

Received August 28, 2019, accepted September 10, 2019, date of publication September 19, 2019, date of current version December 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942358

Extended Local Binary Patterns for Efficient and Robust Spontaneous Facial Micro-Expression Recognition

CHENGYU GUO¹, JINGYUN LIANG¹, GENG ZHAN³, ZHONG LIU¹,
MATTI PIETIKÄINEN², (Fellow, IEEE), AND LI LIU^{1, 2}, (Senior Member, IEEE)

¹College of Systems Engineering, National University of Defense Technology, Changsha 410073, China

²Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland

³School of Electrical and Information Engineering, The University of Sydney, Camperdown, NSW 2006, Australia

Corresponding author: Li Liu (li.liu@oulu.fi)

This work was supported by the National Natural Science Foundation of China under Grant 61872379.

ABSTRACT Facial Micro-Expressions (MEs) are spontaneous, involuntary facial movements when a person experiences an emotion but deliberately or unconsciously attempts to conceal his or her genuine emotions. Recently, ME recognition has attracted increasing attention due to its potential applications such as clinical diagnosis, business negotiation, interrogations, and security. However, it is expensive to build large scale ME datasets, mainly due to the difficulty of inducing spontaneous MEs. This limits the application of deep learning techniques which require lots of training data. In this paper, we propose a simple, efficient yet robust descriptor called Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP) for ME recognition. ELBPTOP consists of three complementary binary descriptors: LBPTOP and two novel ones Radial Difference LBPTOP (RDLBPTOP) and Angular Difference LBPTOP (ADLBPTOP), which explore the local second order information along the radial and angular directions contained in ME video sequences. ELBPTOP is a novel ME descriptor inspired by unique and subtle facial movements. It is computationally efficient and only marginally increases the cost of computing LBPTOP, yet is extremely effective for ME recognition. In addition, by firstly introducing Whitened Principal Component Analysis (WPCA) to ME recognition, we can further obtain more compact and discriminative feature representations, then achieve significantly computational savings. Extensive experimental evaluation on three popular spontaneous ME datasets SMIC, CASME II and SAMM show that our proposed ELBPTOP approach significantly outperforms the previous state-of-the-art on all three single evaluated datasets and achieves promising results on cross-database recognition. Our code will be made available.

INDEX TERMS Micro-expression recognition, local binary pattern, feature extraction.

I. INTRODUCTION

Facial Micro-Expressions (MEs) are spontaneous, involuntary facial movements when a person experiences an emotion but deliberately or unconsciously attempts to conceal his or her genuine emotions [1]–[3]. MEs are more likely to occur in high-risk environments because there are more risks to show true emotions [4]. Recently, automatic facial ME analysis has attracted increasing attention of affective computing researchers and psychologists because of its potential applications such as clinical diagnosis, business negotiation, interrogations, and security [5], [6]. The study

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

of facial MEs is a well established field in psychology, however, it is a relatively new area from the computer vision perspective with many unsolved and challenging problems [7], [8]. There are three main challenges in automatic ME analysis.

A. MEs HAVE A VERY SHORT DURATION, LOCAL AND SUBTLE FACIAL MOVEMENTS

Compared to ordinary facial expressions, the duration of a ME is usually very short, typically being no more than 500 ms [9]. Besides short duration, MEs also have other unique characteristics such as local and subtle facial movements [10]. Because of these unique characteristics, it is very difficult for human beings to recognize MEs.

B. LACK OF LARGE SCALE SPONTANEOUS ME DATASETS

Datasets have played a key role in visual recognition problems, especially in the era of deep learning which requires large scale datasets for training [11]. ME analysis is not an exception. However, another challenging issue faced by automatic facial ME analysis is the lack of benchmark datasets (especially large scale ME datasets) due to the difficulties in inducing spontaneous MEs and labeling them [1], [7]. To the best of our knowledge [12], there are eight ME datasets: USF-HD [13], Polikovsky's database [14], YorkDDT [15], SMIC [1], CASME [16], CASME II [2], CAS(ME)² [17], and SAMM [3]. The first two are posed and not publicly-available. Posed MEs are different from naturally occurring spontaneous MEs significantly. Thus recent works focus on spontaneous ME datasets. All of the datasets are small. Besides, the emotion categories of the collected samples in these datasets are unevenly distributed, because some emotions are easier to elicit hence they have more samples.

C. LACK OF EFFICIENT AND DISCRIMINATIVE FEATURE REPRESENTATIONS

Above challenges make ME analysis much harder and more demanding than ordinary facial tasks. Therefore, the extraction of efficient and discriminative feature representations becomes especially important for automatic ME analysis.

In automatic ME analysis, there are mainly two tasks: ME spotting and ME recognition. The former refers to the problem of automatically and accurately locating the temporal interval of a micro-movement in a video sequence, where extended versions of SMIC [1], CAS(ME)² [17], and SAMM [3] are widely used; while the latter is to classify the ME in the video into one of the predefined emotion categories (such as Happiness, Sadness, Surprise, Disgust, etc), where SMIC [1], CASME II [2], and SAMM [3] are widely adopted. ME recognition is the focus of this paper.

Like ordinary facial expression recognition, ME recognition consists of three steps: preprocessing, feature representation and classification [7]. As we discussed previously, the development of powerful feature representations plays a very important role in ME recognition, and thus has been one main focus of research [18]. Representative feature representation approaches for ME recognition are mainly based on Local Binary Patterns (LBP) [18], [19], Local Phase Quantization (LPQ) [20], Histogram of Oriented Gradients (HOG) [21] and Optical Flow (OF) [22].

Despite these efforts, there is still significant room for improvement towards achieving good performance. The small scale of existing ME datasets and the imbalanced distribution of samples are the primary obstacles to applying existing data hungry deep convolutional neural networks which have brought significant breakthroughs in various visual recognition problems in computer vision due to their ability to learn powerful feature representations directly from raw data. Therefore, state-of-the-art methods for ME recognition are still dominated by traditional handcrafted features like Local

Binary Patterns on Three Orthogonal Planes (LBPTOP) [23], 3D Gradient Oriented Histogram (HOG 3D) [14] and Histograms of Oriented Optical Flow (HOOF) [24].

Due to its prominent advantages such as theoretical simplicity, computational efficiency, and robustness to monotonic grey scale changes, the texture descriptor LBP [25] has emerged as one of the most prominent features for face recognition [26]. Its 3D extension LBPTOP [23] is widely used for facial expression and ME recognition [27]. Many variants of LBP have been proposed to improve robustness, and discriminative power, as summarized in recent surveys [28], [29]. However, most LBP variants [30], [31] have not been explored for ME recognition. In other words, in contrast to LBP-based face recognition, LBPTOP type ME recognition is surprisingly underexplored. Moreover, current state-of-the-art ME features like LBPTOP and its variants LBPSIP [32], LBPMOP [33], STLBP-IP [34], and STRBP [35] suffer from some drawbacks, such as limited representation power of using only one type of binary feature, limited robustness, and increased computational complexity.

In this paper, in order to build more discriminative features that can inherit the advantages of LBP type features without suffering the shortcoming of using filters as complementary features [25] (i.e., the expensive computation cost), we propose a novel binary feature descriptor named Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP) for ME recognition. ELBPTOP is a descriptor that, we argue, nicely balances the three concerns: high distinctiveness, good robustness and low computational cost. In addition, LBPTOP can be considered as a special case of the proposed ELBPTOP descriptor. Our contributions of this paper are summarized as follows.

- Inspired by the unique texture information of human faces and the subtle intensity variations of local subtle facial movements, the novel ELBPTOP encodes not only the first order information, i.e. the pixel difference information between a central pixel and its neighbours (called Center Pixel Difference Vector, CPDV), but also encodes the second order discriminative information in two directions: the radial direction (Radial Pixel Difference Vector, RPDV) and the angular direction (Angular Pixel Difference Vector, APDV). They are named ADLBPTOP and RDLBPTOP respectively. The proposed ELBPTOP is more effective to capture local, subtle intensity changes and thus delivers stronger discriminative power.
- To achieve our goal of being computationally efficient while preserving distinctiveness, we then apply Whitened Principal Component Analysis (WPCA) to get a more compact, robust, and discriminative global descriptor. We are aware of the fact that WPCA has proven to be effective in face recognition. However, we argue that we are the first to apply WPCA to the problem of ME recognition, which has its own unique challenges compared to the extensively studied face recognition problem.

- We provide extensive experimental evaluation on three popular spontaneous ME datasets CASME II, SMIC, and SAMM to test the effectiveness of the proposed approach, and find that our proposed ELBPTOP approach significantly outperforms previous state-of-the-art on all three evaluated datasets. Our proposed ELBPTOP achieves 73.94% on CASMEII, which is 6.6% higher than state-of-the-art on this dataset. More impressively, ELBPTOP increases recognition accuracy from 44.7% to 63.44% on the SAMM dataset.

