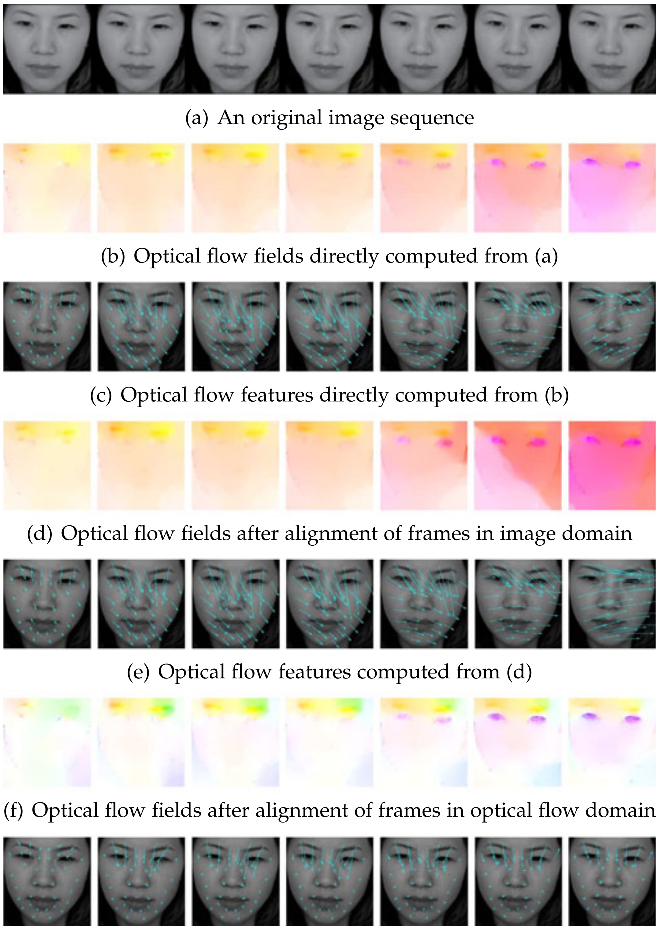
which results in

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \tag{5}$$

where V_x and V_y are the x and y components, respectively, of the velocity or optical flow of $I(x, y, t)$. Thus, between two frames with distance Δt, the optical flow value of a pixel at time t is expressed as a two-dimensional vector:

$$[V_x^t, V_y^t]^T \tag{6}$$

Many methods can be used to compute the optical flow field [19]. In our implementation, we use the method presented in [21].


(a) An original image sequence

(b) Optical flow fields directly computed from (a)

(c) Optical flow features directly computed from (b)

(d) Optical flow fields after alignment of frames in image domain

(e) Optical flow features computed from (d)

(f) Optical flow fields after alignment of frames in optical flow domain

(g) Optical flow features computed from (f)

Fig. 3. The optical flow of an image sequence in the micro-expression labeled by "tense" in CASME II. Due to limited space, we only show seven frames: Second, fourth, sixth, eight, 10th, 12th, 24th. Optical flow fields in (b), (d) and (f) are visualized using the color coding scheme in [31].

The brightness constancy constraint is suitable for the CASME II database, in which the micro-expressions were recorded in a well-controlled laboratory environment and proper illumination was used to remove light flickering. However, for the databases of SMIC and CASME, the illumination changes between image frames may influence the accuracy of optical flow estimation. To address the intensity inconsistency problem, we use the method in [20] to preprocess the image sequence: each image is decomposed into two parts, that is, a structural part and a textural part. It was shown in [20] that the intensity inconsistency due to shadow and shading reflections can only appear in the original image and the structural part, and thus, the computation of optical flow in the textural part can provide an accurate result. One example of the optical flow computation is shown in Figs. 3b, 3d, and 3f.

3.3 Face Alignment in the Optical Flow Domain

In the short duration of a micro-expression, there may be a small rotation and translation of the facial region in the image sequence. To correct this small head movement, for a micro-expression video clip, we use the positions of some feature points in the first frame. Among the 66 facial feature points (Fig. 1, left) detected by using the DRMF method, we choose 13 feature points (Fig. 4) including one at the nose root and the others at the contour of the facial region. These 13 feature points are least affected by the actions of various micro-expressions.

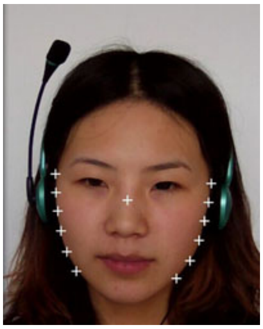


Fig. 4. At the first frame of a micro-expression, 13 feature points (including one at the nose root and the others at the contour of facial region) are detected by the DRMF method to align all subsequent frames.

For each frame f_i $(i \neq 1)$ in a micro-expression video clip, we align f_i with f_1 by computing the optical flow between f_i and f_1. Denote the resulting optical flow field in f_i as O_i. Let

$$\mathbf{P}^1 = \begin{bmatrix} p^1_{x1} & p^1_{x2} & \cdots & p^1_{x13} \\ p^1_{y1} & p^1_{y2} & \cdots & p^1_{y13} \end{bmatrix}^T$$

be the positions of 13 chosen feature points in the first frame f_1. Given O_i, the positions $\mathbf{P}^i$ of 13 feature points in f_i,

$$\mathbf{P}^i = \begin{bmatrix} p^i_{x1} & p^i_{x2} & \cdots & p^i_{x13} \\ p^i_{y1} & p^i_{y2} & \cdots & p^i_{y13} \end{bmatrix}^T,$$

can be determined by

$$\begin{cases} p^i_{xj} = p^1_{xj} + V^i_{xj} \\ p^i_{yj} = p^1_{yj} + V^i_{yj} \end{cases}, \quad j = 1, 2, \ldots, 13 \tag{7}$$

where $[V^i_{xj}, V^i_{yj}]^T$ is an optical flow vector of the jth feature point in O_i (ref. Eq. (6)). Given $\mathbf{P}^1$ and $\mathbf{P}^i$, an affine transformation matrix $\mathbf{T}^i$ can be readily obtained by solving

$$\arg\min_{\mathbf{T}^i} \|\mathbf{P}^i\mathbf{T}^i - \mathbf{P}^1\|_2 \tag{8}$$

