Reading Hidden Emotions: Spontaneous Micro-expression Spotting and Recognition

Xiaobai Li, *Student Member, IEEE*, Xiaopeng Hong, *Member, IEEE,* Antti Moilanen, Xiaohua Huang, *Student Member, IEEE,* Tomas Pfister, Guoying Zhao, *Senior Member, IEEE,* and Matti Pietikäinen, *Fellow, IEEE*

***Abstract*—Micro-expressions (MEs) are rapid, involuntary facial expressions which reveal emotions that people do not intend to show. Studying MEs is valuable as recognizing them has many important applications, particularly in forensic science and psychotherapy. However, analyzing spontaneous MEs is very challenging due to their short duration and low intensity. Automatic ME analysis includes two tasks: ME spotting and ME recognition. For ME spotting, previous studies have focused on posed rather than spontaneous videos. For ME recognition, the performance of previous studies is low. To address these challenges, we make the following contributions: (i) We propose the first method for spotting spontaneous MEs in long videos (by exploiting feature difference contrast). This method is training free and works on arbitrary unseen videos. (ii) We present an advanced ME recognition framework, which outperforms previous work by a large margin on two challenging spontaneous ME databases (SMIC and CASMEII). (iii) We propose the first automatic ME analysis system (MESR), which can spot and recognize MEs from spontaneous video data. Finally, we show that our method achieves comparable performance to humans at this very challenging task, and outperforms humans in the ME recognition task by a large margin.**

***Index Terms*—Micro-expression, facial expression recognition, affective computing, LBP, HOG.**

I. Introduction

FACIAL expressions (FE) are one of the major ways that humans convey emotions. Aside from ordinary FEs that we see every day, under certain circumstances emotions can also manifest themselves in the special form of micro-expressions (ME). An ME is a very brief, involuntary FE which shows the emotion that people try to conceal [1]. In contrast to ordinary FEs, (i) MEs are very short (1/3 to 1/25 second, the precise length definition varies [2], [3]), and (ii) the intensities of involved muscle movements are subtle [4].

The phenomenon was first discovered by Haggard and Isaacs [5] in 1966, who called them micromomentary facial expression. Three years later, Ekman and Friesen also reported finding MEs [6] when they were examining a video of a psychiatric patient, trying to find possible trait of her suicide tendency. Although the patient seemed happy throughout the film, a fleeting look of anguish lasting two frames (1/12s) was found when the tape was examined in slow motion. This feeling of anguish was soon confirmed through a confession from the patient in her another counseling session: she lied to conceal her plan to commit suicide. In the following decades, Ekman and his colleague continued researching MEs [1], [7], [8]. Their work has drawn increasing interests to this topic in both the scientific and non-scientific communities, resulting even in a new internationally broadcasted TV series ('Lie to Me'), with tens of millions of viewers.

A major reason for the considerable interest in MEs is that it is an important clue for lie detection [1], [9]. Spontaneous MEs occur fast and involuntarily, and they are difficult to control through one's willpower [10]. In high-stake situations [7] for example when suspects are being interrogated, an ME fleeting across the face could give away a criminal pretending to be innocent, as the face is telling a different story than his statements. Furthermore, as has been demonstrated in Ekman's research [9], [11], people who perform better at ME recognition tests are also better lie detectors. Due to this, MEs are used as an important clue by police officers for lie detection in interrogations. Ekman also developed a Micro Expression Training Tool (METT) [8] to help improve the ME recognition abilities of law enforcement officers. In addition to law enforcement, ME analysis has potential applications in other fields as well. In psychotherapy, MEs may be used for understanding genuine emotions of the patients when additional reassurance is needed. It can also help border control agents to detect abnormal behavior, and thus to screen potentially dangerous individuals during routine interviews.

However, one considerable obstacle lies in the way of all above-mentioned applications: Detecting and recognizing MEs are very difficult for human beings [11] (in contrast to normal FEs, which we can recognize effortlessly). Study [8] shows that for ME recognition tasks, people without training only perform slightly better than chance on average. This is because MEs are too short and subtle for human eyes to process. The performance can be improved with special training, but it is still far below the 'efficient' level. Moreover, finding and training specialists to analyze these MEs is very time-consuming and expensive.

Meanwhile, in computer vision (CV), many research groups have accumulated experience in analyzing ordinary FEs. Algorithms have been reported to achieve FE recognition performance of above 90% on frontal view FE databases [12]. Given these recent advances, it is reasonable to explore the use of computers for automatically analyzing MEs based on former knowledge gained from FE studies. FE recognition has become popular since Picard proposed the concept of Affective

X. Li, X. Hong, A. Moilanen, X. Huang, G. Zhao and M. Pietikäinen are with the Department of Computer Science and Engineering, University of Oulu, Oulu, Finland.

E-mail: {xiaobai.li, xiaopeng.hong, antti.moilanen, huang.xiaohua, guoying.zhao, mkp}@ee.oulu.fi

T. Pfister is with Department of Engineering Science, University of Oxford, Oxford, UK. E-mail: tp@robots.ox.ac.uk

This research is supported by the Academy of Finland, Tekes and Infotech Oulu.

computing [13] in 1997. But studies of ME have been very rare until now due to several challenges. One challenge is the lack of available databases. It is difficult to gather a large number of spontaneous ME samples, as MEs only occur under certain conditions, and the incidence rate is low. Other challenges include developing methods able to deal with the short durations and the low intensity levels of MEs.

For the problem of spontaneous ME analysis there are two major tasks: (i) *spotting* when the ME occurs from its video context (ME Spotting); and (ii) *recognizing* what emotion the ME represents (ME Recognition). In natural circumstances an ME may occur along with other more prominent motion such as head movements and eye blinks. This makes spotting and recognizing rapid and faint MEs very challenging.

To address these issues in spotting and recognizing spontaneous MEs, we make the following *main contributions*:

(1) We propose a new method for spotting spontaneous MEs. To our best knowledge, this is the first ME spotting method which works on spontaneous videos. Our method is based on feature difference (FD) contrast and peak detection, and it is training free. An initial version of this work was published in [14].

(2) We develop an advanced framework for ME recognition based on our previous work [15], [16]. This new framework achieves much better performance than previous work due to two main improvements: (i) Eulerian video magnification (EVM) method is employed for motion magnification to counter the the low intensity of MEs; and (ii) Three different feature descriptors (Local Binary Pattern (LBP), Histograms of Oriented Gradients (HOG) and Histograms of Image Gradient Orientation (HIGO)) and their combinations on three orthogonal planes are comprehensively investigated for this task.

(3) We provide the first comprehensive exploration of several key issues related to ME recognition. In particular, we draw the following conclusions basing on substantial experimental results: (i) temporal interpolation (TIM) is valuable for ME recognition as it unifies the ME sequence lengths; (ii) combining feature histograms on all three orthogonal planes (TOP) does *not* consistently yield the best performance, as the XY plane may contain redundant information; (iii) gradient-based features (HOG and HIGO) outperform LBP for ordinary color videos, while LBP features are more suitable for near-infrared (NIR) videos; and (iv) using a high speed camera facilitates ME recognition and improves results.

(4) Lastly, we propose an automatic ME analysis system (MESR), which first spots and then recognizes MEs from long videos. This is the first system that has ever been demonstrated to be able to spot and and recognize MEs from spontaneous video data. In experiments on the SMIC database, we show that it achieves comparable performance to the humans.

The remaining parts of this paper are organized as follows. Section II reviews related work; Section III introduces our ME spotting method; Section IV describes our new ME recognition framework; and Section V presents experimental results and discussions for both ME spotting and recognition. Conclusions are drawn in Section VI.

II. Related Work

A. ME databases

Adequate training data is a prerequisite for the development of ME analysis. Several *FE* databases exist, such as JAFFE [17], CK [18], MMI [19] and others.These databases have enhanced the progress of algorithms for conventional FE analysis. However, when it comes to the available ME databases, the options are much more limited.

Eliciting spontaneous MEs is difficult. According to Ekman, an ME may occur under ‘high-stake’ conditions [7], which indicate situations when an individual tries to hide true feelings (because (s)he knows that the consequences for being caught may be enormous). In Ekman’s work he proposed three ways to construct a high-stake situation: (i) asking people to lie about what they saw in videos; (ii) constructing crime scenarios; and (iii) asking people to lie about their own opinions. Different ways were used to motivate participants to lie successfully: for example, participants were informed that assessments of their performance (from professionals) will impact their future career development; or good liars got extra money as rewards. Maintaining these scenarios are demanding for the conductors of the experiments, and the occurrence of MEs is still low even under these conditions. Not everyone could yield an ME [16], [20], [21]. The second challenge comes after data collection: labeling MEs is very challenging and time-consuming even for trained specialists.

