Microexpression Identification and Categorization using a Facial Dynamics Map

Feng Xu, Junping Zhang, James Z. Wang

***Abstract*—Unlike conventional facial expressions, microexpressions are instantaneous and involuntary reflections of human emotion. Because microexpressions are fleeting, lasting only a few frames within a video sequence, they are difficult to perceive and interpret correctly, and they are highly challenging to identify and categorize automatically. Existing recognition methods are often ineffective at handling subtle face displacements, which can be prevalent in typical microexpression applications due to the constant movements of the individuals being observed. To address this problem, a novel method called the Facial Dynamics Map is proposed to characterize the movements of a microexpression in different granularity. Specifically, an algorithm based on optical flow estimation is used to perform pixel-level alignment for microexpression sequences. Each expression sequence is then divided into spatiotemporal cuboids in the chosen granularity. We also present an iterative optimal strategy to calculate the principal optical flow direction of each cuboid for better representation of the local facial dynamics. With these principal directions, the resulting Facial Dynamics Map can characterize a microexpression sequence. Finally, a classifier is developed to identify the presence of microexpressions and to categorize different types. Experimental results on four benchmark datasets demonstrate higher recognition performance and improved interpretability.**

***Index Terms*—Microexpression, optical flow, expression recognition, emotions, face recognition**

I. Introduction

Analyzing and modeling facial expressions has been an important research focus in affective computing for over a decade. It has diverse applications, including face recognition and age or gender estimation [1]–[4]. For example, Rahulamathavan *et al.* [5] tackled the issue of expression recognition in an encrypted space for cloud computing based recognition while preserving privacy. Wang *et al.* [6] introduced thermal image for spontaneous/posed expression determination. 3D imaging has been used for improved recognition because it provides more information than 2D imaging [7]. Zheng *et al.* [8] proposed a multi-view strategy combined with group sparse technique for expression recognition. Recently, researchers have investigated spontaneous expressions that can reflect true emotions [9], [10].

As a type of spontaneous expression, microexpressions have not yet been explored extensively [11]–[15]. Microexpressions are involuntary and instant facial dynamics during the time the subject is experiencing emotions, including when the subject attempts to conceal emotions, especially in high-stake situations [16], [17]. As early as in 1969, Ekman observed microexpresssions when he analyzed an interview video of a patient with depression [18]. The patient who attempted to commit suicide showed brief yet intense sadness but resumed smiling quickly. Such microexpressions only last for a few frames in the standard 25 fps (frames per second) video but are often capable of revealing the actual emotions of subjects even though such leaks are unintentional. Although people can simulate or neutralize normal expressions, microexpressions are generally believed to be distinguishable from forged ones [19]–[22].

As being neither hidden nor forged, microexpressions are useful in affect monitoring and lie detection [20], [23], serving as a vital clue in law enforcement, evidence collection, clinical diagnosis [24], teaching assistance [25], teaching engagement assessment [26], and business negotiation [27].

However, as microexpressions are transient and subtle in essence, professional training is often required for people to spot a microexpression. Even with such training, a low accuracy of human performance has been reported [28].

The two difficulties of microexpression recognition, which originate from its short duration and subtle movement, considerably hinder the potential use of microexpression identification in practical applications. Thus having a system that can identify and categorize microexpressions automatically and accurately is desirable.

Some researchers have developed prototypes in recent years. Polikovsky *et al.* [11] designed a 3D orientation gradient histogram to describe facial area variations across time. Shreve *et al.* [13], [14] extracted strain map from microexpression frames, which indicates deformation extent of facial dynamics. One disadvantage for these prototypes is that they use posed microexpressions rather than spontaneous ones.

Pfister *et al.* [12] utilized a spatiotemporal local texture descriptor LBP-TOP, which is an extension to Local Binary Pattern in three orthogonal planes, to characterize a microexpression over time. However, their pre-processing stage heavily depends on the Active Shape Model (ASM) [29]. In this approach, tiny errors may be accumulated to unmanageable levels in subsequent stages. Wang *et al.* [30] improved the work by using independent color space as an aid for better

F. Xu and J. Zhang have been supported by National Natural Science Foundation of China (No. 61273299) and Ministry of Education of China (No.20120071110035). J. Zhang was visiting The Pennsylvania State University when the manuscript was completed. His visit was supported by the China Scholarship Council and the US National Science Foundation (NSF). J. Z. Wang has been supported by the NSF under Grant No. 1110970.

F. Xu and J. Zhang are with the Shanghai Key Laboratory of Intelligent Information Processing, Key Laboratory for Information Science of Electromagnetic Waves (MoE) and the School of Computer Science, Fudan University, Shanghai, 200433, China. (e-mails: {feng_xu, jpzhang}@fudan.edu.cn)

J. Z. Wang is with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA (e-mail: jwang@ist.psu.edu)

TABLE I
EXISTING DATASETS FOR MICROEXPRESSION RECOGNITION

DATASET		FRAME RATE	#PARTICIPANTS	#MICROEXPRESSIONS	#NON-MICRO.	ELICITATION	TAGGING
SMIC		100	6	76	76	SPONTANEOUS	EMOTION
SMIC2	HS	100	20	164	164	SPONTANEOUS	EMOTION
	VIS	25	10	71	71		
	NIR	25	10	71	71		
CASME		60	35	195	/	SPONTANEOUS	EMOTION/FACS
CASME II		200	35	247	/		
USF-HD		29.7	/	100	181	POSED	MICRO/NON-MICRO
POLIKOVSKY DATASET		200	10	/	/	POSED	FACS

feature extraction. Wang *et al.* [15] viewed a face in video as a 3rd-order tensor. Discriminant Tensor Subspace Analysis (DTSA) and Extreme Learning Machine (ELM) [31] were performed for microexpression recognition.

Throughout the relevant literature, we observe that little attention has been paid to fine-scale face alignment. Researchers either assumed that the alignment had already been coped with or relied on facial landmark locating methods such as the ASM. Furthermore, most existing approaches utilized texture-based features such as the Gabor filter and LBP, and the whole recognition process is carried out in a 'black box' fashion. Under these approaches, the extracted features cannot provide easily interpretable or intuitive information for understanding microexpressions.

To address these issues, we propose a novel approach, named the Facial Dynamics Map (FDM), for distinguishing images with a microexpression from those without a microexpression (or non-microexpression), and for categorizing different microexpressions. First, we estimate facial surface movement between adjacent frames. Next, the movements are extracted in a coarse-to-fine manner, indicating different levels of facial dynamics. With our approach, the issue of head movements can be alleviated. Finally, we use a classifier, such as Support Vector Machines (SVM), for both identification and categorization. We validate the proposed framework with experiments on several benchmark microexpression datasets.

Our *main contributions* can be summarized as follows.

1) We propose a new method, the FDM, to characterize facial dynamics in different granularity. Experimental results show that our method exceeds state-of-the-art approaches.
2) The proposed FDM provides an effective and intuitive understanding of human facial expressions. The algorithm can capture not only subtle facial movements but also relatively large ones.
3) Along with the FDM, we propose an iterative procedure to quickly calculate the principal direction for facial dynamics.
4) We utilize a simple yet effective one-dimensional histogram-based strategy to achieve fine-scale alignment in the pre-processing stage. Experimental results show that the strategy brings a small amount of improvement to the subsequent recognition.

We also give an overview of existing datasets for microexpression recognition and discuss three aspects of these datasets. Other researchers interested in this problem can benefit from this analysis of the datasets.

The rest of this paper is organized as follows. Section II surveys prevailing datasets for microexpression analysis. Section III reviews some related literature. In Section IV, we describe our method along with related pre-processing techniques. Experimental results are reported and analyzed in Section V. We conclude and suggest future directions in Section VI.

II. DATASETS

This section gives a brief overview of several published microexpression datasets. To our knowledge, there are six microexpression datasets published in the literature: (i) the Spontaneous Microexpression Corpus (SMIC) [12], (ii) SMIC2 [32], the successor to SMIC, (iii) the Chinese Academy of Sciences microexpression (CASME) [33], (iv) CASME II [34], the successor to CASME, (v) USF-HD [14], and (vi) the Polikovsky dataset [11].

Among them, SMIC2 consists of three subsets, *i.e.*, HS, VIS, and NIR. They were taken by a high speed camera, a normal visual camera, and a near-infrared camera, respectively.

It is worth noting that to construct a useful microexpression dataset for research, there are three crucial considerations: the frame rate, the tagging approach, and the elicitation protocol. Table I summarizes the features of these datasets that the research community has analyzed over the past decade.

As aforementioned, a microexpression is featured in short duration. A typical microexpression lasts for merely 1/3 to 1/25 of a second. A standard webcam (with a typical *frame rate* of 24 fps) may be unable to capture sufficient information and can even miss some key features. Thus high-speed cameras are preferred as they capture more abundant information that can benefit subsequent recognition. However, the performance of microexpression recognition suffers from low spatial resolutions produced by cameras with very high frame rates.

Most existing datasets thus have a frame rate at or higher than 60 fps as a reasonable compromise between speed and spatial resolution. Three exceptions are SMIC2/VIS, SMIC2/NIR, and USF-HD, each of which is used for assessing the possibility of recognizing microexpressions in normal frame rate.

There are two types of *tagging* systems for microexpressions, *i.e.*, emotion categories and Facial Action Coding System (FACS) [35]. Emotion categories are typically labels including happy, sad, and surprised. FACS encodes facial expressions with Action Units (AUs). A particular AU describes

a component of the facial expression, hence an expression is described by a collection of AUs.