Our method presented above is to align the frames in the optical flow domain. That is, the optical flow in each frame is first computed, and then, this frame is aligned with the first frame by applying an affine transformation determined by the correspondence of optical flows. There is another possible method to align the frames in the image domain. That is, 13 feature points are first detected in all frames by using the DRMF method, and then, each frame is aligned with the first frame by applying an affine transformation determined by the correspondence of feature points. One example of alignment in the optical flow domain is illustrated in Fig. 3f, and one example of alignment in the image domain is illustrated in Fig. 3d. Our optical-flow-domain-based alignment method can lead to a smoother and better face alignment than the image-domain-based method because the optical flow

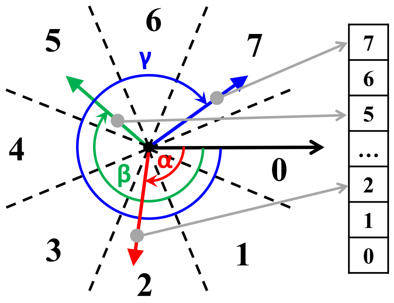


Fig. 5. The histogram of oriented optical flow with a bin number 8. For example, the optical flow vectors of angles α, β and γ belong to the second, fifth and seventh bins.

estimation is more robust than feature point detection in each frame.

3.4 MDMO Feature

In this section, we propose a MDMO feature for micro-expression recognition.

The optical flow between the first frame f_1 and each f_i of subsequent frames after alignment is denoted as $[V_x^i, V_y^i]^T$. We convert the euclidean coordinates $[V_x^i, V_y^i]^T$ into polar coordinates (ρ_i, θ_i), where ρ_i and θ_i are the magnitude and direction of the optical flow vectors, respectively.

In each frame f_i, we consider the optical flow inside each ROI R_i^k, where $i = 2, 3, \ldots, n_f$ is the index of frames and $k = 1, 2, \ldots, 36$ is the index of ROIs. Inside each R_i^k, a histogram of oriented optical flow (HOOF) [22] is computed as follows. Denote the optical flow vector at a location $p \in R_i^k$ as $\mathbf{u}_i^k(p) = (\rho_i^k(p), \theta_i^k(p))$. All of the optical flow vectors $\mathbf{u}_i^k(p)$, $p \in R_i^k$, are classified into eight bins (see Fig. 5) according to their direction $\theta_i^k(p)$.

In contrast to the original HOOF feature, which uses a distribution of optical flow represented by a normalized histogram, we select the bin in which the number of optical flow vectors is maximum and compute a mean vector as the feature vector to characterize ROI R_i^k:

$$\bar{\mathbf{u}}_i^k = \frac{1}{|B_{max}|} \sum_{\mathbf{u}_i^k(p) \in B_{max}} \mathbf{u}_i^k(p) \tag{9}$$

where B_{max} is the set of optical flow vectors falling in the bin with maximum count and $|\cdot|$ denotes the set cardinality.

Note that $\bar{\mathbf{u}}_i^k = (\bar{\rho}_i^k, \bar{\theta}_i^k)$, in which $\bar{\theta}_i^k$ is called the *main direction* of the optical flow in ROI R_i^k. Furthermore, we build a feature Ψ_i for the ith frame by

$$\Psi_i = \left(\bar{\mathbf{u}}_i^1, \bar{\mathbf{u}}_i^2, \ldots, \bar{\mathbf{u}}_i^{36}\right) \tag{10}$$

The dimension of Ψ_i is $36 \times 2 = 72$, where 36 is the number of ROIs.

After extracting the optical flow feature Ψ_i for every frame f_i, a micro-expression video clip is represented by optical flow feature series $\Gamma = (\Psi_1, \Psi_2, \ldots, \Psi_{n_f})$, where n_f is the frame number of the video clip. Considering that the frame number n_f may be various in different clips and the magnitudes of feature vectors in different ROIs may be widely distributed, we use the following normalization.

We define the Cartesian coordinate counterparts of polar coordinates $\bar{\mathbf{u}}_i^k$ and feature Ψ_i as $C(\bar{\mathbf{u}}_i^k) = (x_i^k, y_i^k)$ and $C(\Psi_i) = \left(C(\bar{\mathbf{u}}_i^1), C(\bar{\mathbf{u}}_i^2), \ldots, C(\bar{\mathbf{u}}_i^{36})\right)$, respectively. We use a Cartesian coordinate average to obtain the MDMO feature:

$$C(\overline{\Psi}) = \frac{1}{n_f} \sum_{i=1}^{n_f} C(\Psi_i) \tag{11}$$

Then, we convert $C(\overline{\Psi})$ back into polar coordinates represented by

$$\overline{\Psi} = \left[\left(\bar{\rho}_1, \bar{\theta}_1\right)^T, \left(\bar{\rho}_2, \bar{\theta}_2\right)^T, \ldots, \left(\bar{\rho}_{36}, \bar{\theta}_{36}\right)^T\right] \tag{12}$$

Because the strengths of the main directions in different video clips may be different, we further normalize the magnitudes in $\overline{\Psi}$ by

$$\widetilde{\rho}_k = \frac{\bar{\rho}_k}{\max\{\bar{\rho}_j, j = 1, 2, \ldots, 36\}}, \quad k = 1, 2, \ldots, 36 \tag{13}$$

Note that $\overline{\Psi}$ is a feature for one micro-expression video clip and different video clips have different scaling factors $\max\{\bar{\rho}_j, j = 1, 2, \ldots, 36\}$ in (13). Finally, the normalized MDMO feature for a micro-expression video clip is represented by $\widetilde{\Psi}$:

$$\widetilde{\Psi} = \left[\left(\widetilde{\rho}_1, \bar{\theta}_1\right)^T, \left(\widetilde{\rho}_2, \bar{\theta}_2\right)^T, \ldots, \left(\widetilde{\rho}_{36}, \bar{\theta}_{36}\right)^T\right] \tag{14}$$

One example of the normalized MDMO feature $\widetilde{\Psi}$ is shown in Fig. 3g.

The normalized MDMO feature has a good capacity to recognize micro-expressions, which is demonstrated in the next section, due to the following reasons:

- The optical flow is a classic pattern for characterizing motion in the image sequence. The ROI-based MDMO feature, which uses statistic information of the optical flow in each ROI, considers both local motion information and its spatial location.
- The optical flow feature that we compute is not only insensitive to translation and rotation (by face alignment in the optical flow domain (Section 3.3)), but also robust with respect to illumination variations (by preprocessing the image sequence into a structural part and a textual part (Section 3.2)).