Although our method is simple and handcrafted, the very strong quality results obtained on three popular ME datasets in addition with the low computational complexity prove the efficiency of our approach for ME recognition.

The remainder of the paper is organized as follows. Section II reviews related work in micro-expression recognition and gives a brief outline of LBP and LBPTOP. The main model and more details are represented in Section III, including the proposed ADLBPTOP and the RDLBPTOP descriptors and our ME recognition scheme. Experimental results are presented in Section IV, leading to conclusions in Section V.

II. RELATED WORKS

Feature representation approaches of ME recognition can be divided into two distinct categories: geometric-based and appearance-based [36] methods. Specifically, geometric-based features describe the face geometry such as the shapes and locations of facial landmarks, so they need precise landmarking and alignment procedures. By contrast, appearance-based features describe the intensity and textural information such as wrinkles and shading changes, and they are more robust to illumination changes and alignment error. Thus, appearance-based feature representation methods, including LBPTOP [23], HOG 3D [14], HOOF [24] and deep learning, have been more popular in ME recognition [7].

LBPTOP variants: Since the pioneering work by Pfister *et al.* [6], LBPTOP has emerged as the most popular approach for spontaneous ME analysis, and quite a few variants have been proposed. LBP Six Interception Points (LBPSIP) [32] is based on three intersecting lines crossing over the center point. LBP Mean Orthogonal Planes (LBPMOP) [33] first computes an average plane for three orthogonal planes, and then computes the LBP on the three orthogonal average planes. By reducing redundant information, LBPSIP and LBPMOP achieved better performance. Reference [37] explores two effective binary face descriptors: Hot Wheel Patterns [37] and Dual-Cross Patterns [38] and makes use of abundant labelled micro-expressions. Besides computing the sign of pixel differences, Spatio-Temporal Completed Local Quantized Patterns (STCLQP) [39] also exploits the complementary components of magnitudes and orientations. Decorrelated Local Spatiotemporal Directional Features (DLSTD) [40] uses Robust Principal Component Analysis (RPCA) [41] to extract subtle emotion information and division of 16 Regions of

Interest (ROIs) to utilize the Action Unit (AU) information. Spatio-Temporal Local Radon Binary Pattern (STRBP) [35] uses Radon Transform to obtain robust shape features, while Spatiotemporal Local Binary Pattern with Integral Projection (STLBP-IP) [34] turns to integral projections to preserve shape attributes.

HOOF variants: Histograms of Oriented Optical Flow (HOOF) [24] is one of the baseline methods that makes use of optical flow in ME recognition. Facial Dynamics Map (FDM) [42] describes local facial dynamics by extracting principal OF direction of each cuboid. Similarly, [43] designs Main Directional Mean Optical Flow (MDMO) features that utilize the AU information from partitioning facial area into 36 ROIs. Different from these methods, Consistent Optical Flow Maps [44] estimates consistent OF to characterize facial movements, which are calculated from 25 ROIs and the OF of each ROI could be in multiple directions. Recently, Bi-Weighted Oriented Optical Flow (BI-WOOF) [45] makes use of only the apex frame and the onset frame. The majority of OF-based methods need to partition the face area precisely to make use of AU information. This improves the performance but increases the complexity of preprocessing. Reference [46] calculates the LBPTOP and HOOF fusion features for automatic Necessary Morphological Patches (NMPs) extraction which combines the AU-based method and the feature selection method.

HOG 3D variants: HOG 3D [14] is firstly used to recognize posed MEs and then as a baseline on spontaneous MEs. Its variants, the Histogram of Image Gradient Orientation (HIGO) [47] ignores the magnitude weighting, hence can suppress the influence of illumination. This makes HIGO become one of the most accurate descriptors at present. However, it is worth noting that HOG is an edge-based gradient descriptor. It is sensitive to noise when not being filtered, and the use of low pass filters could lead to the loss of subtle motion change information in ME recognition. Besides, the computation process is time-consuming and cumbersome, resulting in slow speed.

Deep learning methods: Reference [48] adopts a shallow network with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). Other neural networks are explored in Dual Temporal Scale Convolutional Neural Network (DTSCNN) [49], 3D Flow Convolutional Neural Network (3DFCNN) [50], 3D Spatiotemporal Convolutional Neural Networks (3DCNN) [51] and Micro-Expression Recognition algorithm using Recurrent CNNs (MER-RCNN) [52]. These methods achieve some improvements in ME recognition, but they are still significantly below state-of-the-art handcrafted features, mainly due to lack of large scale ME data.

Cross-database ME recognition (CDMER) is a new topic in micro-expression analysis. CDMER considers the large difference of feature distributions existing between the training and testing ME samples in a real scenario to exploit more generalizing approach on different datasets collected by different cameras or under different environments.

Besides, a combination of different datasets increased the number of subjects and samples, which is beneficial to the data-driven methods and deep-learning methods. The fundamental works of 1st Micro-Expression Grand Challenge (MEGC2018) [53], 2nd Micro-Expression Grand Challenge (MEGC2019) [54], and [55] facilitate the development of CDMER. Macro to Micro Transfer Learning [56] utilizes transfer learning to implement CNN from big macro-expression datasets to small ME datasets, ranking top in MEGC2018. Besides transfer learning, [57] adopt two domain adaptation techniques including adversarial training and expression magnification obtain the best results on the full composite database in MEGC2019. Other methods [58]–[63] also show promising performance in cross-database challenges.

A. LBP AND LBPTOP

LBP was firstly proposed in [19], and a completed version was developed in [18]. Later on, it was introduced to face recognition in [26] and its 3D extended version LBPTOP was proposed in [23] with application to facial expression analysis.

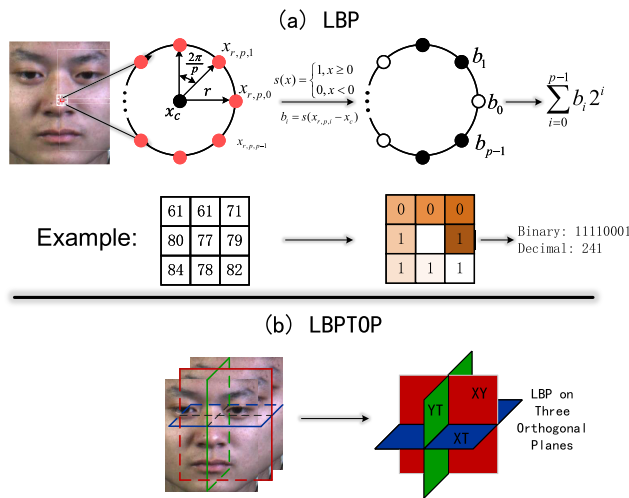


FIGURE 1. (a) LBP pattern: The sample neighborhood is the center pixel x_c with p equally spaced pixels on a circle of radius r . Then the binary code is calculated by comparing the differences between the center pixel and its neighbors. An example is in the figure. (b) The process of LBPTOP.

LBP characterizes the special structure of p pixels, that are evenly distributed in angle on a circle of radius r centered at pixel x_c . In specific, as shown in Figure 1(a), for a central pixel x_c and its p neighboring equally spaced pixels $\{x_{r,p,n}\}_{n=0}^{p-1}$ on the circle of radius r , the LBP pattern is computed via:

$$LBP_{r,p}(x_c) = \sum_{n=0}^{p-1} s(x_{r,p,n} - x_c) 2^n, \quad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0, \end{cases} \quad (1)$$

where $s(\cdot)$ is the sign function. The gray values of points that do not fall exactly in the center of pixels are estimated by interpolation. The decimal value of LBP pattern is given

by the binary sequence of the circular neighborhood, such as $241 = (11110001)_2$ in Figure 1(a). LBP is gray scale invariant and is able to encode important local patterns like lines, edges, and blobs because it measures the differences between the center pixel and its neighbors.

Given an $N \times M$ texture image, a LBP pattern $LBP_{r,p}(x_c)$ can be computed at each pixel c , such that a textured image can be characterized by the distribution of LBP values, representing the whole image by a LBP histogram vector. By altering r and p , one can compute LBP features for any quantization of the angular space and for any spatial resolution.

LBPTOP [23] is the 3D extension of LBP by extracting LBP patterns separately from three orthogonal planes: the spatial plane (XY) similar to the regular LBP, the vertical spatiotemporal plane (YT) and the horizontal spatiotemporal plane (XT), as illustrated in Figure 1(b).

Clearly, LBPTOP encodes temporal changes, and component information. A video can be represented by concatenating LBP on TOP. Despite a little more complex than the static LBP, LBPTOP can achieve real time processing speed depending on the size of the local sampling neighborhood. The dimensionality of LBPTOP is higher than LBP. Since LBPTOP, which extracts features from TOP, becomes popular when extending 2D spatial appearance descriptors to the spatiotemporal domain.

III. PROPOSED APPROACH

In this section, we first introduce the proposed novel binary descriptor ELBPTOP and then present how to use it for ME recognition.

