Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection for Spontaneous Facial Micro-Expression Recognition

Xiaohua Huang, Sujing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng and Matti Pietikäinen

Abstract—Recently, there have been increasing interests in inferring mirco-expression from facial image sequences. Due to subtle facial movement of micro-expressions, feature extraction has become an important and critical issue for spontaneous facial micro-expression recognition. Recent works used spatiotemporal local binary pattern (STLBP) for micro-expression recognition and considered dynamic texture information to represent face images. However, they miss the shape attribute of face images. On the other hand, they extract the spatiotemporal features from the global face regions while ignore the discriminative information between two micro-expression classes. The above-mentioned problems seriously limit the application of STLBP to micro-expression recognition. In this paper, we propose a discriminative spatiotemporal local binary pattern based on an integral projection to resolve the problems of STLBP for micro-expression recognition. Firstly, we revisit an integral projection for preserving the shape attribute of micro-expressions by using robust principal component analysis. Furthermore, a revisited integral projection is incorporated with local binary pattern across spatial and temporal domains. Specifically, we extract the novel spatiotemporal features incorporating shape attributes into spatiotemporal texture features. For increasing the discrimination of micro-expressions, we propose a new feature selection based on Laplacian method to extract the discriminative information for facial micro-expression recognition. Intensive experiments are conducted on three availably published micro-expression databases including CASME, CASME2 and SMIC databases. We compare our method with the state-of-the-art algorithms. Experimental results demonstrate that our proposed method achieves promising performance for micro-expression recognition.

Index Terms—Spontaneous facial micro-expression, spatiotemporal, local binary pattern, integral projection, feature selection

1 Introduction

Micro-expressions amongst nonverbal behavior like gestures and voice have received increasing attention in recent years [1]. In situations in which people are motivated to conceal or suppress their true emotions, their facial expressions may leak despite that they try to conceal them. These leakages can be very useful for true emotion analysis and many of these leakages are manifested in the form of micro-expressions. However, micro-expressions are very short involuntary facial expressions that reveal emotions people try to hide. Generally, they last 1/25 to 1/3 second [2], [3]. It is important to note that due to the visual differences of human beings, not all people reach the same level of ability to detect these facial expressions. Currently only highly trained individuals are able to distinguish them, but even with proper training the recognition accuracy is still less than 50% [4]. Therefore, this poor performance makes an automatic micro-expression recognition system very attractive.

Several earlier studies on automatic facial micro-expression analysis primarily focused on distinguishing facial micro-expressions from macro-expressions [5] [6]. Shreve *et al.* [5] [6] used an optical flow method for automatic micro-expression spotting on their own database. However, their database contains 100 clips of posed micro-expressions, which were obtained by asking participants to mimic some example videos that contain micro-expressions. Polikovsky *et al.* in [7] proposed to use a 3D-gradient orientation histogram for action unit recognition on their collected database. Unfortunately, their work focused on posed micro-expression as well, since in the collection procedure they asked subjects to perform seven basic emotions with low intensity and go back to neutral expression as quickly as possible. Wu *et al.* in [8] combined Gentleboost and a support vector machine classifier to recognize synthetic micro-expression samples from the Micro Expression Training Tool. The significant problem of posed micro-expressions is that they are different from real naturally occurring spontaneous micro-expressions. A study in [2] shows that spontaneous micro-expression occurs involuntarily, and that the producers of the micro-expressions usually do not even realize that they have presented such an emotion. Therefore, methods trained on posed micro-expressions cannot really solve the problem of automatic micro-expression analysis in practice.

Recently, researchers have started to conduct analysis on spontaneous micro-expression recognition, as they can reveal genuine emotions which people try to conceal. However, micro-expression recognition suffers from two critical problems including the short duration and low intensity of micro-expressions,

- *X. Huang, X. Liu, G. Zhao and M. Pietikäinen are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. E-mail: xiaohua.huang@oulu.fi, xliu@ee.oulu.fi, gyzhao@ee.oulu.fi, mkp@ee.oulu.fi*
- *S.J. Wang is with State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Science, Beijing, China. E-mail: wangsujing@psych.ac.cn*
- *X. Feng is with School of Electronic and Information, Northwestern Polytechnic University, Xi'an, China. E-mail: fengxiao@nwpu.edu.cn*

because short duration makes micro-expression difficult to detect and low intensity causes feature extraction algorithms hard to extract the useful information. For the problem of short duration, there are several available schemes to handle the short duration change of micro-expression. In [9], they hypothesized that dynamics in subtle occurring expressions are likely to be sparse, and contains a significantly large number of redundant frames. A principled approach of deconstructing motions into several dynamic modes was utilized to much great effect than standard temporal interpolation model. In [10], they characterized the local movements of a micro-expression by the principal optical flow direction of spatiotemporal cuboids extracted of a chosen granularity. On the other hand, some researchers attempt to extract the discriminative and useful feature from the low-intensity micro-expression [11] [12] [13] [14]. As we know, geometry-based and appearance-based features have been commonly employed to analyze facial expressions [15] [16] [17]. Due to the subtle change of micro-expression, the geometric-based features cannot accurately capture subtle facial movements (*e.g.*, the eye wrinkles) of micro-expression recognition, while appearance-based features describe the skin texture of faces, which can capture subtle appearance changes such as wrinkles and shading changes. An alternative way is to exploit deep features for micro-expression recognition, which is motivated by its application of deep learning in facial expression recognition [18] [19] [20] [21] [22]. An interesting work proposed by Patel *et al.* [23] is to use feature selection to choose the useful information of deep features for micro-expression recognition. However, it is found that their work [23] has still far away from the state-of-the-art work in micro-expression recognition. Additionally, deep learning based features require a large-scale database for training networks such as deep convolutional neural network. Unfortunately, the small sample size of the current available micro-expression databases seriously limits the application of deep learning for facial micro-expression recognition.

Amongst appearance-based features, local binary pattern (LBP) has been commonly used in face recognition [24] and facial expression recognition [25]. Recently, LBP is extended to spatiotemporal domain for texture recognition and facial expression recognition [26], which is named Local binary pattern from three orthogonal planes (LBP-TOP). LBP-TOP has shown its promising performance for facial expression recognition [26]. Therefore, many researchers have actively focused on the potential ability of LBP-TOP for micro-expression recognition. Pfister *et al.* [13] proposed to use LBP-TOP for analyzing spontaneous micro-expression recognition and conducted experiments on spontaneous micro-expression corpous (SMIC) database. The system is the first one to automatically analyze spontaneous facial micro-expressions. It primarily consists of a temporal interpolation model and feature extraction based on LBP-TOP. In [12], Li *et al.* continued implementing LBP-TOP on the full version of SMIC and obtained the recognition result of 48.48%. Meanwhile, Yan *et al.* [14] used the method of [12] as the baseline algorithm on the second version of Chinese Academy of Sciences Micro-expression (CASME2) database. Since then, LBP and its variants have often been employed as the feature descriptors for micro-expression recognition in many other studies. For example, Davison *et al.* [27] exploited LBP-TOP to investigate whether micro-facial movement sequences can be differentiated from neutral face sequences.

However, according to [12] [13] [14] it is observed that there is a gap to achieve a high-performance micro-expression analysis using LBP-TOP, since LBP-TOP attempts to obtain features by exploiting the pixel information of an image. Consequently, many works have attempted to improve the LBP-TOP. Ruiz-Hernandez and Pietikäinen [28] used the re-parameterization of second order Gaussian jet on the LBP-TOP achieving promising micro-expression recognition result on the first version of SMIC database [13]. As well, Wang *et al.* [29] extracted Tensor features from Tensor Independent Colour Space (TICS) for micro-expression recognition, but their results on the CASME2 database showed no improvement comparing with the previous results. Furthermore, Wang *et al.* [30] used Local Spatiotemporal Directional Features with robust principal component analysis for micro-expressions. Recent work in [31] reduced redundant information in LBP-TOP by using six intersection points (LBP-SIP) and obtained better performance than LBP-TOP. Guo *et al.* [32] employed Centralized Binary Patterns from Three Orthogonal Panels with extreme learning machine to recognize micro-expressions. Oh *et al.* [33] employed Riesz wavelet transform to obtain multi-scale monogenic wavelets for micro-expression recognition. Huang *et al.* [34] proposed spatiotemporal completed local binary pattern, namely STCLQP, to utilize magnitude and orientation as additional source and flexible encoding algorithm for improving LBP-TOP on micro-expression recognition. Li *et al.* [35] proposed two spatiotemporal feature descriptors for micro-expression recognition, where they extended histograms of oriented gradient and histograms of image gradient to three orthogonal planes, named HOG-TOP and HIGO-TOP, respectively.

It is noted that the most of previous methods used in micro-expression recognition have two critical problems to be resolved. Firstly, they exploited some variant of LBP-TOP in which dynamic texture information is considered to represent face images. However, they missed the shape attribute of face images. Recent study in [36] suggests that the fusion of texture and shape information can perform better results than only using appearance features for facial expression recognition. Moreover, the work in [37] demonstrated that LBP enhanced by shape information can distinguish an image with different shape from those with the same LBP feature distributions. The method of [37] has been used for to achieve better performance than LBP for face recognition [38] [39]. Secondly, the methods in [12] [13] [14] [34] used a block-based approach for spatiotemporal features. Specifically, they firstly divide a video clip into some blocks and then concatenated features from all blocks into one feature vector. However, we observe that the dimensionality of feature may be huge. On the other hand, the same contribution from all block features would decrease the performance. Normally speaking, all spatial temporal features do not contribute equally. Therefore, in the present paper we aim to develop a new method simultaneously incorporating the shape attribute with LBP and considering the discriminative information for micro-expression recognition.

Image projection techniques are classical methods for pattern analysis, widely used, *e.g.*, in motion estimation [40] and face tracking [41] [42], as they enhance shape properties and increase discrimination of images. Integral projection provides simple and efficient computation in computer vision amongst image projection techniques. It firstly is invariant to a number of image transformations [43]. It is also highly robust to white noise [40]. Then it preserves the principle of locality of pixels and sufficient information in the process of projection. Recently, integral projection is used to incorporate with LBP for achieving promising performance in bone texture classification [37] and face

recognition [38] [39]. However, it was observed that the identity information seriously destroys the discriminative capability of integral projection for describing micro-expressions. As well, the subtle motion information can provide discriminative information to integral projection for describing micro-expressions, which should be taken into account. Our previous work [44] presented image difference-based method to extract the subtle motion information for integral projection. However, the limitation of the work is that we suppose that the first frame is neutral face. Therefore, in this paper, we further extend our previous work [44] by introducing a new method to relax this hypothesis, and then propose a new spatiotemporal feature descriptor, which incorporates shape attributes to dynamic texture information for improving the performance of micro-expression recognition.

For simplicity, for extracting discriminant of feature, we may employ dimensionality reduction methods such as Linear Discriminative Analysis. However, these approaches may fail to work on micro-expression recognition because of small number of classes and high dimensionality of micro-expression. Zhao *et al.* [45] proposed a novel method based on AdaBoost to select the discriminative slices for facial expression recognition. But we observe that AdaBoost did not consider the closeness between two micro-expression samples and is not stable to micro-expression recognition. Recently, Laplacian method [46] is presented to select more compact and discriminative feature for face recognition. It considers the discriminative information and the closeness of two samples through a weighted graph. Therefore, based on the framework of [45], we propose a new method based on Laplacian algorithm to learn the discriminative group-based features for micro-expression analysis.