The difference between our normalized MDMO feature and HOOF features [22] lies in the following aspects:

- The original HOOF feature is a normalized histogram quantized by a number of bins, computed from the entire facial region. As a holistic feature, even using a large number of bins in HOOF cannot effectively distinguish the micro-expressions. If we trivially apply the HOOF feature in each ROI, the resulting combinatorial HOOF feature will have a high dimension (e.g., if there are eight bins in a HOOF, the dimension is $36 \times 8 \times 2 = 576$), which contains unnecessary and redundant dimensions.

TABLE 2
The Best Recognition Rates of Four Features for Micro-Expression Recognition Using Leave-One-Subject-Out (LOSO) Cross-Validation in Three Databases

CASME	MDMO	LBP-TOP	HOOF-whole	HOOF-ROIs
	68.86%	64.07%	49.70%	55.69%
CASME II	MDMO	LBP-TOP	HOOF-whole	HOOF-ROIs
	67.37%	57.16%	42.80%	52.12%
SMIC	MDMO	LBP-TOP	HOOF-whole	HOOF-ROIs
	80.0%	71.40%	51.43%	61.43%

- Our normalized MDMO feature selects the strongest component from a ROI's HOOF feature (i.e., the main direction) and incorporate ROIs' combinatorial information. Therefore, the MDMO feature has a low dimension of $36 \times 2 = 72$ and achieves a good trade-off between the number of dimensions and effectiveness of characterizing micro-expressions.

The experiments presented in Section 4 show that the MDMO feature recognizes micro-expressions better than the original HOOF and combinatorial HOOF features.

Furthermore, our normalized MDMO feature is computationally simple. Similar to the block-based LBP-TOP feature [23] that are widely used in recognition of facial expressions [23] and micro-expressions [13], our feature is also extracted from a small local neighborhood (called region-based in this paper) that uses both local motion information and spatial locations. A significant advantage of the MDMO feature is that it is insensitive to the number of frames in image sequences. As a comparison, the LBP-TOP feature used in [13] requires a sufficient number of frames to extract stable features, and thus, for a short micro-expression with a small number of frames, a temporal interpolation model has to be used with the LBP-TOP feature to interpolate the limited number of frames in a low-dimensional manifold to obtain a sufficient number of artificial frames.

3.5 Micro-Expression Recognition

To recognize micro-expressions, we partition the normalized MDMO feature in Eq.(14) into two parts:

1) The magnitude part represented by $P = (\widetilde{\rho}_1, \widetilde{\rho}_2, \ldots, \widetilde{\rho}_{36})$,
2) the direction part represented by $\Theta = (\overline{\theta}_1, \overline{\theta}_2, \ldots, \overline{\theta}_{36})$.

Then, we introduce one parameter λ to balance the effect of P and Θ, and rewrite the feature into a one-row vector:

$$\overline{\overline{\Psi}} = (\lambda P, (1-\lambda)\Theta) \tag{15}$$

Based on a given micro-expression database, we use SVM with the polynomial kernel $\mathcal{K}(x_i, x_j) = (\gamma x_i^T x_j + coef)^{degree}$, and the optimal value of λ is determined by obtaining the best leave-one-out cross-validation result. The experiment details, including the choice of optimal parameters in the polynomial kernel, are presented in the next section.

The overall recognition algorithm is summarized in Algorithm 1.

Algorithm 1. Micro-Expression Recognition Using the Normalized MDMO Feature. A Micro-Expression DataBase $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ is given, in which ω_i is ith video Clip in the Database Ω

```
1:  I. Data Preprocessing
2:  for each $\omega_i$, $i \in n$ do
3:      for each frame $f_{i,j} \in \omega_i$ do
4:          Locate facial region by the Voila-Jones face detector
            and normalize the facial region;
5:      end for
6:      Detect 66 facial feature points in the first frame using
        the DRMF method;
7:      for each frame $f_{i,j} \in \omega_i$ do
8:          Compute the optical flow on the image sequence
            with normalized facial areas (ref. Fig. 1) using
            the illumination insensitive method presented in
            Section 3.2;
9:          Align the optical flow to remove the possible
            head movement using the method proposed
            in Section 3.3;
10:         Partition the face into 36 regions of interest (ROIs)
            using the method proposed in Section 3.1;
11:     end for
12: end for
13: II. Feature extraction
14: for each video clip $\omega_i$, $i \in n$, after preprocessing do
15:     Compute the main directional mean optical-flow
        feature $\overline{\Psi}_i$ in Eq. (12);
16:     Normalize the magnitudes in $\overline{\Psi}_i$ as in Eq. (13) and
        obtain the normalized MDMO feature $\widetilde{\Psi}_i$ as in Eq. (14);
17:     Pack the feature $\widetilde{\Psi}_i$ into one-row vector $\overline{\overline{\Psi}}_i$ with a
        parameter $\lambda$ as in Eq. (15);
18: end for
19: III. Training and recognition
20: Determine the optimal value of $\lambda$ and optimal parameters
    of polynomial kernel in SVM by obtaining the best leave-
    one-out cross-validation result;
21: Recognition with the optimal value of $\lambda$.
```

4 Experiment

In this section, we compare our proposed MDMO feature with two baseline features in micro-expression recognition, i.e., LBP-TOP [13], [23] and HOOF [22]. We use two versions of the HOOF feature:

- HOOF-whole: This is the original HOOF feature applied to the entire facial region. Four to ten bins are used in the HOOF histogram.
- HOOF-ROIs: This is a combinatorial HOOF feature, created by applying the HOOF feature in each of the 36 ROIs. Four to ten bins are used in each ROI's HOOF histogram, and the dimension of HOOF-ROIs ranges from $36 \times 4 \times 2 = 288$ to $36 \times 10 \times 2 = 720$.

The detailed performance on three databases (CASME, CASME II and SMIC) are presented in the following subsections. The main results are summarized in Table 2.