A. ELBPTOP

LBPTOP has emerged as one of the dominant descriptors for ME recognition. Despite this fact, it has several limitations.

- Currently, LBPTOP [23] usually only exploit the uniform patterns for ME representation. This results in information loss since the proportion of uniform patterns may be too small to capture the variations.
- It encodes the difference between each pixel and its neighboring pixels only. It is common to combine complementary features like Gabor filters to improve discriminative power. However, this brings a significant computational burden.
- A large sampling size is helpful since it encodes more local information and provides better representation power. However, increasing the number of sampling points of LBPTOP increases its feature dimensionality significantly.

The above analysis leads us to propose novel binary type descriptors, which should not be competitive with LBPTOP, but complement and extend a set of binary feature candidates.

We propose to explore the second order discriminative information in two directions of a local patch: the radial differences (RDLBPTOP) and the angular differences (ADLBPTOP), as complement to the differences between a pixel and its neighbors (LBPTOP). The proposed

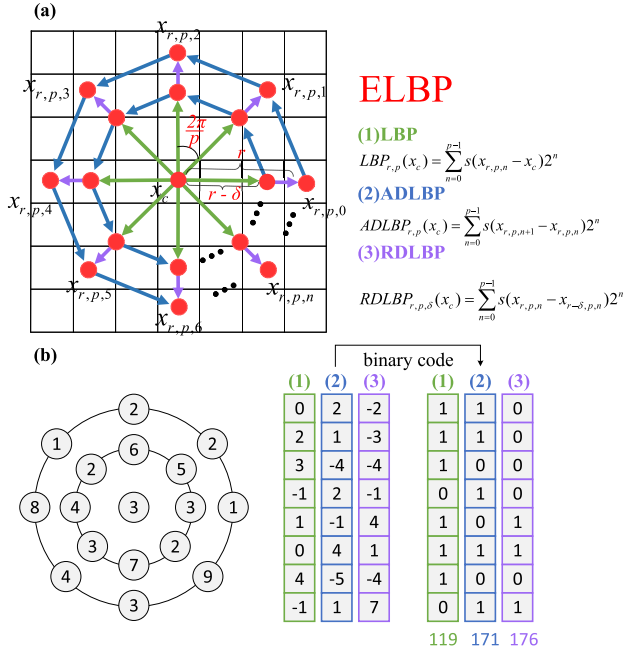


FIGURE 2. (a) A local circularly symmetric neighbor sampling of ELBP. Two circles of $p = 8$ neighbor points are around the central pixel x_c . The radius of the inner circle is $r - \delta$, and the radius of the outer circle is r . (b) An illustration of the process to calculate ELBP pattern.

RDLBPTOP and ADLBPTOP preserve the advantages of LBP, such as a computational efficiency and gray scale invariance.

1) RADIAL DIFFERENCE LOCAL BINARY PATTERN (RDLBP)

As illustrated in Section II, LBP is computed by thresholding the neighboring pixel values on a ring against its center pixel value. It only encodes the relationship between the neighboring pixels on the same ring (i.e. a single scale) and the center one, failing to capture the second order information of neighboring pixels between different rings (different scales). For every pixel in the image, we look at two rings of radii r and $r - \delta$ centered on the pixel x_c and p pixels distributed evenly on each ring, as shown in Figure 2. To produce the RDLBP codes, we first compute the radial differences $\{x_{r,p,n} - x_{r-\delta,p,n}\}_n$ between pixels on the two rings and then threshold them against 0. The formal definition of the RDLBP code is as follows:

$$RDLBP_{r,p,\delta}(x_c) = \sum_{n=0}^{p-1} s(x_{r,p,n} - x_{r-\delta,p,n}) 2^n, \quad (2)$$

where r and $r - \delta$ denote the outer ring and the inner ring respectively. As can be seen from Figure 3, the LBP values of two different pixels can be same in some cases, but for RDLBP, they are totally different. This is because RDLBP encodes radial pixel difference information.

2) ANGULAR DIFFERENCE LOCAL BINARY PATTERN (ADLBP)

LBP also fails to encode the second order information between pixels on the ring. Therefore, ADLBP is composed

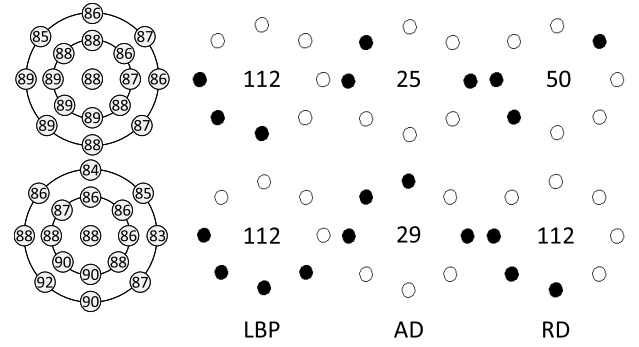


FIGURE 3. The two given patterns in the left would be considered equivalent by LBP. However, the patterns are, in some ways, quite different from one to others. Fortunately, this underlying change properties can be revealed via angular and radial differences.

of neighboring pixel comparisons in angular (like clockwise) direction for all pixels except the center pixel.

Formally, it can be calculated as follows:

$$ADLBP_{r,p}(x_c) = \sum_{n=0}^{p-1} s(x_{r,p,n+1} - x_{r,p,n}) 2^n. \quad (3)$$

Similarly, Figure 3 shows that ADLBP encodes angular difference information, which is different from the original LBP descriptor. It is very compact and provides useful information. We can see that both RDLBP and ADLBP are gray scale invariant and computationally efficient. They can also benefit from rotation invariant extension, uniform extension and 3D extension of LBP.

3) EXTENDED LBP (ELBP)

We use ELBP to represent the combination of all three binary descriptors: LBP, RDLBP, and ADLBP. The three operators LBP, RDLBP and ADLBP can be combined in two ways, jointly or independently. Because the joint way (3D joint histogram) leads to huge dimension, we use the latter way.

For ME recognition, as shown in Figure 1(b), we extend ELBP to ELBPTOP. Most LBPTOP based ME descriptors use uniform LBP patterns and group the nonuniform patterns into one bin. However, this leads to lots of information loss because uniform LBPs may not be the majority of LBPs, as illustrated in Figure 4. This is more obvious in the case of ADLBP, where the nonuniform patterns are the dominant patterns. Therefore, in this paper, we use all 2^p patterns, rather than uniform patterns only.

B. ELBPTOP FOR ME RECOGNITION

In this section, the ME representation is addressed using our proposed ELBPTOP approach to explicitly handle the encountered challenges.

To enhance the discrimination power, we propose to fuse the information extracted by three binary descriptors LBPTOP, RDLBPTOP and ADLBPTOP. The ME feature representation algorithm is illustrated in Figure 6(a). For each binary descriptor LBPTOP (or RDLBPTOP or ADLBPTOP),

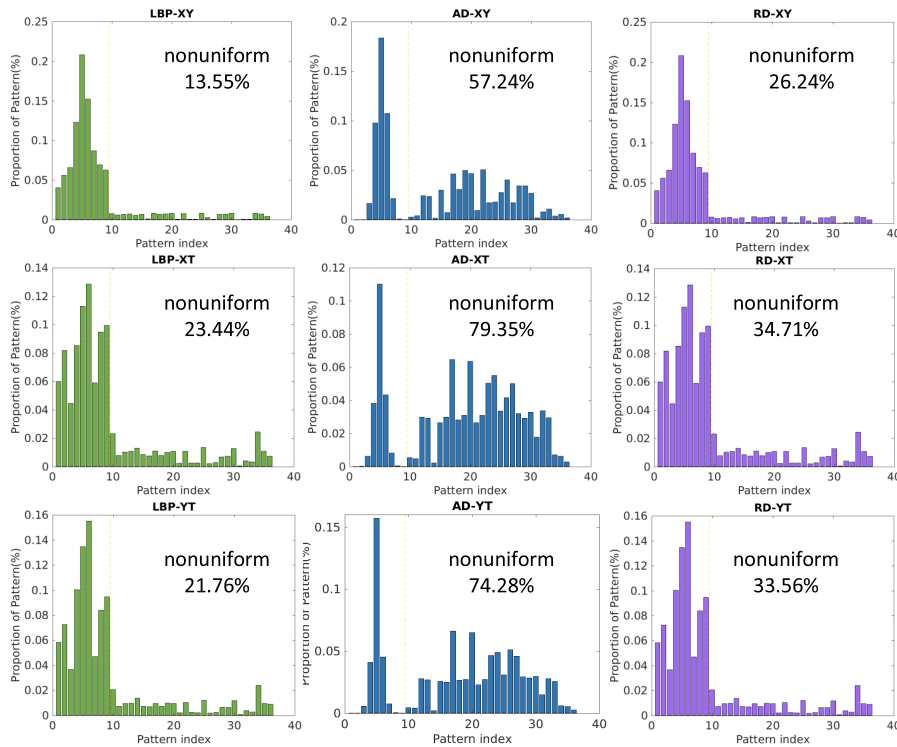


FIGURE 4. Proportions of the uniform¹ LBPs for the ELBP descriptors (LBP, ADLBP, and RDLBP) on three planes (XY, XT, YT) from the CASME II dataset. The first 9 bins of each histogram are the uniform patterns, and others are the nonuniform patterns. We could observe that the uniform patterns may not account for the major proportion of overall patterns. This is especially obvious in the case of ADLBP.