Different from [44], the present work includes three new interesting parts: (1) A robust way is proposed to resolve the problem of integral projection. It can relax the strong hypothesis used in [44] and be flexible to micro-expression recognition; (2) A feature selection approach is presented to automatically select the discriminative group-based feature for enhancing the performance of micro-expression recognition; and (3) More parameter evaluation and algorithm comparison are conducted on three micro-expression databases.

To explain the concepts of our approach, the paper is organized as follows. In Section 2, we explain our method of exploring the spatiotemporal features and discriminative information for micro-expression analysis. The results of applying our method for recognizing micro-expressions are provided in Section 3. Finally we summarize the paper in Section 4.

2 Proposed Methodology

Recently, the combination of the integral projection and texture descriptor was applied to bone texture characterization [37] and face recognition [38]. They demonstrate that the texture descriptor is enhanced by shape information extracted by the integral projection. However, we observe that integral projection mainly represents subject information so that it cannot be directly used to describe the shape attribute of micro-expressions. In this section, we firstly resolve the problem of integral projection and then propose a new spatiotemporal feature descriptor for micro-expression recognition.

2.1 Revisited Integral Projection

2.1.1 Problem Setting

An integral projection aims to generate a one-dimensional pattern through the sum of a given set of pixels along a given direction. Mathematically, given the intensity of a pixel $\mathbf{I}(x, y)$, its integral projection is formulated as:

$$\Re[f](\theta, s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{I}(x, y)\delta(x\cos\theta + y\sin\theta - s)dxdy, \tag{1}$$

where δ is a Dirac's delta function, θ and s are a projection angle and the threshold value, respectively. Essentially, the integral projection can capture the common underlying structure of face images of the same face subject. In other words, it is very relative to face identity [41], [42]. However, it is double that whether integral projection works for micro-expression. We evaluate Equation 1 based on $\theta = 0^\circ$ and $\theta = 90^\circ$ on two following cases: (1) two micro-expression images with the same class from two different persons and (2) two images with the different class from a person. The evaluations are shown in Fig. 1(d), in which the left image to the right image represent the integral projections of Figs. 1(a), 1(b) and 1(c), respectively. It demonstrates that the integral projection cannot provide discriminative information for different micro-expressions, such as the integral projections of Figs. 1(a) and 1(c). In other words, the integral projection fails to extract the shape attribute for micro-expressions. As a result, it is necessary to change integral projection method to obtain the class information for micro-expressions.

2.1.2 Micro-expression Augment for Integral Projection

Due to short duration and low intensity of micro-expression, the micro-expression data are sparse in both temporal and spatial domains [47]. Moreover, for the integral projection, our previous work [44] and Fig. 1 demonstrate that the identity information seriously causes the discriminant of integral projection badly work for describing micro-expressions. Instead, the subtle motion information can provide discriminative information to integral projection for describing micro-expressions. For a micro-expression image $\mathbf{I}_t$, it is assumed to contain $\mathbf{Q}_t$ and $\mathbf{E}_t$, which is implicitly represented as following:

$$\mathbf{I}_t = \mathbf{Q}_t + \mathbf{E}_t, \tag{2}$$

where $\mathbf{E}_t$ includes the subtle motion information of micro-expression at the t-th frame while $\mathbf{Q}_t$ is the other information. In our previous work [44], we used the difference-based image method to extract the subtle motion information $\mathbf{E}_t$ for resolving the problem of integral projection, but it is strongly hypothesized that the first frame should be neural face. However, this hypothesis is not ensured for all micro-expression videos. Therefore, the key problem for Equation 1 is how to extract the robust subtle facial motion information which is discriminant for recognizing micro-expression. For convenience, the subscript t for $\mathbf{I}_t$, $\mathbf{Q}_t$ and $\mathbf{E}_t$ is omitted in discussing the way of obtaining $\mathbf{E}_t$.

For a micro-expression video clip, other information including illumination, pose and subject identity accounts for the great proportion of the whole information in a clip, while the subtle facial motion information is sparse. Equation 2 can be viewed to extract the sparse information for a video clip. As we know, the background modeling is popular in background subtraction for video analysis. The robust principal component analysis (RPCA)

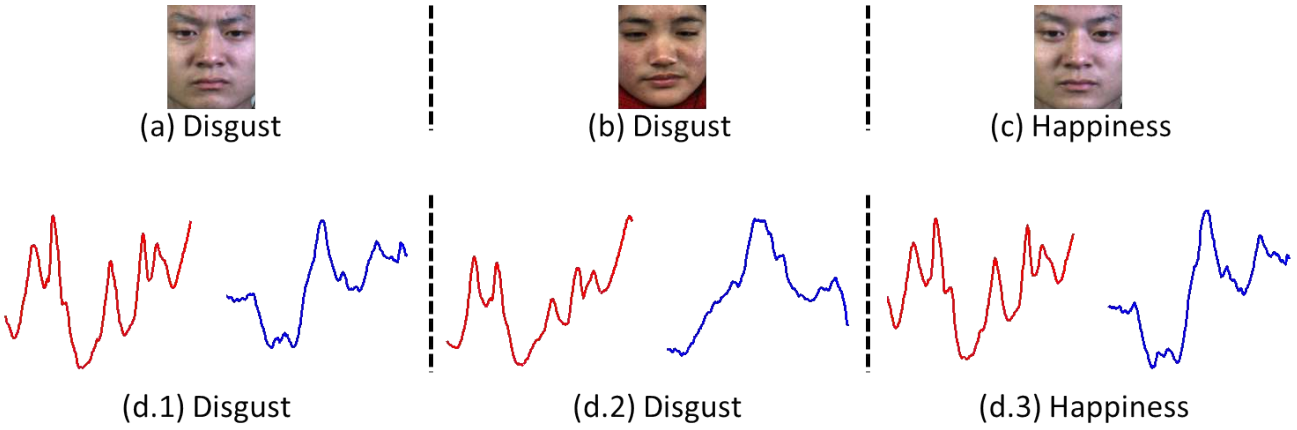

Fig. 1: Integral projection on micro-expression images with 'Disgust' and 'Happiness', where the red and blue colors represent integral projections based on $\theta = 0°$ and $\theta = 90°$, respectively.

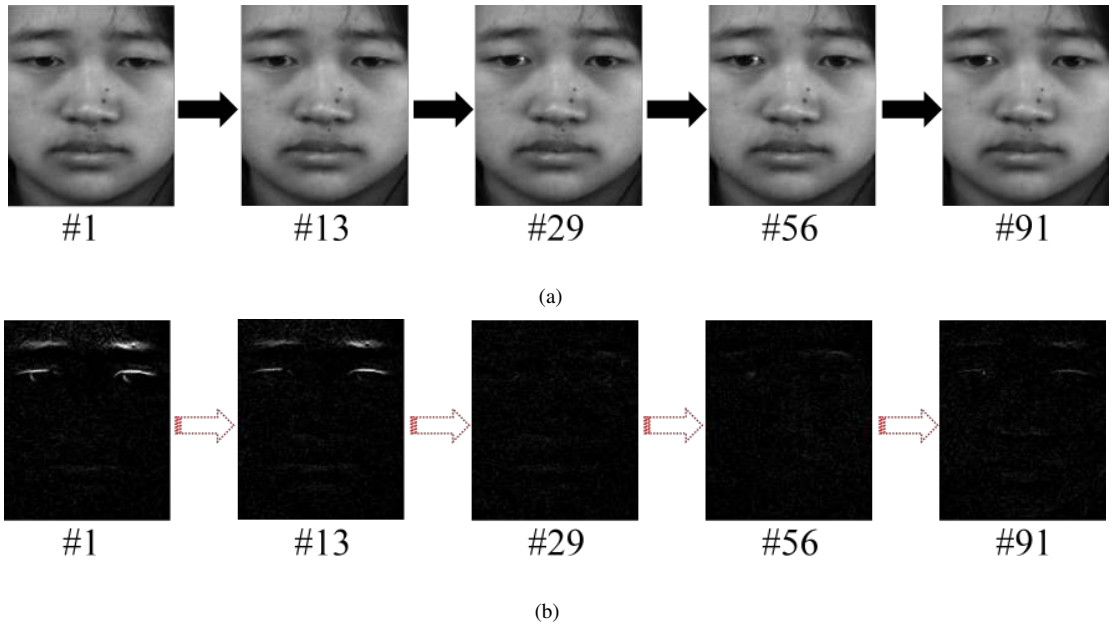

Fig. 2: An example of extracting the subtle facial motion information from a micro-expression video with 'surprise'. (a) indicate the original micro-expression frames sequence, where the number below an image represents the frame index. (b) indicate the extracted subtle facial motion information.

is widely used for face recognition [48] and facial expression [49]. It leverages on the data are characterized by low-rank subspaces. Recently, Wang *et al.* [30] used the RPCA to extract subtle motion of micro-expression. They demonstrated the subtle motion can extract the low intensity and enhance the ability of dynamic texture features. Therefore, we aim to extract this sparse information E using RPCA [48] for the integral projection.

Given a micro-expression video clip, each of its frames is vectorized as a column of matrix $\mathbf{I} \in \Re^D$. As $\mathbf{E}$ includes the derived sparse subtle motion information, the optimization problem of Equation 2 is formulated as follows:

$$[\mathbf{Q}, \mathbf{E}] = \min \text{rank}(\mathbf{Q}) + \| \mathbf{E} \|_0, \text{w.r.t. } \mathbf{I} = \mathbf{Q} + \mathbf{E}, \quad (3)$$

where $\text{rank}(\cdot)$ denotes the rank of matrix and $\| \cdot \|_0$ means L_0 norm. Because of not-convex problem of Equation 3, it is converted into the convex optimization problem as followed:

$$[\mathbf{Q}, \mathbf{E}] = \min \| \mathbf{Q} \|_* + \lambda \| \mathbf{E} \|_0, \text{w.r.t. } \mathbf{I} = \mathbf{Q} + \mathbf{E}, \quad (4)$$

where $\| \cdot \|_*$ denotes the nuclear norm, which is the sum of its singular values. λ is a positive weighting parameter.