In some earlier studies on automatic ME analysis, posed ME data were used to bypass the difficulty of getting spontaneous data. Shreve *et al.* [22], [23] reported collecting a database called USF-HD which contains 100 clips of posed MEs. They first showed some example videos that contain MEs to the subjects, and then asked them to mimic those examples. There are limited details about the collected ME samples in their paper, and it is unclear what emotion categories were included or how the ground truth were labeled. The USE-HD data were used by the authors for automatic ME spotting using an optical flow method. Polikovsky *et al.* [24] also collected a posed ME database by asking subjects to perform seven basic emotions with low intensity and to go back to neutral expression as quickly as possible. Ten subjects were enrolled and the posed MEs were recorded by a high speed camera (200 fps). Their data were labeled with action units (AU) for each frame following the facial action coding system (FACS [25]), and the authors proposed to use a 3D-gradient orientation histogram descriptor for AU recognition.

The biggest problem of posed MEs is that they are different from real, naturally occurring spontaneous MEs. Studies show [6] that spontaneous MEs occur involuntarily, and that the producers of the MEs usually do not even realize that they have presented such an emotion. By asking people to perform expressions as quickly as possible one can obtain posed MEs, but they are different from spontaneous ones in both spatial and temporal properties [2], [4]. So methods trained on posed MEs can not really solve the problem of automatic ME analysis in practice.

Therefore it is important to obtain a large, representative spontaneous database of MEs. To this end, we collected the first spontaneous ME database SMIC [15], [16]. In order to simulate a ‘high-stake’ situation as Ekman suggested, we designed our ME inducing protocol as follows: (i) we used video clips which were demonstrated to be strong emotion inducing materials in previous psychological studies [26], [27] to make sure that subjects will have strong emotional reactions; (ii) we asked the subjects to hide their true feeling and keep a neutral face while watching the video clips. The high-stake situation was created by informing the subjects that they were being monitored through a camera while watching movies, and that they will be punished (filling a very long boring questionnaire) once we spot any emotion leaks on their faces. The ME induction setup is shown in Figure 1.

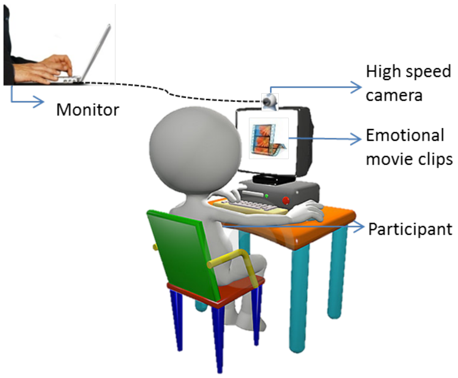

Fig. 1. Experiment setups for the collection of the SMIC database.

By using the above paradigm, we successfully induced spontaneous MEs from 16 out of 20 subjects. The first version of SMIC includes 77 MEs from six subjects [15], and it was later extended to the full version of SMIC which includes 164 MEs from 16 subjects [16]. The full version of SMIC contains three datasets (all with resolution of 640 × 480): (i) a HS dataset recorded by a high speed camera at 100 fps, (ii) a VIS dataset recorded by a normal color camera at 25 fps; and (iii) an NIR dataset recorded by a near infrared camera both at 25 fps. The HS camera was used to record all data, while VIS and NIR cameras were only used for the recording of the last eight subjects’ data. All datasets include ME clips of image sequences from onset to offset. The ME clips were segmented from the original long videos and labeled into three emotion classes: positive, surprise and negative. The labeling was performed by two annotators, first separately and then cross-checked. Only the labels that both annotators agreed on were included.

Yan and colleagues [20] collected another spontaneous ME database using a similar emotion-inducing paradigm. The Chinese Academy of Sciences Micro-expression (CASME) database [20] contains 195 MEs elicited from 19 participants. CASME data were recorded using two cameras: the BenQ M31 camera with frame rate of 60 fps and resolution of 1280 × 720 (CASME-A), and the Point Grey GRAS-03K2C camera with frame rate of 60 fps and resolution of 640 × 480 (CASME-B). One major difference between CASME and SMIC is that CASME has action unit (AU) labels. Each ME clip was first labeled with AUs on its onset, apex and offset frames, and then classified into one of the eight emotion categories: amusement, sadness, disgust, surprise, contempt, fear, repression and tension.

Later Yan *et al.* [21] summarized three main shortcomings in current ME databases: (i)) the number of samples are too small to be sufficient; (ii) the image resolutions are low which means that the face size is not big enough; (iii) the frame rates may be not high enough. As none of the previous ME studies could achieve high performance, Yan *et al.* argued that a new database which provides more ME samples with higher spatial and temporal resolutions would be helpful for achieving progress on this difficult task. So they collected the CASMEII database, which improves on all three aforementioned factors. The new database includes 247 ME samples from 26 subjects. The face videos were recorded at 200 fps, with an average face size of 280 × 340. The ME-inducing paradigm for CASMEII is similar to the one used in SMIC, and the data contains both AU labels and emotion labels of five classes (happiness, disgust, surprise, repression and other).

TABLE I
CURRENT MICRO-EXPRESSION DATABASES. ELICITATION METHODS **P/S**: POSED/SPONTANEOUS.

Database	USF-HD	Polikovsky	SMIC [16]			CASME [20]		CASMEII
	[23]	[24]	HS	VIS	NIR	A	B	[21]
MEs	100	N/A	164	71	71	100	95	247
Subjects	N/A	10	16	8		7	12	26
Fps	30	200	100	25	25	60		200
Elicitation	P	P	S			S		S
Emotions	N/A	7	3			8		5

Table I lists the key features of existing ME databases. In this paper we focus on spontaneous MEs. SMIC and CASMEII databases are used in our experiments as they are the most comprehensive spontaneous ME databases currently publicly available.

B. State of the art of ME spotting research

ME spotting refers to the problem of automatically detecting the temporal interval of an ME in a sequence of video frames. Solutions to similar problems, such as spotting ordinary FEs, eye-blinking, and facial AUs from live videos [28]–[30] have been performed, but only a few studies have investigated automatic spotting of MEs.

Most existing works have tackled the problem in posed ME databases. In Polikovsky’s papers [24], [31], the authors used 3D gradient histograms as the descriptor to distinguish the onset, apex and offset stages of MEs from neutral faces. Although their method could potentially contribute to the problem of ME spotting, two drawbacks have to be considered. First, their method was only tested on posed MEs, which are much easier compared to spontaneous ME data. Second, their experiment was run as a classification test, in which all frames were clustered into one of the four stages of MEs. It means the method could not be applied to online ME spotting.

Wu *et al.* [32] proposed to use Gabor filters to build an automatic ME recognition system which includes the spotting step. They achieved very high spotting performance on the METT training database [8]. However, the METT training samples are fully synthetic: they have been synthesized by inserting one FE image in the middle of a sequence of identical neutral face images. This makes the spotting task significantly easier, as in real conditions the onset and offset of an ME would not be as abrupt, and the context frames would be more dynamic. It is unknown how well the method would perform on real videos.

Shreve *et al.* [22], [23] used an optical strain-based method to spot both macro (ordinary FEs – the antonym for 'micro') and micro expressions from videos. But as in the aforementioned studies, the data used in their experiments are posed. The authors reported being able to spot 77 from the 96 MEs, with 36 false positives on their posed database. The paper also evaluated their method on a small database of 28 spontaneous MEs found in TV or on-line videos, with results of 15 true positives and 18 false positives. But the database was small and not published.

As seen above, most previous ME spotting methods were tested only using posed data, which are different (and easier) compared to spontaneous MEs. In particular, during the recording of posed ME clips, subjects can voluntarily control their behavior according to the given instructions. Therefore one should expect the posed data to be more 'clean', *e.g.* contain restricted head movements, more clear-cut onsets and offsets, less movement in the context frames between two MEs, and so on. In contrast, the situation is more complicated in real videos which contain spontaneous MEs. This is because spontaneous MEs can appear during the presence of an ordinary FE (with either the same or the opposite valence of emotion [4], [33]), and sometimes they overlap with the occurrences of eye blinks, head movements and other motions. Therefore the spotting task is more challenging on spontaneous ME data. As an intermediate step to spot MEs in spontaneous data, several studies [15], [34]–[36] tackled an easier step which was referred to as ME 'detection'. ME detection is a two-class classification problem, in which a group of labeled ME clips are classified against the rest non-ME clips. However, to our knowledge, none of these works present any experiments for spotting spontaneous MEs from long realistic videos.

C. State of the art of ME recognition research

ME recognition is the task where, given a 'spotted' ME (*i.e.* a temporal interval containing an ME), the ME is classified into two or more classes (*e.g.* happy, sad *etc.*). Experiments on this problem are more prominent in the literature, and have been carried out using both posed and spontaneous data.

Some researchers investigated ME recognition on posed ME databases. Polikovsky *et al.* [24], [31] used a 3D gradient descriptor for the recognition of AU-labeled MEs. Wu *et al.* [32] combined Gentleboost and an SVM classifier to recognize synthetic ME samples from the METT training tool.

