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/TMM.2018.2820321, IEEE Transactions on Multimedia

Learning from Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition

Yuan Zong, Xiaohua Huang, Wenming Zheng, *Member, IEEE,* Zhen Cui, *Member, IEEE,* and Guoying Zhao, *Senior Member, IEEE*

***Abstract*—Micro-expression recognition aims to infer genuine emotions which people try to conceal from facial video clips. It is a very challenging task because micro-expressions have very low intensity and short duration, which makes micro-expressions difficult to observe. Recently, researchers have designed various spatiotemporal descriptors to describe micro-expressions. It is notable that for better capturing the low-intensity facial muscle movement, a fixed spatial division grid, 8×8 for example, is commonly used to partition the facial images into a few facial blocks before extracting descriptors. However, it is hard to choose an ideal division grid for different micro-expression samples because the division grids affect the discriminative ability of spatiotemporal descriptors to distinguish micro-expressions. To address this problem, in this paper we design a hierarchical spatial division scheme for spatiotemporal descriptor extraction. By using the proposed scheme, it would not be a problem to determine which division grid is most suitable regarding different micro-expression datasets. Furthermore, we propose a kernelized group sparse learning (KGSL) model to process hierarchical scheme based spatiotemporal descriptors such that they are more effective for micro-expression recognition tasks. To evaluate the performance of the proposed micro-expression recognition method consisting of the hierarchical scheme based spatiotemporal descriptors and KGSL, extensive experiments are conducted on two public micro-expression databases: CASME II and SMIC. Compared with many recent state-of-the-art approaches, our method achieves more promising recognition results.**

***Index Terms*—Micro-expression recognition, spatiotemporal descriptor, hierarchical spatial division, group sparse learning, kernelized group sparse learning**

Manuscript received December 22, 2016; revised June 25, 2017; accepted March 12, 2018. This work was supported by the National Basic Research Program of China under Grant 2015CB351704, the National Natural Science Foundation of China under Grant 61572009, Grant 61772276, and Grant 61602244, the Jiangsu Provincial Key Research and Development Program under Grant BE2016616, China Scholarship Council, the Scientific Research Foundation of Graduate School of Southeast University under Grant YBJJ1774, Academy of Finland, Tekes Fidipro Program, and Infotech Oulu. *(Corresponding author: Wenming Zheng.)*

Y. Zong is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China, and also with the Center for Machine Vision and Signal Analysis, Faulty of Information Technology and Electrical Engineering, University of Oulu, Oulu FI-90014, Finland (e-mail: xhzongyuan@seu.edu.cn).

W. Zheng is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China (e-mail: wenming_zheng@seu.edu.cn).

X. Huang and G. Zhao are with the Center for Machine Vision and Signal Analysis, Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu FI-90014, Finland (e-mail: {xiaohua.huang, guoying.zhao}@oulu.fi).

Z. Cui is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zhen.cui@njust.edu.cn).

I. Introduction

MICRO-EXPRESSION is a subtle, repressed, and involuntary facial expression and occurs when people try to conceal their true underlying emotions [1]. Similar with conventional facial expressions, accurately recognizing micro-expressions has great values in many application fields, e.g., emotion interfaces [2], multimedia entertainment [3], clinical diagnosis [4], interrogation [5], and security field [6]. Consequently, micro-expression recognition has attracted lots of researchers' attention and become a very active research topic. The early researches about micro-expression recognition can be traced back to the work of Ekman [7] in 1969, in which Ekman observed micro-expressions when he was analyzing a video of an interview with a depressed patient. Since then, Ekman's group carried out extensive and deep studies on micro-expressions. Among their works, it is worth mentioning a training tool which is named Micro-Expression Training Tool (METT) [8]. METT is developed to help people learn how to understand micro-expressions. After receiving intensive training of METT, people can recognize seven categories of basic micro-expressions. However, different from conventional facial expressions, micro-expressions last for a very short time and have considerably low intensity [7], [9]. Moreover, according to recent works of [10], [11], it is known that the muscle movements caused by micro-expressions only emerge in a few local and small facial regions. Due to these facts, micro-expressions are actually very difficult to observe and recognize and hence micro-expression recognition is a more challenging task than conventional facial expression recognition [3], [12]. In the work of [4], Frank et al. pointed out that people achieve a low-level result just around 40% recognition accuracy in micro-expression recognition test after receiving METT training, which is far from practical applications. Consequently, it is necessary and urgent to develop a high-quality automatic micro-expression recognition method to assist people to accurately recognize micro-expressions.

As a typical pattern recognition task, micro-expression recognition can be roughly divided into two main parts. One is micro-expression feature extraction, which targets at extracting useful information from a facial video clip to describe the micro-expressions that the video clip contains. The second one is micro-expression classification, which designs a classifier such as support vector machine (SVM) [13] based on features extracted in the first part for micro-expression recognition tasks. In recent years, most researches of micro-expression recognition focus on feature extraction part. It

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/TMM.2018.2820321, IEEE Transactions on Multimedia

is believed that designing reliable micro-expression features, which can effectively describe the subtle changes of micro-expressions, would benefit micro-expression recognition tasks. For example, in the works of [1], [14], [15], local binary pattern from three orthogonal planes (LBP-TOP), which has demonstrated its effectiveness in video based facial expression recognition [16], [17] and other computer vision tasks, was employed for describing micro-expressions and achieved good results. Although these results are seemingly higher than that of observation of human beings, it is still far from a high-quality micro-expression recognition method. Therefore, there have been many researchers to employ different techniques to improve LBP-TOP or design more suitable spatiotemporal descriptors for micro-expression recognition. Ruiz-Hernandez et al. [18] proposed to apply re-parameterization of second order Gaussian jet on LBP-TOP and obtained a promising micro-expression recognition result on SMIC database. Wang et al. [19] extracted the background information using robust principal component analysis (RPCA) [20] and then used its LBP-TOP and Local Spatiotemporal Directional Features (LSDF) for micro-expression recognition. In the work of [21], Wang et al. proposed to use six intersection points to reduce redundant information of LBP-TOP to obtain a LBP-TOP's variant, called LBP-SIP, and experimental results showed that LBP-SIP is better at recognizing micro-expressions than LBP-TOP. Recently, Huang et al. [22] proposed to use integral operation, which is a popular image projection technique, to develop a novel spatiotemporal LBP based descriptor, called Spatiotemporal Local Binary Pattern with Integration Projection (STLBP-IP), for describing micro-expressions. Extensive experiments demonstrated that STLBP-IP outperforms many spatiotemporal descriptors such as LBP-TOP and LBP-SIP in micro-expression recognition tasks. More recently, two non-LBP framework based spatiotemporal descriptors, i.e., Main Directional Mean Optical (MDMO) [23] feature and Facial Dynamics Map (FDM) [24] feature, have been also developed for describing micro-expressions. MDMO and FDM both achieve promising results on various micro-expression databases.

It is worthy to mentioned that in micro-expression recognition tasks, a spatial division method is usually used together with the above mentioned spatiotemporal descriptors to enhance their discriminative ability to distinguish micro-expressions. Specifically, a micro-expression video clip is divided into a few spatial blocks with a preset grid, e.g., 8×8 in advance. Subsequently, the spatiotemporal features such as LBP-TOP are extracted from these divided facial blocks and then concatenated to compose the micro-expression feature vector. Extensive works [1], [14], [15], [18], [19], [21] have demonstrated the effectiveness of this method. However, it is needed to consider how to choose a most suitable preset spatial division grid for different micro-expression samples when we use spatial division method. Unfortunately, in most existing works, the spatial division grid is selected just empirically and there is no detailed explanation for why such a grid is chosen. Nevertheless, to solve this problem, Zhao et al. [25] proposed a boosted multi-resolution method for spatiotemporal descriptors to deal with dynamic facial expression recognition tasks. Wang et al. [10], [11] designed 16 Regions of Interests (ROIs) based on Facial Action Coding System (FACS) [26] for descriptor extraction instead of fixed grid based division method. Similar with the work of Wang et al. [10], [11], Liu et al. [23] also designed a more elaborate set of ROIs whose number reaches 32. Inspired by these works, in this paper we first attempt to explain why the widely-used spatial division method works well from the view of FACS [26]. Then, based on the explanation, we design a hierarchical spatial division scheme for spatiotemporal descriptor extraction. By adopting this scheme, we can use hierarchical spatiotemporal descriptor to describe micro-expressions and do not need to consider how to choose a suitable division grid for different micro-expression databases excessively like conventional spatial division method. Furthermore, we propose a kernelized group sparse learning (KGSL) model, which is derived from group sparse learning (GSL) model, to build the relationship between hierarchical spatiotemporal descriptors and micro-expressions.

Overall, the contributions of this paper mainly include following two parts:

1) We attempt to explain why spatial division method is beneficial for improving the discriminative ability of spatiotemporal descriptors to describe micro-expressions. Meanwhile, based on our explanation, we design a hierarchical division scheme instead of the widely used fixed grid based division method for spatiotemporal descriptor extraction.
2) We propose a KGSL model to process the hierarchical spatiotemporal descriptors to learn a set of importance weights which can not only select the important facial blocks from various facial blocks yielded by hierarchical scheme but also measure their specific contributions to micro-expression recognition.