For solving Equation 4, the iterative thresholding technique can be used to minimize a combination of both the L_0 norm and the nuclear norm, while this scheme converges extremely slowly. Instead, Augmented Lagrange Multipliers (ALM) is more efficient way to solve Equation 4. Specifically, ALM is introduced for solving the following constrained optimization problem:

$$\mathbf{X} = \min f(\mathbf{X}), \text{w.r.t. } h(\mathbf{X}) = 0, \quad (5)$$

where $f : \Re^n \rightarrow \Re$ and $h : \Re^n \rightarrow \Re^m$. The augmented Lagrangian function can be defined as follows:

$$L(\mathbf{X}, \mathbf{Y}, \mu) = f(\mathbf{X}) + \langle \mathbf{Y}, h(\mathbf{X}) \rangle + \frac{\mu}{2} \| h(\mathbf{X}) \|_F^2 . \quad (6)$$

Let $\mathbf{X}$ be $(\mathbf{Q}, \mathbf{E})$, $f(\mathbf{X})$ be $\| \mathbf{Q} \|_* + \lambda \| \mathbf{E} \|_1$, and $h(\mathbf{X})$ be

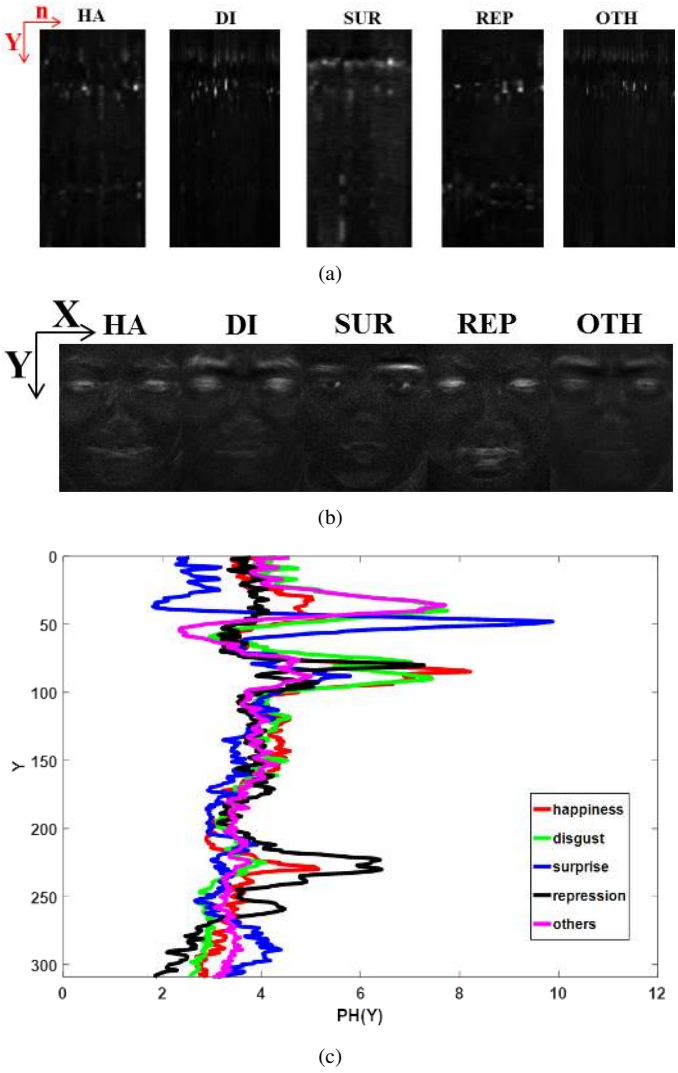

Fig. 3: Revisited integral projection on horizontal direction. (a) The horizontal integral projections of 247 subtle motion images, where n-axis describes the number of facial images and each projection is represented as a single column. (b) The mean face of the subtle motion images for happiness (HA), disgust (DI), surprise (SUR), repression (REP) and others (OTH). (c) The horizontal integral projections of the derived sparse subtle motion information in (b), where x-axis (PH(Y)) means the value of horizontal integral projection, and y-axis represents the height of an image.

$\mathbf{I}-\mathbf{Q}-\mathbf{E}$. Equation 6 is re-written as followed:

$$L(\mathbf{Q},\mathbf{E},\mathbf{Y},\mu) = \| \mathbf{Q} \|_* + \lambda \| \mathbf{E} \|_1 + \langle \mathbf{Y}, \mathbf{I}-\mathbf{Q}-\mathbf{E} \rangle + \frac{\mu}{2} \| \mathbf{I}-\mathbf{Q}-\mathbf{E} \|_F^2 . \quad (7)$$

Equation 7 can be resolved by exact ALM or inexact ALM proposed by Lin et al. [50]. A slight improvement over the exact ALM leads to the inexact ALM, which converges practically as fast as the exact ALM, but the required number of partial SVDs is significantly less. Therefore, we choose inexact ALM to extract the subtle facial motion information $\mathbf{E}$. Fig. 2 shows the subtle motion frames of $\mathbf{E}$ from an micro-expression video clip, in which it is labeled as 'surprise'. From Fig. 2(a), it is difficult for people to perceive the subtle facial movement. However, from Fig. 2(b), we can easily see the obvious movement of eyebrows (highlighted in the red rectangle). As well, it is found that identify information is mostly reduced. These possibly further improve integral projection for describing micro-expression. Based on $\mathbf{E}$ obtained from Equation 7, we only consider the integral projection on $\mathbf{E}$ instead of $\mathbf{I}$, which is formulated as,

$$\Re[f](\theta,s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{E}(x,y)\delta(x\cos\theta + y\sin\theta - s)dxdy. \quad (8)$$

In this paper, we consider the horizontal and vertical directions, because our pre-experimental results show these two directions can better describe the shape of micro-expressions than other direction. Specifically, θ on Equation 8 are 0° and 90° for horizontal and vertical directions, respectively. For evaluating discriminant ability for micro-expressions, we investigate revisited integral projection along horizontal direction in Fig. 3. In Fig. 3, we choose 247 facial images of 5-class micro-expressions from CASME2 [14], in which each image is selected at apex state of micro-expression video clip. In Fig. 3(a), the horizontal integral projections from 247 images capture the various structure of signals for different micro-expressions. Additionally, they obtain the specific structure from such regions of interest of micro-expression as mouth region for happiness expression. Moreover, as seen from Fig. 3(b), the subtle motion image obtained by RPCA well characterizes the specific regions of facial movements for different micro-expressions. For example, disgust expression mostly appears in eyebrows and eyes. Another finding in Fig. 3(c) argues the improved integral projection can preserve the discriminative structure of 1D signals for different micro-expressions. From these observations, the improved integral projection can provide more discriminative information for micro-expressions.

2.2 Spatiotemporal Local Binary Pattern based on Revisited Integral Projection

By introducing subtle motion, the revisited integral projection (RIP) preserves the shape attribute of different micro-expressions and has discriminative ability. But it is not robust to describe the appearance and motion of facial images. As LBP-TOP [26] considers micro-expression video clips from three orthogonal planes, representing appearance and motion information, respectively. We exploit the nature of LBP-TOP to obtain the appearance and motion features from the revisited integral projections.

Based on Equation 8, we can obtain the revisited integral projection signals along horizontal and vertical directions, respectively. For convenience, we denote them as $\mathbf{H}$ and $\mathbf{V}$ for horizontal and vertical directions, respectively. For the sake of simplification, we only discuss the $\mathbf{H}$ for appearance and motion features.

In [37], Houam *et al.* proposed one-dimensional local binary pattern (1DLBP) to describe the appearance information of bone texture image. Specifically, they defined the linear mask of size W as shown in Fig. 4, in which W can be designed as 3, 5, 7 or 9. With the mask, the 1DLBP code is obtained by thresholding the neighborhood values against the central element. The neighbors is assigned the value 1 if they are greater than or equal to the current element and 0 otherwise. Then each binary element of the resulting vector is multiplied by a weight depending on its position. 1DLBP can be summarized as,

$$\text{1DLBP}_\text{W} = \sum_p \delta(\mathbf{H}_p - \mathbf{H}_c)2^p, \quad (9)$$

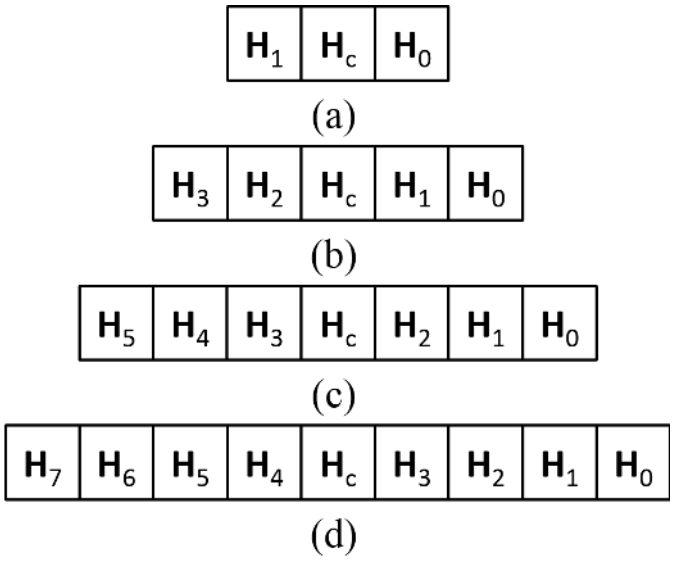

Fig. 4: Linearly symmetric neighbour sets for different values of W: (a) W=3; (b) W=5; (c) W=7; (d) W=9.

where δ is a Dirac's delta, $\mathbf{H}_c$ is the value at the center of the mask, and $\mathbf{H}_p$ is the neighbors of $\mathbf{H}_c$.

2.2.1 Spatial domain

The distribution of the 1D patterns of each frame is modeled by a histogram. It characterizes the frequency of each pattern in the 1D projected signal. This encodes the local and global texture information since the projected signal handles both cues. Fig. 5 shows the procedure to encode the integral projection by using 1DLBP. To describe the appearance of each video clip, the histograms of frames are accumulated. Finally the spatial histogram is represented by histograms $\mathbf{f}^{\mathrm{XYH}}$ and $\mathbf{f}^{\mathrm{XYV}}$ of horizontal and vertical projections.

2.2.2 Temporal domain

Motion features are extracted from horizontal and vertical direction. Firstly, we consider a simple way to extract the motion histogram along horizontal direction, as shown in Fig. 6. We formulate the horizontal integral projections from all difference images in a video clip as a new texture image, which is similar to the YT plane of LBP-TOP. It represents the motion of micro-expression video clip along vertical direction. As seen from Fig. 6, the change of value $\mathbf{S}(z)$ ($z \in [y_1, y_2]$) along the time t definitely shows the motion change of shape of micro-expressions along the vertical direction.

However, the changing rate of micro-expression video clips might be different; it might cause unfair comparison among the motion histograms. Bilinear interpolation is utilized to ensure $\mathbf{S}(z)$ along the time t with the same size T. Here we name this procedure "temporal normalization". Based on the new texture image, a gray-scale invariant texture descriptor, LBP operator [51], which is defined as

$$\mathrm{LBP}_{M,R} = \sum_{m=0}^{M-1} \delta(\mathrm{g}_m - \mathrm{g}_c) 2^m, \tag{10}$$

is exploited to extract the motion histogram, where $\mathbf{g}_c$ is the gray value of the center pixel, $\mathbf{g}_m$ is the gray value of M equally spaced pixels on a circle of radius R at this center pixel. The same procedure is applied to vertical integral projection.

Empirical experiments tell us that the procedure normalizing all images into the same size could produce a promising performance. It also allows us to use the same value of R for motion texture images. So far, the motion histograms, which represent motion change along the vertical (YT) and horizontal (XT) directions, are obtained by the process described above. Here, we denote them as $\mathbf{f}^{\mathrm{YT}}$ and $\mathbf{f}^{\mathrm{XT}}$.

The final feature vector of a micro-expression video clip can be formulated by $[\mathbf{f}^{\mathrm{XYH}}, \mathbf{f}^{\mathrm{XYV}}, \mathbf{f}^{\mathrm{XT}}, \mathbf{f}^{\mathrm{YT}}]$, where this feature preserves shape and texture information. For convenience, we abbreviate Spatiotemporal local binary pattern with revisited integral projection as "STLBP-RIP".