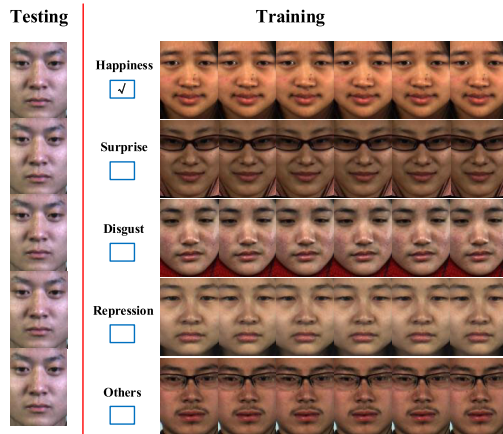


FIGURE 5. Illustration of the ME classification problem. Samples frames are from CASME II [2].

the ME video sequences are represented as the concatenated spatiotemporal histograms of the binary codes. In specific, a video sequence is divided into $m \times q \times l$ blocks, then for every single binary descriptor, the dimension of the histogram

¹For clear illustration, we transform the “full” pattern into “rotation invariant (ri)” pattern [18]. Accordingly, the “uniform (u2)” pattern is transformed into “rotation invariant uniform (riu2)” pattern. Meanwhile, the proportion of the “u2” pattern in the “full” pattern is equal to the proportion of the “riu2” pattern in the “ri” pattern. The transformation has no effect on our conclusion.

is $m \times q \times l \times 2^p$. For instance, if we divide the video sequence into $8 \times 8 \times 2$ blocks, and we choose $p = 8$, the histogram dimension of a single descriptor would be $8 \times 8 \times 2 \times 2^p = 32768$.

An efficient and effective feature representation scheme is equally important for ME recognition as an efficient and good local descriptor. For each binary code (LBPTOP, RDLBPTOP or ADLBPTOP), the dimension of the feature representation for each ME video sequence is $m \times q \times l \times 2^p$, which is in fact very high. This would cause a computational burden for later classification stage. Therefore, to improve efficiency and preserve distinctiveness, Whiten Principal Component Analysis (WPCA) [64], [65] is firstly introduced for dimensionality reduction before feature fusion.

The idea behind WPCA is that discriminative information is equally distributed along all principle components. The whitening transformation is applied to normalize the contribution of each principal component. Specifically, given a feature representation \mathbf{h} , standard PCA is used to get the projected feature $\mathbf{h}_{pca} = \mathbf{W}_{pca}\mathbf{h}$ at first, where \mathbf{W}_{pca} is the projection matrix of with k orthonormal columns. Then, the sorted eigenvectors corresponding to the descending sorted first k principal components are transformed to normalized eigenvectors \mathbf{h}_{wpca} whose variances equal to 1.

In summary, figure 6(a) illustrates the overview of the proposed feature extraction framework. At first, the video sequences are spatially divided into multiple nonoverlapping subblocks, from each of which three sub-region histograms

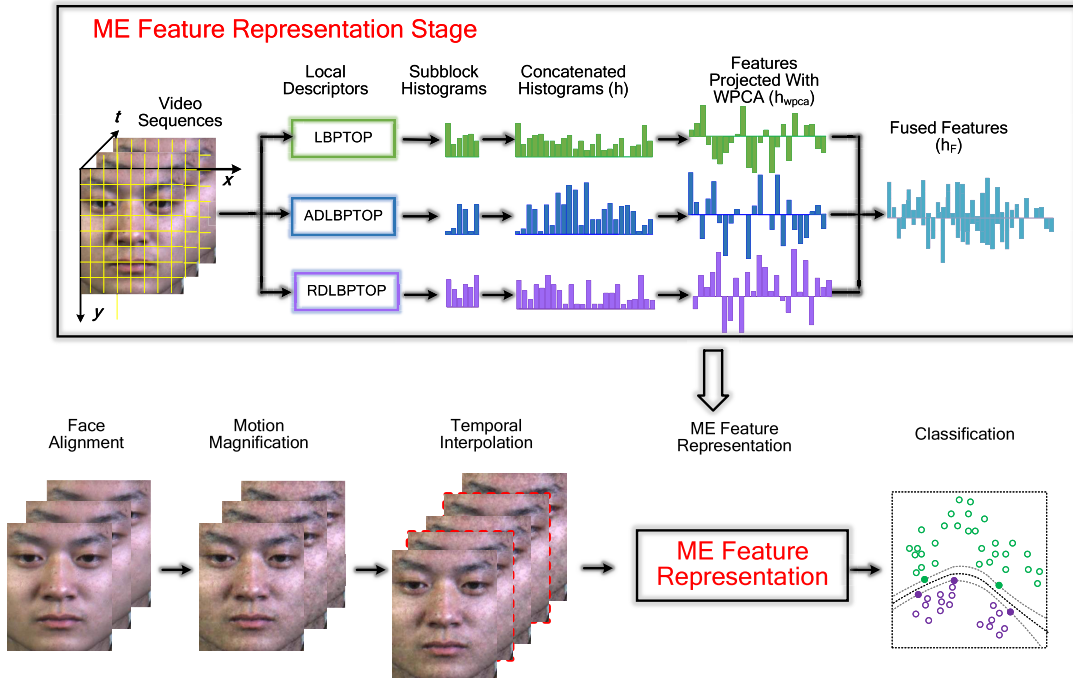


FIGURE 6. Overview of the proposed ME recognition framework.

are extracted via the three proposed binary codes. Each subblock histogram is normalized to sum one. Then, histograms of different subblocks are concatenated and projected by WPCA for dimensionality reduction. Finally, three feature representation vectors with low dimensionality from LBPTOP, RDLBPTOP and ADLBPTOP, are concatenated as a single vector \mathbf{h}_F , which is used for final ME feature representation.

C. THE ME RECOGNITION PIPELINE

The ME recognition problem is illustrated in Figure 6(b). The proposed overall pipeline for ME classification is shown in Figure 5. Following [47], raw ME video sequences are generally processed by the following steps: face alignment, motion magnification, temporal interpolation, feature extraction and classification.

Our main contribution in this work is the feature representation step, which is presented in detail in the previous sections. A very brief introduction of the other involved steps are given below. Readers are requested to referred to [3], [6], [47] for more information.

Face Alignment: For CASME II [2], SMIC [1] datasets, we use the given cropped images so that face alignment is not required. For SAMM [3] dataset, Active Shape Model [66] is used to detect 77 facial landmarks and then all the facial images are normalized using affine transformation and cropped into the same size according to the eye center points and the outermost points.

Motion Magnification: Since local intensity changes and facial movement changes in ME are subtle, effective ME characteristics are difficult to capture. To tackle these issues,

following [47], [67] we use Eulerian Video Magnification (EVM) [68] to magnify the subtle motions in videos. The goal is to consider the time series of intensity values at any spatial location (pixel) and amplify variation in a given temporal frequency band of interest. The filtered spatial bands are then amplified by a given factor α , added back to the original signal, and collapsed to generate the output video.

Temporal Interpolation: To address the issue that ME clips are short and have varied duration, we use the Temporal Interpolation Model (TIM) [69] and the code provided by [6]. The model first seeks a low-dimensional manifold where visual features extracted from the frames of a video can be projected onto a continuous deterministic curve embedded in a path graph. Moreover, it can map arbitrary points on the curve back into the image space, making it suitable for temporal interpolation.

Classification: For classification, we use Linear Support Vector Machine (LSVM) [70] as the classifier. Leave-one-subject-out cross-validation (LOSOCV) method is adopted to determine the penalty parameter c in SVM. For each test subject, LOSOCV is applied to the training samples, where in each fold the samples belonging to one subject are served as validation set and the rest of samples compose the new training set to select the best c and the selected c is used for testing.

IV. EXPERIMENTS

A. DATASETS

Three most popular spontaneous datasets, including CASME II [2], SMIC [1] and SAMM [3], are used to evaluate the performance of the proposed method. The dataset statistics are summarized in Table 1.

TABLE 1. A summary of the different features of the SMIC, CASME II and SAMM datasets.