The rest of the paper is organized as follows. Section II introduces the hierarchical spatial division scheme for spatiotemporal descriptor extraction. Section III describes KGSL model and shows how it works with hierarchical spatiotemporal descriptors for recognizing micro-expressions. In Section IV, we conduct extensive experiments on two commonly used micro-expression databases to evaluate the proposed method consisting of hierarchical spatiotemporal descriptor and KGSL. Finally, the conclusion is drawn in Section V.

II. Hierarchical Spatiotemporal Descriptors for Describing Micro-Expressions

According to the theory of FACS [26], common expressions, e.g., $Angry$, $Disgust$, $Fear$, and $Happy$, can be coded by the combinations of some action units (AUs)[1]. As described previously, Wang et al. [10], [11] and Liu et al. [23] designed a set of ROIs based on FACS, respectively, where each ROI comprises one or more AUs, instead of the aforementioned fixed grid based spatial division method and then extracted spatiotemporal descriptors from these ROIs to describe micro-expressions. They showed ROIs based method

[1]There are 38 elementary components including 32 AUs and 6 action descriptors (ADs) in FACS. FACS can quantify facial movement with a combination of these components.

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/TMM.2018.2820321, IEEE Transactions on Multimedia

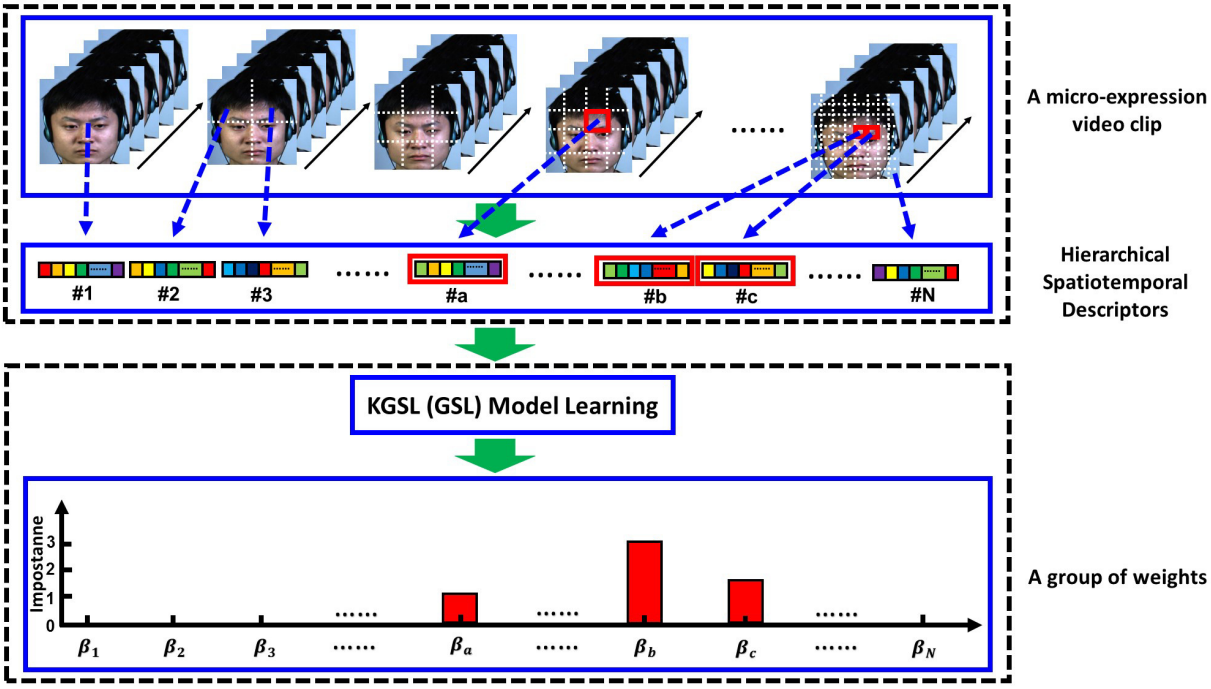

Fig. 1. An illustration of the proposed hierarchical spatiotemporal descriptors + KGSL micro-expression recognition method. The top dashed box describes how to construct a hierarchical spatiotemporal descriptor given a micro-expression video clip. The bottom dashed box presents the key novelty of KGSL model that it can learn a group of weights to measure and quantify the contributions of facial regions to micro-expression recognition. In the example, the weights with the value of 0 mean that their corresponding facial blocks have no contributions to micro-expression recognition. The weights with positive rational numbers reveal different contributions to micro-expression recognition.

is more powerful and effective than fixed grid based spatial division method. Inspired by their works, it is believed that different micro-expressions are also associated with different AU combinations. As well, the features extracted from the AU regions associated with micro-expressions would be more discriminative and effective for micro-expression recognition. Based on this fact, the fixed grid based spatial division method and the ROI method actually share an important point in common that they are exhaustively aiming at locating good facial regions which cover the AU regions strongly associated with micro-expressions. This is why spatiotemporal descriptors, e.g., LBP-TOP, combined with both the above methods would work well in micro-expression recognition. Note that the areas of these good regions are actually hoped to be appropriate, which means the selected regions cannot be either too large or too small. The reason is that large regions may produce noisy information and interfere with the performance of spatiotemporal descriptors, while small regions lead to the loss of useful information. However, as analyzed previously, conventional spatial division method cannot ensure such appropriate regions are obtained because its grid size is fixed.

To solve this problem, we design a hierarchical division scheme which consists of multiple types of gradually denser grids. The grids with different densities would yield facial blocks with different sizes, which guarantees that ideal facial blocks covering the critical AU regions associated with micro-expressions are most probably and appropriately included. Consequently, the hierarchical division scheme is more flexible to all spatiotemporal descriptors than fixed grid based division method since we do not need to try out the best type of grid for different micro-expression datasets. An example of the hierarchical spatial division scheme is depicted in the top dashed box of Fig. 1. It shows how a micro-expression video clip is converted to a hierarchical spatiotemporal descriptor feature vector. More specifically, given a micro-expression video clip $\mathcal{M}_k$, we first divide it into a few non-overlap spatial blocks with equal size according to different types of grids whose numbers of yielded facial blocks are increased gradually, e.g., $\{1 \times 1,\ 2 \times 2,\ 3 \times 3,\ \cdots,\ n \times n\}$, $\{1 \times 1,\ 2 \times 2,\ 4 \times 4,\ \cdots,\ n \times n\}$, respectively. Thus, there will be N blocks, where N is the total number of blocks. For all facial block video clips, spatiotemporal descriptors are extracted, respectively, and denoted by $\{\mathbf{x}_{k,j}\}_{j=1}^{N}$. Note that $\mathbf{x}_{k,j}$ here is a column vector. Then, we concatenate all the spatiotemporal descriptors one by one into a supervector as is illustrated in Fig. 1. The proposed hierarchical spatiotemporal feature vector can be expressed as $\mathbf{x}_k = [\mathbf{x}_{k,1}^T, ..., \mathbf{x}_{k,N}^T]^T$, where T is the transpose operator. In addition, it is worthy to mention that the multi-resolution division method proposed in [25] consisting of four types of grids can be actually treated as a special case of our hierarchical division scheme. It is also noted that the hierarchical division scheme looks like the spatial pyramid proposed in [27]. In fact, this division manner is very popular in many computer vision tasks and in our hierarchical division scheme, we make use of the advantage of this image division approach such that all the divided regions are able to cover all possible micro-expression aware AU regions.

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/TMM.2018.2820321, IEEE Transactions on Multimedia

III. KGSL with Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition

In this section, we will introduce our KGSL model and show how it processes the hierarchical spatiotemporal descriptors and deals with micro-expression recognition tasks. To begin with, we build a GSL model based on hierarchical spatiotemporal descriptors and micro-expressions and then derive KGSL from GSL formulation.

A. GSL Model

Suppose that we have M micro-expression video clips as training samples and their hierarchical spatiotemporal descriptors are extracted according to Section II. We denote them by $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M] \in \mathbb{R}^{d\times M}$, where d is the dimension of hierarchical spatiotemporal descriptors. Instead of using commonly used discrete labels, e.g., 1, 2, and 3, we employ label vector to represent the micro-expression categories for constructing GSL model. More specifically, let $\mathbf{L} = [\mathbf{l}_1, \cdots, \mathbf{l}_M] \in \mathbb{R}^{c\times M}$ be the corresponding label matrix, in which c is the number of micro-expression classes and $\mathbf{l}_k = [l_{k,1}, \cdots, l_{k,c}]^T$ is a column vector whose entries take a binary value 0 or 1 according to the following rule:

$$l_{k,j} = \begin{cases} 1, & \text{if } \mathbf{x}_k \text{ belongs to the } j^{th} \text{ micro-expression class;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

It is notable that label vectors are actually a set of standard orthogonal basis vectors and can span a vector space containing label information. Intuitively, it is suitable to introduce a projection matrix to bridge the feature space and the label space. For this purpose, we build the relationship between label space and feature space by using the following optimization problem:

$$\min_{\mathbf{U}} \|\mathbf{L} - \mathbf{U}^T\mathbf{X}\|_F^2, \tag{2}$$

where $\mathbf{U}$ is the projection matrix. Note that with the use of block matrix trick, $\mathbf{U}^T\mathbf{X}$ can be rewritten as $\sum_{i=1}^N \mathbf{U}_i^T\mathbf{X}_i$, where $\mathbf{X}_i$ is the spatiotemporal descriptors associated with each block video clip as shown in Fig. 1 and $\mathbf{U}_i$ is the corresponding sub-projection matrix of $\mathbf{X}_i$. Thus, replacing $\mathbf{U}^T\mathbf{X}$ by $\sum_{i=1}^N \mathbf{U}_i^T\mathbf{X}_i$, the optimization problem of Eq. (2) is equivalent to the following one:

$$\min_{\mathbf{U}_i} \|\mathbf{L} - \sum_{i=1}^N \mathbf{U}_i^T\mathbf{X}_i\|_F^2, \tag{3}$$