On the other hand, both SMIC and SMIC2 were labeled with emotion categories. In SMIC, emotions are divided into positive and negative categories. Meanwhile emotions are divided into positive, negative, and surprise categories in SMIC2. USF-HD and Polikovsky datasets utilized FACS for tagging. Both emotions and AUs were tagged in CASME and CASME II.

The *elicitation protocol* is important as it determines whether a microexpression is genuine or inauthentic. Posed expressions differ drastically from genuine ones, even if they are controlled to be rapid and subtle.

In SMIC and SMIC2, specifically, videos evoking different kinds of emotions were shown to participants, who were required to suppress their expression to the best of their ability. To ensure this, any participant whose expression was observed correctly by others had to fill out a tedious form as a "penalty". Finally, the content of videos and self reporting were used to determine the ground truth microexpressions. The CASME and CASME II took a similar approach by using videos with high emotional valence to induce microexpressions. Moreover, monetary reward would be reduced if a participant made a noticeable expression.

It is worth pointing out that in a strict sense, expressions in USF-HD and Polikovsky datasets were not microexpressions. In the Polikovsky dataset, ten university students were trained to pose subtle expressions. Those expressions that appeared to be microexpressions were selected manually. In the USF-HD, participants were shown example pictures of microexpressions and were then asked to mimic them.

III. Related Work

This section reviews some existing microexpression recognition methods as well as the optical flow estimation technique, which is closely related to our proposed FDM method.

First, we define some terms to avoid confusion. Generally, there are two subtasks in microexpression recognition, *i.e.*, identification and categorization. *Identification* is to spot whether a microexpression is present in the given facial image sequence. Meanwhile, *categorization* is to determine the type of emotion exhibited after a microexpression is identified. We use the term *recognition* to refer to both identification and categorization because both subtasks are considered classification problems under our framework. Generally speaking, a two-phase strategy consisting of identification and categorization is used to recognize microexpressions in the real world [12].

Though the strategy is suitable to pre-segmented videos, a sliding window is helpful to identify and categorize microexpressions in long-duration videos. For example, the sliding window of a length d uses the last d seconds of video sequence for subsequent recognition. d is set to $\frac{1}{3}$ second, which is the typical length of a microexpression.

A. Existing Recognition Methods

Roughly speaking, microexpression recognition methods can be divided into two major categories, local and holistic methods.

Local methods segment face into multiple subregions, such as forehead, left and right eyes, mouth, etc. After that, the identification and categorization steps are conducted within each subregion. The reported performances are attained by encoding them with AUs [35]. For examples, Polikovsky *et al.* [11] utilized ASM [29] to locate facial landmarks used to split faces into twelve areas. In the pre-processing stage, images are normalized and smoothed, followed by using a carefully designed mask to neutralize unnecessary border pixels in each facial region. Within each area, derivatives in three planes $(X-Y, Y-T, X-T)$ are extracted as a quantitative descriptor of skin movements and discretized into several angular bins. Here, X and Y denote the spatial coordinates and T the temporal coordinates. In the classification stage, a microexpression sequence is further divided into three stages, *i.e.,* onset, apex, and offset, each of which is associated with different weights. Finally, a voting procedure is used to determine the final AU.

After segmenting face into regions with the assistance of ASM, Shreve *et al.* [13], [14] calculated a strain tensor for defining a finite strain magnitude based on optical flow. The tensor is a scalar at each facial pixel, indicating the elastic modulus of each point in face. Local facial movements can be observed by examining the sum of strain magnitude with each pre-segmented facial region and then hand-crafted thresholding strategy is utilized to determine a conventional expression. Similar approaches along with constraints on duration and number of facial regions are employed to determine the microexpressions.

On the other hand, holistic approaches treat the face as a whole for determining the presence and the emotion category of microexpressions. For instances, Pfister *et al.* [12] proposed a holistic approach to identify and categorize microexpressions. Specifically, the method first locates 68 landmark points in the first frame by applying an ASM to each face image in the sequence. These points are used to align faces based on a local weighted mean algorithm [36]. Furthermore, temporal interpolation model is applied for each sequence so that the lengths of the whole sequences are equal. More concretely, the model views the original image sequence as sampling from an underlying curve residing in a low-dimensional space, and the required number of frames can thus be re-sampled from the curve. Generally, the image sequences are regarded as cuboids in the $X-Y-T$ space. Consequently, spatiotemporal changes in it are treated as a 3-D texture in the space. LBP-TOP features [37] that describes the texture are extracted to characterize a microexpression. Finally, three well-studied classifiers, the SVM, Multiple Kernel Learning, and Random Forests, are employed in the classification step.

Besides the aforementioned holistic methods, Wu *et al.* extracted Gabor feature descriptor and utilized GentleSVM as the classifier [38]. Moreover, Wang *et al.* assumed an image sequence as a 3-order tensor $S \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ [15]. Three 2-order tensor $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3$ are learned to project the original tensor to a new form $\tilde{S} = S \times_1 \mathcal{U}_1 \times_2 \mathcal{U}_2 \times_3 \mathcal{U}_3$, in order to minimize within-class distances and maximize between-class distances. The Extreme Learning Machine (ELM) is then used for microexpression classification. We summarize these

existing methods in Table II.

TABLE II
EXISTING METHODS FOR MICROEXPRESSION RECOGNITION

METHOD	HOLISTIC OR LOCAL	ALIGNMENT METHOD	MAJOR FEATURE	DATASET
[13], [14]	LOCAL	VIOLA-JONES DETECTOR	STRAIN MAP	USF-HD
[11]	LOCAL	AAM	3D HOG	POLIKOVSKY
[12]	HOLISTIC	ASM + LWM	LBP-TOP	SMIC
[15]	HOLISTIC	N/A	3D TENSOR	CASME
[38]	HOLISTIC	N/A	GABOR	N/A

While local approaches [11], [13], [14] divide the problem into smaller and seemingly easier ones, the use of AUs to represent a microexpression loses information from the full picture. A cast from AUs notation to emotion is still needed for real applications. In addition, existing methods share the following two weaknesses:

1) They place less emphasis on fine face alignment. They either assume that the face has been well aligned, or heavily rely on the performance of facial landmark location approaches such as ASM. Nevertheless, a fine face alignment is more crucial in microexpression recognition than in conventional expression recognition because microexpression is a subtle movement, and thus a rather small misalignment may result in considerably large performance degradation.
2) Most methods utilize textural features like Gabor filters or LBP descriptors. These features cannot shed light on the mechanism of microexpressions intuitively. A facial-dynamics-based descriptor has the potential to expose the nature of microexpressions.

B. Optical Flow Estimation

Optical flow estimation technique [39] lays the foundation of our facial dynamics descriptor for tracking the movement of mass points in a video clip. Given two consecutive frames in a sequence (I_t and I_{t+1}), the optical flow algorithms aim at finding horizontal and vertical components of the optical flow field U^t and V^t, such that displacements of corresponding points between I_t and I_{t+1} satisfies the following equation:

$$I_t(i,j) = I_{t+1}(i + u^t_{i,j}, j + v^t_{i,j}) \ , \tag{1}$$

where $u^t_{i,j}, v^t_{i,j}$ are elements of U^t and V^t respectively.

The goal is often formulated as an optimization problem, *e.g.*, minimizing the following energy function:

$$\begin{aligned} E(U^t, V^t) = \sum_{i,j} [&\rho_D(I_t(i,j) - I_{t+1}(i + u^t_{i,j}, j + v^t_{i,j})) \\ &+ \lambda(\rho_S(u^t_{i,j} - u^t_{i+1,j}) + \rho_S(u^t_{i,j} - u^t_{i,j+1}) \\ &+ \rho_S(v^t_{i,j} - v^t_{i+1,j}) + \rho_S(v^t_{i,j} - v^t_{i,j+1}))] \ . \end{aligned} \tag{2}$$

Here the data penalty function ρ_D is used to ensuring the consistence of displacements shown in Eq. (1), and the spatial loss function ρ_S is to guarantee a physically meaningful solution. The parameter λ is a trade-off factor. Typically, there are three optional functions: 1) quadratic HS (Horn and Schunck) loss function: $\rho(x) = x^2$; 2) Charbonnier loss function: $\rho(x) = \sqrt{x^2 + \epsilon^2}$; and 3) Lorentzian loss function: $\rho(x) = \log 1 + \frac{x^2}{2\sigma^2}$. It is worth pointing out that there are many approaches to detect the displacements of face motions with varying loss functions, energy functions and optimization algorithms in literature [39]–[41]. For example, Sun *et al.* [39] showed that some techniques such as texture decomposition and median filtering can boost the estimation accuracy.

d

IV. THE FACIAL DYNAMICS MAP

In this section, we introduce our proposed Facial Dynamics Map and the fine-scale within-sequence alignments, as well as the pre-processing approaches.

A flowchart of the proposed method is illustrated in Fig. 1. Classical facial alignment procedures are less effective for handling microexpressions due to the short durations and subtle movements of the microexpressions. Therefore, a fine-scale alignment approach is proposed. Faces are located and cropped out by a common bounding box. Then optical flow estimation technique is used to measure the pixel-level movements. Moreover, a speed-up algorithm based on 1D histogram is conducted to attain a fine-scale alignment within the sequence, followed by extracting a Facial Dynamics Map at each specified granularity level. Finally, the Facial Dynamics Map is fed into a classifier for the identification and categorization of the microexpression.

