Sparse MDMO: Learning a Discriminative Feature for Micro-Expression Recognition

Yong-Jin Liu, *Senior Member, IEEE,* Bing-Jun Li, Yu-Kun Lai, *Member, IEEE*

Abstract—Micro-expressions are the rapid movements of facial muscles that can be used to reveal concealed emotions. Recognizing them from video clips has a wide range of applications and receives increasing attention recently. Among existing methods, the main directional mean optical-flow (MDMO) feature achieves state-of-the-art performance for recognizing spontaneous micro-expressions. For a video clip, the MDMO feature is computed by averaging a set of atomic features frame-by-frame. Despite its simplicity, the average operation in MDMO can easily lose the underlying manifold structure inherent in the feature space. In this paper we propose a sparse MDMO feature that learns an effective dictionary from a micro-expression video dataset. In particular, a new distance metric is proposed based on the sparsity of sample points in the MDMO feature space, which can efficiently reveal the underlying manifold structure. The proposed sparse MDMO feature is obtained by incorporating this new metric into the classic graph regularized sparse coding (GraphSC) scheme. We evaluate sparse MDMO and four representative features (LBP-TOP, STCLQP, MDMO and FDM) on three spontaneous micro-expression datasets (SMIC, CASME and CASME II). The results show that sparse MDMO outperforms these representative features.

Index Terms—Micro-expression, MDMO feature, sparse coding, recognition.

1 Introduction

Micro-expressions are brief and involuntary movements of facial muscles, typically lasting for less than 0.5 seconds [1]. Psychological studies have shown that a person may intentionally conceal her/his genuine emotions, but cannot fake micro-expressions [2]. Recognizing micro-expressions from video clips is useful in many applications, including clinical diagnosis, social interaction and national security.

The choice of features is critical for micro-expression recognition. A few features have been proposed and they can be broadly divided into two classes: *appearance-based* and *optical-flow-based*. Local binary pattern (LBP) is a classic feature in the first class, which has been successfully applied in image-based macro-expression recognition [3]. An extension of LBP, called local binary pattern from three orthogonal planes (LBP-TOP), is proposed in [4] for macro-expression recognition in video clips. Pfister et al. [5] propose a micro-expression recognition method based on LBP-TOP. By introducing local structure information into LBP-TOP, Huang et al. [6] propose spatiotemporal completed local quantization patterns (STCLQP). STCLQP partitions a video clip into blocks and concatenates individual features from all the blocks into an overall feature. However, the dimension of the STCLQP feature may be very large.

The second feature class relies on a robust and accurate optical flow estimation. Histograms of oriented optical flow (HOOF) [7] is an elaborated feature that is originally proposed for human action recognition. To apply HOOF for micro-expression recognition, Liu et al. [8] divide the whole facial area into 36 regions of interest (ROIs) based on the facial action coding system [9], and compute a HOOF feature for each ROI, from which a main direction is determined. The main directions of all the ROIs are consolidated into a 72-dimensional feature vector. Finally, the feature vector is averaged over time, leading to a so-called *main directional mean optical-flow* (MDMO) feature. Facial dynamics map (FDM) [10] is another optical-flow-based feature. Instead of using 36 ROIs as in MDMO, FDM computes a pixel-level alignment for micro-expression sequences. Each sequence is further divided into spatiotemporal cuboids, in which the principal optical flow directions are computed to represent the local facial dynamics. Both MDMO and FDM make use of special properties[1] in micro-expressions to optimize the estimated optical flow, such that it is insensitive to illumination changes.

Our study presented in this paper is inspired by two key observations:

- Due to the characteristics of short duration and low intensity, any features depicting micro-expressions are sparse in both temporal and spatial domains [11];
- The feature data is likely to reside on a low-dimensional manifold embedded in a high-dimensional feature space.

Based on these observations, we propose a sparse MDMO feature that preserves the underlying manifold structure and has more discriminating power than the original

- *Y.-J. Liu and B. Li are with the Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, P. R. China. E-mail: liuyongjin@tsinghua.edu.cn*
- *Y.-K. Lai is with School of Computer Science and Informatics, Cardiff University, UK.*

This work was supported in part by the Natural Science Foundation of China (61725204, U1736220, 61521002) the National Key Research and Development Plan (2016YFB1001202) and Royal Society-Newton Advanced Fellowship (NA150431).

1. For example, FDM assumes that most of facial areas in neighboring frames remain motionless, due to very few facial muscles involved in micro-expressions.

MDMO feature. Sparse representations have been widely used for face and facial expression recognition (e.g., [12], [13]). The classic sparse-representation-based classification (SRC) method [12] directly builds a dictionary $\mathbf{D}$ from the entire training set as the concatenation of all k classes, $\mathbf{D} = \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_k \end{array} \right]$, where $\mathbf{c}_i = \left[\begin{array}{cccc} \mathbf{v}_1^i & \mathbf{v}_2^i & \cdots & \mathbf{v}_{n_i}^i \end{array} \right]$, $\mathbf{v}_j^i$ is a vector representing the jth sample image I_j by packing the grayscale values of all the pixels in I_j column by column, and n_i is the number of samples in the ith class. SRC does not need an explicit feature extraction scheme and can efficiently handle occlusion and corruption in facial images. However, using all the pixel information in training images may lead to a dictionary of a huge size.

Rather than using a fixed dictionary like the one in SRC, many dictionary learning methods have been proposed to learn an effective dictionary from training data (e.g., [13], [14], [15], [16]). Unsupervised dictionary learning, such as K-SVD [14], works well for image restoration, image compression and denoising. For classification tasks, recent studies can be broadly categorized into two classes. One class is supervised dictionary learning which takes full advantage of class labels of training data. Two representative works are discriminative K-SVD (D-KSVD) [13] and label consistent K-SVD (LC-KSVD) [15]. The other class is to consider the local geometric structure in the sparse data. In many image and vision applications, the sample data in a high-dimensional space is observed to lie on or close to a smooth low-dimensional manifold. By building a k-nearest neighbor graph to encode the local manifold structure, the methods in this class learn a sparse representation that explicitly incorporates the graph Laplacian as a regularizer (e.g., GraphSC in [16]). Some other state-of-the-art methods include [17], [18], [19]. Experimental results demonstrate that graph regularized sparse representations have good discriminating power for classification and have a good scalability to large training data such as those in video applications.