As mentioned previously, we hope to select the good divided facial blocks which are exactly associated with micro-expressions and cover more useful regions because only these facial blocks have enormous contributions to micro-expression recognition. As well, since their contributions are probably different, it is a good way to quantify their importance. We give an illustration to describe this idea shown in Fig. 1. To achieve this goal, we introduce a weighted parameter β_i for each divided facial block and impose its l_1 norm regularization with non-negative constraint, i.e., $\sum_{i=1}^N \beta_i$ $(\beta_i \geq 0)$ onto the objective function of Eq. (3). It is clear to see that this regularization term has two expectable benefits. Firstly, it discards the facial blocks having little contributions during model learning. Secondly, it endows a positive rational number to each selected block region to measure its importance. Therefore, our GSL model is finally formulated as follows:

$$\min_{\mathbf{U}_i,\beta_i} \|\mathbf{L} - \sum_{i=1}^N \beta_i\mathbf{U}_i^T\mathbf{X}_i\|_F^2 + \mu\sum_{i=1}^N \beta_i, \text{ s.t. } \beta_i \geq 0, \tag{4}$$

where μ is a trade-off parameter determining the number of nonzero elements of importance weight vector β.

B. KGSL Model

In many recent works [28], [29], [30], [24], [31], nonlinear classifiers, e.g., nonlinear SVM, are preferred for micro-expression recognition problem and the experimental results have demonstrated their advantages over linear ones. Our recent works [22], [32] also suggest that ChiSquare kernel seems more suitable than linear one for modeling some spatiotemporal descriptors when using SVM as classifier. This is the main motivation of why we intend to extend GSL to a kernelized version for processing hierarchical spatiotemporal descriptors instead of directly using GSL. For this purpose, let $\mathbf{X}_i$ and $\mathbf{U}_i$ be mapped into a kernel space $\mathbf{F}$ via a nonlinear mapping operator ϕ as below,

$$\phi : \mathbb{R}^d \to \mathbf{F}, \ \mathbf{X}_i \to \mathbf{X}_i^\phi \text{ and } \mathbf{U}_i \to \mathbf{U}_i^\phi,$$

Then, we can arrive at the following optimization problem by substituting $\mathbf{U}_i^\phi$ and $\mathbf{X}_i^\phi$ into the GSL formulation of Eq. (4):

$$\min_{\mathbf{U}_i^\phi,\beta_i} \|\mathbf{L} - \sum_{i=1}^N \beta_i{\mathbf{U}_i^\phi}^T\mathbf{X}_i^\phi\|_F^2 + \mu\sum_{i=1}^N \beta_i, \text{ s.t. } \beta_i \geq 0, \tag{5}$$

According to the reproduced kernel theory, in such an infinite-dimensional space, each column $\mathbf{u}_{i,j}^\phi$ of $\mathbf{U}_i^\phi$ can be represented as the linear combination of $\mathbf{X}_i^\phi$, i.e., $\mathbf{u}_{i,j}^\phi = \mathbf{X}_i^\phi\mathbf{p}_j$, where $\mathbf{p}_j$ is the linear combination coefficients of $\mathbf{u}_{i,j}^\phi$ with respect to $\mathbf{X}_i^\phi$. Let $\mathbf{U}_i^\phi = [\mathbf{u}_{i,1}^\phi, \cdots, \mathbf{u}_{i,c}^\phi]$, then we have $\mathbf{U}_i^\phi = \mathbf{X}_i^\phi\mathbf{P}$, where $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_c]$. Substituting the new expression of $\mathbf{U}_i^\phi$, we are able to rewrite Eq. (5) as:

$$\min_{\mathbf{P},\beta_i} \|\mathbf{L} - \mathbf{P}^T\sum_{i=1}^N \beta_i\mathbf{K}_i\|_F^2 + \mu\sum_{i=1}^N \beta_i, \text{ s.t. } \beta_i \geq 0, \tag{6}$$

where $\mathbf{K}_i = {\mathbf{X}_i^\phi}^T\mathbf{X}_i^\phi$ is a Gram matrix and can be directly calculated by different preset kernel functions.

It is noted that a reliable and reasonable $\mathbf{p}_j$ needs be sparse because only a few columns of $\mathbf{X}_i^\phi$ can compose $\mathbf{u}_{i,j}^\phi$ linearly instead of the whole. Therefore, it is better to add a l_1 norm regularization of $\mathbf{P}$, i.e., $\|\mathbf{P}\|_1 = \sum_{j=1}^c \|\mathbf{p}_j\|_1$ to the objective function of Eq. (6). Additionally, it can avoid the overfitting problem [33] during the optimization of KGSL model. Thus, we will obtain our KGSL model whose optimization problem can be expressed as:

$$\min_{\mathbf{P},\beta_i} \|\mathbf{L}-\mathbf{P}^T\sum_{i=1}^N \beta_i\mathbf{K}_i\|_F^2+\lambda\|\mathbf{P}\|_1+\mu\sum_{i=1}^N \beta_i, \text{ s.t. } \beta_i \geq 0, \tag{7}$$

where λ is also a trade-off parameter which controls the sparsity of $\mathbf{P}$.

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/TMM.2018.2820321, IEEE Transactions on Multimedia

Algorithm 1: Algorithm for learning the optimal parameters $\mathbf{P}$ and β of KGSL.

Input: training feature matrix $\mathbf{X} = [\mathbf{X}_1^T, \cdots, \mathbf{X}_N^T]^T$, training label matrix $\mathbf{L}$, kernel function $k(\mathbf{x}, \mathbf{y})$, and trade-off parameters λ and μ.
Initialize: $k = 0$, β^k, $\mathbf{K}_i = k(\mathbf{X}_i, \mathbf{X}_i)$.
Output: the optimal $\mathbf{P}$ and β.
while the objective function $< \epsilon$ or iter $k <$ maxIter **do**
1) Fix β^k and compute $\tilde{\mathbf{K}}^{k+1}$:

$$\tilde{\mathbf{K}}^{k+1} = [\beta_1^k \mathbf{K}_1^T, \cdots, \beta_N^k \mathbf{K}_N^T]^T,$$

2) Fix β^k and update $\mathbf{P}^{k+1}$:

$$\mathbf{P}^{k+1} = \arg\min_{\mathbf{P}} \|\mathbf{L} - \mathbf{P}^T \tilde{\mathbf{K}}^{k+1}\|_F^2 + \lambda \|\mathbf{P}\|_1,$$

3) Fix $\mathbf{P}^{k+1}$ and compute $\mathbf{Z}^{k+1}$:

$$\mathbf{Z}^{k+1} = [\mathbf{z}_1, \cdots, \mathbf{z}_N],$$

where $\mathbf{z}_i = \text{vec}({\mathbf{U}_i^{k+1}}^T \mathbf{K}_i)$ and $\text{vec}(\cdot)$ means a matrix is reshaped into a vector by concatenating its columns one by one;
4) Fix $\mathbf{P}^{k+1}$ and update β^{k+1}:

$$\beta^{k+1} = \arg\min_{\beta} \|\tilde{\mathbf{l}} - \mathbf{Z}^{k+1}\beta\|^2 + \mu\|\beta\|_1, \text{ s.t. } \beta \succeq 0.$$

where $\tilde{\mathbf{l}} = \text{vec}(\mathbf{L})$ and $\|\beta\|_1 = \sum_{i=1}^N \beta_i$.
5) $k = k + 1$;
end while

C. Optimization of KGSL

The optimization problem of KGSL model in Eq. (7) can be simply solved via the alternative direction method (ADM) [34], [35], i.e., updating the parameters $\{\beta, \mathbf{P}\}$ alternately until convergence. The solving procedures consist of four important steps, which is summarized in Algorithm 1. It is notable that the non-negative Lasso problem in Step 4 is solved by Liu et al.'s SLEP package [36] in our algorithm implementation. Meanwhile, we use the inexact augmented Lagrangian multiplier (inexact ALM) approach [34], [35] to optimize $\mathbf{P}$ in Step 2. More specifically, we firstly introduce an auxiliary parameter $\mathbf{Q}$ which equals $\mathbf{P}$. Then, the original optimization problem with respect to $\mathbf{P}$ is converted to a constrained one and is reformulated as follows:

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{Q}} \|\mathbf{L} - \mathbf{Q}^T \tilde{\mathbf{K}}\|_F^2 + \lambda\|\mathbf{P}\|_1, \\ &\text{s.t. } \mathbf{P} = \mathbf{Q}. \end{aligned} \tag{8}$$

Thus, the Lagrangian function of Eq.(8) is obtained as:

$$\begin{aligned} L(\mathbf{P}, \mathbf{Q}, \mathbf{Y}, \kappa) = \; & \|\mathbf{L} - \mathbf{Q}^T \tilde{\mathbf{K}}\|_F^2 + \lambda\|\mathbf{P}\|_1 \\ & + tr(\mathbf{Y}^T(\mathbf{P} - \mathbf{Q})) + \tfrac{\kappa}{2}\|\mathbf{P} - \mathbf{Q}\|_F^2, \end{aligned} \tag{9}$$

where $\mathbf{Y}$ is the Lagrangian multiplier matrix and κ is a trade-off parameter.