A. Pre-processing

In our experiments, the datasets we used have performed two pre-processed steps, *i.e.*, 'Facial Landmark Location' and 'Coarse Alignment and Face Cropping'. Specifically, in Facial Landmark Location, 68 facial landmarks have been detected with an ASM model [12]. In Coarse Alignment and Face Cropping, a Local Weight Mean (LWM) transformation [42] is calculated for the first frame of each sequence. The transformation is then applied to all frames in that sequence. The distance between two eyes, noted as δ, is calculated based on facial landmarks. Then a box of $1.8\delta(width) \times 2.2\delta(height)$ is cropped based on facial landmarks [12], [34].

B. Fine-scale Alignment

In our experience, the quality of alignment in microexpression tasks is more critical than in many other applications. Unfortunately, the landmark-based approach alone is not sufficiently accurate for our task at hand, and the subsequent identification and categorization suffer from inaccurate alignment. Consequently, we propose a fine-scale in-sequence alignment, *i.e.,* a pixel-level alignment.

Given two consecutive frames in a video sequence I_t and I_{t+1}, we estimate a pixel-level movement by calculating the optical flow fields matrices U^t, V^t based on the method proposed by [39], so that:

$$I_t(i,j) = I_{t+1}(i + u^t_{i,j}, j + v^t_{i,j}) \ , \tag{3}$$

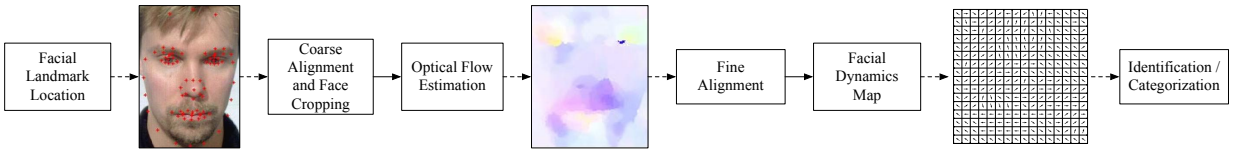


Fig. 1. Flowchart of the facial dynamics recognition process. First, facial landmark points are located. Faces are then aligned and cropped. An optical flow map is extracted for finer alignment. Facial Dynamics Maps are calculated for each clip. The Facial Dynamics Map in this picture is taken from a subject with negative emotion in SMIC2 [32]. It reveals a lip curling which is negligible to typical human eyes.

where scalars $u^t_{i,j}$ and $v^t_{i,j}$ are elements of U^t and V^t, respectively. The vector $(u^t_{i,j}, v^t_{i,j})$ indicates the motion at the location (i, j) between frame I_t and I_{t+1}.

Because microexpressions involve fewer facial muscles than posed or normal ones, we can safely assume that most of the facial area in neighboring frames remain motionless. To achieve a fine alignment, we subtract common movement at each position from the original frame image. Formally, we have Eq. 4 as follows:

$$\begin{aligned} \Delta u^{t*} &= \arg\max_{\Delta u} \Phi(U^t + \Delta u \times \mathbb{I}) \ , \\ \Delta v^{t*} &= \arg\max_{\Delta v} \Phi(V^t + \Delta v \times \mathbb{I}) \ , \\ \mathbb{I}_{ij} &= 1 \ , \end{aligned} \tag{4}$$

where U^t and V^t are the pre-calculated optical flow fields. Given horizontal and vertical displacement correction Δu and Δv, the aligned optical flow fields would be $U^t + \Delta u \times \mathbb{I}$ and $V^t + \Delta v \times \mathbb{I}$. $\Phi(X)$ denotes the number of zeros in matrix X after rounded toward zero.

In the equation, we seek to find optimal displacement correction Δu^{t*} and Δv^{t*} such that most positions in the aligned optical flow map remain zeros. More concretely, Δu^{t*} are optimal displacement correction such that most areas of horizontal optical flow U^t remains 0. Similarly, Δv^{t*} are optimal displacement correction such that most area of vertical optical flow V^t remains 0.

To address the optimization problem, we designed a straightforward algorithm. That is, we construct a 1D histogram h_{U^t}, such that $h_{U^t}(u)$ gives the number of positions with horizontal movement u. Simply letting $\Delta u^{t*} = -\arg\max_u h_{U^t}(u)$ satisfies the requirement. Δv^* is solved similarly in this paper.

The horizontal and vertical components of finely-aligned optical flow fields are $U^t + \Delta u^{t*} \times \mathbb{I}$ and $V^t + \Delta v^{t*} \times \mathbb{I}$, respectively.

To keep the notation uncluttered, we still use U^t, V^t and $u^t_{i,j}, v^t_{i,j}$ to represent the finely-aligned optical flow fields and their elements.

Fig. 2 exemplifies this process in which a holistic movement towards left-bottom corner can be observed. After these movements are removed, the true subtle movements in coordinates $(5,5), (5,6), (6,6), (6,5)$, which are closely related to the microexpression, are revealed.

C. Facial Dynamics Map

After the fine alignment is attained, a more compact representation is needed to characterize the dynamics of the microexpression for better recognition.

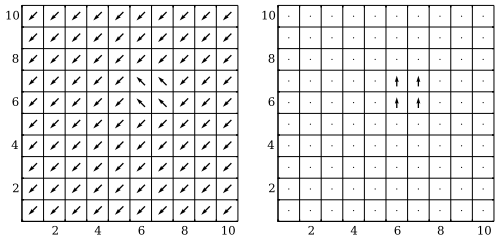


Fig. 2. The process for removing translation. Translation in left figure is detected and removed with a histogram-based strategy. The right figure shows the true dynamics after translation is removed.

Because facial expressions are generated by the movement of facial muscles, two reasonable assumptions on microexpressions can be made:

- Facial surface moves in roughly the same spatial direction when the observation area is small enough. Hence a pixel level movement description for microexpression is redundant.
- Facial surface moves in roughly the same temporal direction when the observation period is short enough. Hence a very high frame rate is redundant.

These assumptions motivate us to propose the Facial Dynamics Map. Roughly speaking, we split the face area into $n \times m$ equally-spaced grids and construct a spatiotemporal cuboid by introducing τ temporally consecutive grids. Here $n \times m$ is noted as *space division*, and τ is noted as *temporal multiplicity*. Therefore, for a video of dimension $X \times Y \times T$, the dimension of each cuboid is $\lfloor \frac{X}{n} \rfloor \times \lfloor \frac{Y}{m} \rfloor \times \tau$, and we have in total $n \times m \times \lfloor \frac{T}{\tau} \rfloor$ cuboids. Such a partition scheme is defined as $(n \times m, \tau)$, which is composed of the *space division* $n \times m$ and the *temporal multiplicity* τ.

To better differentiate from 'global' notation, *e.g.* $u^t_{i,j}, v^t_{i,j}$ in Eq. 3, the subscripts and superscripts are surrounded by brackets if the notation is regarding to 'within-cuboid' operation. For example, $w^{[t]}_{[i,j]} = (u^{[t]}_{[i,j]}, v^{[t]}_{[i,j]})$ is the motion vector at location (i, j) between frame t and $t+1$ within a specified cuboid. Since the dimension of a cuboid is $\lfloor \frac{X}{n} \rfloor \times \lfloor \frac{Y}{m} \rfloor \times \tau$, the subscripts $i \in \{1, \cdots, \lfloor \frac{X}{n} \rfloor\}, j \in \{1, \cdots, \lfloor \frac{Y}{m} \rfloor\}$ and superscript $t \in \{1, \cdots, \tau\}$. Note that while $u^t_{i,j}, v^t_{i,j}$ in Eq. 3 represent the elements of aligned fields in the whole sequence, $i \in \{1, \cdots, X\}, j \in \{1, \cdots, Y\}, t \in \{1, \cdots, T\}$.

Considering the two assumptions above, motion vectors in the same cuboid should be roughly in the same direction when the cuboid is reasonably small. Although abnormal motion vectors can appear in the optical flow estimation, most correctly-estimated motion vectors tend to have similar

directions. This observation inspires us to propose Eq. 5:

$$\xi^* = \arg\max_{|\xi|=1} \sum_{[i,j]} \langle \xi, w_{[i,j]} \rangle, \tag{5}$$

where $w_{[i,j]}$ is selected among a set of τ candidates $\{w_{[i,j]}^{[t]}\}_{t=1}^{\tau}$ and $w_{[i,j]}^{[t]} = (u_{[i,j]}^{[t]}, v_{[i,j]}^{[t]})$. Here $(u_{[i,j]}^{[t]}, v_{[i,j]}^{[t]})$ denote the motion vector at location (i,j) between frame t and $t+1$ within a specified cuboid, and $i \in \{1, \cdots, \lfloor \frac{X}{n} \rfloor\}, j \in \{1, \cdots, \lfloor \frac{Y}{m} \rfloor\}, t \in \{1, \cdots, \tau\}$. The notation of $<,>$ denotes the inner product.

The goal for this is to find a 2-dimensional principal direction ξ that is capable of describing the most obvious movement within the spatiotemporal cuboid. When the magnitude of ξ is fixed, this inner product has a larger value if ξ and $w_{[i,j]}$ have closer directions. Therefore, in each planar position (i,j), Eq. 5 finds one motion vector from τ candidates $\{w_{[i,j]}^{[1]}, \cdots, w_{[i,j]}^{[\tau]}\}$, so that the chosen motion vectors from each planar position have closest direction with ξ^*, which Eq. 5 uses to represent the principal direction of the cuboid.