However, trivially applying existing sparse methods such as K-SVD or GraphSC in micro-expression recognition (MER) does not achieve good performance, since these general sparse models do not consider the special and discriminative structure inherent in MER applications. In this paper, we introduce the classic graph regularized sparse coding (GraphSC) [16] into the MDMO feature with the special consideration to preserve an important manifold structure in MER. The key idea is that in the MDMO feature space, the low-dimensional manifold structure of data points can be depicted by their sparsity. We propose a new distance metric to capture this sparsity. By incorporating this new metric into the unsupervised sparse representation GraphSC, the desired sparse MDMO feature can be efficiently computed by the elegant solver to GraphSC.

The main contributions of this paper are:

- A new distance metric in the MDMO feature space is proposed based on the sparsity of data points. Based on this metric, the manifold structure of data points is revealed.
- The sparse MDMO feature making use of unsupervised learning with sparse coding has the following benefits: it (1) only requires a small amount of training data, (2) fits well with MER due to the limited data availability, and (3) can be efficiently computed by the GraphSC solver [16] with the new metric.
- The proposed sparse MDMO feature remains compact but is substantially more discriminative than MDMO, outperforming state-of-the-art features for MER.

(a) 36 ROIs (b) Optical flow (c) Atomic feature

Figure 1. (a) The facial area is partitioned into 36 regions of interest (ROIs). (b) Optical flow is computed between this frame and the first frame in the video clip. (c) MDMO represents a frame by an atomic optical flow feature, which is a 72-dimensional vector.

Extensive experiments are presented, in which two trivial sparse codings with MDMO, the proposed sparse MDMO and four representative features including MDMO [8], LBP-TOP [4], STCLQP [6] and FDM [10], are evaluated on three spontaneous micro-expression datasets, i.e., SMIC [20], CASME [21] and CASME II [22]. The results show that the proposed sparse MDMO outperforms the existing features.

2 Preliminaries

Since sparse MDMO is based on MDMO [8] and is also closely related to GraphSC [16], we briefly introduce both methods before presenting the sparse MDMO.

2.1 MDMO

Given a micro-expression video clip, i.e., an image sequence $(f_1, f_2, \cdots, f_m)$, the MDMO feature takes the optical flow as the basis, due to its capacity to infer subtle motions by detecting the changing intensity of pixels between two frames. Based on the facial action coding system [9], MDMO partitions the facial area in each frame into 36 regions of interest (ROIs) using 66 facial points (Figure 1(a)). In the first frame f_1, these 66 facial points are detected by discriminative response map fitting (DRMF) method [23]. Optical flow is computed between each frame f_i, $i > 1$, and the first frame f_1 (Figure 1(b)). The facial points in f_i, $i > 1$, are then determined by the optical flow field.

In each frame f_i, $i > 1$, the optical flow vectors in each ROI R_k^i, $k = 1, 2, \cdots, 36$, are categorized into eight orientation bins, and the bin $B_{\max}$ with the maximum number of optical flow vectors is selected. The so-called *main direction* of the optical flow in R_k^i is defined as the average of all optical flow vectors fallen into $B_{\max}$, denoted as $\overline{\mathbf{u}}_k^i = (\overline{\rho}_k^i, \overline{\theta}_k^i)$, where the optical flow vectors are represented in polar coordinates (ρ_i, θ_i), ρ_i and θ_i are the magnitude and the direction. MDMO represents each frame f_i, $i > 1$, by an atomic optical flow feature Ψ_i (Figure 1(c)):

$$\Psi_i = \left(\overline{\mathbf{u}}_1^i, \overline{\mathbf{u}}_2^i, \cdots, \overline{\mathbf{u}}_{36}^i\right)^T \quad (1)$$

The dimension of Ψ_i is $36 \times 2 = 72$, where 36 is the number of ROIs. A micro-expression video clip Γ can be represented by a series of atomic optical flow features

$$\Gamma = (\Psi_2, \Psi_3, \cdots, \Psi_m), \tag{2}$$

where m is the number of frames in the video clip. Finally, the MDMO feature for the video clip Γ is defined as a 72-dimensional vector $\widetilde{\Psi}$, which is a normalized version of $\overline{\Psi}$:

$$\overline{\Psi} = \left[\left(\overline{\rho}_1, \overline{\theta}_1\right)^T, \left(\overline{\rho}_2, \overline{\theta}_2\right)^T, \cdots, \left(\overline{\rho}_{36}, \overline{\theta}_{36}\right)^T \right], \tag{3}$$

where

$$\left(\overline{\rho}_k, \overline{\theta}_k\right) = \frac{1}{m-1} \sum_{i=2}^{m} \overline{\mathbf{u}}_k^i, \ k = 1, 2, \cdots, 36 \tag{4}$$

The 72-dimensional vector $\widetilde{\Psi}$ is represented by:

$$\widetilde{\Psi} = [\alpha \mathbf{P}, (1-\alpha)\mathbf{\Theta}] \tag{5}$$

where $\mathbf{P} = [\widetilde{\rho}_1, \widetilde{\rho}_2, \cdots, \widetilde{\rho}_{36}]$ is a 36-dimensional row vector, $\widetilde{\rho}_k = \frac{\overline{\rho}_k}{\max\{\overline{\rho}_j, j=1,2,\cdots,36\}}$, $k = 1, 2, \cdots, 36$, and $\mathbf{\Theta} = \left[\overline{\theta}_1, \overline{\theta}_2, \cdots, \overline{\theta}_{36}\right]$ is a 36-dimensional row vector. It was shown in [8] that $\alpha \in [0.75, 0.98]$ achieves best results. In all experiments in this paper, we use fixed $\alpha = 0.9$.

2.2 GraphSC

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ be a data matrix, where N is the number of data points, $\mathbf{x}_i$ is a d-dimensional column vector denoting the ith data point. Sparse coding is to find a sparse representation for each data point, based on a dictionary $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_{n_d}] \in \mathbb{R}^{d \times n_d}$, which is an over-complete matrix consisting of n_d basis vectors $\mathbf{d}_j$, $n_d > d$. The sparse coding problem can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{S}} \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda \sum_{i=1}^{N} g(\mathbf{s}_i), \\ &s.t. \ \|\mathbf{d}_i\|^2 \le c, \ i = 1, 2, \cdots, n_d \end{aligned} \tag{6}$$

where $\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_N] \in \mathbb{R}^{n_d \times N}$ is the coefficient matrix, in which each column vector $\mathbf{s}_i$ is a sparse representation for the data point $\mathbf{x}_i$, $\|\cdot\|_F$ denotes the matrix Frobenius norm, g is a function to measure the sparseness of $\mathbf{s}_i$, λ is a weight to balance the reconstruction error and sparsity, and c is a constant imposing a norm constraint for the basis.