Finally, the optimal $\mathbf{P}$ can be learned by minimizing the Lagrangian function of Eq. (9) with respect to different parameters iteratively until convergence. The complete procedures for optimizing $\mathbf{P}$ are summarized in Algorithm 2.

Algorithm 2: Algorithm for solving the optimization problem of step 2 in Algorithm 1.

Input: training feature Gram matrix $\tilde{\mathbf{K}}$, training label matrix $\mathbf{L}$, and trade-off parameter λ.
Initialize: $k = 0$, $\mathbf{P}^k$, $\mathbf{Q}^k$, and $\mathbf{Y}^k$.
Output: the optimal $\mathbf{P}$.
while $\|\mathbf{P}^k - \mathbf{Q}^{k+}\|_\infty < \epsilon$ or iter $k <$ maxIter **do**
1) Fix $\mathbf{P}^k$ and $\mathbf{Y}^k$ and update $\mathbf{Q}^{k+1}$:

$$\begin{aligned} \mathbf{Q}^{k+1} = \arg\min_{\mathbf{Q}} \quad & \|\mathbf{L} - \mathbf{Q}^T\tilde{\mathbf{K}}\|_F^2 + tr({\mathbf{Y}^k}^T(\mathbf{P}^k - \mathbf{Q})) \\ & + \frac{\kappa}{2}\|\mathbf{P}^k - \mathbf{Q}\|_F^2, \end{aligned}$$

which results in

$$\mathbf{Q}^{k+1} = (\tilde{\mathbf{K}}\tilde{\mathbf{K}} + \frac{\kappa^k}{2}\mathbf{I})^{-1}(\tilde{\mathbf{K}}\mathbf{L}^T + \frac{\mathbf{Y}^k + \kappa^k\mathbf{P}^k}{2}),$$

where $\mathbf{I}$ is an identity matrix.
2) Fix $\mathbf{Q}^{k+1}$ and $\mathbf{Y}^k$ and update $\mathbf{P}$:

$$\mathbf{P}^{k+1} = \arg\min_{\mathbf{P}} \frac{\lambda}{\kappa^k}\|\mathbf{P}\|_1 + \frac{1}{2}\|\mathbf{P} - (\mathbf{Q}^{k+1} - \frac{\mathbf{Y}^k}{\kappa^k})\|_F^2,$$

whose solution is:

$$\mathbf{P}_{ij}^{k+1} = \begin{cases} (\mathbf{Q}_{ij}^{k+1} - \frac{\mathbf{Y}_{ij}^k}{\kappa^k}) - \frac{\lambda}{\kappa^k}, & \text{if } (\mathbf{Q}_{ij}^{k+1} - \frac{\mathbf{Y}_{ij}^k}{\kappa^k}) > \frac{\lambda}{\kappa^k}; \\ (\mathbf{Q}_{ij}^{k+1} - \frac{\mathbf{Y}_{ij}^k}{\kappa^k}) + \frac{\lambda}{\kappa^k}, & \text{if } (\mathbf{Q}_{ij}^{k+1} - \frac{\mathbf{Y}_{ij}^k}{\kappa^k}) < \frac{\lambda}{\kappa^k}; \\ 0, & \text{otherwise.} \end{cases}$$

where $\mathbf{P}_{ij}$, $\mathbf{Q}_{ij}$, and $\mathbf{Y}_{ij}$ are the elements of i^{th} row and j^{th} column of their corresponding matrices.
3) Update $\mathbf{Y}^{k+1}$ and κ^{k+1}:

$$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \kappa^k(\mathbf{P}^{k+1} - \mathbf{Q}^{k+1}), \kappa^{k+1} = \min(\rho\kappa^k, \kappa_{max}),$$

4) $k = k + 1$.
end while

D. Micro-Expression Label Prediction Using KGSL

Once the optimal parameters $\hat{\mathbf{P}}$ and $\hat{\beta}_i$ of KGSL model are learned based on training samples, we can estimate the micro-expression category of a given testing micro-expression video clip $\mathcal{M}_t$ by two simple steps below:

1) Solve the following optimization problem to obtain the label vector of a given testing sample:

$$\min_{\mathbf{l}_t} \|\mathbf{l}_t - \hat{\mathbf{P}}^T \sum_{i=1}^N \hat{\beta}_i(\mathbf{K}_t)_i\|_F^2, \tag{10}$$

where $(\mathbf{K}_t)_i = {\mathbf{X}_i^\phi}^T(\mathbf{x}_t)_i^\phi$ and is calculated using the pre-selected kernel function in KGSL model training, $\mathbf{x}_t$ is the feature vector of the testing micro-expression video clip, and $\mathbf{l}_t$ is its corresponding label vector.
2) Assign the micro-expression category to the testing micro-expression video clip according to the criterion described in Eq. (11):

$$\text{micro-expression\_label} = \arg\max_k \{\mathbf{l}_t(k)\}. \tag{11}$$

where $\mathbf{l}_t(k)$ means the k^{th} element of label vector $\mathbf{l}_t$.

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/TMM.2018.2820321, IEEE Transactions on Multimedia

IV. Experiments

A. Databases and Experiment Setting

In this section, we conduct extensive experiments to evaluate the performance of the proposed micro-expression method consisting of hierarchical spatiotemporal descriptors and KGSL on two widely-used micro-expression databases. One is CASME II [14], which is collected by Yan et al. from Institute of Psychology, Chinese Academy of Science. CASME II database contains 247 micro-expression video clips from 26 subjects. These samples are categorized into five micro-expression classes, i.e., $Happy$ (32 samples), $Surprise$ (25 samples), $Disgust$ (64 samples), $Repression$ (27 samples), and $Others$ (99 samples). The other is SMIC [37] collected by Li et al. from University of Oulu. SMIC database consists of 164 samples of 16 participants belonging to 3 different classes, i.e., $Positive$ (51 samples), $Negative$ (70 samples), and $Surprise$ (43 samples), respectively. Particularly, the samples are recorded by a high-speed camera (HS) of 100 fps, a normal visual camera (VIS) of 25 fps, and a near-infrared camera (NIR) of 25 fps, respectively. In this paper, we use the HS samples for the experiments. For CASME II database, we crop and transform the face images of video clips into 308×257 pixels, while for SMIC database, the images are cropped and then transformed to 170×139 pixels.

In the experiments on both databases, leave-one-subject-out (LOSO) strategy is used to calculate recognition accuracies, where in each fold the samples of one subject are used as the test set while the remaining samples are used as training one. After S folds (S denotes the number of subjects), the samples of all subjects have been used as the test set once, and the final recognition rate is then calculated according to $Accuracy = \frac{\sum_{i=1}^{S} T_i}{\sum_{i=1}^{S} N_i} \times 100$, where T_i and N_i are the number of correct predictions and the number of testing samples, respectively, when the samples of i^{th} subject is served as the testing set. Besides, since CASME II and SMIC databases are highly imbalanced [28], [31], [38], which means in the database the number of one type of micro-expression samples is significantly larger or lower than other types of micro-expression samples, we also report the F1-score of all the experimental results such that the actual performance of the micro-expression methods can be reflected. F1-score is calculated according to $F = \frac{1}{c}\sum_{i=1}^{c} \frac{2p_i \times r_i}{p_i + r_i}$, where p_i and r_i mean the precision and recall of the i^{th} micro-expression, respectively, and c is the number of micro-expressions.

For our method, we choose three representative spatiotemporal descriptors, i.e., LBP-TOP [1], [16], LBP-SIP [21], and STLBP-IP [22], to respectively conduct the same experiments according to the above protocol. The detailed setup of our method including the parameters of these descriptors, hierarchical scheme, and the parameters of KGSL are listed as below:

1) Following the suggestion of [22], [32], we choose Chi-Square kernel function for KGSL model throughout the experiments.
2) For the hierarchical division scheme, we choose four types of grids that are gradually denser, i.e., 1×1, 2×2, 4×4, and 8×8, respectively. Thus, we have in total 85 facial blocks given a micro-expression video clip.
3) For LBP-TOP, we set the neighboring radius R and the number of the neighboring points P to be 3 and 8 for LBP operator on three orthogonal planes. The uniform pattern is used in LBP coding. Similar with LBP-TOP, for LBP-SIP, we set its neighboring radius R as 3.
4) There are several parameters for STLBP-IP descriptor including mask size W for spatial domain, radius of LBP R, and neighboring number P of LBP for temporal domain. According to the work of [22], we set W to be 9, while R and P are fixed at 3 and 8, respectively. As well, we use temporal bilinear interpolation method to normalize the temporal texture image as 60 for CASME II and 20 for SMIC, respectively.
5) Leave-one-subject-out cross-validation (LOSOCV) method is adopted to determine the trade-off parameters of KGSL. In our experiments we just select the trade-off parameters for KGSL in the first fold, and then the selected trade-off parameters are fixed in the rest of folds[2]. More specifically, 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. Then, we search the spaces $\lambda \in [1, 15]$ with interval 1 and $\mu \in [0.1, 3]$ with interval 0.1 and fix each parameter combination for KGSL to compute the recognition accuracy of validation set after $S-1$ folds. Finally, the set of parameters corresponding to the best results are selected. The selected parameters of KGSL for hierarchical LBP-TOP, hierarchical LBP-SIP, and hierarchical STLBP-IP are given in Table I.