Several recent studies also reported tests on spontaneous ME databases. In our previous work [15] we were the first to propose a spontaneous ME recognition method. The method combined three steps: (i) a temporal interpolation model (TIM) to temporally 'expand' the micro-expression into more frames, (ii) LBP-TOP feature extraction (after detecting facial feature points with an Active Shape Model (ASM)); and (iii) Multiple kernel learning (MKL) for classification. The method achieved an accuracy of 71.4% (two-class classification) for ME recognition on the first version of the SMIC database. In our later work [16], a similar method was tested on the full version of SMIC and the best recognition result was 52.11% (three-class classification) on the VIS part (25fps RGB frames) of the database. Since then, LBP and its variants have often been employed as the feature descriptors for ME recognition in many other studies. Ruiz-Hernandez and Pietikäinen [34] used the re-parameterization of a second order Gaussian jet to generate more robust histograms, and achieved better ME recognition result than [15] on the first version of SMIC database (six subjects).

Song *et al.* [37] recognized emotions by learning a sparse codebook from facial and body micro-temporal motions. Although the concept of ME was employed in their study, their definition of MEs was wider than that of the current paper, as gestures from body parts (other than face) were also included. They ran experiments on a spontaneous audio-visual emotion database (AVEC2012), but not on any ME database. Wang *et al.* [38] extracted LBP-TOP from a Tensor Independent Colour Space (TICS) (instead of ordinary RGB) for ME recognition, and tested their method on CASMEII database. In their another paper [39], Local Spatiotemporal Directional Features (LSDF) were used together with the sparse part of Robust PCA (RPCA) for ME recognition, achieving an accuracy of 65.4% on CASMEII.

So far most ME recognition studies have considered using LBP-TOP as the feature descriptor. As there is still much room for improvement in the recognition performance, more robust descriptors and machine learning methods need to be explored.

III. Method for ME spotting

In previous work [14] we proposed an ME spotting method that combines appearance-based feature descriptors with Feature Difference (FD) analysis. The method consists of four steps: (i) three facial landmark points (inner eye corners and a nasal spine point) are detected in the first frame and tracked through the video, and each face image is divided into a block structure based on the tracked landmark positions; (ii) appearance-based features are calculated for each block; (iii) the dissimilarity of features for each block of sequential frames within a defined time interval is calculated using the Chi-Squared distance; and (iv) thresholding and peak detection are applied to spot rapid facial movements from the video.

In this work we adopt the same method flow. However, for the second step of feature extraction we employ two kinds of features: the first one is Local Binary Patterns (LBP) [40] and the second one is Histogram of Optical Flow (HOOF) [41]. In the original work [14], only LBP was employed. But as both LBP-based methods and Optical Flow-based methods are demonstrated to be efficient in facial expression recognition research [15], [22], [23], [42], we also experiment with the HOOF feature to compare their performance.

Details about the four steps of the method and how we extract the LBP feature are provided in [14]. HOOF is extracted using the code of Liu *et al.* [41] with default parameters. In particular, we calculate HOOF by obtaining the flow field for each frame (with one frame being the reference flow frame) and then compiling the orientations into a histogram. Two kinds of reference frames are tested: one uses the first frame of the input video as the reference frame throughout the whole video; and the other uses the first frame within the current time window as the reference frame (and therefore changes as the time window slides through the video). We discuss the two options in the experiments. Spotting results vary when using different thresholds. We present and discuss the results by adjusting the percentage parameter (‘p’ in equation 5 of the original paper [14]) in Section V.A.

IV. Method for ME recognition

In this section we present the framework of our proposed method for ME recognition (*i.e.*, given a temporal interval containing a ME, classifying it into two or more classes). An overview of our method is shown in Figure 2. The following subsections discuss the details of the method.

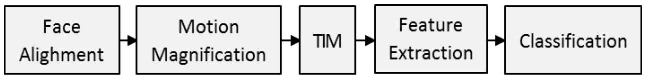

Fig. 2. Framework diagram for the proposed ME recognition method.

A. Face alignment

For ME spotting we do not align faces across different videos, because the method only compares frames within the current input video. However, for ME recognition, training is inevitable to deal with the large inter-class variation. Thus face alignment is needed to minimize the differences of face sizes and face shapes across different video samples. We carry out the following steps for face alignment as a pre-processing step for ME recognition.

Let $V = \{v_1, v_2, \ldots, v_l\}$ denote the whole set of ME clips from a testing database and l is the total number of ME samples. The ith sample clip is given by $v_i = \{I_{i,1}, I_{i,2}, \ldots, I_{i,n_i}\}$, where $I_{i,j}$ is the jth frame and n_i is the frame number of the clip v_i.

First, we select a frontal face image I_{mod} with neutral expression as the model face. 68 facial landmarks of the model face $\psi(I_{\text{mod}})$ are detected using the Active Shape Model [43].

Then the 68 landmarks are detected on the first frame of each ME clip $I_{i,1}$ and normalized to the model face using a Local Weighted Mean (LWM) [44] transformation. The transform matrix $TRAN$ is:

$$TRAN_i = LWM(\psi(I_{\text{mod}}), \psi(I_{i,1})),\ i = 1, \ldots, l, \quad (1)$$

where $\psi(I_{i,1})$ is the coordinates of 68 landmarks of the first frame of the ME clip v_i. Then all frames of v_i were normalized using the same matrix $TRAN_i$. We assume that the rigid head movement could be neglected because the time scope of labeled out ME clips are very short. The normalized image I' was computed as a 2D transformation of the original image:

$$I'_{i,j} = TRAN_i \times I_{i,j},\ j = 1, \ldots, n_i, \quad (2)$$

where $I'_{i,j}$ is the jth frame of the normalized ME clip v'_i.

In the last step, we crop face areas out from normalized images of each ME clip using the rectangular defined according to the eye locations in the first frame $I'_{i,1}$.

B. Motion magnification

One major challenge for ME recognition is that the intensity levels of facial movements are too low to be distinguishable. To enhance the differences of MEs we propose to use the Eulerian video magnification (EVM) method [45] to magnify the subtle motions in videos. The original method was proposed for magnifying either motion or color content of a video. Here we apply it for motion magnification. The main steps of how we utilize EVM to magnify ME sequences are described below. For more details about the EVM method we refer readers to [45].

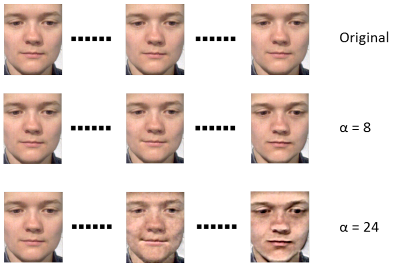

Fig. 3. A ME clip magnified at different α levels.

The input video sequence is first decomposed into different spatial frequency bands using standard Laplacian pyramid method. Then all the spatial frequency bands pass a temporal filter. Different kinds of temporal filters could be selected according to the purpose of applications. Here we use an infinite impulse response (IIR) filter with cut-off frequencies of $[0.4, 3]$Hz, as a filter with broader passband works better for motion magnification. In the next step, the temporal filtered bands are amplified with a given magnification factor α. Bigger values of α lead to larger scale of motion amplification, but also can cause bigger displacement and artifacts especially for the contents at higher spatial frequencies. An example of ME clip magnified at different α levels is shown in Figure 3. In Wu's method [45], they used the spatial wavelength parameter λ to set a constrain for the relationship of α and λ as

$$(1 + \alpha)\delta(t) < \frac{\lambda}{8}, \quad (3)$$

where according to [45], $\delta(t)$ is the displacement function of observed motion in the video along time t. We set $\lambda = 16$ for all magnification process. Effects of using different magnification factor α for ME recognition are explored in our experiment by varying α in ten levels.

It is worth noting that the EVM method facilitates ME recognition as the differences between different categories of MEs are enlarged. However, we do not use it for ME spotting because it magnifies unwanted motions (*e.g.* head movements) at the same time. The issue is discussed in Section V.A.

C. Temporal interpolation model (TIM)

Another difficulty for ME recognition is the short and varied duration. The problem is more evident when the videos are filmed with low frame rate. For example when recording at a standard speed of 25 fps, some MEs only last for four to five frames, which limit the application of some spatial-temporal descriptors, *e.g.*, if we use LBP-TOP we can only use the radius $r = 1$. To counter for this difficulty we propose to use the Temporal interpolation model (TIM) introduced by Zhou *et al.* [46].

The TIM method allows us to interpolate images at arbitrary time positions using very small number of input frames. We use TIM to interpolate all ME sequences into the same length (*e.g.* of 10, 20, ... or 80 frames) for two purposes: (i) up-sampling the clips with too few frames; and (ii) with a unified clip length, more stable performance of feature descriptors can be expected. For more details about the TIM method we refer readers to [46]. The effect of interpolation length is discussed in Section V.B.1.

D. Feature extraction

Several spatial-temporal local texture descriptors (SLTD) have been demonstrated to be effective in tackling the FE recognition problem [47], [48]. Here we employ three kinds of SLTDs in our ME recognition framework to compare their performance. Details of each descriptor are described below, and the comparison of their performance will be discussed in Section V.B.2.