2.3 Enhancing discriminative ability

In general, we divide micro-expression video clip into $m \times n$ blocks in spatial domain. In the k-th block, the feature on spatial domain contains two sub-features $\mathbf{f}_k^{\mathrm{XYH}}$ and $\mathbf{f}_k^{\mathrm{XYV}}$ obtained from horizontal and vertical directions of XY plane, respectively, while the feature on temporal domain consists of $\mathbf{f}_k^{\mathrm{XT}}$ and $\mathbf{f}_k^{\mathrm{YT}}$ extracted from XT and YT planes, respectively. In practice, we concatenate these features into one feature vector for micro-expression recognition. However, all features do not contain equally discriminative information for different micro-expressions. For simplicity, for extracting discriminative of features, we may employ dimensionality reduction methods such as Linear Discriminative Analysis. However, these approaches may fail to work on micro-expression recognition due to few class number and high dimensionality of micro-expression features. Instead, we aim to extract the discriminative features from $\{\mathbf{f}_k^{\mathrm{XYH}}, \mathbf{f}_k^{\mathrm{XYV}}, \mathbf{f}_k^{\mathrm{XT}}, \mathbf{f}_k^{\mathrm{YT}}\}_{k=1}^{m\times n}$ for micro-expression recognition. For convenience, we define one sub-feature from XYH, XYV, XT or YT as a group feature. In our method, we propose a group feature selection on the basis of Laplacian method [46] and pairwise-class micro-expression to extract the discriminative information of STLBP-RIP for micro-expression recognition, since Laplacian method is based on the following observation: two data points are probably related to the same class if they are close to each other. Our method consists of two important steps: (1) formulation of dissimilarity feature and (2) computation of Laplacian scores of group features.

Formulation of dissimilarity feature: Given a micro-expression video clip, we divide it into $m \times n$ blocks in spatial domain. For the k-th block, its feature is represented by $\mathbf{f}_k = [\mathbf{f}_k^{\mathrm{XYH}}, \mathbf{f}_k^{\mathrm{XYV}}, \mathbf{f}_k^{\mathrm{XT}}, \mathbf{f}_k^{\mathrm{YT}}]$. Thus the dissimilarity of the i-th and j-th micro-expression video clips $\mathbf{F}_i$ and $\mathbf{F}_j$ on the k-th block contains the difference between group features, which is defined as followed,

$$d_k(\mathbf{F}_i, \mathbf{F}_j) = [d_k^{\mathrm{XYH}}\ d_k^{\mathrm{XYV}}\ d_k^{\mathrm{XT}}\ d_k^{\mathrm{YT}}], \tag{11}$$

where $d_k^{\mathbf{P}} = \chi^2(\mathbf{f}_{i,k}^{\mathbf{P}}, \mathbf{f}_{j,k}^{\mathbf{P}})$, $\mathbf{P}$ represents one of XYH, XYV, XT and YT, and χ^2 is Chi-square distance metric.

Based on Equation 11, the new feature from any micro-expression-pair samples $\mathbf{g} \in \Re^{m\times n\times 4}$ is formulated as $[d_1^{\mathrm{XYH}}, d_1^{\mathrm{XYV}}, d_1^{\mathrm{XT}}, d_1^{\mathrm{YT}}, \ldots, d_{m\times n}^{\mathrm{XYH}}, d_{m\times n}^{\mathrm{XYV}}, d_{m\times n}^{\mathrm{XT}}, d_{m\times n}^{\mathrm{YT}}]$, in which one dimension describes the dissimilarity of each group feature of two different samples. Its corresponding class information for $\mathbf{g}$ is labeled as followed,

$$c(\mathbf{g}) = \begin{cases} 1 & \text{if } c(\mathbf{F}_i) = c(\mathbf{F}_j) \\ -1 & \text{if } c(\mathbf{F}_i) \neq c(\mathbf{F}_j), \end{cases} \tag{12}$$

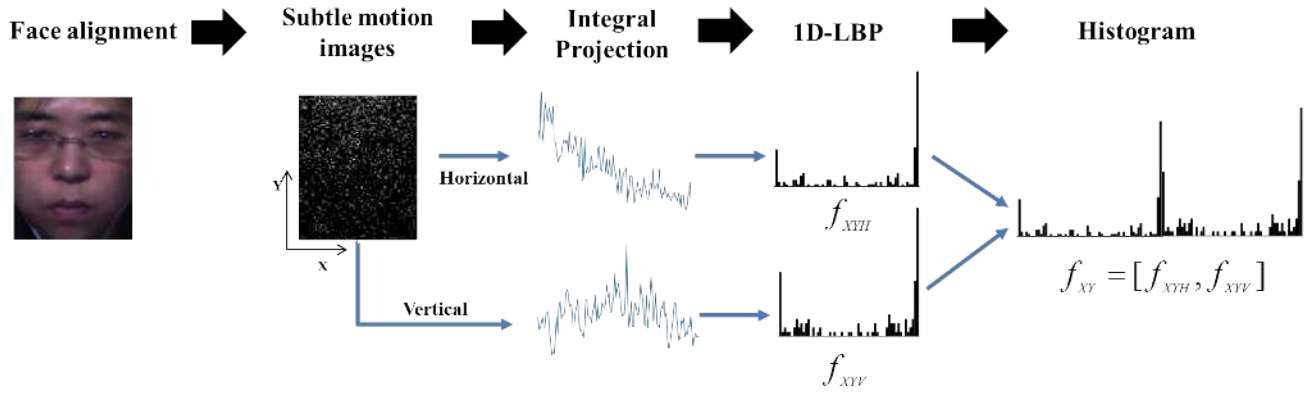

Fig. 5: Procedure of encoding a revisited integral projection on spatial domain. [44]

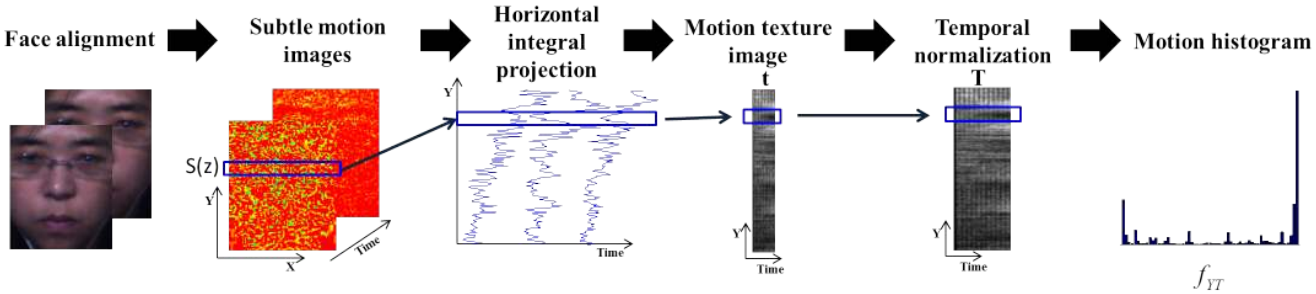

Fig. 6: Motion histogram along YT plane: The blue rectangle represents horizontal integral projection at specific range of Y-axis direction, where t and T are the original and temporal normalization time lengths, respectively. [44]

where $c(\mathbf{F}_i)$ and $c(\mathbf{F}_j)$ are the class label of two micro-expression video clips $\mathbf{F}_i$ and $\mathbf{F}_j$.

Computation of Laplacian scores of group features: We employ one-vs-one class learning strategy to obtain the discriminative group feature for two micro-expression classes. Given samples from the a-th and b-th micro-expression classes, we can formulate the dissimilarity features $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_N]$ and their corresponding labels $\mathbf{C} = [c_1, \ldots, c_N]$, where N is the number of dissimilarity features, $N = N_a(N_a - 1) + N_b N_b$, N_a and N_b are the number of samples with the a-th and b-th micro-expression classes, respectively. We construct a weighted graph $\mathcal{G}$ with edges connecting nearby points to each other, in which W_{uv} evaluates the similarity between the u-th and v-th samples. In our method, we employ the class label and Cosine metric for constructing the weight matrix $\mathbf{W}$, which models the local structure of the data space. The element of weight matrix $\mathbf{W}$ is defined as:

$$\mathbf{W} = \begin{cases} \frac{\mathbf{g}_u \cdot \mathbf{g}_v}{\|\mathbf{g}_u\| \|\mathbf{g}_v\|} & \text{if } c(\mathbf{g}_u) = c(\mathbf{g}_v) \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\cdot$ is a dot product, $\mathbf{g}_u$ and $\mathbf{g}_v$ are the u-th and v-th samples in $\mathbf{G}$, respectively. A reasonable criterion for choosing a good feature is to minimize the following object function:

$$L_r = \frac{\sum_{uv}(g_{r,u} - g_{r,v})^2 \mathbf{W}_{uv}}{Var(\mathbf{g_r})}, \tag{14}$$

where $r \in \{1, \ldots, m \times n \times 4\}$ is dimension index of feature $\mathbf{g}$, $\mathbf{g_r} = [g_r^1, g_r^2, \ldots, g_r^N]$, and $Var(\mathbf{g_r})$ is the estimated variance of the r-th feature.

For a good feature, the bigger $\mathbf{W}_{uv}$, the smaller $(g_{r,u} - g_{r,v})$. As well, by maximizing $Var(\mathbf{g_r})$, we prefer those features with large variance which have more representative power. Thus the Laplacian Score tends to be small. According to [46], $\sum_{uv}(g_{r,u} - g_{r,v})^2 \mathbf{W}_{uv}$ is written as

$$\sum_{uv}(g_{r,u} - g_{r,v})^2 \mathbf{W}_{uv} = 2\mathbf{g_r}^T D \mathbf{g_r} - 2\mathbf{g_r}^T \mathbf{W} \mathbf{g_r}, \tag{15}$$

where $D = diag(\mathbf{W1})$, $\mathbf{1} = [1, \ldots, 1]^T$ and $\mathbf{W}$ is the weight matrix containing $\mathbf{W}_{uv}$. $Var(\mathbf{g_r})$ can be estimated as follows:

$$\begin{aligned} Var(\mathbf{g_r}) &= \sum_u (g_{ru} - \mu_r)^2 \mathbf{D}_{uu} \\ &= \sum_u (g_{ru} - \frac{\mathbf{g_r}^T \mathbf{D1}}{\mathbf{1}^T \mathbf{D1}})^2 \mathbf{D}_{uu}, \end{aligned} \tag{16}$$

after removing the mean from the samples, Equation 16 is rewritten as:

$$Var(\mathbf{g_r}) = \tilde{\mathbf{g_r}}^T \mathbf{D} \tilde{\mathbf{g_r}}. \tag{17}$$

For each feature, its Laplacian score is computed to reflect its locality preserving power. Therefore, the Laplacian Score of the r-th feature as follows:

$$\beta_r = \frac{\tilde{\mathbf{g_r}}^T \mathbf{L} \tilde{\mathbf{g_r}}}{\tilde{\mathbf{g_r}}^T \mathbf{D} \tilde{\mathbf{g_r}}}. \tag{18}$$

Based on Equation 18, the Laplacian score of each group feature is calculated. The group feature with the smallest score have the strongest discriminative ability. We sort them in ascending order, and then choose the first P group features for pairwise micro-expression classes. In this subsection, the group selected features of STLBP-RIP are named as "DiSTLBP-RIP".

3 EXPERIMENTS

In this paper, we develop spatiotemporal local binary pattern based on a revisited integral projection (STLBP-RIP) and its discriminative version (DiSTLBP-RIP). In this section, we evaluate them on the Chinese Academy of Sciences Micro-expression Database (CASME) [11], CASME2 [14] and Spontaneous Micro-expression Corpus (SMIC) [12]. Due to page limitation, we refer readers to our previous work [44] about STLBP-IP which exploits image difference based approach for integral projection, and its experimental results.