Our micro-expression recognition algorithm relied on a face alignment in the optical flow domain (Step 9 in Algorithm 1). Our alignment method (Section 3.3) used 13 facial feature points, including one inner facial point (at nose root)

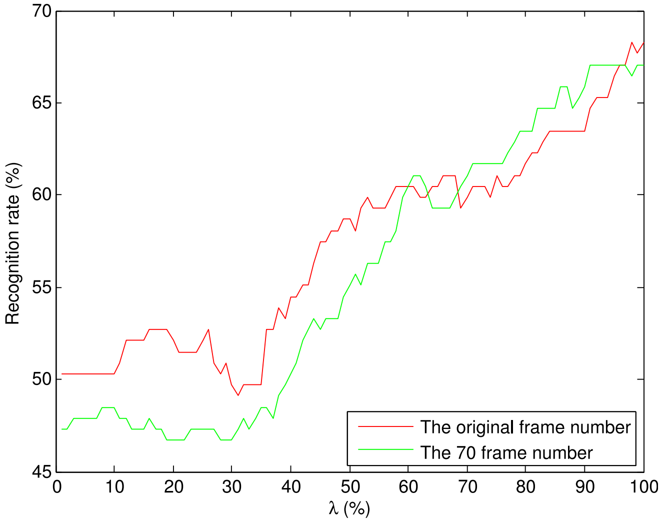


Fig. 6. Using original frame numbers in CASME, by leave-one-subject-out (LOSO) cross-validation, MDMO achieved the best recognition rate 68.26% when $\lambda = 0.98$ and SVM parameters are $\gamma = 0.28$, $coef = 0$ and $degree = 2$. This figure illustrated the recognition rates (red curve) when λ ranged in $[0\%, 100\%]$ with fixed SVM parameters $\gamma = 0.28$, $coef = 0$ and $degree = 2$. We also illustrated the recognition rates (green curve) using normalized frame number 70 with the same SVM parameters, demonstrating that MDMO is insensitive to frame numbers.

and twelve contour points. Both micro-expressions and head movements can disturb the positions of these points and thus affect the stability of face alignment. In Supplemental Material, available online, both qualitative and quantitative studies are presented, showing that our alignment method is very stable.

4.1 Evaluation on CASME

The CASME database [9] contains 195 spontaneous micro-expressions. These micro-expressions were recorded from 20 subjects using a 60 fps camera. Two expert coders were recruited in the work [9] to code the duration and AU combination in these micro-expressions. They independently spotted the onset, apex and offset frames, and arbitrated the disagreement. The reader is referred to [9] for details of the coding and labeling methods.

Among all 195 samples in CASME, several samples in which the 66 facial feature points in the first frame could not be correctly detected by using the DRMF method [24] were removed from our experiment. Then, we used 167 samples from 16 subjects, categorized in four classes: *Positive* (9 samples), *Negative* (48 samples), *Surprise* (15 samples), and *Others* (95 samples).

Optimal SVM parameter setting. LIBSVM [32] with the polynomial kernel $\mathcal{K}(x_i, x_j) = (\gamma x_i^T x_j + coef)^{degree}$ was used in our experiment. Multiclass classification can be performed in LIBSVM. In our application with $k = 4$ classes, $k(k-1)/2 = 6$ classifiers were constructed, each of which was used to train data from two classes; see Section 7 in [32] for further details. As suggested in the LIBSVM manual[1], the default setting of parameters in the polynomial kernel was $\gamma = 1/num\_features = 0.014$, $coef = 0$ and $degree = 3$. To find an optimal set of parameters, we search the spaces $\gamma \in [0, 1]$ with an interval 0.01, $coef \in \{0, 1\}$ and $degree \in [1, 10]$ with an interval of 1. In other words, for each set of parameters, the recognition rate was computed, and the optimal set corresponded to the highest recognition rate. To use the MDMO feature, we also searched the space $\lambda \in [0\%, 100\%]$ with an interval of 1%.

Subject-independent evaluation. Leave-one-subject-out (LOSO) cross-validation was applied for subject-independent evaluation, i.e., in each fold, one subject was used as the test set, and the others were used as the training set. After 16 folds, each subject has been used as the test set once, and the final recognition accuracy was calculated based on all of the results. The experimental results showed that the MDMO feature achieved the best recognition rate (68.26%) in CASME with $\lambda = 0.98$, $\gamma = 0.28$, $coef = 0$ and $degree = 2$. The recognition rates when changing λ in $[0\%, 100\%]$ with fixed SVM parameters of $\gamma = 0.28$, $coef = 0$ and $degree = 2$ are summarized in Fig. 6 (red curve). The result $\lambda = 0.98$ indicated that on CASME, the magnitude part P in the vector form (Eq. (15)) of the MDMO feature plays a dominant role.

Normalized frame number. In the CASME database, the number of frames for the shortest video clip sample is 10, and the number of frames for the longest sample is 68. Pfister et al. [13] proposed using graph embedding to temporally interpolate frames at arbitrary positions and then to obtain a sufficient number of frames. It was argued in [13] that the temporal interpolation of frames can extract more statistically stable LBP-TOP features for micro-expression recognition. In our experiment, to offer a fair comparison with LBP-TOP, the frame numbers of all samples were normalized to 70 by using linear interpolation, and we found that a frame number of more than 70 produced unnecessary redundance, which degraded recognition performance. We applied the MDMO feature on samples with a normalized frame number, and the best recognition rate was 68.86% at $\lambda = 0.91$ (with the optimal SVM parameters $\gamma = 0.22$, $coef = 0$ and $degree = 2$), which is slightly better than the performance (best recognition rate 68.26%) on the original frame numbers. In Fig. 7 (green curve), we present the recognition rates for normalized frame numbers when changing λ in $[0\%, 100\%]$ with fixed SVM parameters $\gamma = 0.22$, $coef = 0$ and $degree = 2$. In Figs. 6 (green curve) and 7 (red curve), we also compare the performance on the original and normalized frame numbers, and the results show that our MDMO feature is insensitive to the number of frames in the sample videos, which is a distinct characteristic of MDMO compared with LBP-TOP.

Comparison with LBP-TOP and HOOF. We applied LBP-TOP to extract features from 36 ROIs in all samples with a normalized frame number of 70. SVM was optimized in the same manner as for the MDMO features. In the setting of LBP-TOP, the radii values in axes X and Y ranged from 1 to 4. To avoid too many combinations of parameters, we chose $R_x = R_y$. The radius value R_t in axis T ranged from 1 to 4. The number of neighboring points in the XY, XT, and YT planes were all set to be 4 or 8. The uniform pattern and basic pattern were used in LBP coding. The recognition rates of LBP-TOP with all above parameters are listed in Table 5 in Appendix A. The results showed that LBP-TOP achieved the best LOSO recognition rate (64.07%), with the optimal parameters $R_x = R_y = 4$ and $R_t = 1$ in uniform