Directly optimizing the function (6) may lead to a solution that ignores the underlying structure in the data set $\mathbf{X}$. To introduce a structure constraint into sparse coding, GraphSC constructs a k-nearest neighbor graph G in the data set $\mathbf{X}$. Each vertex in G is a data point and the edges of G are represented by a weight matrix $\mathbf{W} = \{w_{i,j}\}$,

$$w_{i,j} = \begin{cases} 1, & \mathbf{x}_j \in \mathcal{N}_i \\ 0, & otherwise \end{cases} \tag{7}$$

where $\mathcal{N}_i$ is the set of k-nearest neighbors of $\mathbf{x}_i$. To put the structure constraint represented by the graph G to the sparse representation $\mathbf{S}$, a graph regularization term can be described as:

$$\frac{1}{2} \sum_{i,j} w_{i,j} \|\mathbf{s}_i - \mathbf{s}_j\|_2^2 \tag{8}$$

Eq. (8) can be rewritten in the matrix form using the Laplacian matrix $\mathbf{L}$ as $Tr(\mathbf{S}\mathbf{L}\mathbf{S}^T)$, where $Tr(\cdot)$ is the matrix trace, $\mathbf{L} = \mathbf{\Sigma} - \mathbf{W}$, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_N)$ and $\sigma_i = \sum_{j=1}^{N} w_{i,j}$ is the degree of $\mathbf{x}_i$ in G. GraphSC optimizes the following objective function by incorporating the graph regularization term (Eq. 8) into the original sparse coding (Eq. (6)):

$$\begin{aligned} \min_{\mathbf{D},\mathbf{S}} &\|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + w_g Tr(\mathbf{S}\mathbf{L}\mathbf{S}^T) + \lambda \sum_{i=1}^{N} \|\mathbf{s}_i\|_1 \\ &s.t. \ \|\mathbf{d}_i\|^2 \le c, \ i = 1, 2, \cdots, n_d \end{aligned} \tag{9}$$

where w_g is a regularization parameter and GraphSC chooses $\|\mathbf{s}_i\|_1$ for the function $g(\mathbf{s}_i)$ in Eq. (6). An elegant numerical solver is proposed in [16] to find an optimal solution in Eq. (9) for both $\mathbf{D}$ and $\mathbf{S}$.

3 Sparse MDMO Feature

Observing that any features depicting micro-expressions are sparse [11], we propose a sparse representation for the MDMO feature, which considers the manifold structure of sparse data in the MDMO feature space and therefore is more discriminative than the original MDMO feature.

3.1 Overview

Let $\Pi = \{\Gamma_1, \Gamma_2, \cdots, \Gamma_{n_{clips}}\}$ be a given micro-expression video dataset consisting of n_{clips} video clips. In MDMO, each video clip Γ_i is represented by $m_i - 1$ atomic optical flow features (ref. Eq. (2)), where m_i is the number of frames in Γ_i. We collect all atomic features in Π and consolidate them into a data matrix $\mathbf{X} \in \mathbb{R}^{72 \times N}$:

$$\mathbf{X} = \left[\Gamma_1, \Gamma_2, \cdots, \Gamma_{n_{clips}}\right] = \left[\Psi_1^1, \Psi_2^1, \cdots, \Psi_{m_{n_{clips}}-1}^{n_{clips}}\right] \tag{10}$$

where the number of data points $N = \sum_{i=1}^{n_{clips}} (m_i - 1)$. Whenever there is no risk of confusion, Γ_i is also used as a sub-matrix

$$\Gamma_i = \left[\Psi_1^i, \Psi_2^i, \cdots, \Psi_{m_i-1}^i\right] \in \mathbb{R}^{72 \times (m_i - 1)} \tag{11}$$

Applying the objective function in Eq. (6) with Eq. (10) can learn a sparse representation $\mathbf{S}$ and a dictionary $\mathbf{D}$ for the data set $\mathbf{X}$ depicting in Eq. (10). However, our experiments in Section 4 show that this trivial implementation does not achieve the best performance of sparse representation.

Our key observation is that the data points contained in $\mathbf{X}$ reside on multiple low-dimensional manifolds embedded in $\mathbb{R}^{72 \times N}$. The simple average operation in MDMO (ref. Eq. (4)) can easily lose this underlying manifold structure and most of dynamic details. To overcome this limitation in MDMO, in Section 3.2, we propose a new distance metric in the MDMO feature space based on the sparsity of the data. With the aid of this new metric, the manifold structure in $\mathbf{X}$ is revealed and in Section 3.3, a manifold-structure-preserving sparse coding method is proposed by incorporating the new metric into the GraphSC. Finally in Section 3.4, a temporal pooling is applied to the sparse representation of $\mathbf{X}$, leading to a concise sparse MDMO feature. Both MDMO and sparse MDMO are vector representations, and thus, the same SVM classifier with the polynomial kernel used in [8] can be applied. In Section 4, experimental results are presented, demonstrating that our proposed sparse MDMO outperforms the original MDMO and several other representative micro-expression recognition features. The overview of our proposed sparse MDMO feature is illustrated in Figure 2.