3.1 Database description and protocol

The CASME dataset contains spontaneous 1,500 facial movements filmed with 60 fps camera. Among them, 195 micro-expressions were coded so that the first, peak and last frames were tagged. Referring to the work of [11], we select 171 facial micro-expression videos that contain disgust (44 samples), surprise (20 samples), repression (38 samples) and tense (69 samples) micro-expressions.

The CASME2 database consists of 247 spontaneous facial micro-expression videos with 640×480 spatial resolution. This database was collected by using a 200 fps camera. As well, participants' micro expressions were elicited in a well-controlled laboratory environment with normal illumination. The CASME2 database includes five classes of the micro-expressions in this database: happiness (32 samples), surprise (25 samples), disgust (64 samples), repression (27 samples) and others (99 samples).

The SMIC database contains 164 spontaneous micro-expressions of 16 subjects recorded in a controlled scenario by 100 fps camera with resolution of 640×480. All micro-expression videos are categorized into positive (51 samples), negative (70 samples) and surprise (43 samples) classes.

For three databases, Active Shape Model [52] is used to track the 68 facial landmarks for a facial image. Subsequently, each facial image is aligned to a canonical frame. For the CASME and CASME2 databases, the face images are cropped to 308×257 pixel size, while for the SMIC database, we crop facial images into 170×139. In the experiments, leave-one-subject-out cross validation protocol is implemented, where the samples from one subject are used for testing, the rest for training. For classification, Support Vector Machine (SVM) with Chi-Square Kernel [53] is employed, where the optimal penalty parameter is provided using the three-fold cross validation approach.

3.2 Evaluation of the revisited integral projection

In this scenario, we compare STLBP-RIP with the previous method based on original integral projection (STLBP-OIP) and our previous work (STLBP-IP) [44] on CASME, CASME2 and SMIC databases. For three databases, we conducted a comparison on 7×3 spatial blocks of micro-expression video clip. We set W as 9 and do not use temporal normalization. All comparisons are evaluated using recognition rate.

(1) On CASME database, we list the recognition rate for three spatiotemporal features, where the recognition rates of 35.67%, 54.39% and 56.14% for STLBP-OIP, STLBP-IP and STLBP-RIP, respectively. Comparing with STLBP-OIP and STLBP-IP, STLBP-RIP improves the performance by increasing substantially recognition rate of 18.72% and 1.75%, respectively.

(2) On CASME2 database, STLBP-OIP and STLBP-IP obtain the recognition rate of 42.51% and 52.63%, respectively, while STLBP-RIP achieves the recognition rate of 56.68%. It is seen that comparing with STLBP-OIP and STLBP-IP, STLBP-RIP increases the recognition rates of 10.12% and 4.05%, respectively.

(3) On SMIC database, we obtain the recognition rates of 34.15%, 45.73% and 54.88% for STLBP-OIP, STLBP-IP and STLBP-RIP, respectively. Comparing with STLBP-OIP and STLBP-IP, the performance of micro-expression recognition is considerably increased at the recognition rate of 11.58% and 9.15% by using STLBP-RIP, respectively.

It is seen that the performance is substantially improved by STLBP-RIP comparing with STLBP-OIP. It demonstrates that RPCA can better reduce the influence of subject information for integral projection. The discriminative ability of integral projection can be enhanced by our proposed method. Additionally, we see that STLBP-RIP outperforms STLBP-IP on three databases, as STLBP-RIP uses RPCA to obtain more stable motion information than STLBP-IP.

3.3 Parameter evaluation

The mask size W of 1DLBP, the radius R of LBP and the temporal normalization size T are three important parameters for STLBP-RIP, which determine the computational complexity and classification performance. Additionally, for DiSTLBP-RIP the number of selected group features P decides their performance. In this subsection, we evaluate the effects of W, R, T and P.

The mask size: We evaluate the performance of STLBP-RIP caused by various W on CASME, CASME2 and SMIC databases. In order to avoid bias, and to compare the performance of features on a more general level, spatiotemporal features are extracted by varying block number. It is noted that W relatively controls the feature extracted on spatial domain. So temporal normalization is not considered in comparing the performance of various W. The results of STLBP-RIP on three databases are presented in Table 1. It is found that the performance is boosted with increasing W when we use small block number for three databases. It is explained by that using more neighbors can provide compact and much information for robust binary pattern. But for large block number, the big W decreases the performance, since the feature will be very sparse due to less sampling points along horizontal/vertical integral projection for 1DLBP.

As seen from Table 1, for STLBP-RIP, 7×3, 6×1 and 5×6 are the optimal block number on CASME, CASME2 and SMIC databases, respectively. STLBP-RIP obtains the promising results under $W = 9, 9, 7$ on CASME, CASME2 and SMIC databases, respectively.

The radius of LBP: Based on the designed W and block numbers, we evaluate the effect of $R \in \{1, 2, 3\}$ on CASME, CASME2 and SMIC databases. Results are presented in Table 2. It is found that STLBP-RIP obtains the best recognition when $R = 3$.

The temporal normalization size: Based on the well-designed W, R and block number, we evaluate the influence of T to STLBP-RIP on CASME, CASME2 and SMIC databases. All experiments are conducted under $T \in [0, 60]$, where $T = 0$ means no temporal normalization used for STLBP-RIP. Fig. 7 shows the effect of T to STLBP-RIP on CASME, CASME2 and SMIC databases. It is seen that temporal normalization method boosts the ability of STLBP-RIP on CASME and SMIC databases, while it cannot be helpful to CASME2 database. It may be explained by that the micro-expression video clip on CASME2

TABLE 1: Effects of the mask size (W) to STLBP-RIP on CASME, CASME2 and SMIC databases, where the bold number means the best recognition rate (%).

Block Number	CASME				CASME2				SMIC			
	Mask size W				Mask size W				Mask size W			
	3	5	7	9	3	5	7	9	3	5	7	9
6×1	52.63	51.46	52.05	53.22	59.11	58.70	61.94	**62.75**	49.39	46.95	49.39	48.78
7×3	54.97	54.39	55.56	**56.14**	53.44	51.82	53.04	52.23	49.39	53.05	47.56	54.88
5×5	52.05	53.22	50.88	47.95	51.82	48.58	49.39	51.01	50.00	50.61	55.49	52.44
5×6	54.39	54.39	50.88	52.05	54.66	52.23	52.23	54.25	57.94	56.10	**59.76**	58.54
8×8	43.27	42.69	42.11	43.86	51.42	51.42	43.93	50.20	49.39	50.61	52.44	52.44

TABLE 2: Performance of STLBP-RIP using different radius of LBP (R) on CASME, CASME2 and SMIC databases.

Database	R		
	1	2	3
CASME	50.88	52.05	56.14
CASME2	58.70	59.92	62.75
SMIC	50.00	53.05	59.76

database is recorded by using 200 fps camera, which provides enough temporal information. As examined from Fig. 7, STLBP-RIP obtains the highest performance on CASME and SMIC databases by using $T = 25$.

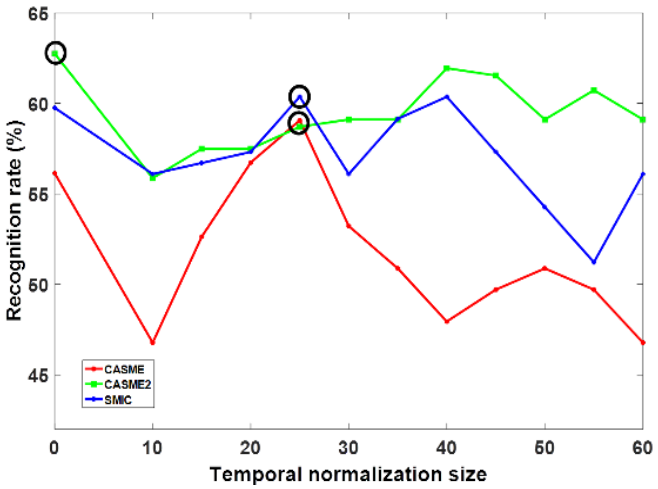

Fig. 7: Influence evaluation of temporal normalization size to STLBP-RIP, where the black circle means the best result.

The number of group features: We evaluate the Laplacian method to STLBP-RIP on CASME, CASME2 and SMIC databases. The effect of the number of group features on three database is presented in Fig. 8.

(1) CASME: For two features, 84 ($7 \times 3 \times 4$) group features are available for feature selection as micro-expression video clip is divided into 7×3 across spatial domain. As shown in Fig. 8(a), the performance of STLBP-RIP is substantially improved with increasing P. It means that more group features can provide discriminative information, but no need to include all group features. It is noted that DiSTLBP-RIP achieves 64.33% with 74 group features. The performance is improved at increased recognition rate of 3.04% comparing with STLBP-RIP.

(2) CASME2: Since we divide facial images into 6×1 blocks in spatial domain, we have 24 group features for STLBP-RIP. From Fig. 8(b), we can see that the recognition rate is increased with a large number of group features. DiSTLBP-RIP achieves 64.78% with 21 group features. Comparing with STLBP-RIP, the performance is improved at increased recognition rate of 2.03%.

(3) SMIC: As we divide facial image into 5×6 for STLBP-RIP, there exists 120 group features. As seen in Fig. 8(c), DiSTLBP-RIP achieves 63.41% with 30 group features. Comparing with STLBP-RIP, the performance is improved at increased recognition rate of 3.04%.

It shows that Laplacian method can enhance the discriminative ability of STLBP-RIP. Moreover, with promising group features, the computational efficiency becomes better, because Laplacian method reduces the dimensionality of spatiotemporal features.

TABLE 3: Comparison with two feature learning strategies on CASME, CASME2 and SMIC databases.

Method	Recognition rate		
	CASME	CASME2	SMIC
AdaBoost [45]	54.97	53.85	53.05
FLD [45]	50.06	61.94	55.49
DiSTLBP-RIP	64.33	64.78	63.41

Zhao *et al.* [45] presented to use AdaBoost and Fisher linear discriminant (FLD) to select the discriminative slices for facial expression recognition. For evaluating our method, we compare Laplacian method with two feature learning strategies [45]. The results are presented in Table 3. We observe that AdaBoost algorithm failed to work for STLBP-RIP on three databases. Instead, FLD substantially improves STLBP-RIP. However, its performance is worse than Laplacian method. Comparisons demonstrate that Laplacian method can enhance the discriminative ability of spatiotemporal feature descriptor better than two feature selection methods presented in [45].

3.4 Algorithm Comparison

In this subsection, we compare STLBP-RIP and DiSTLBP-RIP with the state-of-the-art algorithms on CASME, CASME2 and SMIC databases.