1) LBP on three orthogonal planes: Local binary pattern on three orthogonal planes (LBP-TOP), proposed by Zhao and Pietikäinen [47], is an extension of the original LBP for dynamic texture analysis in spatial-temporal domain. It is one of the most frequently used SLTDs for FE recognition, and also for recent ME recognition studies.

A video sequence can be thought as a stack of XY planes on T dimension, as well as a stack of XT planes on Y dimension, or a stack of YT planes on X dimension. The XT and YT plane textures can provide information about the dynamic process of the motion transitions. Figure 4(a) presents the textures of XY, XT and YT plane around the mouth corner of one ME clip.

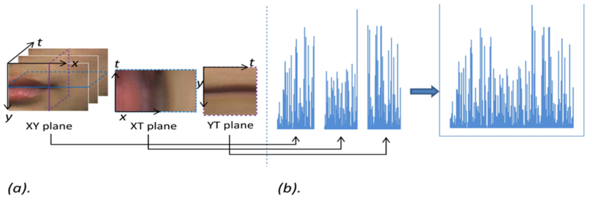

Fig. 4. (a) The textures of XY, XT and YT planes, (b) Their corresponding histograms and the concatenated LBP-TOP feature.

The traditional way of extracting LBP-TOP feature is illustrated in Figure 4(b). To include information from 3D spatial-temporal domain, LBP codes are extracted from every pixel of XY, XT or YT plane to produce their corresponding histograms. The three histograms are then concatenated into a single histogram as the final LBP-TOP feature vector.

It can be seen that the histograms for XY, XT and YT plane represent different information, and previous results indicate that the traditional LBP-TOP feature which uses all three histograms does not always yield the best performance [35]. In our method we will consider five different (combinations of) LBP histograms on the three planes as listed in Table II.

2) HOG and HIGO on three orthogonal planes: Histograms of Oriented Gradients (HOG) [49] is another popular feature descriptor which has been applied in FE recognition [48], [50]. As reviewed in Section II.C, most previous ME recognition studies considered LBP feature. Here we use HOG as the second feature descriptor for ME recognition.

TABLE II
FIVE DIFFERENT COMBINATIONS OF LBP FEATURES ON THREE ORTHOGONAL PLANES AND THEIR CORRESPONDING ABBREVIATIONS.

Abbreviation	**Histogram of which plane**
LBP-TOP	XY + XT + YT
LBP-XYOT	XT + YT
LBP-XOT	XT
LBP-YOT	YT
LBP	XY

First, we formulate the 2D HOG on the XY plane. Provided an image I, we obtain the horizontal and vertical derivatives I_x and I_y using the convolution operation. More specifically $I_x = I * K^T$ and $I_y = I * K$, where $K = [-1\,0\,1]^T$. For each point of the image, its local gradient direction θ and gradient magnitude m are computed as follows:

$$\theta = \arg(\nabla I) = \operatorname{atan} 2(I_y, I_x), \tag{4}$$

$$m = |\nabla I| = \sqrt{I_x^2 + I_y^2}. \tag{5}$$

Let the quantization level for θ be B and $\mathcal{B} = \{1, \dots, B\}$. Note that $\theta \in [-\pi, \pi]$. Thus a quantization function of θ is a mapping $Q : [-\pi, \pi] \to \mathcal{B}$. As a result, the histogram of oriented gradients (HOG) for a local 2D region (*i.e.* a block) or a sequence of 2D regions (*i.e.* a cuboid) $\mathcal{N}$ is a function $g : \mathcal{B} \to R$. More specifically, it is defined by

$$g(b) = \sum_{\mathbf{x} \in \mathcal{N}} \delta(Q(\theta(\mathbf{x})), b) \cdot m(\mathbf{x}), \tag{6}$$

where $b \in \mathcal{B}$ and $\delta(i, j)$ is the Kronecker's delta function as

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j\,. \end{cases} \tag{7}$$

For HOG, each pixel within the block or cuboid has a weighted vote for a quantized orientation channel b according to the response found in the gradient computation.

Second, we introduce the third feature descriptor employed: the histogram of image gradient orientation (HIGO). HIGO is a degenerated variant of HOG: it uses 'simple' vote rather than 'weighted vote' when counting the responses of the histogram bins. In detail, the function h for HIGO is defined as:

$$h(b) = \sum_{\mathbf{x} \in \mathcal{N}} \delta(Q(\theta(\mathbf{x})), b), \tag{8}$$

where b and δ have the same meaning as in Equation (6).

It is worth noting that HIGO depresses the influence of illumination and contrast by ignoring the magnitude of the first order derivatives. For those challenging tasks (*e.g.* recognizing spontaneous MEs recorded in authentic situations) in which the illumination conditions substantially vary, HIGO is expected to have enhanced performance.

The corresponding 3D spatial-temporal version of HOG and HIGO features can be easily obtained by extending the descriptor on the XY plane to the three orthogonal planes, as is done when one extends LBP to LBP-TOP. We explore five different (combinations of) HOG and HIGO from the three orthogonal planes as we explained for the LBP feature. Their abbreviations are defined the same way as in Table II (with 'LBP' changed to'HOG' or 'HIGO').

In order to account for the variations in illumination or contrast, the gradient intensity is usually normalized. In this paper, we use the local L1 normalization for the histogram calculated from each block/cuboid. The concatenations of the normalized block/cuboid histograms are globally normalized (either L1 or L2) to form the final descriptor.

E. Classification

After feature extraction, we use a linear Support Vector Machine (LSVM) [51] as the classifier for ME recognition.

V. Experimental results and discussions

We next present experiments and discuss our results. Section V.A presents results for ME spotting; Section V.B explains ME recognition results in four sub-experiments; Section V.C tests a fully automatic ME analysis system which combines both spotting and recognition process; and Section V.D compares our automatic method to the performance of human subjects on both ME spotting and ME recognition tasks, and shows that our method achieves comparable results to humans in this very challenging task.

A. ME spotting

In this section we report results of ME spotting experiments. We first introduce the testing data, and then describe parameters and measures of the experiment; and finally present spotting results with discussions.

1) Datasets: SMIC and CASMEII databases are used in the spotting experiment. The original SMIC database only includes labeled ME clips that include frames from onset to offset. It is suitable for the ME recognition problem, but the spotting test involves longer video sequences which also include frames before and after the ME span. We re-built an extended version of SMIC (SMIC-E) by extracting longer clips around time points when MEs occur from the original videos. The three datasets in the new SMIC-E database with longer video sequences are denoted as SMIC-E-VIS, SMIC-E-NIR and SMIC-E-HS accordingly. The SMIC-E-VIS and SMIC-E-NIR dataset both include 71 long clips of average duration of 5.9 seconds. The SMIC-E-HS dataset contains 157 long clips of an average length of 5.9 seconds. Three samples from the original SMIC-HS dataset are not included in the SMIC-E-HS due to data loss of the original videos. Unlike SMIC, the original CASMEII database provides long video clips that include extra frames before and after the ME span, so we are able to use the clips for ME spotting as they are. The average duration of CASMEII clips is 1.3 seconds. One video clip in CASMEII was excluded because its duration is too short.

2) Parameters: The ME interval N is set to correspond to a time window of about 0.32 seconds ($N = 9$ for SMIC-E-VIS and SMIC-E-NIR, $N = 33$ for SMIC-E-HS, $N = 65$ for CASMEII). For LBP, uniform mapping is used, the radius r is set to $r = 3$, and the number of neighboring points p is set to $p = 8$ for all datasets. For HOOF, the parameters were left to the default values as in [41]. We tested two different approaches for selecting the reference frame. In the first approach, the reference frame was fixed to be the first frame of the video. In the second approach, the reference frame was set to be the TF, *i.e.* the reference frame follows the CF in a temporal sliding window manner. Results showed that the first approach always yields better performance than the second one – therefore in the following we report results using the first approach. We also tried using motion magnification in ME spotting in another prior test. We magnified videos before extracting features. However, it turned out that both true and false peaks were enlarged and the spotting accuracy was not improved. Therefore motion magnification is not used in the following ME spotting experiments.

After the peak detection all the spotted peak frames are compared with ground truth labels to tell whether they are true or false positive spots. With a certain threshold level, if one spotted peak is located within the frame range of $[\text{onset}-(N-1)/4, \text{offset}+(N-1)/4]$ of a labeled ME clip, the spotted sequence will be considered as one true positive ME; otherwise the N frames of spotted sequence will be counted as false positive frames. We define the true positive rate (TPR) as the percentage of frames of correctly spotted MEs, divided by the total number of ground truth ME frames in the dataset. The false positive rate (FPR) is calculated as a percentage of incorrectly spotted frames, divided by the total number of non-ME frames from all the long clips. Performance of ME spotting is evaluated using receiver operating characteristic (ROC) curves with TPR as the y axis and the FPR as the x axis.

3) Results: We performed the spotting experiments on CASMEII and the three datasets of SMIC-E. The spotting results on each dataset are presented in Figure 5. A ROC curve is draw for each of the two feature descriptors on one dataset. Points of the ROC curves in Figure 5 are drawn by varying the percentage parameter (the 'p' as in equation 5 of the original paper [14]) from 0 to 1 with step size of 0.05.