The advantage of this step is that not only can it eliminate redundant information in consecutive optical flow fields, it can also help avoid optical flow errors. Take for example τ motion vectors, $w_{[i,j]}^{[1]}, \cdots, w_{[i,j]}^{[\tau]}$, at a fixed planar coordinate in consecutive optical flow fields, they describe approximately the same movement according to the second assumption. It is possible that incorrect optical flow estimations appear due to noise or lighting condition changes. However, it is less likely that $w_{[i,j]}^{[1]}, \cdots, w_{[i,j]}^{[\tau]}$ are all wrong as τ increases. As long as the most obvious movements are correctly detected, our scheme is able to select them because correct movements are alike and hence have a larger sum of inner product with optimal ξ^*. As a result, wrong and isolated movements do not bring negative effects to microexpression recognition.

It is clear that when τ is equal to 1, Eq. 5 degenerates to summing up within each grid of every frame, leaving only one optical flow field for our scheme. As for $\tau > 0$, it is difficult to find the optimal direction ξ^* since we need to iterate over $O(N^{\tau+1})$ possible combination of motion vectors to find the principal optical flow direction, where N is the number of motion vectors within a cuboid. Inspired by [43], an iterative scheme is proposed to estimate the optimal principal direction as shown in Algorithm 1. First, we initialize ξ as a unit vector $(1,0)$. Next, for every position within a grid, we select the motion vector which has the closest direction with ξ. This step is re-formulated as selecting the motion vector w that has maximum inner product with suboptimal ξ in that round. After that, ξ is updated according to all selected vectors. The process is repeated until ξ converges or a maximum iteration is reached.

After obtaining the approximately optimal direction $\xi^* = (\tilde{u}, \tilde{v})$, we compute its angle as follows (the distance is 1 because we restrict ξ^* to be a unit vector):

$$\theta = \arctan\left(\tilde{u}^2, \tilde{v}^2\right). \tag{6}$$

θ is then quantized into k discrete bins as illustrated in Fig. 3, *i.e.*, a bin number $\hat{\theta} \in \{1, .., k\}$ is used to describe a single cuboid in the video. The parameter k is empirically set to ten in our work. Features within each grid are linked to form a 3-order tensor representation for a facial expression as $M = \{\hat{\theta}_{i,j}^t\}$. Here t denotes the frame index, and i and j are column and row indices, respectively. It is worth pointing out that as shown in the experiments section, it is unnecessary to partition the input optical flow in a very fine manner.

Algorithm 1 Iterative Computing Principal Directions

Require:
Motion vectors within given cuboid $w_{[i,j]}^{[t]} = (u_{[i,j]}^{[t]}, v_{[i,j]}^{[t]})$
$i \in I, j \in J, t \in \{1, \cdots, \tau\}$
$I = \{1, \cdots, \lfloor \frac{X}{n} \rfloor\}, J = \{1, \cdots, \lfloor \frac{Y}{m} \rfloor\}$
Output:
Principal direction ξ
initiate $\xi = (1, 0)$
repeat
  for all $(i,j) \in I \times J$ **do**
    $w \in \{w_{[i,j]}^{[t]}\}$,
    $max\_v\_candidate[i,j] = \arg\max_w < \xi, w >$
  end for
  $\xi = Sum(max\_v\_candidate)$
  $\xi = \xi / ||\xi||_2$ {ξ is updated according to all selected vectors}
until ξ converges or reaches maximum iterations
return ξ

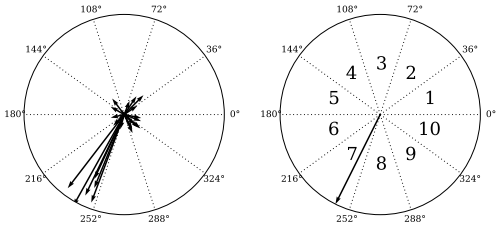


Fig. 3. Facial Dynamics Map quantization process. The principal direction is calculated using all optical flow vectors within the same cuboid and quantized into ten bins.

Finally, a pseudocode of extracting the Facial Dynamics Map is provided in Algorithm 2.

After the extraction process, we obtain a Facial Dynamics Map M which is 3-order tensor of dimension $n \times m \times \lfloor \frac{T}{\tau} \rfloor$. It is then re-shaped to a $nm\lfloor \frac{T}{\tau} \rfloor \times 1$ vector and used as the input for the SVM classifier with RBF kernel. SVM parameters are tuned through grid search [44].

The convergence of Algorithm 1 depends on the input optical flow fields. Although there is no theory to guarantee the convergence, it does converge very quickly in most cases, which we will show empirically in Section V-E.

Note that when introducing the temporal multiplicity, some frame information can be lost. For example, a total of 9 optical flow fields are extracted from a sequence interpolated to 10 frames. Given $\tau = 4$, it means that the 9 optical fields are divided into three parts, *i.e.*, the first 4 optical flow fields, the next 4 optical flow fields, and the 9th optical flow field. We iteratively compute principal directions for the first two parts and discard the 9th optical flow field. Similar procedures are performed under different τ's.

Algorithm 2 Facial Dynamics Map

Require:
The finely aligned optical flow fields of a microexpression clip: $W = \{U^1, V^1, \cdots, U^{T-1}, V^{T-1}\}$
Output:
The Facial Dynamics Map M
1: $\{Cuboid^t_{i,j}\} = Partition(W, n, m, \tau), i \in \{1, \cdots, n\}, j \in \{1, \cdots, m\}, t \in \{1, \cdots, \lfloor \frac{T}{\tau} \rfloor\}$
{partition optical flow clip into spatiotemporal cuboids, each of dimension $\lfloor \frac{X}{n} \rfloor \times \lfloor \frac{Y}{m} \rfloor \times \tau$.}
2: **for all** $Cuboid^t_{i,j}$ **do**
3: $\xi^t_{i,j} = Principal\_Direction(Cuboid^t_{i,j})$ {apply Alg. 1}
4: $\theta^t_{i,j} = Angle\_Of(\xi^t_{i,j})$
5: $\hat{\theta}^t_{i,j} = Discretize(\theta^t_{i,j})$
6: **end for**
7: $M = \{\hat{\theta}^t_{i,j}\}$
8: **return** M

D. Parallel Estimation of Optical Flow Fields

Given a sequence of optical flow fields, our iterative algorithm for the FDM finishes in a few rounds.

The main computational bottleneck stems from the former optical flow estimation stage, which requires time-consuming optimization. Given an image sequence of length T, we should calculate $T-1$ optical flow fields $(U^1, V^1), (U^2, V^2), \cdots, (U^{T-1}, V^{T-1})$ between consecutive frames. Note that each optical flow field (U^t, V^t) only depends on the tth and $(t+1)$-th frames. Thus optical flow fields $(U^1, V^1), (U^2, V^2), \cdots, (U^{T-1}, V^{T-1})$ can be extracted concurrently on multi-processor machines. By specifying a number of parallel threads, we can significantly reduce the time required for processing.

V. Experiments

We now report the evaluation of our method using four benchmark microexpression datasets, *i.e.*, the SMIC and the SMIC2 [12], [32], the CASME I [33] and the CASME II [34], and discuss the influence of parameter setting in detail.

A. Experiment Setup and Notations

Among the four datasets, each of SMIC and SMIC2 consists of two subtasks, identification and categorization. For the categorization subtask, SMIC includes two kinds of emotions, namely positive and negative emotions. Meanwhile, SMIC2 is composed of surprise, positive and negative emotions. Note that we can have more kinds of microexpressions. For example, positive microexpressions can be further categorized into happy, excited, and so on, while negative microexpressions can be divided into disgust, anger, fear, and sadness. However, for the two benchmark microexpression datasets, we use the emotional label released by the authors for fair comparison.

Furthermore, SMIC2 includes three subsets, HS, VIS, and NIR. The HS subset of SMIC2 and SMIC were filmed at 100 fps, and the VIS and NIR subsets of SMIC2 were filmed at 25 fps. Unlike other subsets and SMIC, which are recorded by normal visual cameras, the NIR subset was attained by using a near-infrared camera. Table III describes the numbers of sequences in SMIC and in each subtask of SMIC2 in detail. The original spatial resolution of images in SMIC and SMIC2 is 640×480. Fig. 4 shows several cropped face samples from these datasets. It is evident that these changes in the sequences are very subtle, even for human eyes.

CASME I consists of eight kinds of expressions, namely contempt, disgust, fear, happiness, repression, sadness, surprise, and tenseness. CASME II consists of seven kinds of expressions, namely disgust, fear, happiness, repression, sadness, surprise, and others. Fig. 4 shows 2 samples from CASME I and CASME II, respectively.

Note that the sequence ID of SMIC/SMIC2 offers information about subject and emotion. For instance, smic-s3-notmicro-48-7 indicates a sequence from subject 3 in SMIC which does not show a microexpression, and 48-7 together forms the in-person ID; hs-s3-sur-01 indicates the sequence comes from SMIC2/HS, and it is the 1st sample of surprise by subject 3. As for CASME I/II, the sequence ID is alike but does not contain emotion. For instance, s17-EP05-04 indicates a sequence from subject 17, and EP05-04 together gives the in-person ID. The copies of CASME I/II are slightly different from that described in the original paper, which might be attributed to a new labeling. Description in our paper is thus based on the copies we obtained.

TABLE III
Numbers of samples in SMIC and three tasks of SMIC2. Surp.: Surprise, Pos.: Positive, Neg.: Negative.