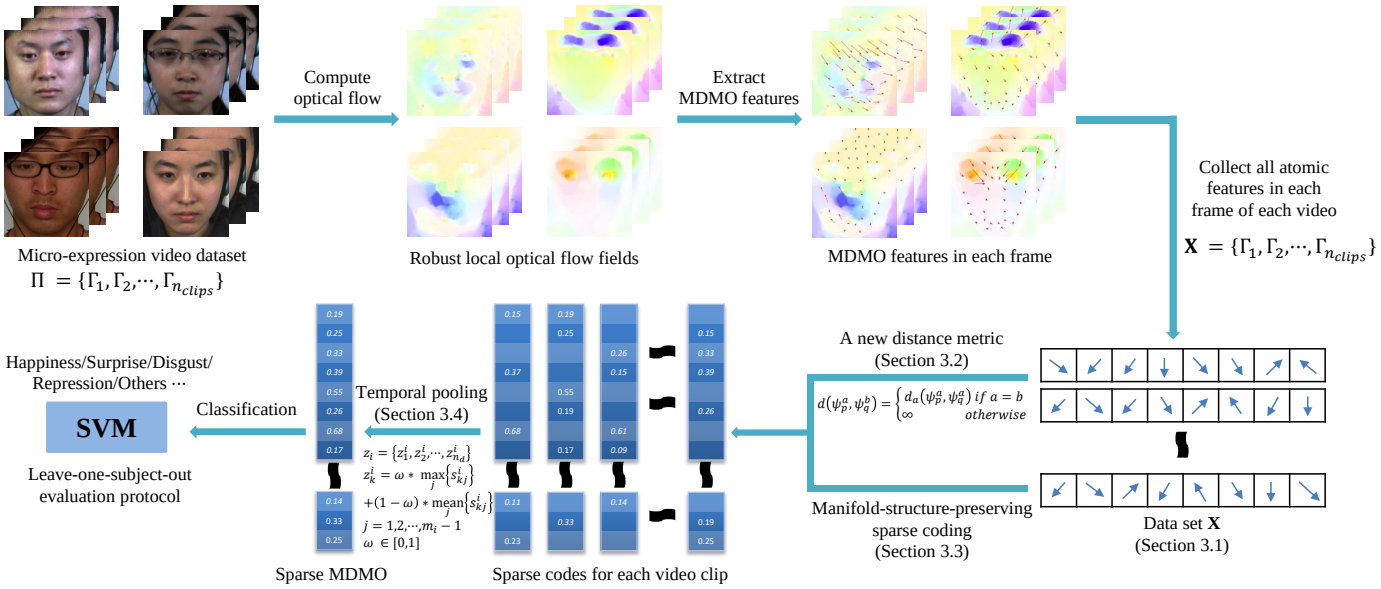


Figure 2. Overview of the proposed MER system with sparse MDMO features.

3.2 A new distance metric

The structure inherent in the data set $\mathbf{X} \in \mathbb{R}^{72\times N}$ (Eq. (10)) is two-fold:

- in the MDMO feature space, the sample points from each micro-expression category form a low-dimensional manifold, and
- different micro-expression categories contribute to multiple low-dimensional manifolds.

To reveal the above manifold structure, we propose a local distance metric, which is defined in a subspace for each video clip $\Gamma_i \in \Pi$ (ref. Eq.(13)), such that in the same micro-expression category, distance is smaller for more relevant features. We then assemble local distance metrics into a global distance metric (ref. Eq.(14)), such that the distance between sample points from different micro-expression categories is infinity.

To define the local distance metric, For each data point Ψ_j^i in the sub-matrix Γ_i (Eq. (11)), we solve the following sparse representation problem:

$$\min \|\boldsymbol{\gamma}_j\|_1 \ \ s.t.\ \Psi_j^i = \mathbf{B}_j^i \boldsymbol{\gamma}_j \tag{12}$$

where $\mathbf{B}_j^i$ is the basis matrix containing all the remaining points $\left[\Psi_1^i, \cdots, \Psi_{j-1}^i, \Psi_{j+1}^i, \cdots, \Psi_{m_i-1}^i\right]$ in Γ_i. Eq. (12) can be solved by the LARS-Lasso method [24]. Note that the kth element $\boldsymbol{\gamma}_j(k)$ in the vector $\boldsymbol{\gamma}_j$, which corresponds to the point Ψ_k^i, represents the contribution of Ψ_k^i to reconstruct Ψ_j^i. Obviously, the higher $\boldsymbol{\gamma}_j(k)$ is, the more similar Ψ_j^i and Ψ_k^i are. Since we use $\boldsymbol{\gamma}_j(k)$ for similarity measure, a nonnegativity constraint is imposed for all $\boldsymbol{\gamma}_j(k)$. Let $\Upsilon_i = \left[\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_{m_i-1}\right]$. We normalize Υ_i by setting $\overline{\Upsilon}_i = \Upsilon_i / \|\Upsilon_i\|_F = \left[\overline{\boldsymbol{\gamma}}_1, \overline{\boldsymbol{\gamma}}_2, \cdots, \overline{\boldsymbol{\gamma}}_{m_i-1}\right]$. Note that $0 \le \overline{\boldsymbol{\gamma}}_j(k) \le 1$, $\forall j, k$.

Based on the similarity measure $\overline{\boldsymbol{\gamma}}_j(k)$, we propose the local distance metric $d_i(\cdot)$ for points in Γ_i as:

$$d_i(\Psi_j^i, \Psi_k^i) = \begin{cases} 0 & \text{if } j = k \\ \frac{1}{\frac{\overline{\boldsymbol{\gamma}}_j(k) + \overline{\boldsymbol{\gamma}}_k(j)}{2} + 1} & \text{otherwise} \end{cases} \tag{13}$$

The bias 1 in the denominator is added to handle the case that both $\overline{\boldsymbol{\gamma}}_j(k)$ and $\overline{\boldsymbol{\gamma}}_k(j)$ are zero. Then for all data points in $\mathbf{X}$ (Eq. (10)), the global metric is defined by

$$d(\Psi_p^a, \Psi_q^b) = \begin{cases} d_a(\Psi_p^a, \Psi_q^a) & \text{if } a = b \\ \infty & \text{otherwise} \end{cases} \tag{14}$$

Property 1. *The distance $d(\cdot)$ defined in Eq. (14) is a metric.*

Proof. We consider the case that two points p_1 and p_2 are in the same set Γ_i and the other case can be proved trivially. First, by definition, we have

- $d(p_1, p_2) = d(p_2, p_1)$ and
- $d(p_1, p_2) = 0$, if and only if $p_1 = p_2$.

Second, if $p_1 \neq p_2$, we have $0.5 \le d(p_1, p_2) \le 1$, since $0 \le \overline{\boldsymbol{\gamma}}_j(k) \le 1$, $\forall j, k$, in Eq. (13). Lastly, we show the triangle inequality. Since $0.5 \le d(p_1, p_2) \le 1$ and $0.5 \le d(p_2, p_3) \le 1$, we have $1 \le d(p_1, p_2) + d(p_2, p_3) \le 2$, and thus, $d(p_1, p_3) \le d(p_1, p_2) + d(p_2, p_3)$. That completes the proof. □

3.3 Manifold-structure-preserving sparse coding

GraphSC (ref. Eq. (9)) relies on a weight matrix $\mathbf{W}$ that encodes a k-nearest neighbor graph G. To explore the manifold structure of data points in $\mathbf{X}$, we build the graph G using the metric defined in Eq. (14). Accordingly, the resulting weight matrix $\mathbf{W}$ has a block diagonal structure:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & & \\ & \mathbf{W}_2 & & \\ & & \ddots & \\ & & & \mathbf{W}_{n_{clips}} \end{bmatrix} \tag{15}$$

where the sub-matrix $\mathbf{W}_i$ is constructed by applying the k-nearest neighbor method to the video clip Γ_i using the metric in Eq. (13). We apply the optimization scheme in [16] to solve Eq. (9) with the weight matrix in Eq. (15), from which the optimal dictionary $\mathbf{D}$ and the manifold-structure-preserving sparse representation $\mathbf{S}$ are obtained.