From the figure, we observe that more MEs are correctly spotted when we drop the threshold value, but with the expense of higher FPR. The area under the ROC curve (AUC) for each curve is calculated and listed in Table III. Higher AUC value indicates better spotting performance.

TABLE III
AUC values of the ME spotting experiments using LBP and HOOF as feature descriptors on CASMEII and three datasets of SMIC-E.

	SMIC-E-HS	SMIC-E-VIS	SMIC-E-NIR	CASMEII
LBP	83.32%	84.53%	80.60%	92.98%
HOOF	69.41%	74.90%	73.23%	64.99%

Figure 5 and Table III show that LBP outperforms HOOF for the proposed ME spotting method, as its AUC values are higher and the FPRs are lower. For spotting on the three datasets of SMIC-E, best performance is achieved on SMIC-E-VIS dataset. By using LBP, our proposed method can spot about 70% of MEs with only 13.5% FPR, and the AUC is 84.53%. We observe that the majority of the false positives are eye blinks (discussed in detail below). On CASMEII, the advantage of LBP feature is more obvious (AUC of 92.98%). We hypothesize that the reason why a higher AUC is achieved on CASMEII is that CASMEII contains shorter video clips than SMIC-E (so the spotting task is easier).

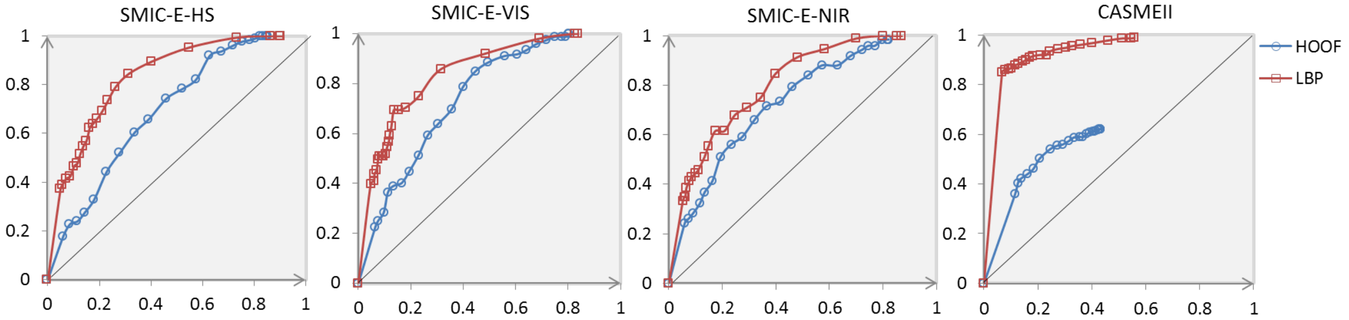

Fig. 5. ROC curves for ME spotting on CASMEII and three SMIC-E datasets. The x axis shows the false positive rate (FPR), and the y axis shows the true positive rate (TPR).

To the best of our knowledge this is the first report of ME spotting on spontaneous ME databases, so there are no result to compare with. Some works have explored ME spotting on posed ME datasets, however unfortunately these datasets have not been made publicly available. The current results show that spontaneous MEs can be spotted by comparing the feature differences of the CF and the AFF, and that LBP is more efficient than HOOF. Spotting MEs in spontaneous videos is significantly more difficult than previous work done on posed videos, as random motions could interfere as noise.

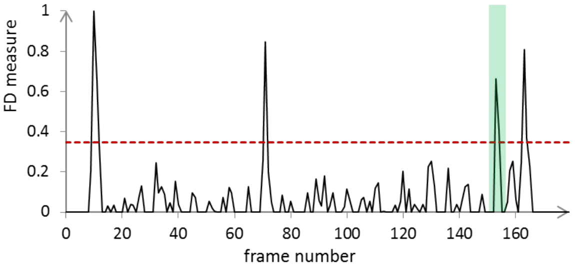

Fig. 6. An example of ME spotting result showing a true spot of ME (the green peak), with three false spots of eye blinks.

Discussion of failure cases*:* Upon detailed examinations of the spotting result, we found that a large portion of false spots were caused by eye movements, such like eye blinks or eye-gaze changes. One such example is shown in Figure 6 (the spotting result on a video from SMIC-E-VIS). Four peaks were spotted when the threshold level was set to the dashed line. We compared the peaks to the ground truth labels and found that only the third peak (around frame 153) is a true positive spot of ME, while the other three peaks were caused by eye blinks. As we report results on spontaneous facial videos, it is expected that there will be many eye blinks in the videos. The proposed method detects transient movements on the scale of the whole face, so eye movements could also be detected as big peaks. In contrast, in posed databases eye blinks are significantly less common as they can be voluntarily concealed. Therefore this issue has never been addressed in previous ME spotting works.

We tried two ways to rule out eye movements: (i) by excluding the eye regions during feature extraction; and (ii) by using an eye-blink-detector to exclude eye blink peaks from the results. Both approaches helped to reduce the FPRs. However, both approaches also caused a decrease in TPR at the same time. This is due to many spontaneous MEs involving muscles around eye regions. Furthermore, the onsets of many MEs (about 50%, as we have empirically measured) also temporally overlap with eye blinks. Thus neither approach turned out to be good enough. We plan to perform more comprehensive investigations about this issue in future work.

B. *ME recognition*

In this section we report results of ME recognition experiments on two spontaneous ME databases: the full version of SMIC and CASMEII. ME clips (including only frames from ME onset to offset) with raw images are used. We use leave-one-subject-out protocol in all of our ME recognition experiments.

As described in Section IV, our ME recognition method consists of three main components: motion magnification, interpolation and feature extraction. Furthermore, SMIC database contains multimodal data (*e.g.* NIR, VIS and HS). We therefore carry out four sub-experiments to evaluate the effect of each of these factors. Experiment 1 evaluates the effect of temporal interpolation; Experiment 2 compares different feature descriptors; Experiment 3 compares the result using different modalities in the datasets; and Experiment 4 evaluates our motion magnification method and provides a comparison to state of the art.

1) Experiment 1: Effect of the interpolation length: The aim of this first sub-experiment is to explore how the sequence length of interpolation affects the accuracy of ME recognition.

Parameters*:* In order to control the effect from other factors and focus on TIM, we skip the motion magnification step, and use only LBP-TOP as the feature descriptor (with a group of fixed parameters). After face alignment, TIM is applied to interpolate ME sequences into eight different lengths (10, 20, 30, . . . , 80), and then LBP-TOP features (8 × 1 blocks, $r = 2$, $p = 8$) are extracted. We test on SMIC-HS, SMIC-VIS and SMIC-NIR datasets. The average sequence length of the original ME samples is 33.7 frames for SMIC-HS and 9.66 frames for SMIC-VIS and SMIC-NIR. Training and testing is done using linear SVM (penalty parameter $c = [0.1, 1, 2, 10, \ldots, 100]$).

Results*:* Figure 7 shows results on the three datasets at different sequence lengths. The best performance for all three datasets is achieved with interpolation to 10 frames (TIM10). We analyzed this finding from two aspects:

(1) The left side of the figure shows that TIM10 leads to higher performance than no TIM (without interpolation). For

the VIS and NIR datasets, TIM10 hardly alters the sequence lengths, as the average length of the no TIM sequences (original ME samples) is 9.66 frames. For the HS dataset, TIM10 is a down-sampling process, as the average length of original sequences is 33.7 frames. We therefore hypothesize that TIM10 performs better than no TIM because the input frames with unified sequence length improve the performance of the feature descriptor.

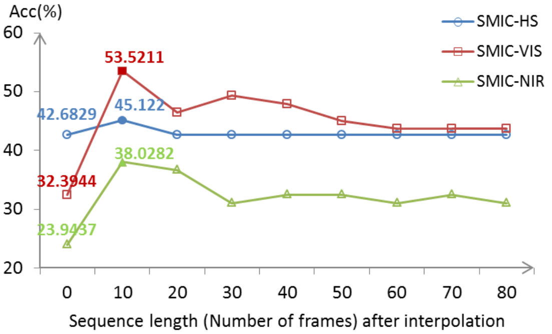

Fig. 7. ME recognition accuracy with different TIM length, using LBP-TOP as feature descriptor. The x-axis shows the frame numbers of ME sequences after TIM interpolation, and the y-axis shows accuracies. Points at $x = 0$ show accuracy of original sequences with no TIM.

(2) The right side of the figure shows that longer interpolated sequences (TIM20 to TIM80) do not lead to improved performance compared to TIM10. With the current datasets using TIM method, it appears that interpolation to 10 frames is enough. A possible explanation for the performance here is that in longer sequences, changes along the time-dimension are diluted, which makes feature extraction more challenging. This finding is supported by previous reported in [15] and [16].

TABLE IV
COMPUTATION TIME COMPARISON FOR DIFFERENT TIM LENGTHS.