TABLE I
THE SELECTED TRADE-OFF PARAMETERS OF KGSL FOR HIERARCHICAL VERSION OF LBP-TOP, LBP-SIP, AND STLBP-IP ON CASME II AND SMIC DATABASES, RESPECTIVELY.

Spatiotemporal Descriptor	CASME II	SMIC
Hierachical LBP-TOP	$\lambda = 10$, $\mu = 0.1$	$\lambda = 1$, $\mu = 2.9$
Hierachical LBP-SIP	$\lambda = 1$, $\mu = 1.8$	$\lambda = 8$, $\mu = 1$
Hierachical STLBP-IP	$\lambda = 9$, $\mu = 1.7$	$\lambda = 3$, $\mu = 1.4$

B. Comparison with the Fixed Grid Based Division Method

In order to show whether our proposed method can boost the performance of spatiotemporal descriptors in dealing with micro-expression recognition tasks, we first compare our method with the widely-used fixed grid based division method. We choose three grids including 2×2, 4×4, and 8×8, respectively to apply on LBP-TOP, LBP-SIP, and STLBP-IP, and use Chi-Square kernel SVM as the classifier to conduct the

[2]Note that since most of existing micro-expression methods report the results using the fixed model parameter, to be consistent we just select the trade-off parameter once (for the first fold). In fact, this strategy may be bias to the first fold in the experiments. A good suggestion is selecting the trade-off parameters several times (for several folds) and then averaging the results to serve as the final experimental result. We also give the results of STLBP-IP based on this suggestion in the last line of Table II, where we randomly select the parameters for three folds.

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/TMM.2018.2820321, IEEE Transactions on Multimedia

TABLE II
EXPERIMENTAL RESULTS ON CASME II AND SMIC DATABASES IN TERMS OF RECOGNITION ACCURACY AND F1-SCORE. THE VALUES IN BRACKET BEHIND THE SPATIOTEMPORAL DESCRIPTORS ARE THE SPATIAL DIVISION GRID SIZES.

Method	CASME II		SMIC	
	Accuracy (%)	F1-score	Accuracy (%)	F1-score
LBP-TOP (2×2) + SVM	36.30	0.3361	45.73	0.4606
LBP-TOP (4×4) + SVM	40.49	0.3770	47.56	0.4840
LBP-TOP (8×8) + SVM	40.90	0.3693	45.73	0.4690
Hierarchical LBP-TOP + SVM	40.89	0.3759	47.56	0.4861
Hierarchical LBP-TOP + KGSL	**45.75**	**0.4230**	**52.44**	**0.4937**
LBP-SIP (2×2) + SVM	32.39	0.2156	35.98	0.3505
LBP-SIP (4×4) + SVM	39.68	0.3394	41.46	0.4096
LBP-SIP (8×8) + SVM	**45.73**	**0.4249**	42.07	0.4222
Hierarchical LBP-SIP + SVM	44.13	0.4047	39.63	0.3985
Hierarchical LBP-SIP + KGSL	42.91	0.3410	**43.29**	**0.4254**
STLBP-IP (2×2) + SVM	46.15	0.4022	40.24	0.4047
STLBP-IP (4×4) + SVM	51.42	0.4654	48.17	0.4891
STLBP-IP (8×8) + SVM	55.06	0.4966	54.27	0.5467
STLBP-IP (8×9) + SVM [22]	59.51	N\A	57.93	N\A
Hierarchical STLBP-IP + SVM	55.47	0.5034	52.44	0.5353
Hierarchical STLBP-IP + KGSL	**63.97**	**0.6125**	**60.37**	**0.6125**
Hierarchical STLBP-IP + KGSL*	**63.83±0.62**	**0.6110±0.0075**	**60.78±0.35**	**0.6126±0.0040**

* The result is the average of several results under the fixed parameters selected for several folds.

TABLE III
COMPARISON BETWEEN ROIS BASED METHODS AND OURS ON CASME II DATABASE, WHERE THE SPATIOTEMPORAL DESCRIPTOR IS LBP-TOP.

Method	Protocol	Accuracy(%)	F1-score
Liu et al.'s ROIs scheme + SVM [23]	LOSO	47.10	N\A
Hierachical scheme + SVM	LOSO	46.96	0.4116
Hierachical scheme + KGSL	LOSO	**53.04**	**0.4841**
Wang et al.'s ROIs scheme + SVM [11]*	LOVO	55.14	N\A

* The result is obtained based on leave-one-video-out (LOVO) protocol and P of LBP-TOP is set as 4.

exactly same experiments previously introduced. For STLBP-IP on fixed grid setting, besides the above three types of grids, we include the best result reported in [22] to comparison as well. In addition, we also use ChiSquare kernel SVM for the proposed hierarchical spatiotemporal descriptors to conduct the experiments.

The experimental results on both CASME II and SMIC databases are depicted in Tables II, respectively. As Table II shows, it can be seen that with SVM as classier, the proposed hierarchical spatiotemporal descriptors perform very closely to the fixed grid division based descriptors and have no obvious advantages. This is because that by using multiple grids, although facial regions suitably covering the beneficial AUs can be included, more useless facial regions which have no or less contributions to micro-expression recognition are introduced as well. The motivation of KGSL model is hence to deal with this problem. Clearly, by using KGSL model, three hierarchical spatiotemporal descriptors almost have the markedly higher recognition accuracy and F1-score on both CASME II and SMIC databases than those achieved by various fixed grid based division methods, which indicates that our method (combining hierarchical scheme and KGSL) can boost the performance of spatiotemporal descriptors in dealing with micro-expression recognition tasks. Particularly, for some originally well-performing spatiotemporal descriptors, e.g., STLBP-IP, our method can still further enhance it dramatically. In the case of STLBP-IP, we can see that by using the fixed grid based division method and SVM, STLBP-IP achieves recognition accuracies of 59.51% and 57.93% on CASME II and SMIC databases, respectively, which are considerably competitive compared with most of recent state-of-the-art results (refer to Tables V and VI in what follows). But by combining our hierarchical scheme and KGSL model, STLBP-IP obtains promising increases of 4.46% and 2.44% recognition accuracies on CASME II and SMIC databases, respectively. In a word, it is concluded that our method is able to make most spatiotemporal descriptors be more competitive and powerful in dealing with micro-expression recognition tasks.

C. Comparison with the ROIs Based Method

We also compare the proposed method with the ROIs based methods [10], [11], [23], which are designed according to FACS theory. We directly take the result of ROIs based uniform LBP-TOP from their works, in which the parameters of LBP-TOP are also set as $R = 3$ and $P = 8$, for comparison.[3]

[3]Since Wang et al. [10], [11] claimed that $P = 4$ is more suitable for their method and only gave the results of $P = 4$, we take the result under this parameter setting.

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/TMM.2018.2820321, IEEE Transactions on Multimedia

TABLE IV
THE RECENT STATE-OF-THE-ART MICRO-EXPRESSION RECOGNITION METHODS AND THEIR MAIN EMPLOYED TECHNIQUES.

Method	Preprocessing Method	Spatiotemporal Descriptor	Division Scheme	Classifier
Le Ngo et al. [28]	Temporal Interpolation	LBP-TOP	5×5 for CASME II 8×8 for SMIC	AdaBoost + STM
Oh et al. [39]	No	Riesz Wavelet	5×5 for CASME II	SVM (Linear)
Wang et al. [30]	Wiener Filter	LBP-MOP for CASME II LBP-TOP for SMIC	5×5 for CASME II 8×8 for SMIC	SVM (Gaussian & Linear)
Park et al. [40]	Adaptive Motion Magnification	LBP-TOP	N\A	SVM
Xu et al. [24]	Linear Interpolation	FDM	N\A	SVM (Gaussian)
Le Ngo et al. [38]	Motion Magnification	LBP-TOP	5×5 for CASME II 8×8 for SMIC	SVM (Linear)
Le Ngo et al. [31]	Sparse Sampling	LBP-TOP	5×5	SVM (Linear)
Huang et al. [32]	Gaussian Filter	STCLQP	8×8	SVM (Linear)
Kim et al. [41]	Data Augmentation	CNN & RNN	No	N\A
Liong et al. [42]	Gaussian Filtering & Noise Block Removal	OSW-LBP-TOP	8×8 for SMIC	SVM (Linear)
Hong et al. [43]	Temporal Interpolation	2Standmap	N\A	SVM (Linear)

Following the experiments on CASME II in [23], we recategorize the micro-expression samples into four classes, i.e., $Positive$ (original $Happy$), $Negative$ (original $Disgust$), $Surprise$, and $Others$ (original $Others$ and $Repression$), and use Hierarchical LBP-TOP with SVM and KGSL with ChiSquare kernel respectively to conduct the experiments under the LOSO protocol. It should be pointed out that since Liu et al. [23] report the best result with the optimal model parameters, to offer a fair comparison here, we follow their grid search strategy and search the parameters for KGSL from a preset parameter grids ($\lambda \in [1, 15]$ with interval 1 and $\mu \in [0.1, 3]$ with interval 0.1) to report the best result corresponding to the optimal parameters. The best result of KGSL corresponds to the optimal parameters $\lambda = 2$ and $\mu = 0.6$. The comparison results are given in Table III. From Table III, we can find that by using SVM as classifier, our hierarchical scheme is competitive against Liu et al.'s ROIs based method as the experimental results shows (46.96% v.s. 47.10%). More importantly, it is interesting to see that together with the proposed KGSL model, the performance of hierarchical LBP-TOP can be improved and our method promisingly outperforms Liu et al.'s method [23] (53.04% vs. 47.10%). In addition, it can be also seen that the recognition accuracy achieved by our method is even at the same level with the result of Wang et al.'s method [10], [11] under the leave-one-video-out (LOVO) protocol.