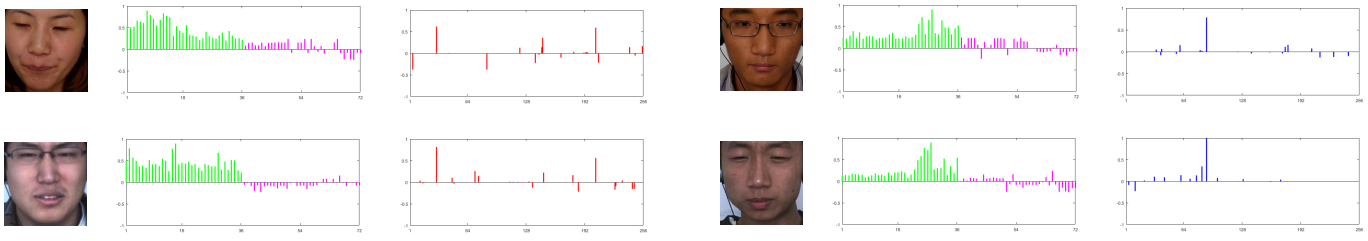


Figure 3. Two representative micro-expressions (disgust and repression), each of which contains two example video clips from the CASME dataset. For each of four video clips, its original MDMO and sparse MDMO features are illustrated. Note that sparse MDMO features show higher similarity for the same micro-expression than the original MDMO features.

3.4 Temporal pooling

Given the sparse representation $\mathbf{S}$ for the video set Π, a micro-expression video clip $\Gamma_i \in \Pi$ can be represented by a coefficient matrix $\mathbf{S}_i = \{s_{kj}^i\} \in \mathbb{R}^{n_d \times (m_i - 1)}$, where n_d is the size of the dictionary. To maintain the simplicity of MDMO in sparse MDMO, biologically-inspired pooling operations [25] can be applied, which further make the feature invariant to small translations and thus more robust. A pooling function replaces a pool of scale values $\{s_{kj}^i\}$, $j = 1, 2, \cdots, m_i - 1$, by a summary statistic; e.g., the max and the average poolings are two representative functions. We use a mixed pooling strategy to obtain a vector representation $\mathbf{z}_i$ for the video clip Γ_i:

$$\begin{aligned} &\mathbf{z}_i = \{z_1^i, z_2^i \cdots, z_{n_d}^i\}, \\ &z_k^i = \omega \max_j\{s_{kj}^i\} + (1-\omega)\text{mean}_j\{s_{kj}^i\}, \\ &\quad j = 1, 2 \cdots, m_i - 1 \end{aligned} \tag{16}$$

where the parameter $\omega \in [0, 1]$ is optimized in Section 4. We call the vector $\mathbf{z}_i$ the *sparse MDMO feature* for the video clip Γ_i. Figure 3 illustrates four examples of sparse MDMO features with the comparison to the original MDMO features, showing that sparse MDMO features have higher similarity for the same micro-expression than the original MDMO features.

3.5 Computational complexity

The computational complexity of computing sparse MDMO features includes three parts:

- compute and collect all atomic optical flow features in Π, which takes $O(n_{clips} k_f m_p)$ time [8], [26];
- build the k-nearest neighbor graph G using the metric (14), which takes $O(n_{clips} k_f^{2.2})$ [24];
- GraphSC optimization and temporal pooling take $O(n_{clips} k_f n_d)$ time [16], [27];

where n_{clips} is the number of video clips in the dataset, k_f is the number of frames in each clip, m_p is the number of pixels in each frame and n_d is the size of dictionary.

4 Experiments

We implement the proposed sparse MDMO feature in MATLAB R2016a and the source code is available[2]. Three sparse versions of MDMO feature are compared:

- BasicSC-MDMO: this feature is trivially obtained by optimizing the objective function in Eq. (6) with the data matrix $\mathbf{X}$ in Eq. (10) and the function $g(\mathbf{s}_i) = \|\mathbf{s}_i\|_1$, followed by the temporal pooling in Section 3.4.
- GraphSC-MDMO: this feature is trivially obtained by optimizing the objective function in Eq. (9) with the data matrix $\mathbf{X}$ in Eq. (10) and using the Euclidean distance metric to construct the k-nearest neighbor graph G, followed by the temporal pooling in Section 3.4.
- Sparse MDMO, this feature is obtained by optimizing the objective function in Eq. (9) with the data matrix $\mathbf{X}$ in Eq. (10) and using the manifold-structure-preserving distance metric in Eq. (14) to construct the k-nearest neighbor graph G, followed by the temporal pooling in Section 3.4.

We compare these sparse features with four representative MER features, including two (LBP-TOP [4] and STCLQP [6]) from the appearance-based class and two (MDMO [8] and FDM [10]) from the optical-flow-based class. In particular, we implement two versions of LBP-TOP feature:

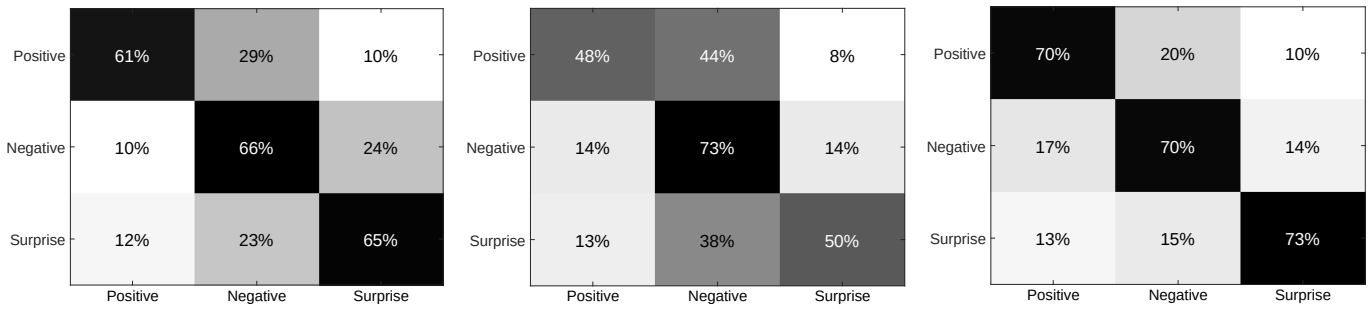
- LBP-TOP: it is the original LBP-TOP feature applied to the entire facial region;
- LBP-TOP-ROIs: it is a combinatorial LBP-TOP feature, constructed by applying the LBP-TOP feature in each of 36 ROIs and consolidating them into one feature.