3.4.1 CASME Database

In this scenario, we firstly compare our method with LBP-TOP, completed local binary pattern from three orthogonal planes (CLBP-TOP) [54], local ordinary contrast pattern from three orthogonal planes (LOCP-TOP) [55], spatiotemporal local monogenic binary pattern(STLMBP) [56], LBP-SIP [31], spatiotemporal cuboids descriptor (Cuboids) [57] and spatiotemporal completed local quantized pattern (STCLQP) [34]. Following the parameters setup of [34], we re-implement all comparative methods on CASME database using SVM based on linear kernel, where we divided micro-expression video clip into 8×8 blocks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2713359, IEEE Transactions on Affective Computing

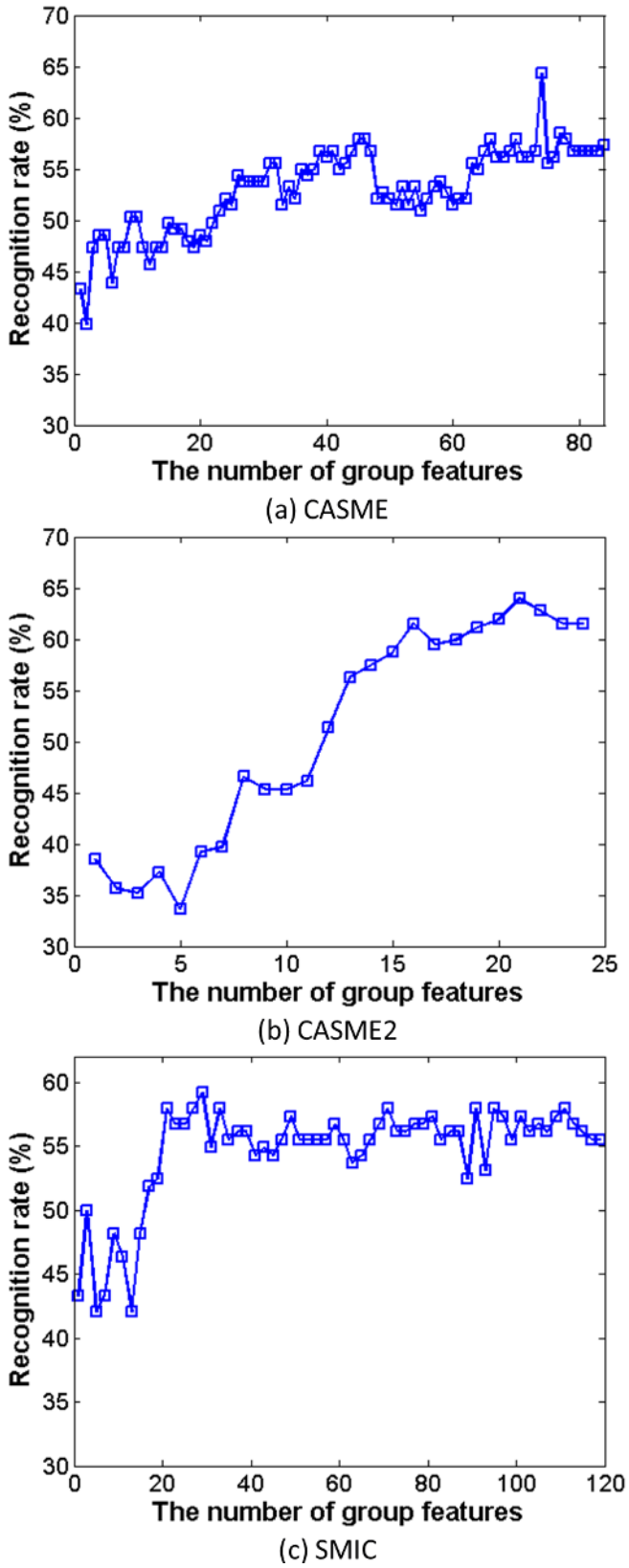

Fig. 8: The influence of the number of group features to STLBP-RIP for (a) CASME, (b) CASME2 and (c) SMIC databases.

Results on recognition rate are reported in Table 4. As seen from the table, LBP-TOP only achieves the recognition rate of 37.43%. LOCP-TOP works worst among all methods. In the comparison algorithms, we see that STCLQP obtains the promising performance on all comparison methods, followed closedly by STLMBP and CLBP-TOP, and more distantly by Cuboids. The reason may be that STCLQP provides more useful information than LBP-TOP and LBP-SIP for micro-expression recognition, as STCLQP extracts completed information through sign, magnitude and orientation. But our proposed method STLBP-RIP works better than STCLQP, which is increased by 1.75%. As well, Laplacian score method further boosts STLBP-RIP, which reaches the best recognition rate of 64.33% over all methods.

TABLE 4: Micro-expression recognition rates (%) in CASME database. The text in bold means our proposed methods.

Methods	Block Number	Recognition rate
LBP-TOP [26]	8×8	37.43
LBP-SIP [31]	8×8	36.84
STCLQP [34]	8×8	57.31
CLBP-TOP [54]	8×8	45.31
STLMBP [56]	8×8	46.20
LOCP-TOP [55]	8×8	31.58
Cuboids [57]	-	33.33
STLBP-RIP	7×3	**59.06**
DiSTLBP-RIP	7×3	**64.33**

Furthermore, we compare STLBP-RIP and DiSTLBP-RIP with the state-of-the-art works on CASME database. The comparative results are reported in Table 5. Although the experimental setups in compared algorithms are different, they still give an indication of the discriminative power of each approach. As we can see, MDMO [58] achieves better performance than ours. It may be explained by that Liu *et al.* simplified the micro-expression categorization proposed by [11] (Disgust, Surprise, Repression and Tense) into four general types (Positive, Negative, Surprise and Others). Additionally, STLBP-RIP and DiSTLBP-RIP obtain more considerable performance than HOFF-ROI [58] by increasing recognition rate of 3.37% and 8.65%, respectively. As well, STLBP-RIP and DiSTLBP-RIP outperform FDM by 2.29% and 8.19%, respectively. The comparisons demonstrate that STLBP-RIP and DiSTLBP-RIP can obtains the decent and comparative results.

The confusion matrix of five micro-expressions is shown in Fig. 9, where we compare our methods with STCLQP. It is found that STLBP-RIP performs better on recognizing Surprise and Tense classes, while it works worse than STCLQP on recognizing Disgust and Repression. As seen from Fig. 9(c), recognition performance on Digust and Repression classes is significantly improved by considering discriminative group features for STLBP-RIP. Additionally, we see that Repression and Tense classes are very hard to DiSTLBP-RIP, as they are falsely classified into opposite class. Perhaps it is because Repression and Tense samples are quite similar on CASME database. They are more difficult to be recognized than Disgust and Surprise.

3.4.2 CASME2 Database

We compare the recognition rate of our method with the baseline algorithm [14], LBP-TOP [26], LBP-SIP [31], LOCP-TOP [55]. The parameter setup for each method is described as followed:

(1) Following experimental setup of [14], we implement LBP-TOP on 5×5 facial blocks, using radius 3 for LBP operator for three orthogonal planes. For classification, we employ linear-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2713359, IEEE Transactions on Affective Computing

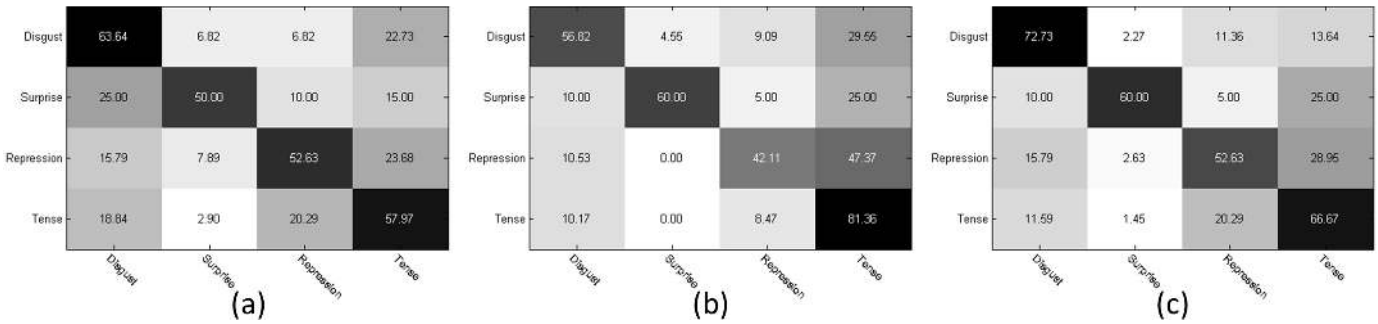

Fig. 9: The confusion matrix of (a) STCLQP, (b) STLBP-RIP, and (c) DiSTLBP-RIP for four micro-expression categorizations on CASME database.

TABLE 5: Performance comparison with the state-of-the-art methods on CASME database, where LOSO and LOO represent leave-one-subject-out and leave-one-sample-out cross validation protocols, respectively. The text in bold shows the results of our methods.

Methods	Protocol	Task	Recognition rate (%)
FDM [10]	LOSO	Contempt, Disgust, Fear, Happiness, Repression, Sadness, Surprise	56.14
MPCA [11]	20-fold	Disgust, Surprise, Repression and Tense	41.01
HOOF-whole [58]	LOSO	Positive, Negative, Surprise and Others	49.70
HOOF-ROI [58]	LOSO	Positive, Negative, Surprise and Others	55.69
MDMO [58]	LOSO	Positive, Negative, Surprise and Others	68.26
DTCM+RF [59]	LOO	Positive, Negative, Surprise and Ambiguous	64.95
STLBP-RIP	LOSO	Disgust, Surprise, Repression and Tense	**59.06**
DiSTLBP-RIP	LOSO	Disgust, Surprise, Repression and Tense	**64.33**

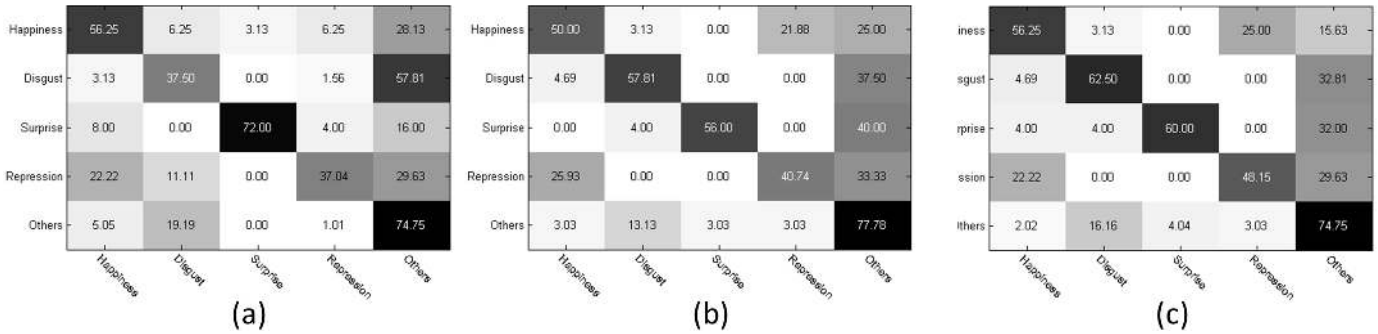

Fig. 10: The confusion matrix of (a) STCLQP, (b) STLBP-RIP and (c) DiSTLBP-RIP for micro-expression categorizations on CASME2 database.

TABLE 6: Comparison under micro-expression recognition rate on CASME2 database. The text in bold means our proposed methods and * means that we directly took the results from their works.

Methods	Block Number	Recognition rate (%)
Baseline [14]	5×5	38.87
LBP-TOP [26]	8×8	39.68
LBP-SIP [31]	8×8	40.08
STCLQP* [34]	8×8	58.39
HIGO-TOP* [35]	8×8	55.87
HOG-TOP* [35]	8×8	57.49
LOCP-TOP [55]	8×8	42.11
STLBP-RIP	6×1	**62.75**
DiSTLBP-RIP	6×1	**64.78**

kernel based SVM [53]. For convenience, we name this method as Baseline.