D. Comparison with the State-of-the-art Results

In this section, we compare the best result achieved by our method (Hierarchical STLBP-IP + KGSL) in Table II with recent state-of-the-art results on both two databases. To make the readers have access to glance through these methods [28], [39], [30], [40], [24], [38], [31], [32], [41], [42], [43], we summarize their main employed techniques including preprocessing method, spatiotemporal descriptor, division scheme and classier in Table IV. For the detail of these methods, the readers can further refer to the corresponding references. In addition, since some works among them report the best results with the optimal model parameters instead of the parameters selected by cross-validation method, we also report the results of our methods with the optimal parameters by using the parameter grid search strategy in the comparison.

TABLE V
COMPARISON BETWEEN OUR METHOD (HIERARCHICAL STLBP-IP + KGSL) WITH SOME STATE-OF-THE ART METHODS ON CASME II DATABASE.

Method	Accuracy (%)	F1-score
Le Ngo et al. [28]	43.78	0.3337
Oh et al. [39]	46.15	0.4307
Wang et al. [30]	45.75	N\A
Park et al. [40]	51.91	N\A
Xu et al. [24]	41.96	0.2972
Le Ngo et al. [38]	51.00	0.4700
Le Ngo et al. [31]	49.00	0.5100
Kim et al. [41]	60.98	N\A
Ours	**63.97**	**0.6125**
Huang et al. [32]*	58.39	0.5836
Ours**	**65.18**	**0.6254**

* The result is obtained with the optimal parameter set searched from a preset parameter space.
** The result is obtained with the optimal parameter set ($\lambda = 8, \mu = 2.5$) searched from a preset parameter space.

1) Comparison results on CASME II database: We take the results on CASME II database under the LOSO protocol achieved by the above methods [28], [39], [30], [40], [24], [38], [31], [32] for comparison. The comparison results are depicted in Table V. From Table V, we can see that our method achieves the highest recognition accuracy and F1-score among all the micro-expression recognition methods and performs substantially better than other methods. Compared with the best results of theirs (60.98% recognition accuracy reported by Kim et al. [41] and 0.5100 F1-score reported by Le Ngo et al. [31]), our method obtains a more promising performance with increase of 2.99% recognition accuracy and 0.1025 F1-score.

2) Comparison results on SMIC database: Among the above comparison methods, the works of [42], [29], [28], [30], [24], [31], [43], [32] employ SMIC database for eval-

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/TMM.2018.2820321, IEEE Transactions on Multimedia

TABLE VI
COMPARISON BETWEEN OUR METHOD (HIERARCHICAL STLBP-IP + KGSL) WITH SOME STATE-OF-THE ART METHODS ON SMIC DATABASE.

Method	Accuracy (%)	F1-score
Liong et al. [42]	53.66	N\A
Liong et al. [29]	53.56	N\A
Le Ngo et al. [28]	44.34	0.4731
Wang et al. [30]	51.83	N\A
Xu et al. [24]	54.88	0.5380
Le Ngo et al. [31]	58.00	0.6000
Hong et al. [43]	57.90	N\A
Ours	**60.37**	**0.6125**
Huang et al. [32]*	64.02	0.6381
Ours**	**66.46**	**0.6577**

* The result is obtained with the optimal parameter set searched from a preset parameter space.
** The result is obtained with the optimal parameter set ($\lambda = 11, \mu = 2.8$) searched from a preset parameter space.

uation experiments by using the LOSO protocol. We list their achieved best results in Table VI. As illustrated in this table, it is clear to see that our method has the best performance in terms of both recognition accuracy (60.37%) and F1-score (0.6125) among all the methods as well. Meanwhile, it can be seen that Huang et al. [32]'s reported result reaches 64.02% recognition accuracy. By using the same parameter search strategy, our KGSL method can achieve the recognition accuracy of 66.46%, which is higher than theirs. Overall, according to the results, we are able to reach a conclusion that our method is better at dealing with the micro-expression recognition tasks on SMIC database than these recent well-performing methods.

3) Statistical significance analysis: We also conduct statistical significance analysis for the experimental results. Firstly, we perform one-sample t-test for testing the null hypothesis that the average recognition accuracy of all the state-of-the-art methods (first EIGHT methods in Table V for CASME II and first SEVEN methods in Table VI for SMIC) is equal to the recognition accuracy of our method on two micro-expression databases. The significance level α is set to 0.05. For CASME II and SMIC databases, we obtain $p = 0.0001846$ and $p = 0.0075$, respectively, which indicates that compared with the state-of-the-art methods, the improvements achieved by our method are statistically significant.

Secondly, we perform two-sample t-test for the comparison between our method and the best-performing comparison method (Huang et al. [32]). To this end, we choose the results of Huang et al. [32] and our method under the optimal parameter setting (corresponding to the last two methods in Table V and Table VI) and calculate the accuracy of each fold for both two methods according to their predictions in the experiments on CASME II and SMIC, respectively. The detailed results for all the folds are given in Table VII. Then, according to the results of all the folds, we are able to obtain $p = 0.2228$ and $p = 0.4595$ for the comparison on CASME II and SMIC, respectively. It can be from the p-values seen that the improvements achieved by our method have no statistical significance compared with the results of Huang et al. [32].

TABLE VII
THE ACCURACY (%) OF EACH FOLD FOR THE PREDICTIONS OF HUANG ET AL. [32] AND OURS ON CASME II AND SMIC, RESPECTIVELY.

Fold	CASME II		SMIC	
	Huang et al.*	Ours	Huang et al.*	Ours
01	44.44	100.00	33.33	66.67
02	38.46	69.23	83.33	66.67
03	85.71	57.14	69.23	58.97
04	100.00	100.00	47.37	57.89
05	57.89	31.58	100.00	100.00
06	60.00	80.00	50.00	50.00
07	100.00	44.44	69.23	53.85
08	33.33	100.00	25.00	0.00
09	76.92	61.54	85.71	100.00
10	100.00	100.00	88.89	77.78
11	70.00	70.00	70.00	90.00
12	50.00	75.00	80.00	80.00
12	100.00	50.00	75.00	75.00
14	75.00	50.00	85.71	71.43
15	33.33	100.00	50.00	100.00
16	25.00	50.00	40.91	68.18
17	20.59	50.00	-	-
18	66.67	100.00	-	-
19	66.67	53.33	-	-
20	81.82	72.73	-	-
21	50.00	50.00	-	-
22	100.00	100.00	-	-
23	58.33	75.00	-	-
24	75.00	75.00	-	-
25	28.57	42.86	-	-
26	31.25	75.00	-	-

* The results achieved by Huang et al.[32] in all the folds are provided by Dr. Xiaohua Huang from University of Oulu, Finland.

E. Evaluation on Hierarchical Scheme with Different Grid Combinations

In the above experiments, we just employs multiple grids with sizes $\{1\times1, 2\times2, 4\times4, 8\times8\}$ (TYPE-I) for hierarchical scheme. In this section, we evaluate the performance of the proposed hierarchical scheme with other grid combinations. For this purpose, we select three grid combinations including $\{1\times1, 2\times2, 3\times3, 4\times4\}$ (TYPE-II), $\{1\times1, 2\times2, 3\times3, 4\times4, 5\times5\}$ (TYPE-III), and $\{1\times1, 2\times2, 3\times3, \cdots, 8\times8\}$ (TYPE-IV), for hierarchical scheme with KGSL to conduct the experiments. The kernel function of KGSL is ChiSquare. We denote the above four types of grid combinations by TYPE-I, TYPE-II, TYPE-III, and TYPE-IV, respectively. The experimental results are given in Table VIII, where the number in brackets in the first column is the facial block number yielded by the corresponding type of grid combination. Note that for convenience, we just employ the parameter grid search strategy as the experiments in Section IV-C for KGSL with different grid combinations and report the best results with the optimal parameter set. As shown in this table, we can find that the hierarchical STLBP-IP with TYPE-I and TYPE-IV grid combinations achieves more promising results in terms of both accuracy and F1-score than other two types, which indicates that for CASME II and SMIC databases, dense grid (8×8) seems be more able to accurately cover micro-expression aware AU regions than sparse grid (e.g., 4×4

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/TMM.2018.2820321, IEEE Transactions on Multimedia

TABLE VIII
EXPERIMENTAL RESULTS OF THE PROPOSED HIERARCHICAL SCHEME USING DIFFERENT TYPES OF GRID COMBINATIONS ON CASME II AND SMIC DATABASES.