Feature	SMIC-HS [1]	CASME II [2]	SAMM [3]
No. of Samples	164	247	159
No. of Subjects	16	26	29
Resolution	640 × 480	640 × 480	2040 × 1088
Facial Area	160 × 130	340 × 280	400 × 400
FPS	100	200	200
FACS Coded	NO	Yes	Yes
Classes	3	5	7

SMIC [1]: SMIC consists of 164 sample video clips of 16 subjects belonging to 3 different classes, *e.g.*, Positive (51 samples), Negative (70 samples) and Surprise (43 samples). The SMIC data has three versions: a high-speed camera (HS) version at 100 fps, a normal visual camera (VIS) version at 25 fps and a near-infrared camera (NIR) version 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. The emotion classes are only based on participants' self-reports. In this paper, we use the HS samples for experiments, and the resolution of average face size is 160 × 130.

CASME II [2]: CASME II contains 247 ME video clips from 26 subjects. All samples are recorded by a high speed camera at 200 fps. The resolution of samples is 640 × 480 pixels and the cropped area has 340 × 280 pixels. These samples are categorized into five ME classes: Happiness (32 samples), Surprise (25 samples), Disgust (64 samples), Repression (27 samples) and Others (99 samples). Different from SMIC, CASME II has AU labels following Facial Action Coding System (FACS). These classes are used in the whole parameter evaluation and they are used for comparison in Table 7. To remove the bias of human reporting, [71] reorganized the classes based on AU instead of original estimated emotion classes. Performance on the reorganized objective classes are also reported in Table 8.

SAMM [3]: SAMM database contains 159 ME video clips from 29 subjects. All samples are recorded by a high speed camera at 200 fps. The resolution of samples is 2040 × 1088 pixels and the cropped facial area has about 400 × 400 pixels. These samples are categorized into seven AU based objective classes. Classes I-VI are linked with Happiness (24 samples), Surprise (13 samples), Anger (20 samples), Disgust (8 samples), Sadness (3 samples), and Fear (7 samples). Class VII (84 samples) relates to contempt and other AUs that have no emotional link in FACS [72]. We carry on experiment on SAMM with classes I-V and the results are shown in Table 8.

B. IMPLEMENTATION DETAILS

We conduct two set of experiments: (1) Single database experiment involving SMIC and CASME II with their original estimated emotion classes, and CASME II and SAMM with the reorganized objective classes I-V. (2) Cross-database

experiments involving SMIC, CASME II, and SAMM following the Guidelines of the 1st MEGC [53] and the 2nd MEGC [54].

Most of the methods adopt leave-one-subject-out (LOSO) strategy for evaluation. For each fold, all samples from one subject are used as a testing set and the rest for training. A few works [40], [46] use leave-one-sample/video-out (LOVO) protocol, in which one sample is used as a testing set and the rest for training. Some works use their own protocols, such as random sampling of test partition [37], [51], five-fold [52] and others [49]. Leave one subject out (LOSO) strategy is used for evaluation in all the experiments. Mean accuracy, F1-score, Weighted F1-score, and Unweighted Average Recall (UAR) are used to measure the performance. Mean accuracy is obtained by averaging accuracies of subjects. F1-score is defined as $F = \frac{1}{c} \sum_{i=1}^c \frac{2p_i \times r_i}{p_i + r_i}$, where p_i and r_i are the precision and recall of the i th ME class, respectively, and c is the number of classes. Weighted F1-scores are weighted by the number of samples in the corresponding classes before averaging. UAR is the "balanced" accuracy (averaging the accuracy of each class without consideration of the number of samples per class).

Parameters: For block division parameters ($m \times q \times l$), $8 \times 8 \times 2$ is for CASME II and SMIC, and $5 \times 5 \times 2$ is for SAMM. For EVM [67], we choose the second-order bandpass filter with cutoff frequencies $\omega_1 = 0.4$, $\omega_2 = 0.05$ and spatial frequency cutoff $\lambda_c = 16$. Magnification value α is set to 20 for CASME II and SAMM, while $\alpha = 8$ is chosen for SMIC. TIM [69] is used to interpolate all ME sequences into the same length 10 according to [47]. Values of the number of neighboring pixels p , outer ring radius r and inner ring radius $r - \delta$ can be found in tables. The WPCA dimension is $v - 1$, where v is the number of video clips of each dataset, *e.g.*, 163 for SMIC and 246 for CASME II.

TABLE 2. ME recognition accuracy (%) of single descriptors on SMIC using two different encoding schemes: full patterns and uniform patterns. p , r and $r - \delta$ indicates the number of neighboring points, the outer ring and the inner ring respectively. All experiments are conducted without WPCA and EVM.

Method	full patterns		uniform patterns	
	Acc. (%) / F1-score	(r, p, δ)	Acc. (%) / F1-score	(r, p, δ)
LBPTOP	52.07 / 0.48	(3,8)	49.85 / 0.45	(3,8)
ADLBPTOP	53.11 / 0.51	(3,8)	49.89 / 0.47	(3,8)
RDLBPTOP	53.26 / 0.50	(3,8,2)	46.80 / 0.46	(3,8,2)

C. PARAMETER EVALUATION

The effect of encoding scheme: Table 2 compares the performance of two encoding schemes, full patterns (all 2^p patterns) and uniform patterns, on SMIC. Results on single binary descriptor without WPCA are reported. From table 2, we can see those histogram representations generated by the full patterns significantly outperform the uniform patterns, on all binary descriptors by a large margin (2.22% to 6.46% in accuracy, and 0.03 to 0.04 in F1-score), clearly demonstrating the insufficiency of the uniform patterns for representing ME videos. As a result, we conduct rest experiments using the full patterns encoding scheme.

TABLE 3. ME recognition accuracy (%) of different binary descriptors on SMIC with or without WPCA. ($m \times q \times l$) is set to $8 \times 8 \times 2$. Experiments are conducted without EVM.

Method	original (h)			WPCA (h ^{PCA})		
	Acc. (%) / F1-score	(r, p, δ)	Dim.	Acc. (%) / F1-score	(r, p, δ)	Dim.
LBPTOP	51.09 / 0.47	(2,8)	98304	52.29 / 0.49	(2,8)	163
ADLBPTOP	55.11 / 0.53	(2,8)	98304	58.45 / 0.54	(2,8)	163
RDLBPTOP	52.61 / 0.49	(2,8,1)	98304	52.61 / 0.49	(2,8,1)	163

The effect of WPCA: Table 3 illustrates the effect of WPCA dimensionality reduction on SMIC. Clearly, the accuracy and F1-score of all descriptors is consistently improved by WPCA. Besides, due to much lower feature dimensionality (163 compared with 98304), WPCA could lead to great computational saving. Therefore, further experiments are conducted using WPCA.

TABLE 4. ME recognition accuracy (%) of the single binary descriptors on SMIC and CASME II under various parameter settings. ($m \times q \times l$) is set to $8 \times 8 \times 2$. The WPCA dimension k for SMIC is 163, and 246 for CASME II. Experiments are conducted without EVM.

dataset	ADLBPTOP _{r,p}			LBPTOP _{r,p}			RDLBPTOP _{r,p,δ}		
	Acc. / F1-score	(r, p)		Acc. / F1-score	(r, p)		Acc. / F1-score	(r, p, δ)	
SMIC	62.27 / 0.58	(1,8)		52.19 / 0.53	(1,8)		52.55 / 0.53	(1,8,0)	
	58.45 / 0.54	(2,8)		52.29 / 0.49	(2,8)		52.61 / 0.49	(2,8,1)	
	53.11 / 0.51	(3,8)		52.07 / 0.48	(3,8)		50.67 / 0.46	(3,8,1)	
	53.11 / 0.51	(3,8)		52.07 / 0.48	(3,8)		53.26 / 0.50	(3,8,2)	
	47.95 / 0.49	(4,8)		54.50 / 0.53	(4,8)		55.97 / 0.50	(4,8,1)	
	47.95 / 0.49	(4,8)		54.50 / 0.53	(4,8)		52.89 / 0.48	(4,8,2)	
	47.95 / 0.49	(4,8)		54.50 / 0.53	(4,8)		55.45 / 0.50	(4,8,3)	
	48.35 / 0.34	(1,8)		50.15 / 0.38	(1,8)		49.14 / 0.35	(1,8,0)	
CASME II	56.45 / 0.39	(2,8)		52.79 / 0.37	(2,8)		50.89 / 0.38	(2,8,1)	
	44.36 / 0.35	(3,8)		50.92 / 0.36	(3,8)		49.49 / 0.35	(3,8,1)	
	44.36 / 0.35	(3,8)		50.92 / 0.36	(3,8)		55.10 / 0.37	(3,8,2)	
	47.23 / 0.35	(4,8)		49.49 / 0.29	(4,8)		40.64 / 0.31	(4,8,1)	
	47.23 / 0.35	(4,8)		49.49 / 0.29	(4,8)		43.19 / 0.28	(4,8,2)	
	47.23 / 0.35	(4,8)		49.49 / 0.29	(4,8)		45.60 / 0.32	(4,8,3)	

Evaluation of single binary descriptor: To explore the characteristics of different binary descriptors, we conduct experiments under various (r, p, δ) settings. As shown in Table 4, the radius r has great impacts on the performance of the three descriptors. The best accuracy often exceeds the second best by a large gap. Therefore, the choice of the best radius r is of great importance. It's the same for F1-score. In some cases, the highest accuracy and F1-score do not appear on the same parameters. In the following experiments, we choose the parameter setting with the highest accuracy. Similarly, δ is very important for the performance of RDLBP. Comparing the best results of ADLBPTOP, LBPTOP and RDLBPTOP, we can find that the proposed ADLBPTOP and RDLBPTOP outperform LBPTOP on both SMIC and CASME II in accuracy, which shows the importance of radial and angular difference information. Especially, ADLBPTOP performs much better than LBPTOP in accuracy and F1-score (3.66% and 8.27% higher in accuracy, 0.01 and 0.05 higher in F1-score on two datasets respectively).