All aforementioned features are compared on three spontaneous micro-expression datasets, including SMIC [20], CASME [21] and CASME II [22]. In our experiments, leave-one-subject-out (LOSO) cross validation is applied for subject-independent evaluation, i.e., in each fold, one subject is used as the test set, and the others are used as the training set. After n folds, where n is the number of subjects in the dataset, each subject has been used as the test set once, and the final recognition accuracy was calculated based on all of the results. In addition to LOSO cross validation, other commonly used metrics including precision, recall and F_1 rate are also evaluated. LIBSVM [28] with the polynomial kernel is used for multiclass classification, i.e., for a dataset with k classes, $k(k-1)/2$ classifiers are constructed, each of which is used to train data from two classes.

2. http://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm

Feature	Dictionary size n_d								
	SMIC			CASME			CASME II		
	$n_d = 128$	$n_d = 192$	$n_d = 256$	$n_d = 128$	$n_d = 192$	$n_d = 256$	$n_d = 128$	$n_d = 192$	$n_d = 256$
BasicSC-MDMO	63.46%	68.59%	67.95%	70.86%	70.86%	62.91%	58.05%	62.29%	60.59%
GraphSC-MDMO	65.38%	66.67%	67.95%	72.19%	72.19%	72.19%	56.78%	63.56%	60.17%
Sparse MDMO	66.67%	69.23%	70.51%	74.83%	73.51%	74.83%	59.75%	63.56%	66.95%
MDMO	58.97%			56.29%			51.69%		

Table 1
In sparse MDMO representations, the average LOSO recognition rates in three spontaneous micro-expression datasets are not sensitive to different dictionary sizes; their results are all better than the results from the original MDMO feature.



(a) STCLQP (b) MDMO (c) Sparse MDMO

Figure 4. The confusion matrices of STCLQP, MDMO and sparse MDMO on the SMIC datset. MDMO has the best performance for recognizing *negative* micro-expression. Sparse MDMO has the best performance for recognizing the other two micro-expressions and has the best average performance over three classes.

4.1 Parameter setting

The three sparse versions of MDMO feature have the same vector representation as the original MDMO feature. So they can be used in the same recognition process as that in MDMO. The parameters in our proposed method are specified as follows:

- the size of dictionary n_d, the regularization parameter w_g, the sparsity balance weight λ and the pooling parameter ω: these parameters are optimized by five-fold cross validation with the candidate values $\{128, 192, 256\}$ for n_d, $\{0.01, 0.1, 1, 10, 100\}$ for w_g, $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ for λ and $\{0.01i\}_{i=0}^{100}$ for ω.
- the number of nearest neighbors k: following [16], we set $k = 5$ empirically;

In practice, our sparse representations are not sensitive to these parameters. Table 1 summarizes the results of three sparse MDMO representations with three dictionary sizes $n_d = 128$, 192 and 256 on three datasets. As a comparison, the results of the original MDMO feature are also summarized in Table 1. The results show that for all three dictionary sizes, the sparse MDMO representations are all better than the original MDMO feature. The results summarized in Table 2 demonstrate that the performance of sparse MDMO is stable with the parameter k in the range of 3 to 7.

Furthermore, we use the optimal parameters specified in [8] for MDMO and LBP-TOP-ROIs, the optimal parameters specified in [6] for STCLQP and LBP-TOP, and the optimal parameters specified in [10] for FDM.

Dataset	the number of nearest neighbors				
	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
SMIC	70.51%	69.87%	70.51%	70.51%	71.79%
CASME	74.83%	74.17%	74.83%	74.17%	73.51%
CASME II	64.83%	65.68%	66.95%	63.98%	63.16%

Table 2
The average LOSO recognition rates of sparse MDMO in three spontaneous micro-expression datasets are stable with the number of nearest neighbors k in the range of $[3, 7]$.