TYPE	CASME II				SMIC			
	Accuracy (%)	F1-score	Optimal Parameters	Runtime (s)	Accuracy (%)	F1-score	Optimal Parameters	Runtime (s)
I (85)	65.18	0.6254	$\lambda = 8$, $\mu = 2.5$	213.23	66.46	0.6577	$\lambda = 11$, $\mu = 2.8$	61.96
II (30)	55.06	0.5294	$\lambda = 3$, $\mu = 0.3$	125.20	56.10	0.5592	$\lambda = 6$, $\mu = 4.9$	33.07
III (55)	57.49	0.5254	$\lambda = 1$, $\mu = 1.7$	160.73	57.93	0.5785	$\lambda = 2$, $\mu = 0.5$	46.00
IV (204)	63.56	0.6206	$\lambda = 4$, $\mu = 0.8$	386.86	63.41	0.6272	$\lambda = 12$, $\mu = 1.2$	120.45

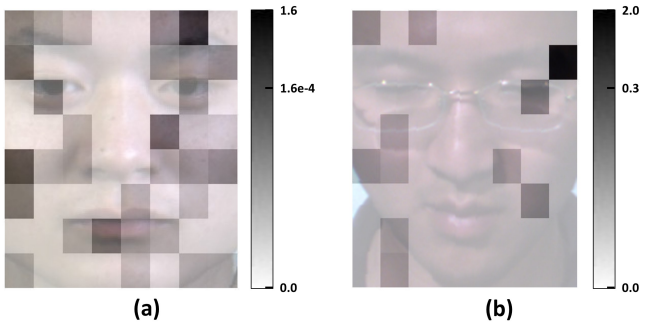

Fig. 2. Examples for the visualization of learned importance parameters β_i. (a) is the result of CASME II database, where the parameters of KGSL are $\lambda = 8$ and $\mu = 2.5$. The selected sample belongs to $Happy$ category and comes from Subject #$sub01$. (b) is the result of SMIC database, where the parameters of KGSL are $\lambda = 11$ and $\mu = 2.8$. The selected sample belongs to $Negative$ category and is from Subject #$s18$. The spatiotemporal descriptor is STLBP-IP. In these two examples, only some of facial blocks yielded by 8×8 grid are selected by KGSL.

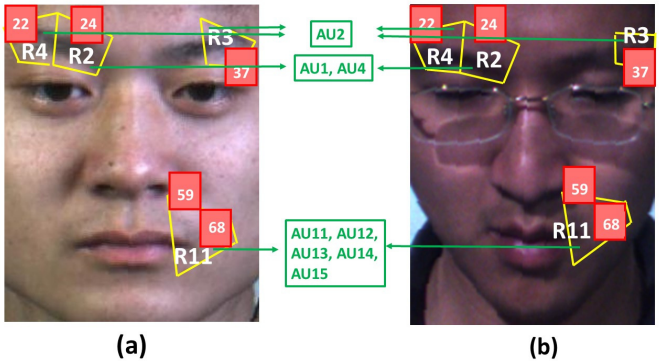

Fig. 3. Some common selected facial blocks between (a) CASME II and (b) SMIC databases and their associations with AUs and Wang et al.'s ROIs [10], [11], where the number in the selected facial blocks is the block index illustrated in Fig. 1.

and 5×5). Meanwhile, it can also be seen that TYPE-I slightly outperforms TYPE-IV for both CASME II and SMIC databases. It may be explained by the fact that TYPE-IV employs a finer granularity of grid combination and bring more useless information compared with TYPE-I. Consequently, the performance of the proposed hierarchical scheme using TYPE-IV degrades compared with TYPE-I.

We also investigate the complexity difference among hierarchical schemes using these four different types of grid combinations. To this end, we list their corresponding time consumption of the experiments on CASME II and SMIC databases, respectively, in Table VIII. Note that the computer for experiments has an Intel Core i5-4570 CPU with 3.20 GHz and an 8GB RAM. The computing software is MATLAB 2016b. From the results of Table VIII, we can see that the time consumption of the proposed hierarchical scheme is very sensitive to the facial block numbers yielded by its preset spatial grids. Theoretically, the time cost of the proposed hierarchical scheme would increase as the increase of the facial block numbers. In practice, however, the facial block numbers should not be too large in order to avoid the redundancy of the facial information. Moreover, it is notable that too small size of facial block may not cover the useful facial regions and hence the feature extracted from these facial blocks will bring irrelevant information to the micro-expression recognition target. Consequently, it can be found that TYPE-I hierarchical scheme is a satisfactory choice in practice by considering the balance between the performance and time complexity.

F. Weighted Parameter Visualization for KGSL

As described previously, the regularization term with respect to β of KGSL model is designed for quantifying the contribution of each block yielded by hierarchical division scheme to micro-expression recognition. To see how it reflects the contributions of different facial blocks to micro-expression recognition, we visualize the learned β of KGSL based on both databases. For this purpose, we draw a weighted parameter visualization example with STLBP-IP as spatiotemporal descriptor for CASME II and SMIC databases, respectively, where the testing samples in CASME II case belong to Subject #$sub01$ and in SMIC example the testing subject is Subject #$s18$. The visualization results are shown in Fig. 2, where the left transparent heat map describes the learned weights β_i of CASME II example and the right one corresponds to the SMIC example.

From the visualization results, it is interesting to see that for both two databases, facial blocks on 1×1, 2×2, and 4×4 grids are all not selected. It is demonstrated that spatiotemporal descriptors extracted based on sparse division grids have no or very low contributions to distinguishing micro-expressions. That is why in the most existing works, they achieve satisfactory performance by adopting dense division grids, e.g., 6×6, 6×8, and 8×8. It should be pointed out that not all facial blocks yielded by a dense grid, e.g., 8×8 in this case, contribute to distinguishing micro-expressions. Only the facial blocks locating at the AU regions associated with micro-expressions, e.g., AUs around eyebrows (AU1, AU2, AU4) and AUs near lip corners (AU13, AU14) in this example

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/TMM.2018.2820321, IEEE Transactions on Multimedia

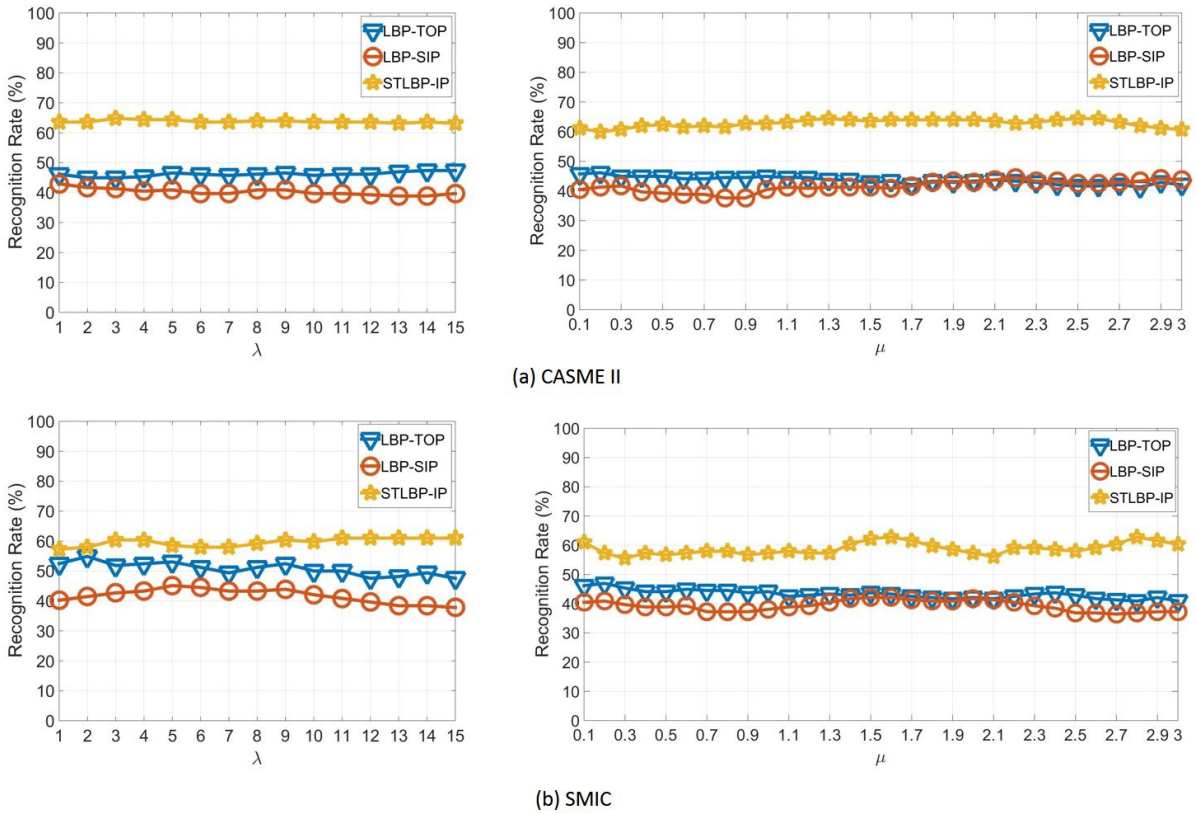

Fig. 4. Results of trade-off parameter sensitive experiments for the proposed KGSL model with different spatiotemporal descriptors (LBP-TOP, LBP-SIP, and STLBP-IP), where (a) corresponds to the results of CASME II database and (b) is the results of SMIC database.

as shown in Fig. 3, are selected. From Fig. 3, it is also interesting to see that these mentioned AUs are mostly the elements composing the ROIs designed by Wang et al. [10], [11]. Our example may provide a support for the fact that different micro-expressions are associated with different AU combinations like facial expressions, which is discussed in Section II.