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

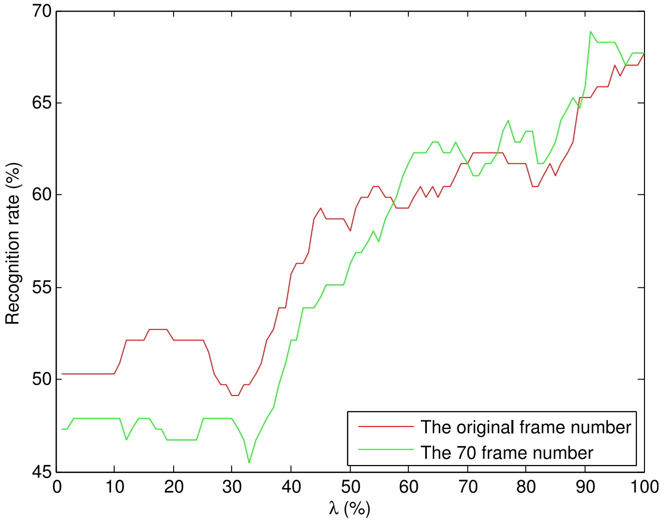


Fig. 7. Using normalized frame number 70 in CASME, by leave-one-subject-out (LOSO) cross-validation, MDMO achieved the best recognition rate 68.86% when $\lambda = 0.91$ and SVM parameters are $\gamma = 0.22$, $coef = 0$ and $degree = 2$. This figure illustrated the recognition rates (green curve) when λ ranged in $[0\%, 100\%]$ with fixed SVM parameters $\gamma = 0.22$, $coef = 0$ and $degree = 2$. We also illustrated the recognition rates (red curve) using original frame numbers with the same SVM parameters, demonstrating that MDMO is insensitive to frame numbers.

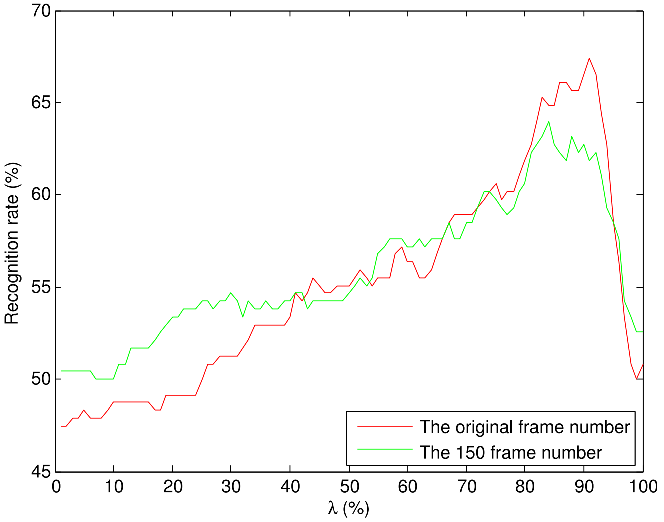


Fig. 8. Using original frame numbers in CASME II, by leave-one-subject-out (LOSO) cross-validation, MDMO achieved the best recognition rate 67.37% when $\lambda = 0.91$ and SVM parameters are $\gamma = 0.5$, $coef = 1$ and $degree = 2$. This figure illustrated the recognition rates (red curve) when λ ranged in $[0\%, 100\%]$ with fixed SVM parameters. We also illustrated the recognition rates (green curve) using normalized frame number 150 with the same SVM parameters, demonstrating that MDMO is insensitive to frame numbers.

pattern with 8 neighboring points. The best recognition rate (64.07%) of LBP-TOP is smaller than that of MDMO (68.86%). We also compare the MDMO feature with the features of HOOF-whole and HOOF-ROIs. The results of these two HOOF features on samples of normalized frame numbers with different parameters are summarized in Table 6 in Appendix A. The results showed that the best HOOF feature is HOOF-ROIs, and the best recognition rate (55.69%) is achieved with a bin number of 8. However, this best rate (55.69%) of HOOF-ROIs is considerably smaller than that of MDMO (68.86%).

Subject-dependent evaluation. In some previous works (e.g., [33], [34]), leave-one-video-out (LOVO) cross validation was used for subject-dependent evaluation, i.e., in each fold, one sample video clip was used as the test set, and the others were used as the training set. After 167 folds, each sample has been used as the test set once, and the final recognition rate was calculated based on all of the results. Except for replacing LOSO with LOVO, all of the other settings were the same as those in LOSO. Detailed experimental results of LOVO on MDMO, LBP-TOP and HOOF are summarized in Supplemental Material, available online. These results show that:

- using the original frame numbers, MDMO achieved the best LOVO recognition rate (73.65%) at $\lambda = 0.96$ with the optimal SVM parameters $\gamma = 0.28$, $coef = 1$ and $degree = 2$;
- using normalized frame numbers,
  - MDMO achieved the best LOVO recognition rate (75.45%) at $\lambda = 0.87$ with the optimal SVM parameters $\gamma = 0.38$, $coef = 0$ and $degree = 1$;
  - LBP-TOP achieved the best LOVO recognition rate (73.05%) with the optimal parameters $R_x = R_y = 4$, $R_t = 1$ in a uniform pattern with 4 neighboring points;
  - HOOF-whole achieved the best LOVO recognition rate (56.89%) with a bin number of 5;
  - HOOF-ROIs achieved the best LOVO recognition rate (65.27%) with a bin number of 4.

We concluded that in the CASME database, MDMO is better than LBP-TOP, HOOF-whole and HOOF-ROIs in both subject-independent and subject-dependent evaluations.

4.2 Evaluation on CASME II

The CASME II database [7] contains 246 spontaneous micro-expressions, recorded using a high-speed 200 fps camera in an elaborate environment that had proper illumination without flickering light. These micro-expressions were recorded from 26 subjects and were selected from nearly 2,500 elicited facial movements. The AU coding and labelling methods used were similar to those for CASME, and the reader is referred to [7] for details.

In CASME II, the samples in which the facial feature points in the first frame cannot be correctly detected by using the DRMF method [24] were removed. Then, we used 236 samples from 26 subjects, categorized into four classes: *Positive* (31 samples), *Negative* (65 samples), *Surprise* (21 samples), and *Others* (119 samples).

Subject-independent evaluation. The same search strategy of finding an optimal SVM parameter setting in CASME was used in CASME II. Via LOSO cross-validation, the MDMO feature achieved the best recognition rate (67.37%) in CASME II at $\lambda = 0.91$ with the optimal SVM parameters $\gamma = 0.5$, $coef = 1$ and $degree = 2$. The recognition rates when changing λ in $[0\%, 100\%]$ and fixing the SVM parameters at the optimal setting are summarized in Fig. 8 (red curve). All of these results were obtained using the original frame numbers in CASME II.