TIM length	no	10	20	30	40	50	60	70	80
Time (s)	87.8	22.1	49.6	77.2	104.9	132.9	160.0	188.3	216.6

We also evaluate the running time for different TIM lengths. These results are shown in Table IV. The higher interpolation frame rate requires higher computation time (and also occupies more storage space, if the frames are stored) but does not consistently lead to improved accuracy.

To conclude, we see evidence that TIM process is necessary and valuable for ME recognition, and that 10 frames is enough. We therefore set TIM length as 10 for all the following experiments.

2) Experiment 2: Comparison of three features: In this sub-experiment, we evaluate the performance of three features (LBP, HOG and HIGO) for the ME recognition task. Multiple combinations of histograms on three orthogonal planes are considered for the three features respectively (see Table II).

Parameters*:* After face alignment, the sequences are interpolated into 10 frames using TIM method (the step of motion magnification is skipped in this experiment for later discussion in sub-experiment 4). For feature extraction, three kinds of features are extracted from evenly divided blocks of each interpolated sequence. In order to avoid bias, and to compare the performance of features on a more general level, multiple groups of features are extracted by varying several parameters. For the LBP feature, we vary the radius r, neighbor points p and the number of divided blocks; for the HOG and HIGO features, the number of bins b is fixed as $b = 8$, and we vary the number of divided blocks. For the classification step, training and testing is done using linear SVM with leave-one-subject-out protocol. Experiments are conducted on CASMEII and the three datasets of SMIC.

Results*:* The results for ME recognition using three feature descriptors on each dataset are shown in Table V. For each feature descriptor, we list results obtained using five combinations from three orthogonal planes. For each combination, the best result with its corresponding parameters are listed. We discuss the first four columns of result in this section. The last column of results for SMIC-subHS dataset will be discussed in Experiment 3.

Two conclusions can be drawn from the results of the first four columns of Table V:

(1) TOP (three orthogonal planes) does not consistently yield the best performance for ME recognition. Sometimes better results are achieved from using only XOT, YOT or XYOT (while the XY plane always yields lowest performance). This finding is consistent for all three features on all four test datasets. This indicates that the dynamic texture along the T dimension represents the most important information for ME recognition. On the other hand, the XY histogram seems to contain redundant information, thus making classification more difficult (because the XY plane contains information about the facial appearance rather than motion). Similar findings were also reported in [35].

(2) The gradient-based features HOG and HIGO outperform LBP for ME recognition on ordinary RGB data (CASMEII, SMIC-HS and SMIC-VIS). The best results are obtained on SMIC is 76.06% using HIGO-XOT. Further comparison between the two gradient based features shows that HIGO performs better than HOG. One possible explanation for this is that HIGO is invariant to the magnitude of the local gradients, which varies significantly across subjects due to different muscle moving speeds. However, for infrared data (SMIC-NIR) this trend is different: LBP performs best. The textures recorded by a near infrared camera are very different from RGB videos, as NIR textures are less affected by illumination. This is consistent with another study [52] which also reported that LBP feature is suitable for NIR data .

3) Experiment 3: Comparison of datasets recorded with different cameras: In this sub-experiment, we compare the ME recognition performance using different recording instruments (the SMIC dataset includes ME samples recorded with three cameras). In sub-experiment 2, the SMIC-VIS dataset led to best performance on SMIC. However, SMIC-HS contains more ME samples than the SMIC-VIS and SMIC-NIR. To make the comparison fair, we form a SMIC-subHS dataset containing the same 71 ME samples as SMIC-VIS and SMIC-NIR, and run the same test as we did in sub-experiment 2.

Results*:* The results are shown in the rightmost column in Figure V. By comparing the results of SMIC-VIS, SMIC-NIR and SMIC-subHS which contain the same samples we observe that:

TABLE V
ME RECOGNITION RESULTS USING LBP, HOG AND HIGO AS FEATURE DESCRIPTORS ON EACH DATASET. LSVM IS EMPLOYED AS THE CLASSIFIER USING LEAVE-ONE-SUBJECT-OUT PROTOCOL. THE HIGHEST ACCURACIES FOR EACH FEATURE ON EACH DATASET ARE MARKED IN BOLD FONT. (p, r) INDICATES THE NEIGHBOR POINTS p AND RADIUS r OF LBP FEATURE; b IS THE NUMBER OF BINS OF HIGO AND HOG FEATURES.

		CASMEII			SMIC-HS			SMIC-VIS			SMIC-NIR			SMIC-subHS		
		block	Acc. (%)	(p, r) or b	block	Acc. (%)	(p, r) or b	block	Acc. (%)	(p, r) or b	block	Acc. (%)	(p, r) or b	block	Acc. (%)	(p, r) or b
LBP-	TOP	8×8×2	**55.87**	(8,2)	8×8×2	51.83	(8,2)	5×5×1	**70.42**	(8,2)	5×5×1	**64.79**	(8,3)	8×8×2	76.06	(8,2)
	XYOT	8×8×4	**55.87**	(8,2)	8×8×2	56.10	(8,1)	5×5×1	**70.42**	(8,2)	8×8×2	**64.79**	(8,2)	8×8×2	**77.46**	(8,2)
	XOT	8×8×4	55.06	(8,2)	8×8×2	**57.93**	(8,1)	5×5×1	**70.42**	(8,2)	8×8×1	54.93	(8,2)	5×5×2	**77.46**	(8,2)
	YOT	5×5×4	54.85	(8,1)	8×8×2	50.61	(8,1)	5×5×2	**70.42**	(8,1)	8×8×4	**64.79**	(8,2)	8×8×2	76.06	(8,2)
	LBP	8×8×2	44.53	(8,2)	8×8×2	43.29	(8,2)	5×5×1	67.61	(8,2)	8×8×4	50.70	(8,2)	8×8×2	64.69	(8,2)
HIGO-	TOP	8×8×2	55.87	8	6×6×2	59.15	8	4×4×2	69.01	8	6×6×2	53.52	8	4×4×2	**80.28**	8
	XYOT	8×8×2	55.47	8	6×6×2	59.76	8	6×6×2	71.83	8	6×6×1	52.11	8	4×4×2	78.87	8
	XOT	8×8×2	53.44	8	4×4×2	**65.24**	8	6×6×2	**76.06**	8	6×6×1	47.89	8	4×4×2	78.87	8
	YOT	8×8×2	**57.09**	8	6×6×2	58.54	8	4×4×2	71.83	8	6×6×2	**59.15**	8	4×4×2	78.87	8
	HIGO	8×8×2	42.51	8	2×2×8	50.61	8	4×4×1	60.56	8	6×6×2	35.21	8	4×4×1	64.79	8
HOG-	TOP	8×8×2	**57.49**	8	2×2×2	**57.93**	8	2×2×2	67.61	8	2×2×8	**63.38**	8	4×4×2	**80.28**	8
	XYOT	8×8×2	**57.49**	8	2×2×2	51.83	8	6×6×2	**71.83**	8	2×2×2	60.56	8	6×6×6	71.83	8
	XOT	8×8×2	51.01	8	4×4×8	**57.93**	8	4×4×2	**71.83**	8	6×6×2	56.34	8	2×2×2	69.01	8
	YOT	8×8×2	56.68	8	2×2×2	51.22	8	6×6×2	67.61	8	2×2×8	59.15	8	6×6×2	69.01	8
	HIGO	8×8×2	40.49	8	2×2×2	52.44	8	6×6×2	54.93	8	2×2×2	53.52	8	6×6×2	60.56	8

(1) SMIC-subHS dataset yields the highest accuracy of 80.28%. This demonstrates that the worse performance on the SMIC-HS dataset was due to it including more (possibly distracting) samples. By comparing results of SMIC-subHS and SMIC-VIS, we observe that camera recording at higher frame rate does facilitate the ME recognition as claimed in [21]. So using a high speed camera is beneficial for automatically analyzing ME.

(2) ME recognition performance on the SMIC-NIR dataset is the lowest. Compared to faces in RGB videos, faces recorded by a near-infrared camera are smoother and lack skin textures. Our initial motivation for adding an NIR camera in the recording of the SMIC dataset was because NIR camera is less affected by illumination and shadows. When comparing the three kinds of features explored here, LBP works better with NIR data than the other two features. But the overall performance achieved on the NIR dataset is unsatisfactory. Other methods for ME recognition on NIR data need to be explored in the future.

4) Experiment 4: Motion magnification: In the above three sub-experiments we skipped the motion magnification step of our method to untangle the effect of different components. In this sub-experiment, we focus on motion magnification. We demonstrate that using our EVM motion magnification further improves ME recognition performance.

Parameters*:* For this sub-experiment we apply all steps in the diagram of Figure 2. After face alignment, we magnify the ME clips at ten different levels with α = 1, 2, 4, 8, 12, 16, 20, 24 and 28. Then the magnified clips are interpolated with TIM10, and the feature extraction and classification procedure are the same as we did in sub-experiment 2 and 3. The sub-experiment is carried on SMIC and CASMEII databases using LBP, HOG and HIGO features.