Evaluation of EVM and parameter p : Evaluation of the number of neighboring pixels p and the effect of EVM are summarized in Table 5. Note that all the results are reported with their best radii. We can see that EVM can generally increase the recognition accuracy and F1-score, sometimes significantly (such as for ADLBPTOP and LBPTOP).

TABLE 5. ME recognition accuracy using different numbers of neighbors p as well as with or without EVM. The parameters of ($m \times q \times l$) and the WPCA dimensions k are the same as Table 4.

Method	SMIC		CASME II	
	Acc. (%) / F1-score	(r, p, δ)	Acc. (%) / F1-score	(r, p, δ)
ADLBPTOP	62.27 / 0.58	(1,8)	56.45 / 0.39	(3,8)
ADLBPTOP+EVM	63.73 / 0.61	(1,8)	69.12 / 0.64	(3,8)
	54.61 / 0.56	(4,4)	70.20 / 0.69	(2,4)
LBPTOP	54.50 / 0.53	(4,8)	52.97 / 0.37	(2,8)
LBPTOP+EVM	60.83 / 0.60	(3,8)	67.08 / 0.58	(4,8)
	65.16 / 0.58	(3,4)	71.55 / 0.65	(3,4)
RDLBPTOP	55.97 / 0.51	(4,8,1)	55.10 / 0.37	(3,8,2)
RDLBPTOP+EVM	61.04 / 0.58	(4,8,3)	67.62 / 0.66	(3,8,2)
	62.57 / 0.56	(4,4,3)	69.24 / 0.64	(3,4,1)

Table 5 also indicates that for each single ELBPTOP descriptor, the performance achieved by $p = 4$ is better than that by $p = 8$, with ADLBPTOP on SMIC being an exception.

TABLE 6. ME recognition accuracy (%) of three binary descriptors on different combinations of planes. The parameters ($m \times q \times l$) and k are the same as Table 4. Experiments are conducted with EVM.

Method		SMIC		CASME II	
		Acc. (%) / F1-score	(r, p, δ)	Acc. (%) / F1-score	(r, p, δ)
ADLBP	TOP	63.73 / 0.63	(1,8)	69.12 / 0.64	(3,8)
	XYOT	55.91 / 0.57	(1,8)	64.12 / 0.60	(3,8)
	XOT	55.92 / 0.58	(1,8)	61.47 / 0.58	(3,8)
	YOT	60.22 / 0.58	(1,8)	62.87 / 0.57	(3,8)
	XY	55.69 / 0.51	(1,8)	56.46 / 0.41	(3,8)
LBP	TOP	60.83 / 0.59	(3,8)	67.08 / 0.61	(4,8)
	XYOT	60.47 / 0.60	(3,8)	65.04 / 0.60	(4,8)
	XOT	57.24 / 0.58	(3,8)	61.38 / 0.56	(4,8)
	YOT	55.47 / 0.57	(3,8)	67.65 / 0.59	(4,8)
	XY	45.97 / 0.45	(3,8)	60.26 / 0.43	(4,8)
RDLBP	TOP	61.04 / 0.58	(4,8,3)	67.62 / 0.65	(3,8,2)
	XYOT	57.84 / 0.59	(4,8,3)	68.85 / 0.66	(3,8,2)
	XOT	56.06 / 0.56	(4,8,3)	62.91 / 0.60	(3,8,2)
	YOT	58.76 / 0.55	(4,8,3)	66.56 / 0.63	(3,8,2)
	XY	48.85 / 0.47	(4,8,3)	57.14 / 0.43	(3,8,2)

Evaluation of orthogonal planes: Table 6 illustrates the performance of three binary features (LBP, ADLBP, RDLBP) on five combinations of planes. TOP, XYOT, XOT, YOT and XY are abbreviations for XY+XT+YT, XT+YT, XT, YT and original spatial plane XY respectively. It can be observed that TOP and XYOT generally yields the best performance, which indicates that the dynamic information along the time dimension represents the most important information for ME recognition. In contrast, the results on XY plane are almost the worst. This is possibly because the XY plane contains much redundant information about the facial appearance. Maybe not all areas in the facial area contain useful discriminative information for ME recognition.

Feature Fusion: In order to find a good fusion of LBP*, RDLBP*, and ADLBP* (here, * represents one of TOP, XYOT, XOT, YOT and XY), we test all 215 ($6^3 - 1$) possible feature fusion schemes on SMIC and CASME II. All results are shown in Figure 7 in descending order. We can see that the highest accuracy is achieved by combining the three type of binary codes. The best results on SMIC-HS is 69.06%, given by $ADLBPTOP_{1,8} + LBPTOP_{3,4} + RDLBPXOT_{4,4,3}$, and on CASME II is 73.94%, given by $ADLBPTOP_{2,4} + LBPTOP_{3,4} + RDLBPXOT_{3,4,1}$.

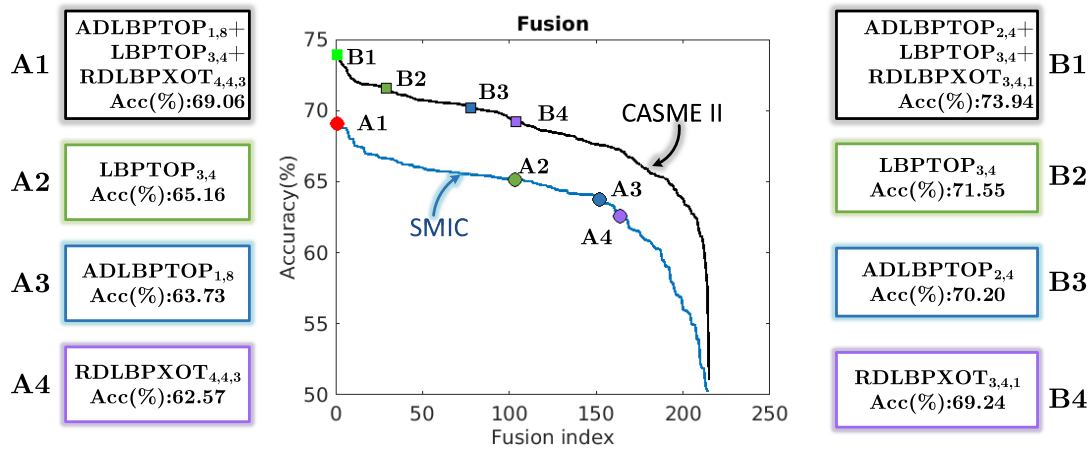


FIGURE 7. ME recognition accuracy(%) of different feature fusion schemes on SMIC and CASME II. In the boxes, we show the accuracy of the best fused descriptor and three single binary descriptors.

TABLE 7. Comparison between ELBPTOP and previous state-of-the-art methods on CASME II (with original classes) and SMIC.

Method	CASME II		SMIC	
	Acc. (%)	F1-score	Acc. (%)	F1-score
LBPTOP [1]	—	—	48.78	—
LBPMOP [33]	44.13	—	50.61	—
FDM [42]	45.93 ³	0.41 ¹	54.88 ³	0.54 ¹
LBPSIP [32]	46.56	—	44.51	—
3DFCNN [50]	59.11	—	55.49	—
STCLQP [39]	58.39 ²	0.58	64.02 ²	0.64
STLBP-IP [34]	59.51	—	57.93	—
CNN+LSTM [48]	60.98	—	—	—
BiWOOF + Phase [73]	62.55 ³	0.65	68.29 ³	0.67
Hierarchical STLBP-IP [74]	63.83 ³	0.61	60.78 ³	0.61
STRBP [35]	64.37	—	60.98	—
Discriminative STLBP-IP [75]	64.78	—	63.41	—
OF Maps [44]	65.35	—	—	—
HIGOTOP [47]	67.31	—	68.29	—
ELBPTOP	73.94	0.69	69.06	0.62

¹ The F1-score here is different, which is defined as $F = \frac{2p \times r}{p+r}$, where p and r are the average precision and recall of all the ME class.

² Mean recognition rate, which is obtained by averaging accuracies of classes.