Normalized frame number. In the CASME II database, the frame number of the shortest sample is 24 and that of

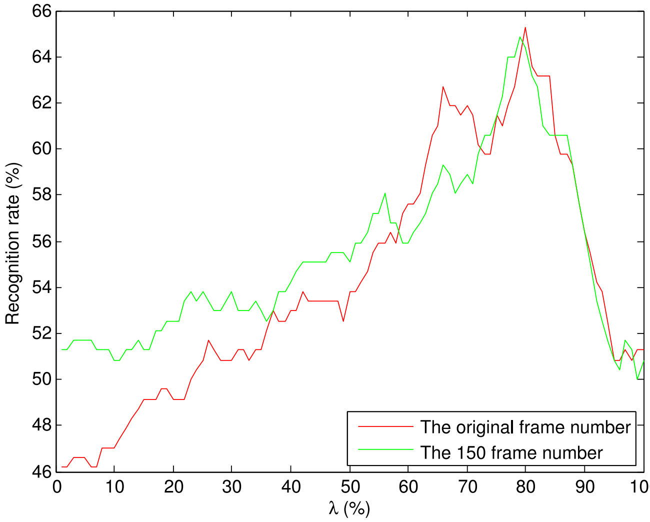


Fig. 9. Using normalized frame number 150 in CASME II, by leave-one-subject-out (LOSO) cross-validation, MDMO achieved the best recognition rate 64.83% when $\lambda = 0.79$ and SVM parameters are $\gamma = 0.24$, $coef = 1$ and $degree = 2$. This figure illustrated the recognition rates (green curve) when λ ranged in $[0\%, 100\%]$ with fixed SVM parameters. We also illustrated the recognition rates (red curve) using original frame numbers with the same SVM parameters, demonstrating that MDMO is insensitive to frame numbers.

the longest sample is 146. To offer a fair comparison with LBP-TOP, which achieved the best performance in a temporal interpolation model [13], we normalized the frame numbers of all samples in CASME II to 150 via linear interpolation. We applied the MDMO feature on samples of normalized frame numbers, and the best recognition rate was 64.83% at $\lambda = 0.79$ (with the optimal SVM parameters $\gamma = 0.24$, $coef = 1$ and $degree = 2$), which is smaller than the performance (best recognition rate of 67.37%) on the original frame numbers. In Fig. 9 (green curve), we illustrate the recognition rates on normalized frame numbers when changing λ in $[0\%, 100\%]$ and fixing the SVM parameters at the optimal setting. In both Figs. 8 (green curve) and 9 (red curve), we also compare the performance on the original and normalized frame numbers, and the results consistently show that our MDMO feature is insensitive to the number of frames in the sample videos.

Comparison with LBP-TOP and HOOF. The LBP-TOP parameter setting used was the same as in CASME. The uniform pattern and basic pattern were used in LBP coding. The recognition rates of LBP-TOP with different parameters are summarized in Table 7 in Appendix A. We also compare the MDMO feature with the features of HOOF-whole and HOOF-ROIs. The results of these two HOOF features on samples of normalized frame number are summarized in Table 8 in Appendix A. The results showed that

- LBP-TOP achieved the best LOSO recognition rate (57.16%), with the optimal parameters $R_x = R_y = 4$ and $R_t = 2$ in uniform pattern with 8 neighboring points;
- the best HOOF feature is HOOF-ROIs and the best recognition rate (52.12%) is achieved with a bin number of 5;
- both the best recognition rates of LBP-TOP (57.16%) and HOOF (52.12%) were considerably smaller than the best recognition rate of MDMO (67.37%).

TABLE 3
The Confusion Matrix of MDMO on the CASME II Database at the Best Recognition Rate, by Leave-One-Subject-Out (LOSO) Cross-Validation

		Ground truth			
		Positive	*Negative*	*Surprise*	*Others*
Prediction	*Positive*	45.16%	3.08%	4.76%	7.61%
	Negative	6.45%	53.84%	4.76%	10.92%
	Surprise	3.23%	3.08%	66.67%	0.80%
	Others	45.16%	40.00%	23.81%	80.67%

TABLE 4
The Confusion Matrix of LBP-TOP on the CASME II Database at the Best Recognition Rate, by Leave-One-Subject-Out (LOSO) Cross-Validation

		Ground truth			
		Positive	*Negative*	*Surprise*	*Others*
Prediction	*Positive*	19.35%	1.54%	4.76%	9.24%
	Negative	3.23%	32.31%	14.29%	15.13%
	Surprise	12.90%	1.54%	19.05%	4.20%
	Others	64.52%	64.61%	61.90%	71.43%

We further compared the confusion matrices of MDMO (Table 3) and LBP-TOP (Table 4) when they both obtained the best recognition rates. The results showed that compared to LBP-TOP, MDMO had a better recognition capacity in all four classes.

Subject-dependent evaluation. Detailed experimental results of LOVO cross validation on MDMO, LBP-TOP and HOOF are summarized in Supplemental Material. These results showed that

TABLE 5
Micro-Expression Recognition Rates (%) of LBP-TOP in CASME with Respect to the Number n of Neighboring Points, by Leave-One-Subject-Out (LOSO) Cross-Validation

Parameters	Uniform pattern		Basic pattern	
	$n = 4$	$n = 8$	$n = 4$	$n = 8$
$R_x = R_y = 1, R_t = 1$	56.89%	55.68%	56.29%	55.68%
$R_x = R_y = 1, R_t = 2$	52.51%	52.96%	52.51%	53.52%
$R_x = R_y = 1, R_t = 3$	53.57%	52.80%	53.15%	55.07%
$R_x = R_y = 1, R_t = 4$	52.65%	51.44%	52.65%	54.07%
$R_x = R_y = 2, R_t = 1$	58.08%	56.89%	58.08%	56.29%
$R_x = R_y = 2, R_t = 2$	51.65%	51.84%	51.65%	53.63%
$R_x = R_y = 2, R_t = 3$	52.51%	51.14%	52.51%	53.88%
$R_x = R_y = 2, R_t = 4$	55.22%	51.84%	55.22%	52.73%
$R_x = R_y = 3, R_t = 1$	56.28%	57.49%	56.29%	56.29%
$R_x = R_y = 3, R_t = 2$	52.70%	52.73%	52.70%	54.52%
$R_x = R_y = 3, R_t = 3$	53.68%	51.81%	53.68%	53.21%
$R_x = R_y = 3, R_t = 4$	53.56%	55.90%	53.56%	52.32%
$R_x = R_y = 4, R_t = 1$	56.29%	**64.07%**	56.89%	55.68%
$R_x = R_y = 4, R_t = 2$	51.39%	54.75%	50.98%	53.21%
$R_x = R_y = 4, R_t = 3$	53.27%	53.83%	52.99%	53.37%
$R_x = R_y = 4, R_t = 4$	53.67%	57.03%	53.67%	50.92%