Dataset	Identification		Categorization		
	Micro.	Non-Micro.	Surp.	Pos.	Neg.
SMIC	76	76	N/A	18	17
HS	164	164	43	51	70
VIS	71	71	20	28	23
NIR	71	71	20	28	23

TABLE IV
Numbers of samples in CASME I and CASME II. C: Contempt, D: Disgust, F: Fear, H: Happiness, R: Repression, Sd: Sadness, Sp: Surprise, T: Tenseness

Dataset	C	D	F	H	R	Sd	Sp	T	Other
CASME I	1	44	2	9	38	6	20	69	N/A
CASME II	N/A	63	2	32	27	7	25	N/A	99

We compare FDM with two state-of-the-art algorithms, *i.e.*, LBP-TOP method proposed by [12] and Discriminant Tensor Subspace Analysis (DTSA) proposed by [15]. DTSA projects face sequences to a low-dimensional space, followed by employing Extreme Learning Machine (ELM) for classification. Note that neither our method nor the LBP-TOP-based one has any limitations on the use of the classifier. Therefore, we utilize the well-known SVM classifier with RBF kernel as the basis classifiers for FDM and LBP-TOP. The parameters of SVM are tuned using grid search [44].

The cross-validation approach we use is the leave-one-subject-out evaluation protocol for different methods. This approach allows us to evaluate the generalization of different methods in recognizing unseen microexpressions in the testing phase. Take SMIC for example. The dataset consists of six

Fig. 4. Samples of cropped faces from SMIC [12], SMIC2 [32], CASME I [33] and CASME II [33]. From top to bottom: Row 1 shows a negative microexpression from SMIC2/VIS [vis-s11-ne-02]. Row 2 shows a positive microexpression from SMIC2/NIR. The original sequence contains 13 images [nir-s11-po-03]. The first 8 images are shown. Row 3 shows a surprise microexpression from SMIC2/HS [hs-s3-sur-01]. The original sequence contains 25 images. Every one of three faces are shown here for clarity. Row 4 shows a non-micro expression from SMIC [smic-s3-notmicro-48-7]. The original sequence contains 34 images. Four faces are shown here. Row 5 shows an example of disgust from CASME I [s1-EP07-1]. The sample contains 10 frames. The subject shows a nose wrinkle. Row 6 shows an example of repression from CASME II [s17-EP05-04]. The original sample contains 66 frames. Here we show 1 frame of every 8 frames for clarity. The subject shows a chin raise.

participants; we train our methods on clips from five of them. Tests are conducted on the remaining one.

Finally, all correctly classified sequences are counted for in the final accuracy as follows:

$$\text{Accuracy} = \frac{\sum_i (tp_i + tn_i)}{\sum_i (tp_i + tn_i + fp_i + fn_i)} , \tag{7}$$

where tp, tn, fp and fn mean true positive, true negative, false positive and false negative, respectively.

One disadvantage of overall accuracy is that the index favors classes with larger number of samples over classes with smaller one. To overcome this issue, we also use macro-mean $F1_M$ score as an important index, which is the based on the macro-mean precision and macro-mean recall [45]:

$$F1_M = \frac{2 \cdot \text{Precision}_M \cdot \text{Recall}_M}{\text{Precision}_M + \text{Recall}_M} , \tag{8}$$

where $\text{Precision}_M = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{tp_i}{tp_i + fp_i}$ and $\text{Recall}_M = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{tp_i}{fn_i + fp_i}$. ℓ is the number of classes in a particular classification problem. Precision_M and Recall_M calculate precision and recall respectively, treating all classes equally. This measure eliminates the influence of basis number of different classes. For example, there is only 1 contempt sample and 2 fear samples in CASME I. It is a great temptation for the classifier to classify these samples to larger classes, such as disgust or repression, because it does not harm the accuracy too much. However, this can be reflected in the macro-mean F1 score.

In the pre-processing stage, we use linear interpolation instead of Temporal Interpolation Model because in a high frame-rate scenario, linear interpolation suffices to characterize the motion pattern with less error. In linear interpolation, pixels at the same planar location in a video clip are linearly interpolated to the specified number of frames. LIT{#frames} denotes a microexpression sequence obtained by linear interpolation. For example, LIT15 is an image sequence interpolated to 15

frames with linear interpolation.

The platform in this experiment has four AMD Opteron 6378 Processors, each of which has 16 cores, with 512 GB RAM. The operating system is Debian 3.2.51, and the Matlab version is R2013b.

B. Standard Experiments

We perform comprehensive comparison on all four datasets by temporally interpolating each image sequence to 10, 15, and 20 frames for both FDM and LBP-TOP using linear interpolation. Moreover, video cubes are divided into $\mathcal{X} \times \mathcal{Y} \times \mathcal{T}$ blocks along the spatial and temporal domains for the LBP-TOP-based method. Here $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{T}$ denote the number of horizontal, vertical, and temporal blocks, respectively. In our experiments, we use the suggested parameters in the original paper. That is, four specifications of $\mathcal{X} \times \mathcal{Y} \times \mathcal{T}$ used for test include $5 \times 5 \times 1$, $5 \times 5 \times 2$, $8 \times 8 \times 1$, and $8 \times 8 \times 2$. We also use the original implementation of [46] for LBP-TOP. As for our FDM-based method, space divisions including 4×4, 8×8, 16×16, and 32×32 have been used. Several temporal multiplicity τ including 2, 3, and 4 have been tested.

In the DTSA-based method, face images are first re-scaled to 64×64 pixels, and then each sequence is temporally interpolated to 64 frames. Assume that each sequence is viewed as a 3-order tensor $S \in \mathbb{R}^{64\times64\times64}$. Three 2-order tensors $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3 \in \mathbb{R}^{64\times L}$ are iteratively and alternatively learned for projecting the original tensor onto a lower dimensional tensor subspace $\mathbb{R}^{L\times L\times L}$ as follows:

$$\tilde{S} = S \times_1 \mathcal{U}_1 \times_2 \mathcal{U}_2 \times_3 \mathcal{U}_3 \ , \tag{9}$$

where $\times_1 \mathcal{U}_1$ means that S is projected into a subspace by multiplying projection matrix $\mathcal{U}_1$. An iterative process is taken to learn those 2-order tensors as suggested in [15]. Here the maximum iteration number is set to 100. For computational simplicity, we choose the final reduced dimension L from a set $\{10, 20, 30, 40, 50, 60\}$. We implement the method strictly according to [15]. As for ELM algorithm, we employ a sigmoid function as the activation one, and vary the number of neurons from 500 to 4000, as in [15].

To make a fair comparison, we report the best results of each method in Fig. 5. Detailed accuracy and $F1_M$ scores of each subtask are shown in Table V.

When compared with LBP-TOP and DTSA, our FDM-based method achieves the best identification and categorization performances in all tasks. One possible reason is that FDM is highly related to the nature of expression, *i.e.,* facial dynamics, whereas LBP-TOP views expression as a spatiotemporal texture. It is worth mentioning that overall accuracy reported in DTSA [15] is higher than reported here. The reasons stem from two different experiment setups:

- They use only a subset of the CASME I for experiment, excluding classes with small number of samples.
- The evaluation protocols are different. In DTSA [15], they randomly selected m samples of each class as training samples and used the rest of samples for testing. We approach training in a person-independent style, restricting that no person in the testing set shows up in the training set. This approach is important because in real situations, it is less likely that the subject will show up in the training set.

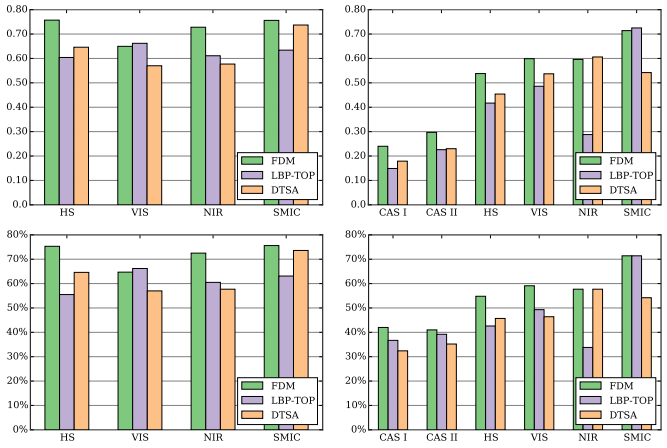


Fig. 5. $F1_M$ score and Accuracy comparison for three methods. From top left to bottom right: $F1_M$ on identification, $F1_M$ on categorization, Accuracy on identification, and Accuracy on categorization. CAS I and CAS II denote CASME I and CASME II, respectively. Both FDM and LBP-TOP use linear interpolation.

Therefore, an additional experiment, where classes with a small number of training sets are truncated and the person-independent evaluation protocol is left untouched, is performed. The results under this setting are reported as in Table VI. As the table shows, our algorithm still outperforms the DTSA method. The overall accuracy of DTSA is very close to the ones reported in the original paper, implying that the determining factor of the performance is the training set, including the number of classes and quality of each set. The person-independent setting is less influential but still makes a difference.

To analyze when FDM has worse performance, we show a negative microexpression sequence of SMIC that is misclassified as positive microexpression, as in Fig. 6. From the original sequence containing 27 images, we select every one of six images for illustration. After the sequence is interpolated to ten frames, we use $(16 \times 16, 3)$ to extract the FDM as shown in Fig. 6 (b). We also extract optical flow fields between five selected images as illustrated in Fig. 6 (c). In the five frames, the participant curls his lip, indicating a negative microexpression. When examining the corresponding optical flow fields, however, we notice that the nose area is wrinkled. This omission appears less frequently in the negative microexpression class. Therefore, it is possible that the classifier could misclassify the negative expression.