³ Overall recognition rate, which is the number of correctly classified samples over the total samples.

As it can be seen from Figure 7, that the fused feature increases the accuracy by 3.90%, 5.33% and 6.49% respectively compared with using LBPTOP, ADLBPTOP or RDLBPTOP alone on SMIC. Similarly, the accuracy is improved by 2.39%, 3.74% and 4.70% on the three binary codes respectively on CASME II. The strong performance improvement shows that the fused approach indeed captures complementary information.

D. COMPARATIVE EVALUATION

1) SINGLE DATABASE RESULTS

We compare the best results achieved by our ELBPTOP with the baseline method and recent and relevant works on CASME II and SMIC with their original estimated emotion classes in Table 7, and on CASME II and SAMM with the reorganized objective classes in Table 8. Since the performance with different protocols is quite different, we only

TABLE 8. Comparison between ELBPTOP and previous state-of-the-art methods on SAMM and CASME II (with reorganized classes).

Method	SAMM		CASME II	
	Acc. (%)	F1-score	Acc. (%)	F1-score
LBPTOP [71]	44.70	0.35	67.80	0.51
HOOF [71]	42.17	0.33	69.64	0.56
HOG 3D [71]	34.16	0.22	69.53	0.51
ELBPTOP	63.44	0.48	79.55	0.66

TABLE 9. The results of composite database evaluation according to MEGC 2018.

Method	F1-score	Weighted F1-score
HOG 3D [76]	0.27	0.44
ELRCN [58]	0.39	0.52
LBPTOP [76]	0.40	0.52
HOOF [76]	0.40	0.53
Transfer learning [56]	0.64	0.73
ELBPTOP	0.64	0.71

compare the methods using the same LOSO strategy. For the same method, results with LOSO are usually lower than those with other protocols (LOVO, k-fold, and so on).

From Tables 7 and 8, we can observe that our proposed approach consistently gives the best results on all three datasets, significantly outperforming the state-of-the-art. As illustrated in Table 7, it is clear that our proposed method produces the highest accuracy (73.94%) and the highest F1-score (0.69), which is 6.63% higher in accuracy and 0.04 higher in F1-score than the second best on CASME II (with original classes). In Table 8, our method also surpasses all other methods on CASME II (with reorganized classes) significantly, improving a margin of 9.91% in accuracy and 0.10 in F1-score. The effectiveness of our method is further demonstrated by the large improvement on SAMM, with an increase from 44.70% to 63.44% (a margin of 18.74%). The strong performance on all ME datasets clearly proves that our proposed ELBPTOP is effective for ME recognition.

TABLE 10. The results of composite database evaluation according to MEGC 2019.

Method	Full		SMIC		CASME II		SAMM	
	F1-score	UAR	F1-score	UAR	F1-score	UAR	F1-score	UAR
LBPTOP [23]	0.59	0.58	0.20	0.53	0.70	0.74	0.40	0.41
Bi-WOOF [45]	0.63	0.62	0.57	0.58	0.78	0.80	0.52	0.51
CapsuleNet [63]	0.65	0.65	0.58	0.59	0.71	0.70	0.62	0.60
OFF-ApexNet [59]	0.72	0.71	0.68	0.67	0.88	0.87	0.54	0.54
Dual-Inception Network [61]	0.73	0.73	0.66	0.67	0.86	0.86	0.59	0.57
STSTNet [60]	0.74	0.76	0.68	0.70	0.84	0.87	0.66	0.68
EMR with Adversarial Training [57]	0.79	0.78	0.75	0.75	0.83	0.82	0.78	0.72
ELBPTOP	0.71	0.69	0.65	0.66	0.89	0.88	0.49	0.49

2) CROSS DATABASE RESULTS

To test the generalization of our method, we also conduct cross database experiments introduced in MEGC2018¹ and MEGC2019² Composite Database Evaluation (CDE) are used to test the performance. Following MEGC2018, all samples from CASME II and SAMM with their reorganized objective classes I-V are combined into a single composite database. There are total of 47 subjects (26 from CASME II and 29 from SAMM) and 253 samples (185 from CASME II and 68 from SAMM). The results are shown in Table 9. It can be seen from the table that our method achieves the best F1-score and the second weighted F1-score, confirming the generalization of our method.

Following MEGC2019, all samples from CASME II and SAMM and SMIC are combined into a single composite database, and the original emotion classes are grouped into three main classes: negative, positive and surprise. There are total of 68 subjects and 442 samples. Results in Table 10. show that our method is extremely powerful on the CASME II, achieving the highest UAR and F1-score. But there is still a need for further exploration on SMIC and SAMM. We infer that the following factors affect the performance on these two datasets: (1) Different pre-processing methods and cropping areas on SAMM. (2) Big differences in age and ethnicity in SAMM. (3) The lower frame rate and lower resolution on SMIC. These factors make the optimal parameters on each data set inconsistent, which in turn affects performance.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a simple, efficient and robust descriptor ELBPTOP for ME recognition. ELBPTOP consists of three complementary binary descriptors: LBPTOP and two novel ones RDLBPTOP and ADLBPTOP, which explore the local second order information along radial and angular directions contained in ME video sequences. For dimension reduction, WPCA is used to obtain efficient and discriminative features. Extensive experiments on three benchmark spontaneous ME datasets, SMIC, CASME II and SAMM have shown that our proposed approach surpasses state-of-the-art by a large margin in single database recognition, and also achieve more promising results on

cross-database recognition.

It is worth noting that there are some difficulties for micro-expression analysis: (1) Lack of standard evaluation protocol. Different evaluation protocols, performance metrics, number of samples, and emotion classes are chosen by different researchers. It raises the barriers of entry to this topic and increases difficulties for a fair comparison. (2) Lack of large scale spontaneous ME datasets. Small sample size and uneven distribution are still the key to restriction the acquisition of effective features and application to real life. Especially, there are single emotion class in some subject, making it more difficult to obtain features that are distinguishable from expressions rather than distinguishing from subjects.

Hand-crafted features confirm that effective discriminant characteristics can be learned. And in the current micro-expression field, many of the deep learning methods are based on hand-crafted features. But some hyper parameters need to be artificially selected, which restricts the performance in cross database problem to some extent. In our future work, we plan to design data-driven methods to learn binary codes directly from data for ME recognition. In addition, in many works, the AU information is very useful but we have not used it in this paper. We will design a better area division algorithm to utilize the AU information.

REFERENCES

- [1] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.
- [2] 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, 2014, Art. no. e86041.
- [3] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 116–129, Jun. 2018.
- [4] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Acad. Sci.*, vol. 1000, no. 1, pp. 205–221, 2003.
- [5] Q. Wu, X.-B. Shen, and X.-L. Fu, "Micro-expression and its applications," *Adv. Psychol. Sci.*, vol. 18, no. 9, pp. 1359–1368, 2010.
- [6] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1449–1456.
- [7] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: Databases, methods, and challenges," *Frontiers Psychol.*, vol. 9, p. 1128, Jul. 2018.
- [8] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*. Cham, Switzerland: Springer, 2016, pp. 63–100.

¹<http://www2.docm.mmu.ac.uk/STAFF/m.yap/FG2018Workshop.htm>

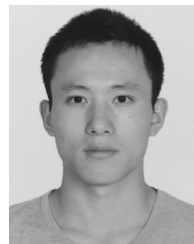
²<https://facial-micro-expressiongc.github.io/MEGC2019/>