TABLE 6
Micro-Expression Recognition Rates (%) of Two HOOF Features in CASME with Respect to Different Bin Numbers n, by Leave-One-Subject-Out (LOSO) Cross-Validation

HOOF-whole feature						
$n=4$	5	6	7	8	9	10
42.52%	49.10%	47.90%	49.70%	49.70%	42.52%	42.52%
HOOF-ROIs feature						
$n=4$	5	6	7	8	9	10
55.09%	53.89%	54.49%	54.49%	**55.69%**	55.09%	55.09%

TABLE 8
Micro-Expression Recognition Rates (%) of Two HOOF Features in CASME II with Respect to Different Bin Numbers n, by Leave-One-Subject-Out (LOSO) Cross-Validation

HOOF-whole feature						
$n=4$	5	6	7	8	9	10
41.10%	42.80%	41.10%	42.80%	42.80%	41.10%	41.10%
HOOF-ROIs feature						
$n=4$	5	6	7	8	9	10
47.88%	**52.12%**	51.69%	50.85%	50.42%	47.88%	47.88%

- using the original frame numbers, MDMO achieved the best LOVO recognition rate (71.61%) at $\lambda = 0.91$ with the optimal SVM parameters $\gamma = 0.46$, $coef = 1$ and $degree = 2$;
- using normalized frame numbers,
  - MDMO achieved the best LOVO recognition rate (70.34%) at $\lambda = 0.76$ with the optimal SVM parameters $\gamma = 0.36$, $coef = 0$ and $degree = 2$;
  - LBP-TOP achieved the best LOVO recognition rate (61.86%) with the optimal parameters $R_x = R_y = 4$, $R_t = 3$ in a uniform pattern with 4 neighboring points;
  - HOOF-whole achieved the best LOVO recognition rate (47.03%) with a bin number of 7;
  - HOOF-ROIs achieved the best LOVO recognition rate (58.90%) with a bin number of 5.

We concluded that in the CASME II database, MDMO is consistently better than LBP-TOP, HOOF-whole and HOOF-ROIs in both the subject-independent and subject-dependent evaluations.

4.3 Evaluation on SMIC

The SMIC database [8] was built by recording 20 subjects, in which 164 spontaneous micro-expressions were selected from 16 subjects. All data were recorded by using a high-speed camera of 100 fps, and half of them were also recorded by using a 25 fps visible light camera and 25 fps near infrared camera. We used the first version of SMIC, the same material (i.e., 17 positive and 18 negative samples from six subjects) in [13], such that a consistent comparison could be made with LBP-TOP, whose performance on SMIC was analyzed in [13].

TABLE 7
Micro-Expression Recognition Rates (%) of LBP-TOP in CASME II with Respect to the Number n of Neighboring Points, by Leave-One-Subject-Out (LOSO) Cross-Validation

Parameters	Uniform pattern		Basic pattern	
	$n=4$	$n=8$	$n=4$	$n=8$
$R_x = R_y = 1, R_t = 1$	42.80%	50.00%	42.80%	50.00%
$R_x = R_y = 1, R_t = 2$	44.92%	52.10%	45.63%	50.85%
$R_x = R_y = 1, R_t = 3$	44.72%	49.85%	44.72%	52.76%
$R_x = R_y = 1, R_t = 4$	47.16%	49.74%	48.41%	51.93%
$R_x = R_y = 2, R_t = 1$	47.03%	46.19%	46.19%	49.15%
$R_x = R_y = 2, R_t = 2$	47.78%	50.50%	47.68%	53.72%
$R_x = R_y = 2, R_t = 3$	46.15%	50.38%	46.77%	54.43%
$R_x = R_y = 2, R_t = 4$	48.26%	55.61%	48.26%	55.43%
$R_x = R_y = 3, R_t = 1$	41.95%	48.31%	42.37%	48.31%
$R_x = R_y = 3, R_t = 2$	47.62%	51.30%	47.47%	54.43%
$R_x = R_y = 3, R_t = 3$	44.02%	47.10%	44.17%	51.99%
$R_x = R_y = 3, R_t = 4$	47.74%	49.71%	46.48%	52.82%
$R_x = R_y = 4, R_t = 1$	42.37%	49.58%	42.37%	47.03%
$R_x = R_y = 4, R_t = 2$	43.67%	**57.16%**	43.67%	52.72%
$R_x = R_y = 4, R_t = 3$	45.17%	51.46%	45.17%	50.82%
$R_x = R_y = 4, R_t = 4$	50.78%	55.53%	44.33%	53.07%

Here, we only present subject-independent evaluation results obtained via LOSO cross-validation. The results of subject-dependent evaluation obtained via LOVO cross validation were similar, and the performance was consistent with those in CASME and CASME II. The MDMO feature achieved the best recognition rate (80.0%) in SMIC at $\lambda = 0.80$ with the optimal SVM parameters $\gamma = 0.1$, $coef = 1$ and $degree = 4$. As a comparison, the best recognition rates of LBP-TOP, HOOF-whole and HOOF-ROIs were 71.40 , 51.43 , and 61.43 percent. We concluded that MDMO consistently had the best performance on SMIC.

5 Conclusion

In this paper, we proposed a simple yet effective MDMO feature for micro-expression recognition. The MDMO is a ROI-based optical flow feature, which makes use of both local statistic motion information (i.e., the mean of all optical flow vectors in a ROI falling into a bin with the maximum count) and its spatial location (i.e., the ROI to which it belongs). The feature dimension of MDMO is small, i.e., $36 \times 2 = 72$, where 36 is the number of ROIs. To obtain reliable optical flow vectors, we proposed an alignment method in the optical flow domain to remove noise induced by small head movements in micro-expression video clips. Experimental results on three spontaneous micro-expression databases (CASME, CASME II and SMIC) showed that compared to two baseline features, i.e., LBP-TOP and HOOF, MDMO consistently has the best performance in both subject-independent and subject-dependent evaluations.

Appendix

In this appendix, some detailed experimental results were presented.

In Section 4.1, the LBP-TOP and HOOF features were compared to our proposed MDMO feature on the CASME database. The recognition rates of LBP-TOP and two HOOF features with different parameters were summarized in Tables 5 and 6, respectively.