DTSA does not perform very well in most situations in terms of overall accuracy. Yet, its performance outperforms LBP-TOP in quite a few datasets (identification of HS and SMIC, as well as all categorization problem except SMIC) when using the $F1_M$ score. This shortcoming happens because DTSA treats all classes equally and tries to maximize the distances between them. Thus it may suffer from the size of different classes. Furthermore, the overall accuracy is affected

TABLE V
RESULTS UNDER STANDARD SETTING.

(a) $F1_M$ score

Task		FDM	LBP-TOP	DTSA
CASME I	Categorization	**0.2401**	0.1257	0.1790
CASME II	Categorization	**0.2972**	0.2312	0.2306
HS	Identification	**0.7571**	0.6097	0.6464
	Categorization	**0.5380**	0.4108	0.4540
VIS	Identification	0.6496	**0.6630**	0.5704
	Categorization	**0.5997**	0.4780	0.5371
NIR	Identification	**0.7281**	0.6343	0.5776
	Categorization	0.5967	0.3083	**0.6069**
SMIC	Identification	**0.7566**	0.6460	0.7376
	Categorization	**0.7145**	0.6934	0.5425

(b) Accuracy

Task		FDM	LBP-TOP	DTSA
CASME I	Categorization	**42.02%**	36.50%	32.45%
CASME II	Categorization	**41.96%**	39.22%	35.29%
HS	Identification	**75.30%**	55.49%	64.63%
	Categorization	**54.88%**	43.90%	45.73%
VIS	Identification	64.79%	**66.20%**	57.04%
	Categorization	**59.15%**	49.30%	46.48%
NIR	Identification	**72.54%**	63.38%	57.75%
	Categorization	**57.75%**	36.62%	**57.75%**
SMIC	Identification	**75.66%**	64.47%	73.68%
	Categorization	**71.43%**	68.57%	54.29%

TABLE VI
$F1_M$ SCORE (TOP) AND ACCURACY (BOTTOM) COMPARISON. CASME I INCLUDES DISGUST, REPRESSION, SURPRISE, AND TENSENESS; CASME II INCLUDES DISGUST, HAPPINESS, REPRESSION, SURPRISE, AND OTHERS.

Dataset	FDM (our method)		LBP-TOP		DTSA	
CASME I	**0.4912**	$(8 \times 8, 2)$, LIT10	0.2647	$5 \times 5 \times 1$, LIT20	0.4539	$L = 40$
CASME II	**0.4053**	$(8 \times 8, 3)$, LIT20	0.3262	$8 \times 8 \times 1$, LIT15	0.2860	$L = 40$

Dataset	FDM (our method)		LBP-TOP		DTSA	
CASME I	**56.14%**	$(4 \times 4, 3)$, LIT10	40.35%	$5 \times 5 \times 1$, LIT20	46.20%	$L = 40$
CASME II	**45.93%**	$(16 \times 16, 2)$, LIT15	40.65%	$8 \times 8 \times 2$, LIT20	36.18%	$L = 10$

by imbalance across class sizes because the measure implicitly assumes that each class should have the same size of samples. As a result, the reported overall accuracy using DTSA is not as good as $F1_M$ score.

We show the confusion matrices of each method in Fig. 7. The figure in coordinate (x, y) (x is the vertical tick and y is the horizontal tick) indicates the portion of samples belonging to category x but being classified as y. Under an ideal condition, the diagonal of each matrix dominates other elements, *i.e.*, $(x, x) > (x, y)$, $y \neq x$.

From the confusion matrices it can be seen that : 1) In some hard cases, for example, categorization of NIR under LBP-TOP and categorization of SMIC under DTSA is biased to classify most samples to a class with a large number of samples. Although the accuracies using these approaches are high, it may be ineffective when the distribution of real application differs from that of testing datasets. This phenomenon is well-revealed via confusion matrix or $F1_M$ score. 2) According to the correctly recognized ratio along the diagonal of confusion matrices, FDM attains the best performance among the three algorithms.

C. Effectiveness of Fine Alignment

This subsection is devoted to justifying the effectiveness of our fine alignment algorithm. After the optical flow is extracted, we perform a fine alignment along with the FDM extraction stage. The process is to eliminate small movements of the person between consecutive frames in the sub-pixel level.

We report the results with and without fine alignment in Fig. 8. It shows that fine alignment improves our algorithm; justifying the fine alignment positively affects recognition results. However, the improvement brought by fine alignment is less significant than that by parameter tuning. Therefore, it is important to tune partition parameters for real-world applications.

D. The Influence of Parameters

We investigate the influence of both spatial grid partition and different temporal multiplicity to microexpression recognition in this subsection. Based on the scheme $(n \times m, \tau)$, we select $\tau \in 2, 3, 4$, and $n \in \{4, 8, 16, 32\}$ while forcing $n = m$. For those two parameter, the larger they are, the smaller the resulting spatiotemporal cuboids are. Therefore, our method catches more detailed information. In space division graphs, for each $n \times m$, we mark the best result obtained from a parameter no larger than $n \times m$. The temporal multiplicity graphs are in the same case.

From the best results shown in Fig. 9, we can see the influence of two different parameter combinations as follows:

1) *Space division ($n \times m$).* We notice that a considerable portion of best results come from space division being less than 16×16. As for the rest, the gap is quite small. This finding motivates us to look into the nature of microexpression. Although it cannot be controlled by human beings themselves, microexpression is still generated by facial muscles. Therefore, the granularity of partitions gets finer than facial muscle, doing little help for recognition, as the number of grid partitions becomes large.
2) *Temporal multiplicity (τ).* $\tau = 3$ is enough for most situations. Although a larger τ may slightly enhance the performance, selecting a small value can save computational time. Therefore, $\tau = 4$ is a good trade-off between performance and time cost.

E. Complexity Analysis

We analyze the time complexity of the three methods in obtaining feature representation. The time consumption of FDM-based method has two parts. For convenience, we investigate them separately according to the optical flow estimation and the iterative procedure for the computation of principal direction.

Fig. 10 shows the time consumption of our method in both sequential and parallel execution. It is obvious that both sequential and parallel versions are sensitive to the number of frames because the number of frames directly determines the workload of optical flow estimation. However, the parallel technique significantly accelerates the computational process.

Algorithm. 1 is used to iteratively find principal directions defined by Eq. 5. Although no theory guarantees its convergence, the algorithm converges very quickly in most cases. We randomly select 60 interpolated sequences from these 6 datasets and iteratively find the principal direction with $\tau \in \{2, 3, 4\}$. Fig. 10 shows percentage of cuboids that converged in corresponding number of iterations. Up to 90% of cuboids converged in 3 iterations. However, a small portion of cuboids failed to converge within the maximum iteration we set (100 in this paper). The percentage is 0.500% when $\tau = 2$, 0.833% when $\tau = 3$, and 0.834% when $\tau = 4$. Those cuboids can be viewed as small noise in feature representation, as they only account for a very small portion of the representation.

Regarding temporal complexity, we investigate the effect of both the temporal multiplicity and the space division in Fig. 10. A sequence interpolated to ten frames is used. Two factors affect the time consumption:

- Space division ($n \times m$) affects the time consumption in two ways. First, it decides how many motion vectors to search iteratively for a principal direction. We observe a high consumption when the space division is small. Second, the space division also determines the number of cuboids, from each of which a principal direction is in need. Therefore, the number of motion vectors decreases,

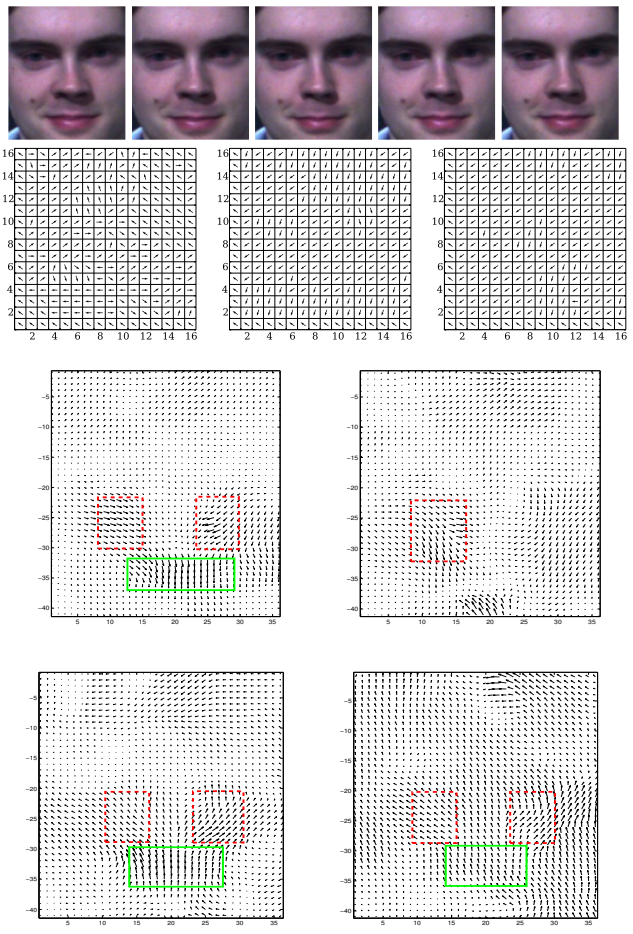


Fig. 6. Samples of categorization failure. The participant shows a negative microexpression but the classifier recognizes it as positive. The original sequence contains 27 images. Every one of six are selected for illustration as in **Top row**. The sequence is interpolated to ten frames, we use $(16 \times 16, 3)$ to extract the FDM, which is shown in **Middle row**. We also attach to optical flow fields in the **Bottom two rows**, which is extracted between five images selected. Green boxes indicate the mouth movement. Red dashed boxes indicate the movement around nose.