TABLE VI
COMPARISON TO STATE-OF-THE-ART IN ME RECOGNITION. OUR RESULTS ARE PRESENTED WITH AND WITHOUT MAGNIFICATION.

	SMIC-HS	SMIC-VIS	SMIC-NIR	CASMEII
LBP	57.93%	70.42%	64.79%	55.87%
LBP+Mag	60.37%	78.87%	**67.61%**	60.73%
HOG	57.93%	71.83%	63.38%	57.49%
HOG+Mag	61.59%	77.46%	64.79%	63.97%
HIGO	65.24%	76.06%	59.15%	57.09%
HIGO+Mag	**68.29%**	**81.69%**	**67.61%**	**67.21%**
HIGO+Mag*	**75.00%***	**83.10%***	**71.83%***	**78.14%***
Li [16]	48.8%	52.1%	38.0%	N/A
Yan [21]	N/A	N/A	N/A	63.41%*
Wang [39]	71.34%*	N/A	N/A	65.45%*
Wang [53]	64.02%*	N/A	N/A	67.21%*
Liong [54]	53.56%	N/A	N/A	N/A
Liong [55]	50.00%	N/A	N/A	66.40%*

* *results achieved using leave-one-sample-out cross validation.*

Results*:* The results are shown in Figure 8. One curve is drawn for each feature on each dataset. The best performance achieved at each magnification level is presented in the curves. The figure shows that, compared to the baseline with no magnification ($\alpha = 1$), ME recognition performance is generally improved when motion magnification is applied. This finding is consistent for all three kinds of features on all four testing datasets. The level of improvement fluctuates with the change of α value. The curves are rainbow-shaped, and the best performance is generally achieved when the motion is magnified in the range $[8, 16]$. Magnification at lower levels might not be enough to reveal the ME motion progress; on the other hand, magnification at higher levels degrades the performance because too many artifacts are induced as shown in Figure 3.

The final results of each feature on each dataset are summarized in Table VI. We observe that: (i) motion magnification substantially increases ME recognition performance; (ii) results with motion magnification support our former findings in Experiment 2 that HIGO feature outperforms HOG and LBP for color camera recorded datasets, while for the NIR dataset, HIGO and LBP work equally well.

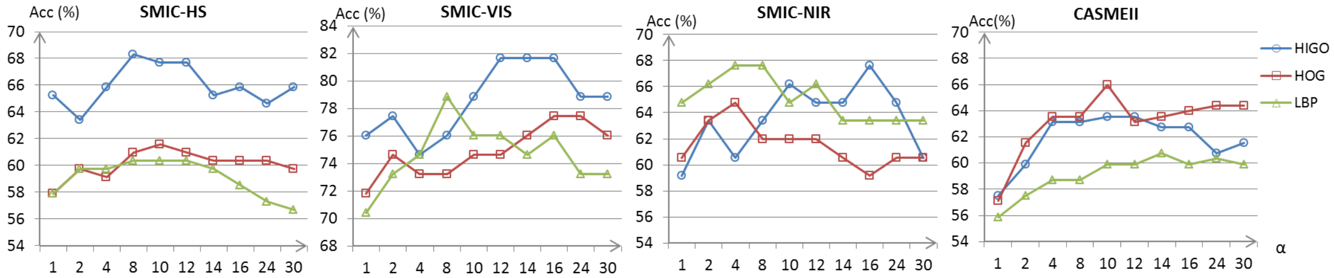

Fig. 8. ME recognition results on SMIC and CASMEII databases at ten different motion magnification levels. The x-axis shows level of the magnification factor α (1 indicates original sequence with no magnification), and the y-axis shows the corresponding recognition accuracy.

Comparison to state of the art*:* The best performance of our method is with motion magnification, TIM10 and HIGO features. By combining these three components, we achieved 81.7% accuracy for ME recognition on the SMIC-VIS dataset, 68.29% on the SMIC-HS dataset, 67.61% on SMIC-NIR dataset, and 67.21% on CASMEII. We list state-of-the-art results for these datasets in the table for comparison. We performed all experiments using the leave-one-subject-out validation protocol, while some of the reference results (for CASMEII, all reference results) were achieved using leave-one-sample-out validation protocol, which is much easier. For direct comparison with those results, we also added one row of results using *HIGO+Mag with leave-one-sample-out protocol*.

Previous work generally use SMIC-HS and CASMEII, while the lower frame rate versions of SMIC (SMIC-VIS and SMIC-NIR) are less explored. For SMIC-VIS and SMIC-NIR datasets, compared to baseline results reported in [16], we achieve an improvement of almost 30%. For SMIC-HS and CASMEII, more reference results are listed, and our results are consistently better regardless of the evaluation protocols. Based on these comparisons, our proposed framework outperforms all previous methods on all the four ME datasets.

C. An automatic ME analysis system (MESR) combining Spotting and Recognition

Previous work has focused purely on either ME spotting or ME recognition, always considering these two tasks separately. However, in reality, these two tasks have to be *combined* to detect MEs in arbitrary long videos. We propose a complete ME analysis system (MESR) which first spots MEs in long videos, and then classifies the emotion category of spotted MEs. The flow of the proposed MESR method is shown in Figure 9. This MESR system works subject-independently (each input video is treated as an 'unknown' test sample, and the classifier is trained on labeled MEs of the other subjects).

Parameters*:* Given a long video clip as the input for ME analysis, the MESR system first finds the locations at which an ME might occur by following the four steps of our ME spotting method described in Section III. LBP is selected for FD analysis. The indexes of spotted frames are fed back to the original videos to excerpt short sequences for ME recognition. In the recognition process of MESR system, we use raw data and perform face alignment (instead of directly using spotted sequences) to register faces to the same model (which makes classification of the MEs easier). According to previous findings, we set the magnification level as $\alpha = 4$, interpolation length as TIM10, use HIGO-XYOT as the feature descriptor, and linear SVM as the classifier for 3-classes classification.

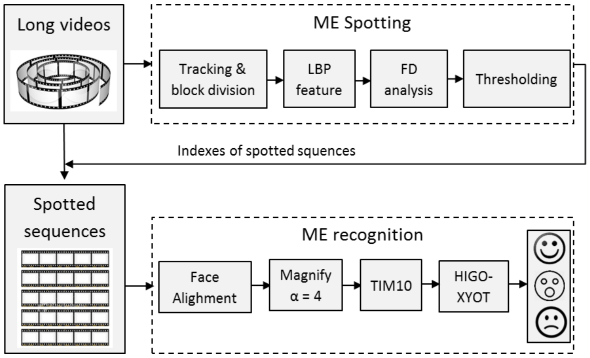

Fig. 9. Framework diagram for our automatic ME analysis system (MESR).

Results*:* We test the MESR system on the SMIC-E-VIS dataset. The output of the spotting process can be varied by adjusting the threshold value, and a higher true spot rate is consistently associated with a higher false spot rate. In our experiments, we select the result at TPR $=$ 74.86% (corresponding FPR $=$ 22.98%, $\tau = 0.15$) and all spotted sequences are all fed into the ME recognition component. For the correctly spotted ME sequences, the ME recognition component achieves 56.67% accuracy for emotion recognition. The recognition accuracy drops comparing to the results in Experiment 4, which is expected as in the previous experiments all the MEs are hand-labeled, and thus the onset and offset time points are accurate. In contrast, in the MESR system we use automatically spotted sequences which do not always locate the MEs precisely, and include some non-ME frames. The overall performance of the MESR system is a multiplication of the two, *i.e.* $Acc_{MESR} = 74.86\% \times 56.67\% = 42.42\%$.

We next compare the performance of our automatic ME analysis system to the performance of human subjects, and show that our automatic method performs comparatively to humans at this difficult task.

D. Human test results

ME spotting and recognition is very challenging for ordinary people, as they are not 'tuned in' to these very brief (and often low-intensity) facial expressions.

In this section we experiment on human subjects on two tasks, and compare the results to our automatic method. We show that our method achieves comparable performance to humans at these difficult tasks (and outperforms humans at ME recognition).

1) Human test of ME recognition: The first experiment concerns ME recognition. 15 subjects (average age 28.5 years) were enrolled in the experiments (ten males, five female). All subjects signed consents which allow their data to be used for academic research and publications. 71 ME samples from the SMIC-VIS dataset were used as the experiment videos. Before test started, the definition of an ME was explained, and three sample videos, each containing one class of ME, were shown to the subject. During the experiment, ME clips were shown on a computer screen one by one, and after each clip subjects selected which emotion category (positive, negative or surprise) the ME belongs to by clicking the corresponding button. The mean accuracy of the 15 human subjects was 72.11% (standard deviation, STD=7.22%). In contrast, the best accuracy of our ME recognition method is 81.69%.

This shows that our proposed ME recognition method outperforms human subjects in the ME recognition task.