In Section 4.2, the LBP-TOP and HOOF features were compared to the MDMO feature on the CASME II database. The recognition rates of LBP-TOP and two HOOF features with different parameters were summarized in Tables 7 and 8, respectively.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments, which improved this paper. The authors also thank Mr. Bingjun Li for preparing the experimental results. This work was supported in part by the National Natural Science Foundation of China (61322206, 61521002, 61379095, 61375009), the Beijing Natural Science Foundation (4152055), the Open Projects Program of National Laboratory of Pattern Recognition (201306295), TNList Cross-discipline Foundation, the Academy of Finland and Infotech Oulu.

REFERENCES

[1] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation and Emotion*, vol. 35, no. 2, pp. 181–191, 2011.

[2] S. Weinberger, "Airport security: Intent to deceive," *Nature*, vol. 465, no. 7297, pp. 59–69, 2009.

[3] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Proc. Methods of Research in Psychotherapy*, 1966, pp. 154–165.

[4] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," Psychiatry, vol. 32, no. 1, pp. 88–106, 1969.

[5] P. Ekman, "METT: Micro expression training tool," *CD-ROM. Oakland*, 2003.

[6] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," presented at The Ann. Meeting of the Int. Commun. Assoc., 2009.

[7] 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*, vol. 9, no. 1, p. e86041, 2014.

[8] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recog. (FG)*, 2013, pp. 1–6.

[9] W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, vol. 136, pp. 82–87, 2014.

[10] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[11] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *J. Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[12] S. Porter and L. ten Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychological Sci.*, vol. 19, no. 5, pp. 508–514, 2008.

[13] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 1449–1456.

[14] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *Proc. 2011 Int. Conf. Automatic Face & Gesture Recog. and Workshops*, 2011, pp. 51–56.

[15] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in *Proc. IEEE Third Int. Conf. Crime Detection and Prevention (ICDP)*, 2009, pp. 1–6.

[16] 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 *Proc. 10th IEEE Int. Conf. and Workshops Automatic Face and Gesture Recog. (FG 2013)*, 2013, pp. 1–7.

[17] S. Afzal and P. Robinson, "Natural affect data-collection & annotation in a learning context," in *Proc. Third Int. Conf. Affective Comput. and Intell. Interaction and Workshops (ACII 2009)*, 2009, pp. 1–7.

[18] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 19, no. 7, pp. 757–763, Jul. 1997.

[19] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 115–137, 2014.

[20] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for tv-l 1 optical flow," *Statistical and Geometrical Approaches to Visual Motion Anal.*, Berlin, Germany: Springer, 2009, pp. 23–45.

[21] T. Senst, V. Eiselein, and T. Sikora, "Robust local optical flow for feature tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1377–1387, May 2012.

[22] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2009, pp. 1932–1939.

[23] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[24] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recog. (CVPR)*, 2013, pp. 3444–3451.

[25] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2006, pp. 929–938.

[26] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 1034–1041.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, vol. 1, 2001, pp. I-511–I-518.

[28] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models – their training and application," *Comput. Vis. and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[29] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

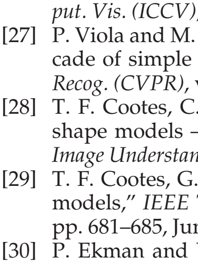
[30] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto," 1978.

[31] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.

[32] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

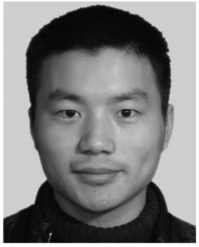
[33] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recog. (CVPR)*, 2011, pp. 137–144.

[34] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Apr. 2012.



Yong-Jin Liu received the BEng degree from Tianjin University, China, in 1998, and the MPhil and PhD degrees from the Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. He is currently an associate professor with the Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China. His research interests include computational geometry, multimedia, computer graphics and computer-aided design. He is a member of the IEEE and a member of IEEE Computer Society. For more information, visit http://cg.cs.tsinghua.edu.cn/people/~Yongjin/yongjin.htm.

Jin-Kai Zhang received the BEng degree in the Department of Computer Science and Technology, Tsinghua University, China, in 2012. He is working toward the master degree in the same department at Tsinghua University. His research interests include micro-expression recognition and multimedia computing.



Wen-Jing Yan received the PhD degree from Institute of Psychology, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an assistant professor in the Department of Psychology in Wenzhou University, China. His interests include facial expression and deception. His research interests include interdisciplinary research on facial expression and affective computing.

Su-Jing Wang received the Master's degree from the Software College of Jilin University, China, in 2007, and the PhD degree from the College of Computer Science and Technology of Jilin University in 2012. He is a postdoctoral researcher in Institute of Psychology, Chinese Academy of Sciences. He is One of Ten Selectees of the Doctoral Consortium at International Joint Conference on Biometrics 2011. He was named *Chinese Hawkin* by the Xinhua News Agency. His current research interests include pattern recognition, computer vision, and machine learning. He serves as an associate editor of Neurocomputing (Elsevier). He is a member of the IEEE. For more information, visit http://sujingwang.name.

Guoying Zhao received the PhD degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently an associate professor with the Center for Machine Vision Research, University of Oulu, Finland, where she has been a researcher since 2005. In 2011, she was selected to the highly competitive Academy Research Fellow position. She has authored or co-authored more than 100 papers in journals and conferences, and has served as a reviewer for many journals and conferences. She has lectured tutorials at ICPR 2006, ICCV 2009, and SCIA 2013, and authored/edited three books and two special issues in journals. She was a co-chair of the International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA) at ECCV2008, ICCV2009, and CVPR2011, ECCV2014 workshop on "Spontaneous Facial Behavior Analysis: Long term continuous analysis of facial expressions and micro-expressions" and ACCV 2014 workshop on "RoLoD: Robust local descriptors for computer vision" and a special session for IEEE International Conference on Automatic Face and Gesture Recognition 2013 (FG13). Her current research interests include image and video descriptors, gait analysis, dynamic-texture recognition, facial-expression recognition, human motion analysis, and person identification. She is a senior member of the IEEE.

Xiaolan Fu received the BS and MS degrees in psychology from Peking University, Peking, China, in 1984 and 1987, respectively, and the PhD degree from the Institute of Psychology, Chinese Academy of Sciences, Beijing, China, in 1990. Currently, she is the director of the Institute of Psychology, Chinese Academy of Sciences and vice director of State Key Laboratory of Brain and Cognitive Science. Her research interests include visual and computational cognition, including attention and perception, learning and memory, and affective computing. She serves as an associate editor of PsyCH Journal and journal of Protein & Cell. She is a member of the IEEE.