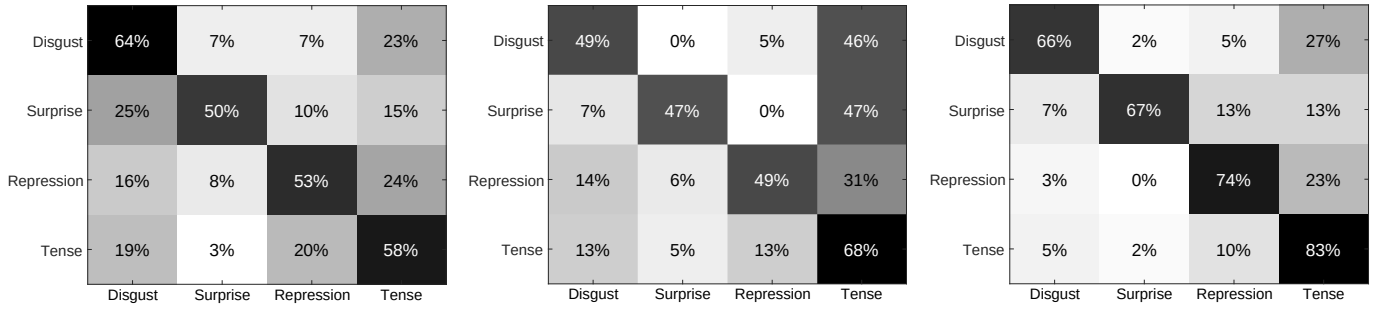
4.2 Experimental results

Evaluation on SMIC. The SMIC dataset [20] has three subsets and we take the largest subset SMIC-HS in our experiment. SMIC-HS contains 164 spontaneous micro-expression video clips recorded from 16 subjects in three classes: positive, negative and surprise. All video data were recorded by a high speed camera of 100 fps with 640×480 resolution. We follow [29] to optimize the normalized frame number to 20 for each clip using the temporal interpolation model (TIM). The LOSO recognition rates of eight features, averaged over three classes, are summarized in Table 3. The results show that in previously existing features, the best appearance-based feature is STCLQP, whose average recognition rate is 64.02%, and the best optical-flow feature is MDMO, whose average recognition rate is 58.97%. The sparse representations of the MDMO feature improve the performance and the sparse MDMO is the best sparse MDMO feature (70.51%), demonstrating that manifold-preserving sparse coding with our proposed new metric (Eq. (14)) achieves good discriminating power for classification. We further compare the confusion matrices of STCLQP, MDMO and sparse MDMO in Figure 4. The results show that MDMO has the best performance for recognizing *negative* micro-expression; while sparse MDMO has the best performance for recognizing the other two micro-expressions and has the best average performance over three classes.

Evaluation on CASME. The CASME dataset [21] contains 195 spontaneous micro-expression video clips recorded from 20 subjects in seven classes. Since the three classes

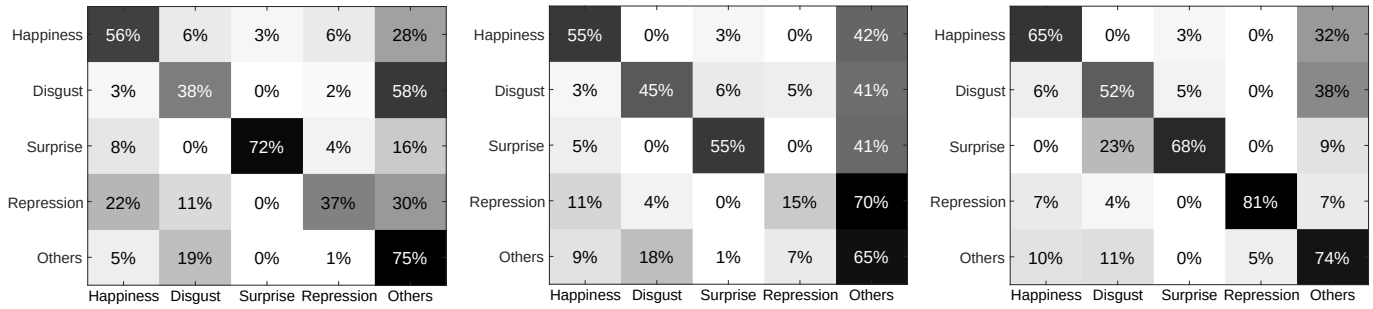
Feature	SMIC				CASME				CASME II			
	LOSO	P	R	F_1	LOSO	P	R	F_1	LOSO	P	R	F_1
LBP-TOP	53.66%	53.62%	53.69%	53.65%	37.43%	36.35%	30.14%	32.96%	46.46%	41.52%	30.87%	35.41%
LBP-TOP-ROIs	51.28%	50.54%	49.38%	49.95%	53.64%	56.64%	45.87%	50.69%	44.49%	40.88%	30.28%	34.79%
STCLQP	64.02%	64.69%	64.06%	64.37%	57.31%	56.30%	56.06%	56.18%	58.39%	59.95%	55.18%	57.47%
FDM	54.88%	55.63%	52.74%	54.17%	56.14%	57.36%	52.82%	54.99%	45.93%	43.32%	29.63%	35.19%
MDMO	58.97%	60.08%	56.91%	58.45%	56.29%	58.17%	53.09%	55.51%	51.69%	52.24%	47.33%	49.66%
BasicSC-MDMO	68.59%	70.03%	68.08%	69.04%	70.86%	70.31%	65.63%	67.89%	62.29%	64.82%	58.56%	61.53%
GraphSC-MDMO	67.95%	68.03%	68.87%	68.44%	72.19%	76.87%	68.35%	72.36%	63.56%	65.01%	62.34%	63.64%
Sparse MDMO	**70.51%**	**70.09%**	**70.73%**	**70.41%**	**74.83%**	**77.59%**	**72.54%**	**74.98%**	**66.95%**	**69.81%**	**68.42%**	**69.11%**

Table 3
Average LOSO recognition rates, precision (P), recall (R) and F_1 metrics of different features in three spontaneous micro-expression datasets. The best result for each dataset is shown in bold.



(a) STCLQP (b) MDMO (c) Sparse MDMO

Figure 5. The confusion matrices of STCLQP, MDMO and sparse MDMO on the CASME dataset. Sparse MDMO has the best performance for recognizing each of four micro-expressions.



(a) STCLQP (b) MDMO (c) Sparse MDMO

Figure 6. The confusion matrices of STCLQP, MDMO and sparse MDMO on the CASME II dataset. STCLQP has the best performance for recognizing *surprise* and *others* micro-expressions. Sparse MDMO has the best performance for recognizing all other three micro-expressions.

of happiness, fear and sadness contain very few samples, we chose the remaining four classes in our experiment: disgust, surprise, repression and tense. All video data in CASME were recorded by a 60 fps camera with 1280×720 resolution. We follow [29] to optimize the normalized frame number to 64 for each clip using TIM. The LOSO recognition rates of eight features, averaged over four classes, are summarized in Table 3. The results show that in previously existing features, the best appearance-based feature is STCLQP (57.31%), and the best optical-flow feature is MDMO (56.29%). All three sparse representations of the MDMO feature improve the performance and the best sparse MDMO feature is sparse MDMO (74.83%). These results are consistent with those in SMIC. We further compare the confusion matrices of STCLQP, MDMO and sparse MDMO in Figure 5. The results show that sparse MDMO has the best performance for recognizing each of four micro-expressions.

Evaluation on CASME II. The CASME II dataset [22] contains 246 spontaneous micro-expression video clips recorded from 26 subjects in five classes: happiness, surprise, disgust, repression and others. All video data in CASME II were recorded by a high speed camera of 200 fps with 640×480 resolution. The cropped facial area in each video frame is only of 170×140. We follow [29] to optimize the normalized frame number to 90 for each clip using TIM. The LOSO recognition rates of eight features, averaged over five classes, are summarized in Table 3. The results show that in existing features, the best appearance-based feature is STCLQP (58.39%), and the best optical-flow feature is MDMO (51.69%). Consistent with SMIC and CASME, on CASME II, all three MDMO-based sparse representations improve the performance and the best sparse MDMO feature is sparse MDMO (66.95%). We further compare the confusion matrices of STCLQP, MDMO and sparse MDMO in Figure 6. The results show that STCLQP has the

best performance for recognizing *surprise* and *others* micro-expressions, while sparse MDMO has the best performance for recognizing other three micro-expressions and has the best average performance over five classes.