(2) We implement the framework of [12] based on LBP-TOP [26], LBP-SIP [31] and LOCP-TOP [55] as a comparison. Features are extracted on 8×8 facial blocks. According to [12], we firstly use temporal interpolation method [66] to interpolate all videos into 15 frames. Then we implement spatiotemporal features, where the radius and number of neighbors are 3 and 8, respectively. Support Vector Machine (SVM) with Chi-Square Kernel [53] is used, where the optimal penalty parameter is provided using the three-fold cross validation approach.

Comparative performance are presented in Table 6, where the results of STCLQP [34], HIGO-TOP [35] and HOG-TOP [35] are directly extracted from their works. As can be seen, STLBP-RIP is shown to outperform the re-implementation of [14]. Its recognition rate is increased by 23.88%. Comparing with LOCP-TOP and LBP-SIP, STLBP-RIP increases the performance by 20.64% and 22.67% for micro-expression recognition, respectively. These results demonstrate that STLBP-RIP achieves better performance than LOCP-TOP and LBP-SIP. This is explained by STLBP-RIP preserves the shape for texture descriptor by using the improved

TABLE 7: Performance comparison with the state-of-the-art methods on CASME2 database, where LOSO and LOO represent leave-one-subject-out and leave-one-sample-out cross validation protocols, respectively. The text in bold indicates the results of our methods.

Methods	Protocol	Task	Recognition rate (%)
LBP-TOP [14]	LOO	Happiness, Surprise, Disgust, Repression and Others	63.41
TICS [29]	LOO	Happiness, Surprise, Disgust, Repression and Others	61.76
LSDF [30]	LOO	Happiness, Surprise, Disgust, Repression and Others	65.44
LBP-SIP [31]	LOO	Happiness, Surprise, Disgust, Repression and Others	67.21
CUDA based DCNN [60]	LOO	Happiness, Surprise, Disgust, Repression and Tense	64.9
Local LBP-TOP+Local OF statistics [61]	LOO	Happiness, Surprise, Disgust, Repression and Others	65.50
OSW-LBP-TOP [62]	LOO	Happiness, Surprise, Disgust, Repression and Others	66.40
Sparse sampling [9]	LOSO	Happiness, Surprise, Disgust, Repression and Others	49
FDM [10]	LOSO	Happiness, Surprise, Disgust, Repression, Fear, Sadness and Others	45.93
CNN+SFS [23]	LOSO	Happiness, Surprise, Disgust, Repression and Others	47.3
Monogenic Riesz Wavelet [33]	LOSO	Happiness, Surprise, Disgust, Repression and Others	46.15
STCLQP [34]	LOSO	Happiness, Surprise, Disgust, Repression and Others	58.39
HIGO-TOP [35]	LOSO	Happiness, Surprise, Disgust, Repression and Others	55.87
HOG-TOP [35]	LOSO	Happiness, Surprise, Disgust, Repression and Others	57.49
MDMO [58]	LOSO	Positive, Negative, Surprise and Others	67.37
Local LBP-TOP+RF [61]	LOSO	Happiness, Surprise, Disgust, Repression and Others	43.92
AdaBoost+STM [63]	LOSO	Happiness, Surprise, Disgust, Repression and Others	43.78
MMFL [64]	LOSO	Happiness, Surprise, Disgust, Repression and Others	57.61
CNN+LSTM [65]	LOSO	Happiness, Surprise, Disgust, Repression and Others	60.98
STLBP-RIP	LOSO	Happiness, Surprise, Disgust, Repression and Others	**62.75**
DiSTLBP-RIP	LOSO	Happiness, Surprise, Disgust, Repression and Others	**64.78**

integral projection. Additionally, STLBP-RIP achieves better than STCLQP, HOG-TOP and HIGO-TOP in the recogntiion rate. Comparing with LBP-TOP, DiSTLBP-RIP obtains a significant improvement on micro-expression recognition, since DiSTLBP-RIP extracts the discriminative ability of STLBP-RIP by using feature selection.

Table 7 shows a comparison to some other dynamic analysis approaches using the recognition rates given in each paper. It should be noted that the results are not directly comparable due to different experimental setups and so forth, but they still give an indication of the discriminative power of each approach. It is seen that LBP-SIP [31] obtained the best recognition rate of 67.21% among the algorithms under leave-one-sample-out cross validation protocol. However, it is observed from Table 6 LBP-SIP had the recognition rate of 40.08%. It demonstrates that our methods perform better than all the algorithms under leave-one-sample-out cross validation protocol. Additionally, under the leave-one-subject-out cross validation protocol, our approaches outperform the other methods in almost all cases. Algorithm comparisons on CASME2 database indicate that STLBP-RIP and DiSTLBP-RIP is promisingly competitive to all methods.

The confusion matrices of STCLQP and our methods are shown in Fig. 10. Comparing with STCLQP, STLBP-RIP achieves better performance on three micro-expression classes (Disgust, Repression and Other), while STLCQP outperforms it on recognizing Surprise and Happiness. Our another method, DiSTLBP-RIP outperforms STLCQP at the most of micro-expression classes except Surprise. Unfortunately, DiSTLBP-RIP makes falsely classification of Surprise to Other class. It may be explained that Other class includes some confused micro-expressions similar to Surprise class. From these comparisons, we see that DiSTLBP-RIP has a promising ability to recognize five micro-expressions on CASME2 database, followed by STLBP-RIP and STCLQP.

TABLE 8: Micro-expression recognition rates (%) in SMIC database. The text in bold means our proposed methods and * means that we directly extracted the result from their work.

Methods	Block number	Recognition rate
Baseline* [12]	8×8	48.78
LBP-TOP [26]	5×6	42.07
LBP-SIP [31]	5×6	43.29
STCLQP* [34]	8×8	64.02
HIGO-TOP* [35]	6×6	59.15
HOG-TOP* [35]	2×2	57.93
LOCP-TOP [55]	5×6	43.90
Cuboids [57]	-	37.08
LCRF [67]	-	33.54
GCRF [67]	-	32.93
DOL [68]	-	20.12
STLBP-RIP	5×6	**60.37**
DiSTLBP-RIP	5×6	**63.41**

3.4.3 SMIC Database

For SMIC database, we compare our methods with the commonly used spatiotemporal features [26], [31], [55] and feature descriptor based on temporal model [67], [68]. In our implementation, we used temporal interpolation method (TIM) to normalize each video into 10 frames. As comparison, we use LBP-TOP, LOCP-TOP and LBP-SIP on 5×6 facial blocks for micro-expression recognition. We employ spatiotemporal cuboids feature of [57] for comparison, where we use k-nearest-neighbor (KNN) classification. We use the same parameter setup to [57]. Finally, we employ LBP features with conditional random field (LCRF) [67], geometric features with CRF (GCRF) [67], dense optical flow with hidden markov model (DOL) [68], for comparison.

The comparison results are reported in Table 8. The temporal models with appearance and shape features (LCRF [67], GCRF [67], DOL [68]) work poorly in micro-expression recognition. Among the temporal model, LCRF [67] gets the best one

of 33.54% for micro-expression recognition. Among all comparative algorithms, STCLQP [34] obtains the best recognition rate, followed by baseline algorithm provided by [12]. Comparing with LBP-TOP [26], STLBP-RIP increases the recognition rate of 18.3% for micro-expression recognition. Comparing with LOCP-TOP [55], the micro-expression recognition performance is increased by 16.47% for STLBP-RIP. These results demonstrate that STLBP-RIP achieves better performance than geometric features and three spatiotemporal features. Additionally, DiSTLBP-RIP further improves the performance of micro-expression recognition.

TABLE 9: Performance comparison with the state-of-the-art methods on SMIC database, where LOSO and LOO represent leave-one-subject-out and leave-one-sample-out cross validation protocols, respectively. The text in bold means the results of our methods.

Methods	Protocol	Recognition rate (%)
FDM [10]	LOSO	54.88
Baseline [12]	LOSO	48.78
CNN+SFS [23]	LOSO	53.60
STCLQP [34]	LOSO	64.02
HOG-TOP [35]	LOSO	57.93
HIGO-TOP [35]	LOSO	59.15
OSW-LBP-TOP [62]	LOSO	53.05
AdaBoost+STM [63]	LOSO	44.34
MMFL [64]	LOSO	62.33
OS+Wiener filter [69]	LOSO	53.56
DLSTD [30]	LOO	67.68
CUDA based DCNN [60]	LOO	65.85
STLBP-RIP	LOSO	**60.37**
DiSTLBP-RIP	LOSO	**63.41**

Finally, we compare STLBP-RIP and DiSTLBP-RIP with the state-of-the-art works on SMIC database. Table 9 shows the comparative results on SMIC database, where we straightforwardly extracted the results and protocols from their works. Under the subject independent protocol, i.e., leave-one-subject-out, we can see that the STCLQP [34] achieves the highest recognition rate among all methods, followed by our DiSTLBP-RIP, because STLBP-RIP and DiSTLBP-RIP only used sign information for micro-expression recognition, while STCLQP exploits magnitude, orientation and sign information. As well, STLCQP is very restricted by the codebook learning procedure. Instead, STLBP-RIP and DiSTLBP-RIP have simple but efficient way for micro-expression recognition. With easier leave-one-sample-out protocol, the works of [30], [60] obtained the recognition rate of 67.68% and 65.85%, respectively. From comparative results, we can see that STLBP-RIP and DiSTLBP-RIP can obtain the promising and competitive performance.

4 Conclusion

In the paper, we have shown the new spatiotemporal local binary pattern improved by a revisited integral projection for micro-expression recognition. Additionally, we propose the discriminative method to boost the performance of this spatiotemporal feature descriptor. The novel feature and its discriminative one can achieve the state-of-the-art performance on three facial micro-expression databases. Specifically, we firstly develop a revisited integral projection method to preserve the shape property of micro-expressions and then enhance discrimination of the features for micro-expression recognition. Furthermore, we have presented to use local binary pattern operators to further describe the appearance and motion changes from horizontal and vertical integral projections, well suited for extracting the subtle micro-expressions. Based on Laplacian method, discriminative group features are explored for further enhancing discriminative capability of STLBP-RIP. Experiments on three facial micro-expression databases demonstrate our methods outperform the state-of-the-art methods on micro-expression recognition.

Recently, the deep learning has become popular in computer vision, also in affective computing. Several works on [23], [60], [65] attempted to use the deep learning framework for micro-expression recognition. However, it is seen that their performance is still far from that of the conventional feature descriptors. This gap motivates us to develop promising deep learning methodology for improving the performance of micro-expression recognition. In future, we will focus on the deep learning and its combination with the conventional feature descriptors.