- [9] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *J. Nonverbal Behav.*, vol. 37, no. 4, pp. 217–230, 2013.
- [10] S. Porter and L. T. Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychol. Sci.*, vol. 19, no. 5, pp. 508–514, 2008.
- [11] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From BoW to CNN: Two decades of texture representation for texture classification," *Int. J. Comput. Vis.*, vol. 127, no. 1, pp. 74–109, 2019.
- [12] W. Merghani, A. K. Davison, and M. H. Yap, "A review on facial micro-expressions analysis: Datasets, features and metrics," 2018, *arXiv:1805.02397*. [Online]. Available: <https://arxiv.org/abs/1805.02397>
- [13] M. Shreve, S. Godavathy, D. Goldgof, and S. Sarkar, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *Proc. Face Gesture*, Mar. 2011, pp. 51–56.
- [14] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention*, Dec. 2009, pp. 1–6.
- [15] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *J. Nonverbal Behav.*, vol. 33, no. 1, pp. 59–69, 2009.
- [16] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Sep. 2013, pp. 1–7.
- [17] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)²: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 424–436, Oct./Dec. 2017.
- [18] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [19] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.
- [20] S. ul Hussain and B. Triggs, "Visual recognition using local quantized patterns," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012.
- [21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.
- [22] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
- [23] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
- [24] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1932–1939.
- [25] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," *Pattern Recognit.*, vol. 62, pp. 135–160, Feb. 2017.
- [26] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
- [27] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, and W. Matusik, "Learning-based video motion magnification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 633–648.
- [28] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 765–781, Nov. 2011.
- [29] A. Fernández, M. X. Álvarez, and F. Bianconi, "Texture description through histograms of equivalent patterns," *J. Math. Imag. Vis.*, vol. 45, no. 1, pp. 76–102, 2013.
- [30] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, and D. Hu, "Extended local binary patterns for face recognition," *Inf. Sci.*, vol. 358, pp. 56–72, Sep. 2016.
- [31] L. Liu, S. Lao, P. W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Median robust extended local binary pattern for texture classification," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1368–1381, Mar. 2016.
- [32] Y. Wang et al., "LBP with six intersection points: Reducing redundant information in LBP-top for micro-expression recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014.
- [33] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition," *PLoS ONE*, vol. 10, no. 5, 2015, Art. no. e0124674.
- [34] X. Huang, S.-J. Wang, G. Zhao, and M. Pietikäinen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Jun. 2015, pp. 1–9.
- [35] X. Huang and G. Zhao, "Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern," in *Proc. Int. Conf. Frontiers Adv. Data Sci. (FADS)*, Oct. 2017, pp. 159–164.
- [36] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
- [37] X. Ben, X. Jia, R. Yan, X. Zhang, and W. Meng, "Learning effective binary descriptors for micro-expression recognition transferred by macro-information," *Pattern Recognit. Lett.*, vol. 107, pp. 50–58, May 2018.
- [38] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, Mar. 2016.
- [39] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, Jan. 2016.
- [40] S.-J. Wang et al., "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014.
- [41] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.
- [42] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affective Comput.*, vol. 8, no. 2, pp. 254–267, Apr. 2017.
- [43] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affective Comput.*, vol. 7, no. 4, pp. 299–310, Oct. 2016.
- [44] B. Allaert, I. M. Bilasco, and C. Djeraba, "Consistent Optical Flow Maps for full and micro facial expression recognition," in *Proc. VISAPP*, vol. 5. Setúbal, Portugal: SCITEPRESS-Science and Technology Publications, 2017.
- [45] S. Liong, J. See, K. Wong, and R. C. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Process., Image Commun.*, vol. 62, pp. 82–92, Mar. 2018.
- [46] Y. Zhao and J. Xu, "An improved micro-expression recognition method based on necessary morphological patches," *Symmetry*, vol. 11, no. 4, p. 497, 2019.
- [47] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 563–577, Oct./Dec. 2018.
- [48] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 382–386.
- [49] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers Psychol.*, vol. 8, p. 1745, Oct. 2017.
- [50] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1331–1339, 2019.
- [51] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," 2019, *arXiv:1904.01390*. [Online]. Available: <https://arxiv.org/abs/1904.01390>
- [52] Z. Xia, X. Feng, X. Hong, and G. Zhao, "Spontaneous facial micro-expression recognition via deep convolutional network," in *Proc. 8th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2018, pp. 1–6.
- [53] M. H. Yap, J. See, X. Hong, and S.-J. Wang, "Facial micro-expressions grand challenge 2018 summary," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 675–678.

- [54] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019—The second facial micro-expressions grand challenge," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
- [55] Y. Zong, W. Zheng, X. Hong, C. Tang, Z. Cui, and G. Zhao, "Cross-database micro-expression recognition: A benchmark," in *Proc. Int. Conf. Multimedia Retr.*, 2019, pp. 354–363.
- [56] M. Peng, Z. Wu, Z. Zhang, and T. Chen, "From macro to micro expression recognition: Deep learning on small datasets using transfer learning," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 657–661.
- [57] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–4.
- [58] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 667–674.
- [59] S.-T. Liong, Y. S. Gan, W.-C. Yau, Y.-C. Huang, and T. L. Ken, "Off-ApexNet on micro-expression recognition system," 2018, *arXiv:1805.08699*. [Online]. Available: <https://arxiv.org/abs/1805.08699>
- [60] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," 2019, *arXiv:1902.03634*. [Online]. Available: <https://arxiv.org/abs/1902.03634>
- [61] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
- [62] M. Peng, C. Wang, T. Bi, T. Chen, X. Zhou, and Y. Shi, "A novel apex-time network for cross-dataset micro-expression recognition," 2019, *arXiv:1904.03699*. [Online]. Available: <https://arxiv.org/abs/1904.03699>
- [63] N. Van Quang, J. Chun, and T. Tokuyama, "CapsuleNet for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.
- [64] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
- [65] H. V. Nguyen, L. Bai, and L. Shen, "Local Gabor binary pattern whitened PCA: A novel approach for face recognition from single image per person," in *Proc. Int. Conf. Biometrics*. Berlin, Germany: Springer, 2009.
- [66] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.
- [67] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21665–21690, 2017.
- [68] H.-Y. Wu, M. Rubinstein, E. Shih, J. V. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, 2012.
- [69] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *Proc. CVPR*, Jun. 2011, pp. 137–144.
- [70] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [71] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *J. Imag.*, vol. 4, no. 10, p. 119, 2018.
- [72] P. Ekman and W. V. Friesen, *Facial Action Coding System: Investigator's Guide*. Mountain View, CA, USA: Consulting Psychologists Press, 1978.
- [73] S.-T. Liong and K. Wong, "Micro-expression recognition using apex frame with phase information," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 534–537.
- [74] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3160–3172, Nov. 2018.
- [75] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 32–47, Jan./Mar. 2017.
- [76] W. Merghani, A. Davison, and M. Yap, "Facial micro-expressions grand challenge 2018: Evaluating spatio-temporal features for classification of objective classes," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 662–666.



CHENGYU GUO received the B.Sc. degree in mathematics and applied mathematics from Shanghai Jiao Tong University (SJTU), China, in 2017. She is currently pursuing the M.Sc. degree in control science and engineering with the National University of Defense Technology. Her current research interests include computer vision and deep learning.



JINGYUN LIANG received the B.Sc. degree in information system engineering from the National University of Defense Technology (NUDT), China, in 2016, where he is currently pursuing the M.Sc. degree in control science and engineering. His current research interests include computer vision and deep learning.



GENG ZHAN received the B.E. degree from the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, in 2017. He is currently pursuing the M.Phil. degree with the School of Electrical and Information Engineering, The University of Sydney. His current research interests include computer vision and machine learning.



ZHONG LIU received the Ph.D. degree in management science from the National University of Defense Technology (NUDT), Changsha, China, in 2000. He is currently a Professor with NUDT. He is also a Vice-Dean with the College of Systems Engineering, NUDT, where he is a Senior Advisor with the Research Center for Computational Experiments and Parallel Systems. His current research interests include planning systems, computational organization, and intelligent systems.



MATTI PIETIKÄINEN received the D.Sc. degree in technology from the University of Oulu, Finland. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory, University of Maryland. He is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. He has authored about 350 refereed articles in international journals, books, and conferences. His articles have over 51,000 citations in Google Scholar (H-index

of 76), and eight of these have over 1,300 citations. His current research interests include local binary pattern (LBP) methodology, texture-based image and video analysis, and facial image analysis. He was the President of the Pattern Recognition Society, Finland, from 1989 to 1992, and was named its Honorary Member, in 2014. From 1989 to 2007, he served as a member of the Governing Board of International Association for Pattern Recognition (IAPR), and became one of the Founding Fellows of the IAPR, in 1994. In 2014, his research on LBP-based face description was received the Koenderink Prize for fundamental contributions in computer vision. He was a recipient of the prestigious IAPR King-Sun Fu Prize 2018 for fundamental contributions to texture analysis and facial image analysis. He was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *Pattern Recognition*, and the IEEE TRANSACTIONS ON FORENSICS AND SECURITY journals. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE and *Image and Vision Computing* journals, and a Guest Editor for special issues of IEEE TPAMI and *International Journal of Computer Vision*. He is a Fellow of the IEEE for the contributions to texture and facial image analysis for machine vision.



LI LIU received the B.Sc. degree in communication engineering, the M.Sc. degree in photogrammetry and remote sensing, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2003, 2005, and 2012, respectively. During the Ph.D. study, she spent more than two years as a Visiting Student with the University of Waterloo, Canada, from 2008 to 2010. She joined the Faculty of NUDT, in 2012,

where she is currently an Associate Professor with the College of System Engineering. From 2015 to 2016, she spent ten months with the Multimedia Laboratory, The Chinese University of Hong Kong. From 2016 to 2018, she was a Senior Researcher with the Machine Vision Group, University of Oulu, Finland. Her current research interests include facial behavior analysis, texture analysis, image classification, object detection and recognition. She was a Co-Chair of International Workshops, including CVPR2016, ICCV2017, ECCV2018, and CVPR2019. Her articles have currently over 2080+ citations in Google Scholar. She was a Guest Editor of special issues for IEEE TPAMI and IJCV. She currently serves as an Associate Editor of *The Visual Computer* journal.

...