Besides, by comparing two visualization results, we can find that the number of selected facial blocks of CASME II are larger than that of SMIC. We think there may be two possible reasons. Firstly, the micro-expression categorizations of two databases are totally different. CASME II has five micro-expressions including $Happy$, $Surprise$, $Disgust$, $Repression$, and $Others$, while SMIC samples are divided into three micro-expressions, i.e., $Positive$, $Negative$, and $Surprise$. Therefore, the related facial regions that are discriminative to these two types of micro-expression categorizations have different numbers and locations. Secondly, the micro-expressions induced by the stimuli materials collected by CASME II owners may have more facial muscle movements than SMIC database and hence more facial blocks of CASME II samples are selected by KGSL model. In addition, although the selected facial blocks of two databases have different numbers, it is clear that there are some shared blocks locating at the AU regions as mentioned above. It is believed that the AUs associated with these shared facial block regions are very important and contributing to distinguishing micro-expression independently, which means for either different databases or different categorizations for micro-expression samples, they play a crucial role in micro-expression recognition tasks.

G. Parameter Sensitivity of KGSL model

In this section, we investigate the parameter sensitivity of the proposed KGSL model. As Eq. (7) shows, the KGSL model has two important trade-off parameters, i.e., λ and μ. To see whether KGSL model with different spatiotemporal descriptors is robust to their selection, we conduct additional experiments using KGSL with hierarchical LBP-TOP, LBP-SIP, and STLBP-IP on CASME II and SMIC databases, respectively, by fixing the values of one parameter of KGSL while changing the other one. More specifically, the changing space and interval of λ and μ are set as follows: $\lambda \in [1:1:15]$ and $\mu \in [0.1:0.1:3]$, while the fixed λ and μ in the experiments are set as same as the values selected by CV method, which are shown in Table I. Experimental results are given in Fig. 4, where Fig. 4 (a) corresponds to the results of CASME II database and Fig. 4 (b) is the results of SMIC database. From the results, it is clear to see that the performance of KGSL with different spatiotemporal descriptors changes slightly within the parameter space for both CASME II and SMIC databases. This indicates that the proposed KGSL model is robust to its trade-off parameters.

V. Conclusion

In this paper, we have designed a hierarchical spatial division scheme for spatiotemporal descriptors to better describe facial micro-expressions, which is motivated by the inner relationship between FACS and micro-expressions. Meanwhile, based on the fact that the facial regions corresponding to some AUs contribute to distinguishing different micro-expressions, we have further proposed a novel model called KGSL to process hierarchical scheme based spatiotemporal descriptors. We conduct extensive experiments on two publicly available spontaneous micro-expression databases, i.e., CASME II and SMIC, to evaluate the performance of the proposed method. The experimental results show that our method can effectively enhance spatiotemporal descriptors in dealing with micro-expression recognition tasks. By combining our method with STLBP-IP, we achieve promising results that are better than recent state-of-the-art results on both databases.

In addition, something important is worth pointing out. Firstly, the proposed method in this paper is in fact also suitable for dealing with macro-expression recognition problem because it has been demonstrated that selecting AU-aware facial regions benefits distinguishing macro-expressions as well [25], [35]. Secondly, it is clear that AU regions associated with micro-expressions are very important for describing micro-expressions. Consequently, it is a good choice to investigate how to better benefit from these AU regions for micro-expression recognition tasks. In the future, we will focus on developing an AU related feature learning method which can effectively utilize the discriminative information for micro-expressions from AU regions.

References

[1] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 1449–1456.

[2] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[3] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1543–1552, 2013.

[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 *the Annual Meeting of the International Communication Association. Sheraton New York, New York City*, 2009.

[5] M. G. Frank, C. J. Maccario, and V. Govindaraju, "Behavior and security," *Protecting Airline Passengers in the Age of Terrorism. Greenwood Pub Group, Santa Barbara, California*, pp. 86–106, 2009.

[6] M. OSullivan, M. G. Frank, C. M. Hurley, and J. Tiwana, "Police lie detection accuracy: The effect of lie scenario," *Law and Human Behavior*, vol. 33, no. 6, pp. 530–538, 2009.

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

[8] P. Ekman, "Mett. micro expression training tool," *CD-ROM. Oakland*, 2003.

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

[10] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proceedings of the 22nd International Conference onPattern Recognition (ICPR)*. IEEE, 2014, pp. 4678–4683.

[11] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, "Micro-expression recognition using color spaces," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6034–6047, 2015.

[12] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[13] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[14] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[15] A. K. Davison, M. H. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: an investigation on spatio-temporal descriptors," in *Proceedings of the 13th European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 111–123.

[16] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[17] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2181–2191, 2012.

[18] J. A. Ruiz-Hernandez and M. Pietikäinen, "Encoding local binary patterns using the re-parametrization of the second order gaussian jet," in *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[19] S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, and C.-G. Zhou, "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *Proceedings of the 13th European Conference on Computer Vision Workshops (ECCV Workshops)*. Springer, 2014, pp. 325–338.

[20] 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 *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 2080–2088.

[21] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Proceedings of the 12th Asian Conference on Computer Vision (ACCV)*. Springer, 2014, pp. 525–537.

[22] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proceedings of the 14th IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2015, pp. 1–9.

[23] 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 Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2016.

[24] F. Xu, J. Zhang, and J. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, 2016.

[25] 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.

[26] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, 1978.

[27] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE computer society conference on Computer vision and pattern recognition*, vol. 2. IEEE, 2006, pp. 2169–2178.

[28] A. C. Le Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Proceedings of the 12th Asian Conference on Computer Vision (ACCV)*. Springer, 2014, pp. 33–48.

[29] S.-T. Liong, R. C.-W. Phan, J. See, Y.-H. Oh, and K. Wong, "Optical strain based recognition of subtle emotions," in *Proceedings of the 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, 2014, pp. 180–184.

[30] 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, p. e0124674, 2015.

[31] A. C. Le Ngo, J. See, and C.-W. R. Phan, "Sparsity in dynamics of spontaneous subtle emotion: Analysis & application," *IEEE Transactions on Affective Computing*, 2016.

[32] 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, 2016.

[33] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
[34] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.
[35] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014.
[36] J. Liu and J. Ye, "Efficient euclidean projections in linear time," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*. ACM, 2009, pp. 657–664.
[37] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.
[38] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan, and J. See, "Eulerian emotion magnification for subtle expression recognition," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1243–1247.
[39] Y.-H. Oh, A. C. Le Ngo, J. See, S.-T. Liong, R. C.-W. Phan, and H.-C. Ling, "Monogenic riesz wavelet representation for micro-expression recognition," in *Proceedings of the 20th IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 1237–1241.
[40] S. Y. Park, S. H. Lee, and Y. M. Ro, "Subtle facial expression recognition using adaptive magnification of discriminative facial motion," in *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 2015, pp. 911–914.
[41] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 382–386.
[42] S.-T. Liong, J. See, R. C.-W. Phan, A. C. Le Ngo, Y.-H. Oh, and K. Wong, "Subtle expression recognition using optical strain weighted features," in *Proceedings of the 12th Asian Conference on Computer Vision (ACCV)*. Springer, 2014, pp. 644–657.
[43] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, and M. Pietikäinen, "Capturing correlations of local features for image representation," *Neurocomputing*, vol. 184, pp. 99–106, 2016.

Yuan Zong received the B.S. and M.S. degrees in electronics engineering from Nanjing Normal University, Nanjing, China, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, China. From 2016 to 2017, he was working as a Visiting Student with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His research interests include affective computing, pattern recognition, and computer vision.

Xiaohua Huang received the B.S. degree in communication engineering from Huaqiao University, Quanzhou, Fujian, 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 since 2006. He has been a scientist researcher in the Center for Machine Vision and Signal Analysis at University of Oulu, Finland 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.

Wenming Zheng (M'08) received the B.S. degree in computer science from Fuzhou University, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004. Since 2004, he has been with the Research Center for Learning Science, Southeast University. He is currently a Professor with the School of Biological Sciences and Medical Engineering, Southeast University, and the Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University. His research interests include neural computation, pattern recognition, machine learning, and computer vision. He is an Associated Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and Neurocomputing and a member of the associate editors-in-chief of The Visual Computer.

Zhen Cui (M'14) received the B.S., M.S., and Ph.D. degrees from Shandong Normal University, Sun Yat-sen University, and Institute of Computing Technology (ICT), Chinese Academy of Sciences in 2004, 2006, and 2014, respectively. He was a Research Fellow in the Department of Electrical and Computer Engineering at National University of Singapore (NUS) from 2014 to 2015. He also spent half a year as a Research Assistant on Nanyang Technological University (NTU) from Jun. 2012 to Dec. 2012. Currently, he is a Professor of Nanjing University of Science and Technology, China. His research interests cover computer vision, pattern recognition and machine learning, especially focusing on deep learning, manifold learning, sparse coding, face detection/alignment/recognition, object tracking, image super resolution, emotion analysis, etc.

Guoying Zhao (SM'12) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. In 2011, she was selected to the highly competitive academy research fellow position. She was a Nokia Visiting Professor in 2016. She is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where she has been a Senior Researcher since 2005 and an Associate Professor since 2014. She has authored or co-authored over 160 papers in journals and conferences. She has authored/edited three books and seven special issues in journals. Her research has been reported by Finnish TV programs, newspapers, and MIT Technology Review. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, gait analysis, dynamictexture recognition, human motion analysis, and person identification. She is a Co-Publicity Chair for FG2018. She has served as area chairs for several conferences. She was a Co-Chair of many international workshops at ECCV, ICCV, CVPR, ACCV, and BMVC. She has lectured tutorials at ICPR 2006, ICCV 2009, and SCIA 2013. She is an Associate Editor of Pattern Recognition, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and Image and Vision Computing journals. Her papers have currently over 6900 citations in Google Scholar (h-index 35).