References

[1] 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.
[2] P. Ekman and W. Friesen, "Detecting deception from emotional and unemotional cues," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.
[3] X. Shen, Q. Wu, and X. Fu, "Effects of the duration of expressions on the recognition of microexpressions," *Journal of Zhejiang University Science B*, vol. 3, no. 13, pp. 221–230, 2012.
[4] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *International Communication Association*, 2009.
[5] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," in *Proc. WACV*, 2009, pp. 1–6.
[6] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-exprssion spotting in long videos using spatio-temporal strain," in *Proc. AFGR*, 2011, pp. 51–56.
[7] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in *ICDP*, 2009.
[8] Q. Wu, X. Shen, and X. Fu, "The machine knows what you are hiding: an automatic micro-expression recognition system," in *Affective Computing and Intelligence Interaction*, 2011, pp. 152–162.
[9] A. Le, S. Liong, J. See, and R. Phan, "Sparsity in dynamics of spontaneous subtle emotion: analysis & application," *IEEE Trans. on Affective Computing*, no. 99, pp. –, 2016.
[10] F. Xu, J. Zhang, and J. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. on Affective Computing*, no. 99, pp. –, 2016.
[11] W. Yan, Q. Wu, Y. Liu, S. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. AFGR*, 2013, pp. 1–7.
[12] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. AFGR*, 2013.
[13] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. ICCV*, 2011, pp. 1449–1456.
[14] W. Yan, X. Li, S. Wang, G. Zhao, Y. Liu, Y. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLOS ONE*, vol. 9, no. 1, pp. 1–8, 2014.
[15] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," *IEEE Trans. on PAMI*, vol. 31, no. 1, pp. 39–58, 2009.
[16] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. on Cybernetics*, vol. 44, no. 2, pp. 161–174, 2014.
[17] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. on Cybernetics*, vol. 45, no. 8, pp. 1499–1510, 2015.
[18] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, 2015, pp. 2983–2991.

[19] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *ICMI*, 2015, pp. 435–442.

[20] A. Majumder, L. Behera, and V. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Trans. on Cybernetics*, no. 99, pp. 1–12, 2016.

[21] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Trans. on Cybernetics*, vol. 46, no. 4, pp. 916–929, 2016.

[22] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural network," *IEEE Trans. on Cybernetics*, vol. 34, no. 3, pp. 1588–1595, 2004.

[23] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression," in *Proc. ICPR*, 2016, pp. 2259–2264.

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

[25] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary pattern: a comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[26] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary pattern with an application to facial expressions," *IEEE Trans. on PAMI*, vol. 29, no. 6, pp. 915–928, 2009.

[27] A. Davison, M. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: an investigation on spatio-temporal descriptors," in *ECCV workshop on SFBA*, 2014.

[28] J. Ruiz-Hernandez and M. Pietikäinen, "Encoding local binary patterns using re-parameterization of the second order gaussian jet," in *Proc. AFGR*, 2013.

[29] S. Wang, W. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proc. ICPR*, 2014.

[30] S. Wang, W. Yan, G. Zhao, and X. Fu, "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *ECCV workshop on SFBA*, 2014.

[31] Y. Wang, J. See, R. Phan, and Y. Oh, "LBP with six interaction points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Proc. ACCV*, 2014.

[32] Y. Guo, C. Xue, Y. Wang, and M. Yu, "Mirco-expression recognition based on cbp-top feature with elm," *International Journal for Light and Electro Optics*, vol. 127, no. 4, p. 2404, 2009.

[33] X. He, D. Cai, and P. Niyogi, "Monogenic riesz wavelet representation for micro-expression recognition," in *IEEE DSP*, 2015, pp. 1237–1241.

[34] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, no. 175, pp. 564–578, 2016.

[35] 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. on Affective Computing*, pp. –, 2017.

[36] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 833–851, 2008.

[37] L. Houam, A. Hafiane, A. Boukrouche, E. Lespessailles, and R. Jennane, "One dimensional local binary pattern for bone texture characterization," *Pattern Analysis and Applications*, vol. 17, pp. 179–193, 2014.

[38] A. Benzaoui and A. Boukrouche, "Face recognition using 1DLBP texture analysis," in *FUTURE COMPUTING*, 2013, pp. 14–19.

[39] A. Benzaoui, "Face recognition using 1DLBP, DWT and SVM," in *CEIT*, 2015, pp. 1–6.

[40] D. Robinson and P. Milanfar, "Fast local and global projection-based methods for affine motion estimation," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 35–54, 2003.

[41] G. Mateos, "Refining face tracking with integral projection," in *Proc. AVBPA*, 2003, pp. 360–368.

[42] G. Mateos, A. Ruiz-Garcia, and P. Lopez-de Teruel, "Human face processing with 1.5d model," in *Proc. AMFG*, 2007, pp. 220–234.

[43] G. Carcia-Mateos, A. Ruiz, and P. L. de Teruel, "Face detection using integral projection models," in *Joint IAPR International Workshops SSPR 2002 and SPR 2002 Windsor*, 2002, pp. 644–653.

[44] X. Huang, S.-J. Wang, G. Zhao, and M. Pietikäinen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proc. ICCV Workshop*, 2015, pp. 1–9.

[45] G. Zhao and M. Pietikäinen, "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition," *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1117–1127, 2009.

[46] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. NIPS*, 2005.

[47] W. Yan, Q. W. an J. Liang, Y. 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.

[48] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: exact recovery of corrupted low-rank matrices via context optimization," in *Proc. NIPS*, 2009, pp. 2080–2088.

[49] X. Mao, Y. Xue, Z. Li, K. Huang, and S. Lv, "Robust facial expression recognition based on rpca and adaboost," in *Proc. WIAMIS*, 2009.

[50] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. NIPS*, 2011.

[51] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texutre classification with local binary pattern," *IEEE Trans. on PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

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

[53] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, vol. 2, pp. 27:1–27:27, 2011.

[54] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Proc. ICCV*, 2011, pp. 868–875.

[55] C. Chan, B. Goswami, J. Kittler, and W. Christmas, "Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication," *IEEE TIFS*, vol. 2, no. 7, pp. 602–612, 2012.

[56] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Spatiotemporal local monogenic binary patterns for facial expression recognition," *IEEE Signal Processing Letters*, vol. 5, no. 19, pp. 243–246, 2012.

[57] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. VSPETS*, 2005, pp. 65–72.

[58] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. on Affective Computing*, no. 99, p. 1, 2016.

[59] Z. Lu, Z. Luo, H. Zheng, J. Chen, and W. Li, "A delaunay-based temporal coding model for micro-expression recognition," in *Proc. ACCV workshop*, 2014, pp. 698–711.

[60] V. Mayya, R. Pai, and M. Pai, "Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences," in *Proc. ICACCI*, 2016, pp. 699–703.

[61] S. Zhang, B. Feng, Z. Chen, and X. Huang, "Micro-expression recognition by aggregating local spatio-temporal patterns," in *Proc. MMM*, 2017, pp. 638–648.

[62] S. Liong, J. See, R. Phan, A. Nego, Y. Oh, and K. Wong., "Subtle expression recognition using optical strain weighted features," in *Proc. ACCV workshop*, 2014, pp. 644–657.

[63] A. Ngo, R. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Proc. ACCV*, 2014, pp. 33–48.

[64] J. He, J. Hu, X. Lu, and W. Zheng, "Mult-task mid-level feature learning for micro-expression recognitionn," *Pattern Recognition*, vol. 66, pp. 44–52, 2017.

[65] D. Kim, W. Baddar, and Y. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representation," in *Proc. Multimedia*, 2016, pp. 382–386.

[66] Z. Zhou, G. Zhao, Y. Guo, and M. Pietikäinen, "An image-based visual speech animation system," *IEEE TCSVT*, vol. 22, no. 10, pp. 1420–1432, 2012.

[67] S. Jain, C. Hu, and J. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Proc. ICCV*, 2011, pp. 1642–1649.

[68] G. Shin and J. Chun, "Spatio-temporal facial expression recognition using optical flow and hmm," *Software Engineering, Artificial Intelligence, Network, and Parallel/Distributed Computing*, pp. 27–38, 2008.

[69] S. Liong, R. Phan, J. See, Y. Oh, and K. Wong, "Optical strain based recognition of subtle emotions," in *Proc. ISPACS*, 2014, pp. 180–184.

Xiaohua Huang received the B.S. degree in communication engineering from Huaqiao University, Quanzhou, China in 2006. He received his Ph.D degree in Computer Science and Engineering from University of Oulu, Oulu, Finland in 2014. He was a research assistant in Southeast University, Nanjing, China in 2006-2012. He has been a scientist researcher in the Center for Machine Vision and Signal Analysis at University of Oulu since 2015. He has authored or co-authored more than 20 papers in journals and conferences, and has served as a reviewer for journals and conferences. His current research interests include facial expression recognition, micro-expression analysis, group-level emotion recognition, multi-modal emotion recognition and texture classification.

Sujing Wang (M'12) received the Master's degree from the Software College of Jilin University, Changchun, China, in 2007. He received the Ph.D. degree from the College of Computer Science and Technology of Jilin University in 2012. He was a postdoctoral researcher in Institute of Psychology, Chinese Academy of Sciences from 2012 to 2015. He is now an Assistant Researcher in Institute of Psychology, Chinese Academy of Sciences. He has published more than 40 scientific papers. He is One of Ten Selectees of the Doctoral Consortium at International Joint Conference on Biometrics 2011. He was called as *Chinese Hawkin* by the Xinhua News Agency. His current research interests include pattern recognition, computer vision and machine learning. He serves as an associate editor of Neurocomputing (Elsevier).

Xin Liu (M'16) received the Ph.D. degree in information and communication engineering from Xian Jiaotong University, Xian, China in 2016. He is currently a Researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His research interests include human behavior analysis, image restoration, and object detection.

Guoying Zhao (SM12) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently an Associate Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where she has been a senior researcher since 2005. In 2011, she was selected to the highly competitive Academy Research Fellow position. She has authored or co-authored more than 160 papers in journals and conferences. Her papers have currently over 5700 citations in Google Scholar (h-index 33).She has served as area chairs for FG 2017 and WACV 2017 and is associate editor for Pattern Recognition and Image and Vision Computing Journals. She has lectured tutorials at ICPR 2006, ICCV 2009, and SCIA 2013, authored/edited three books and six special issues in journals. Dr. Zhao was a Co-Chair of 12 International Workshops at ECCV, ICCV, CVPR and ACCV, and two special sessions at FG13 and FG15. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, gait analysis, dynamic-texture recognition, human motion analysis, and person identification. Her research has been reported by Finnish TV programs, newspapers and MIT Technology Review.

Xiaoyi Feng received the B.S. degree in signal processing from Northwest University, Xi'an, China, in 1991, and the Ph.D. degree in Navigation, Guidance and Control from Northwestern Polytechnical University, Xi?an, China, in 2001. In 1994, she joined the School of Electronics and Information, Northwestern Polytechnical University, as an Assistant Professor, and she is currently a Professsor in this school. She was nominated as one of the ?New Century Excellent Talents? and sponsored by the ?Program for New Century Excellent Talents in University? by the Ministry of Education in 2008. Her interests include image processing, computer vision, and affective computing.

Matti Pietikäinen received his Doctor of Science in Technology degree from the University of Oulu, Finland. He is currently a professor, Scientific Director of Infotech Oulu and Director of Center for Machine Vision Research at the University of Oulu. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory at the University of Maryland. He has made pioneering contributions, e.g. to local binary pattern (LBP) methodology, texture-based image and video analysis, and facial image analysis. He has authored over 340 refereed papers in international journals, books, and conferences. His research is frequently cited, and its results are used in various applications around the world. Dr. Pietikinen was Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence,IEEE Transactions on Forensics and Security and Pattern Recognition journals, and currently serves as Associate Editor of Image and Vision Computing journal. He was President of the Pattern Recognition Society of Finland from 1989 to 1992, and was named its Honorary Member in 2014. From 1989 to 2007 he served as Member of the Governing Board of International Association for Pattern Recognition (IAPR), and became one of the founding fellows of the IAPR in 1994. He is IEEE Fellow for contributions to texture and facial image analysis for machine vision. In 2014, his research on LBP-based face description was awarded the Koenderink Prize for Fundamental Contributions in Computer Vision.