2) Human test of ME spotting & recognition: In the second experiment, we asked subjects to first spot whether there are any MEs in a video, and then indicate their emotion category (if there are any). Another 15 subjects (average age is 25.8 years) were enrolled (12 male, three female). All subjects signed consents which allow using their data for academic research and publications. 71 long clips from the SMIC-E-VIS dataset, plus five neutral clips which do not include any MEs, were used as the experiment videos. Before test started, the definition of an ME was explained, and three example videos, each containing one class of ME, were shown to the subject. During the test, clips were shown on a computer screen one by one. The subject first reported how many MEs (s)he spotted from the clip by clicking a button (either '0', '1', '2' or 'more than 2'). If a button other than '0' was clicked, the subject was asked to further select which emotion category (positive, negative and surprise) the spotted ME(s) show. The subjects were informed that multiple MEs occurring near each other present the same emotion, so only one emotion category could be selected (even if multiple MEs were spotted) for one clip.

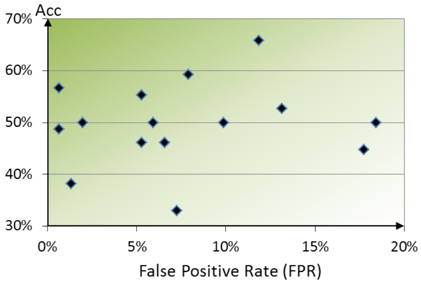

Fig. 10. Performance of human subjects on ME spotting & recognition.

We calculate an accuracy and a FPR for each subject, and then compute means and STDs for the group. For each subject, we count the number of MEs that were correctly spotted and recognized. The accuracy is a percentage of that number divided by 71. FPR is calculated as a percentage of false spots divided by the maximum possible false spots (152). Results are shown as a scatter plot in Figure 10. Each point represents one subject, and darker area indicates better performance. As shown, the ME recognition ability varies across individuals. The mean accuracy of the human subjects is 49.74% (STD=8.04%) and the mean FPR is 7.31% (STD=5.45%).

We note that due to practical reasons, there are in fact two mismatches in the experimental protocols which made the task *easier* for humans to perform compared to our automatic method. The first is that subjects cannot respond quickly enough to locate each ME, so we only ask them for the *number of spots*. The second is that we allowed subjects to replay videos for multiple times because they felt the task was too difficult. These two 'mismatches' in experimental protocols made the task significantly easier for human subjects compared to our automatic method. However, the human subjects could still only recognize half of the MEs.

In previous subsection our automatic MESR system achieved an accuracy of 42.42%, which is a comparable performance (within one STD) to the mean accuracy of the human subjects. In comparison to the humans, our method's main shortage is the false spot rate (22.98%). This could be further reduced *e.g.* by excluding non-ME fast movements such as eye blinks. As the first fully automatic system for ME analysis, our method is a very promising start. More explorations will be done in the future.

VI. Conclusions

In this work we focused on the study of *spontaneous* MEs, which are much more difficult to analyze than posed expressions explored in previous work. We proposed novel methods for both ME *spotting* and ME *recognition*.

For ME spotting, we are the first to propose a method able to spot MEs from spontaneous long videos. The method is based on feature difference (FD) comparison. Two kinds of features (LBP and HOOF) are employed, and LBP is shown to outperform HOOF on two databases.

For ME recognition, we proposed a new framework where motion magnification is employed to counter the low intensity of MEs. We validated the new framework on SMIC and CASMEII databases, and showed that our method outperforms state of the art on both databases. We also drew many interesting conclusions about the respective benefits of our method's components.

Finally, and importantly, we proposed the first automatic ME analysis system in the literature (which first spots and then recognizes ME). Our method is the first system that has ever been tested on a hard spontaneous ME dataset, containing natural MEs. It outperforms humans at ME recognition by a significant margin, and performs comparably to humans at the combined ME spotting & recognition task. This method has many high-impact applications, particularly in lie detection, law enforcement and psychotherapy.

In future we plan to: (i) explore in more detail how MEs could be separated from other non-emotional rapid movements such as eye blinks; (ii) examine alternative features (*e.g.* deep learning methods) for ME recognition; and (iii) examine ME spotting and recognition methods on more challenging data recorded in realistic situations.

References

[1] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[2] 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," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[3] D. Matsumoto and H. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation and Emotion*, pp. 1–11, 2011.

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

[5] E. Haggard and K. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," *Methods of research in psychotherapy. New York: Appleton-Century-Crofts*, pp. 154–165, 1966.

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

[7] M. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies." *Journal of Personality and Social Psychology*, vol. 72, no. 6, p. 1429, 1997.

[8] P. Ekman, "Microexpression training tool (METT)," *San Francisco: University of California*, 2002.

[9] ——, "Lie catching and microexpressions," *The Philosophy of Deception*, pp. 118–133, 2009.

[10] ——, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Macmillan, 2007.

[11] P. Ekman, M. O'Sullivan, and M. G. Frank, "A few can catch a liar," *Psychological Science*, vol. 10, no. 3, pp. 263–266, 1999.

[12] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *PAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[13] R. Picard, *Affective Computing*. MIT Press, 1997.

[14] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *ICPR*. IEEE, 2014, pp. 1722–1727.

[15] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *ICCV*. IEEE, 2011, pp. 1449–1456.

[16] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *FG*. IEEE, 2013, pp. 1–6.

[17] M. Kamachi, M. Lyons, and J. Gyoba. (1998) The japanese female facial expression (JAFFE) database. [Online]. Available: http://www.kasrl.org/jaffe.html

[18] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *FG*. IEEE, 2000, pp. 46–53.

[19] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*. IEEE, 2005, pp. 5–pp.

[20] 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 *FG*. IEEE, 2013, pp. 1–7.

[21] 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.

[22] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," in *WACV*. IEEE, 2009, pp. 1–6.

[23] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *FG*. IEEE, 2011, pp. 51–56.

[24] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *ICDP*. IET, 2009, pp. 1–6.

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

[26] J. Coan and J. Allen, *Handbook of Emotion Elicitation and Assessment*. Oxford University Press, USA, 2007.

[27] J. Gross and R. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[28] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang, "Spontaneous emotional facial expression detection." *Journal of Multimedia*, vol. 1, no. 5, pp. 1–8, 2006.

[29] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," *Universal Access in the Information Society*, vol. 11, no. 4, pp. 409–419, 2012.

[30] S. Liwicki, S. Zafeiriou, and M. Pantic, "Incremental slow feature analysis with indefinite kernel for online temporal video segmentation," in *ACCV*. Springer, 2013, pp. 162–176.

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

[32] Q. Wu, X. Shen, and X. Fu, "The machine knows what you are hiding: an automatic micro-expression recognition system," *Affective Computing and Intelligent Interaction*, pp. 152–162, 2011.

[33] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 59–69, 2009.

[34] J. A. Ruiz-Hernandez and M. Pietikäinen, "Encoding local binary patterns using the re-parametrization of the second order gaussian jet," in *FG*. IEEE, 2013, pp. 1–6.

[35] A. K. Davison, M. H. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: An investigation on spatio-temporal descriptors," in *ECCV Workshop*, vol. 8926, 2014.

[36] S. Yao, N. He, H. Zhang, and O. Yoshie, "Micro-expression recognition by feature points tracking," in *COMM*. IEEE, 2014, pp. 1–4.

[37] Y. Song, L.-P. Morency, and R. Davis, "Learning a sparse codebook of facial and body microexpressions for emotion recognition," in *ICMI*. ACM, 2013, pp. 237–244.

[38] 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 *ICPR*. IEEE, 2014, pp. 4678–4683.

[39] S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, and C.-G. Zhou, "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *ECCV Workshop*, vol. 8925, 2014.

[40] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[41] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[42] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *PAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.

[43] T. Cootes, C. Taylor, D. Cooper, J. Graham *et al.*, "Active shape models-their training and application," *CVIU*, vol. 61, no. 1, pp. 38–59, 1995.

[44] A. Goshtasby, "Image registration by local approximation methods," *IVC*, vol. 6, no. 4, pp. 255–261, 1988.

[45] H.-Y. Wu, M. Rubinstein, E. Shih, J. V. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world." *ACM Trans. Graph.*, vol. 31, no. 4, p. 65, 2012.

[46] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *CVPR*. IEEE, 2011, pp. 137–144.

[47] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[48] Z. Li, J.-i. Imai, and M. Kaneko, "Facial-component-based bag of words and phog descriptor for facial expression recognition," in *SMC*. IEEE, 2009, pp. 1353–1358.

[49] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1. IEEE, 2005, pp. 886–893.

[50] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *PR Letters*, vol. 32, no. 12, pp. 1598–1603, 2011.

[51] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *TIST*, vol. 2, no. 3, p. 27, 2011.

[52] G. Zhao, X. Huang, M. Taini, S. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *IVC*, 2011.

[53] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *ACCV*, 2014.

[54] S.-T. Liong, R.-W. Phan, J. See, Y.-H. Oh, and K. Wong, "Optical strain based recognition of subtle emotions," in *ISPACS*, Dec 2014, pp. 180–184.

[55] S.-T. Liong, J. See, R. C.-W. Phan, A. Le Ngo, Y.-H. Oh, and K. Wong, "Subtle expression recognition using optical strain weighted features," in *ACCV Workshop*, 2014.