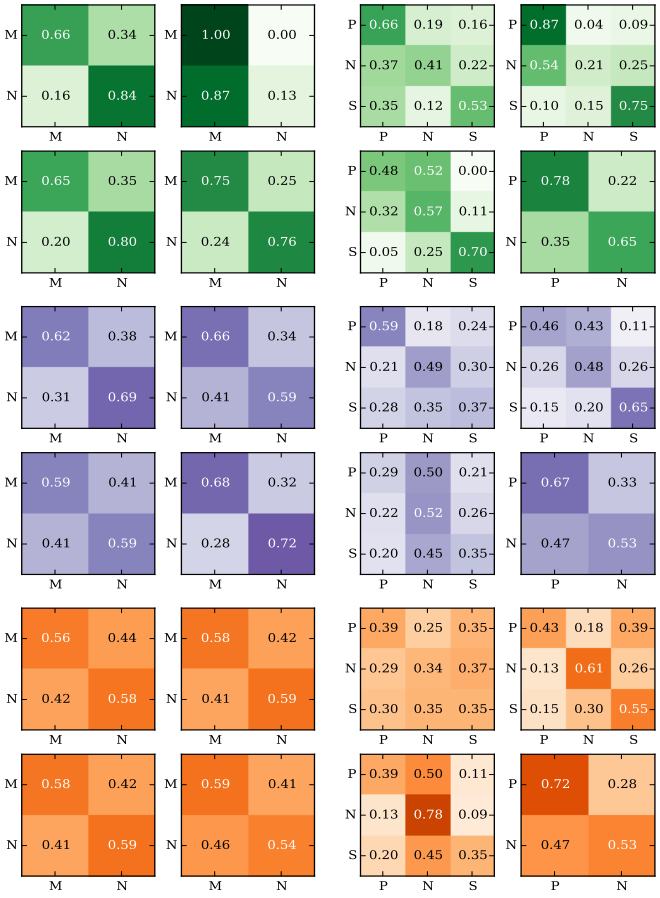


Fig. 7. Confusion matrices of each method under optimal parameters. From top to bottom, there are FDM, LBP-TOP and DTSA confusion matrices per two rows, respectively. In each method, four blocks on the left side are the identification confusion matrices and four blocks on the right side are the categorization confusion matrices. Order within each part: upper left is for HS; upper right is for VIS; bottom left is for NIR; and bottom right is for SMIC. Abbreviation symbols shown in the matrix are: In identification, M: micro, N: non-micro; In Categorization, P: positive, N: negative, S: surprise.

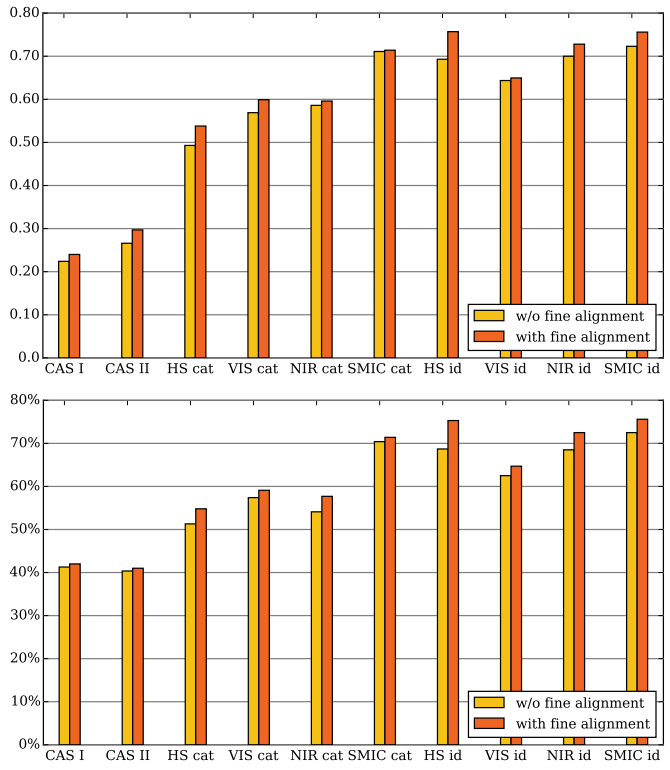


Fig. 8. Effectiveness of alignment. The picture shows best results obtained with and without fine alignment. 'cat' and 'id' denote 'categorization' and 'identification', respectively.

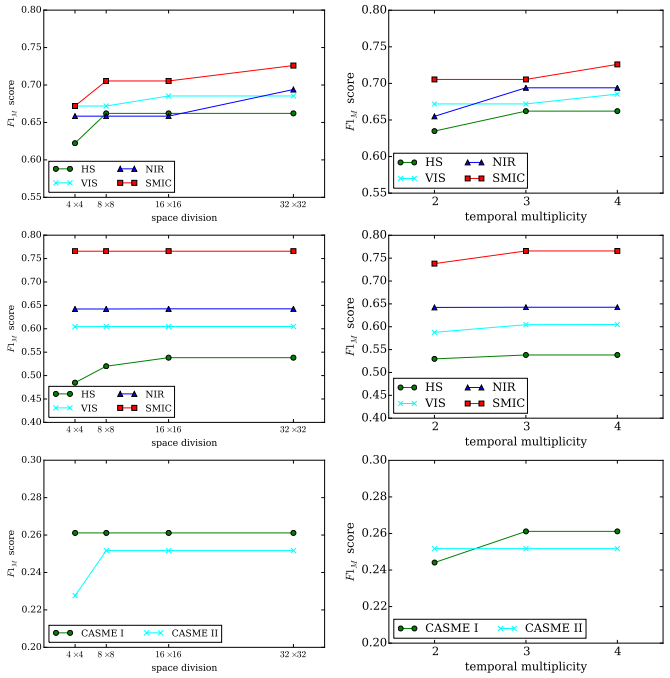


Fig. 9. Parameter's effect on the recognition results. Row 1: identification task on SMIC/SMIC2; row 2: categorization task in SMIC/SMIC2; row 3: categorization task on CASME I/II. Each data point indicates the best result obtained for a parameter no larger than the specified coordinates.

and the iterative procedure terminates quicker as the space division increases. The large space division, however, requires more cuboids to process. Thus we observe an increase in computational time as the space division increases.

- Larger temporal multiplicity τ introduces more candidate motion vectors for the principal direction, and thus increases the time consumption. The case where $\tau = 1$ is extremely fast because it only involves addition. The time consumption of $\tau = 4$ is less than that of $\tau = 3$. This is because $\tau = 3$ calculates three batches ($\lfloor \frac{10-1}{3} \rfloor$) of principal directions, but $\tau = 4$ only calculates two batches ($\lfloor \frac{10-1}{4} \rfloor$).

The time consumption of LBP-TOP-based method and DTSA-based method is shown in Table VII. The LBP-TOP method is quite stable to the number of frames.

As for DTSA, the time consumption is related to three factors: the maximum iteration number, the dimension of the target tensor (the L parameter), and the number of training sample. The maximum iteration number is set to 20. Because DTSA takes all samples for the final representation, there is no time consumption reported for a single sample. We show the amortized time consumption for single sample (*i.e.*, the total time cost divided by the number of samples).

TABLE VII
TIME CONSUMPTION OF LBP-TOP AND DTSA

(a) Typical Runtime of LBP-TOP in seconds

LBP-TOP	TIM10	TIM15	TIM20
$5 \times 5 \times 1$	26.43	27.13	26.70
$5 \times 5 \times 2$	23.47	24.00	23.87
$8 \times 8 \times 1$	24.34	24.94	24.84
$8 \times 8 \times 2$	22.33	22.82	22.72

(b) Runtime of DTSA in identification of NIR in seconds

L	10	20	30	40	50	60
Time	2.92	7.34	16.98	21.55	23.46	38.81

Strictly speaking, time consumption of different methods are incommensurable due to different parameter settings. We can roughly compare time consumption, however, based on the datasets in our experiments. We have found that DTSA is fast if the dimension of the target tensor is low. However, this may sacrifice some accuracy because it requires the selection of L to determine the most suitable parameter. The time consumption of LBP-TOP is stable and acceptable. Our serial version FDM is relatively slow due to the time consumption of the optical flow estimation phase, but with parallelization, the time consumption can be greatly reduced. Even when combined with a large temporal multiplicity, our FDM can finish the tasks in time comparable with state-of-the-art methods.