We note that micro-expression labels in [8] are grouped in a non-standard manner; e.g., four classes of positive, negative, surprise and other are used for CASME II in [8], but in the original CASME II dataset [22], five classes of happiness, surprise, disgust, repression and other are used. In this paper, to facilitate comparison with state-of-the-art methods, we use the standard labels from the original datasets, and thus the results of MDMO and LBP-TOP are different from [8].

Statistical significance. The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures and is used to detect differences in treatments across multiple test attempts [30]. We performed the Friedman test for statistical significance measurement, since it is suitable for comparing multiple algorithms on different datasets. To reduce randomness, 10-fold cross validation was run 10 times on all the three datasets. We compare sparse MDMO with original MDMO, BasicSC-MDMO and GraphSC-MDMO. The differences between these features were all statistically significant ($p < 0.001$). In statistics, the Nemenyi test is a post-hoc test intended to find the groups of data that differ after performing the Friedman test [31]. In our case, a follow-up Nemenyi test showed that the mean ranking of original MDMO, BasicSC-MDMO, GraphSC-MDMO and sparse MDMO were 3.613, 2.235, 2.278 and 1.873 (where 1 is the best and 4 is the worst), and the improvement of sparse MDMO over GraphSC-MDMO ($p < 0.001$), and all three sparse features over the original MDMO (all with $p < 0.001$) were statistically significant.

4.3 Comparison with other state of the art

We further compare sparse MDMO with [29], which uses an appearance-based feature called *HIGO* and has state-of-the-art performance (LOSO recognition rate 65.24% for SMIC and 57.09% for CASME II). Their results show that even TIM and Eulerian motion magnification can significantly improve the HIGO performance (68.29% for SMIC and 67.21% for CASME II), sparse MDMO still has comparable performance (70.51% for SMIC and 66.95% for CASME II).

Different from designing elaborate artificial features for MER, recently blossoming deep learning methods can automatically learn effective features in a multi-layer style. Some pioneering works [32], [33], [34] have applied deep learning methods to MER. However, their performances (LOSO recognition rates are 47.3%, 60.98% and 59.47% for CASME II, respectively) are inferior to the performance of sparse MDMO (66.95% for CASME II).

5 Conclusion

In this paper, we propose an effective sparse representation that learns a discriminative feature called *sparse MDMO* for spontaneous micro-expression recognition. To introduce sparsity into the original MDMO feature, we construct a data set $\mathbf{X}$ that contains all the atomic optical flow features in video frames. We further propose a new distance metric (Eq.(14)) in the MDMO feature space, such that the underlying manifold structure inherent in $\mathbf{X}$ can be revealed. By incorporating this new metric into the classic GraphSC scheme, an efficient sparse representation for micro-expression recognition is built and the concise sparse MDMO feature is obtained by applying temporal pooling to this sparse representation. Experimental results on three spontaneous micro-expression datasets (SMIC, CASME and CASME II) show that sparse MDMO outperforms the state-of-the-art features including LBP-TOP, STCLQP, MDMO and FDM.

References

[1] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[2] M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements," *Scientific Reports*, vol. 6, p. 22049, 2016.

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

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

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

[6] 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.

[7] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1932–1939.

[8] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affective Computing*, vol. 7, no. 4, pp. 299–310, 2016.

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

[10] F. Xu, J. Zhang, and J. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[11] A. C. L. Ngo, J. See, and C.-W. R. Phan, "Sparsity in dynamics of spontaneous subtle emotion: Analysis & application," *IEEE Trans. Affective Computing*, DOI:10.1109/TAFFC.2016.2523996, 2017.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[13] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2691–2698.

[14] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[15] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.

[16] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Processing*, vol. 20, no. 5, pp. 1327–1336, 2011.

[17] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph laplacian sparse coding, and applications," *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 92–104, 2013.

[18] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview Hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50–60, 2014.

[19] W. Liu, Z.-J. Zha, Y. Wang, K. Lu, and D. Tao, "p-Laplacian regularized sparse coding for human activity recognition," *IEEE Trans. Industrial Electronics*, vol. 63, no. 8, pp. 5120–5129, 2016.
[20] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.
[21] W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, vol. 136, pp. 82–87, 2014.
[22] 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.
[23] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3444–3451.
[24] S. Han, H. Huang, H. Qin, and D. Yu, "Locality-preserving l1-graph and its application in clustering," in *ACM Symp. Applied Computing*, 2015, pp. 813–818.
[25] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *International Conference on International Conference on Machine Learning (ICML)*, 2010, pp. 111–118.
[26] T. Senst, V. Eiselein, and T. Sikora, "Robust local optical flow for feature tracking," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1377–1387, 2012.
[27] T. Ge, K. He, and J. Sun, "Product sparse coding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 939–946.
[28] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.
[29] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affective Computing*, DOI:10.1109/TAFFC.2017.2667642, 2017.
[30] A. Field, *Discovering Statistics using IBM SPSS Statistics*. 4th ed., SAGE Publications Ltd, 2013.
[31] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*. 3rd ed., Wiley, 2013.
[32] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2258–2263.
[33] 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 (MM'16)*, 2016, pp. 382–386.
[34] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," in *arXiv:1705.01842v2*, 2017.

Yong-Jin Liu received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2004. He is an associate professor with BNRist, Department of Computer Science and Technology, Tsinghua University, China. His research interests include computational geometry, computer graphics and computer-aided design. He is a senior member of the IEEE and a member of ACM. For more information, visit http://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm

Bing-Jun Li received his BEng degree from Beijing University of Posts and Telecommunications, China, in 2014. Currently, he is now a master student at Department of Computer Science and Technology, Tsinghua University, China. His research interests include computer vision and image processing.

Yu-Kun Lai is a Senior Lecturer at School of Computer Science and Informatics, Cardiff University, UK. He received his B.S and PhD degrees in Computer Science from Tsinghua University, in 2003 and 2008 respectively. His research interests include computer vision, computer graphics and geometric computing. For more information, visit https://users.cs.cf.ac.uk/Yukun.Lai/