VI. CONCLUSIONS AND FUTURE WORK

We proposed the Facial Dynamics Map for microexpression identification and categorization. It estimates the movement between consecutive frames with the optical flow estimation algorithm. A fine-scale alignment is performed with a fast optimization method. The aligned optical flow fields are divided

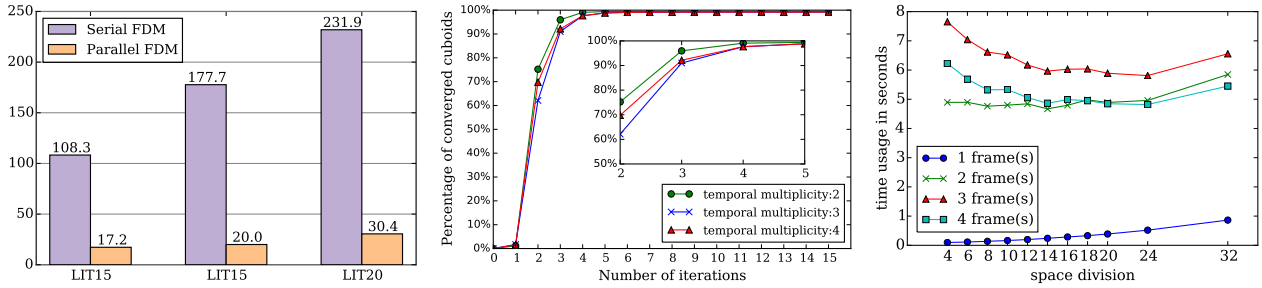


Fig. 10. Left: Time consumption of optical flow estimation in seconds. Middle: Convergence of the iterative principal direction algorithm. Each data point indicates percentage of cuboids that converged in the corresponding number of iterations. Right: Time consumption of iterative principal direction in seconds. A sequence interpolated to ten frames is used.

into a collection of cuboids. Within each cuboid, we extract the principal direction of each cuboid to characterize facial dynamics of microexpressions. We also proposed an iterative method to accelerate the searching process. We design the parallel algorithm for FDM and achieve high computational efficiency. By comparing the FDM with two state-of-the-art methods on four publicly available microexpression datasets (SMIC, SMIC 2, CASME and CASME II), we experimentally justify the effectiveness of the proposed FDM. Also, FDM provides a physically meaningful interpretation for microexpression recognition.

In the future, we will investigate how to refine the efficiency of estimating pixel-level movement, which is important in real-time microexpression recognition. Furthermore, motion distance information that may better describe facial dynamics deserves in-depth study. Finally, we will consider extending the framework of building blocks in our algorithm to real time recognition of microexpressions in long-duration videos.

ACKNOWLEDGMENT

The authors would like to thank two reviewers and associate editor for their comments and constructive suggestions.

REFERENCES

[1] S. Baluja and H. A. Rowley, "Boosting sex identification performance," *International Journal of Computer Vision*, vol. 71, no. 1, pp. 111–119, 2007. 1

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013. 1

[3] G. Toderici, S. M. OMalley, G. Passalis, T. Theoharis, and I. A. Kakadiaris, "Ethnicity- and gender-based subject retrieval using 3-d face-recognition techniques," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 382–391, 2010. 1

[4] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, 2014. 1

[5] Y. Rahulamathavan, R. C.-W. Phan, J. A. Chambers, and D. J. Parish, "Facial expression recognition in the encrypted domain based on local fisher discriminant analysis," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 83–92, 2013. 1

[6] S. Wang, Z. Liu, Z. Wang, G. Wu, P. Shen, S. He, and X. Wang, "Analyses of a multimodal spontaneous facial expression database," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 34–46, 2013. 1

[7] M. Hayat and M. Bennamoun, "An automatic framework for textured 3d video-based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 301–313, 2014. 1

[8] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014. 1

[9] S. Taheri, V. M. Patel, and R. Chellappa, "Component-based recognition of faces and facial expressions," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 360–371, 2013. 1

[10] M. K. Abd El Meguid and M. D. Levine, "Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 141–154, 2014. 1

[11] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expression detection in hi-speed video based on facial action coding system (FACS)," *IEICE Transactions on Information and Systems*, vol. 96, no. 1, pp. 81–92, 2013. 1, 2, 3, 4

[12] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *IEEE International Conference on Computer Vision*, 2011. 1, 2, 3, 4, 7, 8

[13] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro- and micro-expression spotting in video using strain patterns," in *IEEE Workshop on Applications of Computer Vision*, 2009. 1, 3, 4

[14] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2011. 1, 2, 3, 4

[15] S.-J. Wang, H.-L. Chen, W.-J. Yan, Y.-H. Chen, and X. Fu, "Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine," *Neural Processing Letters*, vol. 39, no. 1, pp. 25–43, 2014. 1, 2, 3, 4, 7, 9

[16] P. Ekman and W. Friesen, "Nonverbal leakage and clues to deceptions," DTIC Document, Tech. Rep., 1969. 1

[17] P. Ekman, "Lie catching and microexpressions," in *The philosophy of deception*. Oxford Press, 2009, pp. 118–133. 1

[18] ——, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. WW Norton & Company, 2009. 1

[19] M. Bartlett, G. Littlewort, M. Frank, and K. Lee, "Automatic decoding of facial movements reveals deceptive pain expressions," *Current Biology*, vol. 24, no. 7, pp. 738–743, 2014. 1

[20] S. Porter and L. t. Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychological Science*, vol. 19, no. 5, pp. 508–514, 2008. 1

[21] H. Jascha, "Exhibition: how faces share feelings," *Nature*, vol. 452, no. 7186, pp. 413–413, 2008. 1

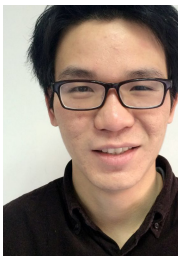
[22] P. Ekman, E. T. Rolls, D. I. Perrett, and H. D. Ellis, "Facial expressions of emotion: An old controversy and new findings [and discussion]," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 335, no. 1273, pp. 63–69, 1992. 1

[23] D. M. Bernstein and E. F. Loftus, "How to tell if a particular memory is true or false," *Perspectives on Psychological Science*, vol. 4, no. 4, pp. 370–374, 2009. 1

[24] T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," *British Journal of Clinical Psychology*, vol. 45, no. 4, pp. 579–583, 2006. 1

[25] L. Dacre Pool and P. Qualter, "Improving emotional intelligence and emotional self-efficacy through a teaching intervention for university students," *Learning and Individual Differences*, vol. 22, no. 3, pp. 306–312, 2012. 1

[26] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014. 1
[27] F. Salter, K. Grammer, and A. Rikowski, "Sex differences in negotiating with powerful males," *Human Nature*, vol. 16, no. 3, pp. 306–321, 2005. 1
[28] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *The Annual Meeting of the International Communication Association*, 2009. 1
[29] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995. 1, 3
[30] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *International Conference on Pattern Recognition*, 2014. 1
[31] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012. 2
[32] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013. 2, 5, 7, 8
[33] 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 *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013. 2, 7, 8
[34] 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. 2, 4, 7
[35] P. Ekman and W. V. Friesen, *Facial action coding system*. Consulting Psychologists Press, 1977. 2, 3
[36] G. Ardeshir, "Image registration by local approximation methods," *Image and Vision Computing*, vol. 6, no. 4, pp. 255–261, 1988. 3
[37] 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*, vol. 29, no. 6, pp. 915–928, 2007. 3
[38] Q. Wu, X. Shen, and X. Fu, "The machine knows what you are hiding: an automatic micro-expression recognition system," in *International Conference on Affective Computing and Intelligent Interaction*, 2011. 3, 4
[39] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014. 4
[40] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*. Springer, 2009, pp. 23–45. 4
[41] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1744–1757, 2012. 4
[42] A. Goshtasby, "Image registration by local approximation methods," *Image and Vision Computing*, vol. 6, no. 4, pp. 255–261, 1988. 4
[43] J. Zhang, Q. Wang, L. He, and Z.-H. Zhou, "Quantitative analysis of nonlinear embedding," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 1987–1998, 2011. 6
[44] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003. 6, 7
[45] C. J. Van Rijsbergen, *Information Retrieval (2nd ed.)*. London: Butterworths, 1979. 8
[46] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002. 9



Feng Xu received the B.S. degree from the School of Computer Science of Fudan University, China, in 2013. He is currently pursuing the Masters degree in Computer Science at Fudan University. His research interests include machine learning, data mining, computer vision, and applications in emotion analysis.

Junping Zhang (M'05) received a B.S. degree from Xiangtan University, China, in 1992 and obtained an M.S. degree from Hunan University, Changsha, China, in 2000. In 2003, Dr. Zhang received a Ph.D. degree from the Institution of Automation, Chinese Academy of Sciences. In 2006, he joined Fudan University as an Associate Professor and became a Professor in the School of Computer Science in 2011. His research interests include machine learning, intelligent transportation systems, image processing and biometric authentication. Dr. Zhang is an associate editor of IEEE Intelligent Systems and IEEE Transactions on Intelligent Transportation Systems.

James Z. Wang received the bachelor's degree in mathematics and computer science *summa cum laude* from the University of Minnesota, the MS degree in mathematics and the MS degree in computer science, both from Stanford University, and the PhD degree in medical information sciences from Stanford University. He has been a faculty member at The Pennsylvania State University since 2000 where he is a Professor and the Faculty Council Chair in the College of Information Sciences and Technology. His main research interests are automatic image tagging, image retrieval, computational aesthetics, and computerized analysis of paintings. He was a visiting professor at the Robotics Institute at Carnegie Mellon University (2007-2008) and a program manager at the National Science Foundation (2011-2012). In 2008, he served as the lead guest editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence Special Section on Real-world Image Annotation and Retrieval. He has been a recipient of a US National Science Foundation Career award and the endowed PNC Technologies Career Development Professorship.