Micro-expression Recognition using Robust Principal Component Analysis and Local Spatiotemporal Directional Features

Su-Jing Wang[1,3], Wen-Jing Yan[1], Guoying Zhao[2], Xiaolan Fu[1], and Chun-Guang Zhou[3]

[1] State Key Lab of Brain and Cognitive Science,
Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
{wangsujing, yanwj, fuxl}@psych.ac.cn
[2] Center for Machine Vision Research, University of Oulu, Finland
gyzhao@ee.oulu.fi
[3]College of Computer Science and Technology, Jilin University,
Changchun 130012, China.
cgzhou@jlu.edu.cn

Abstract. One of important cues of deception detection is micro-expression. It has three characteristics: short duration, low intensity and usually local movements. These characteristics imply that micro-expression is sparse. In this paper, we use the sparse part of Robust PCA (RPCA) to extract the subtle motion information of micro-expression. the local texture features of the information are extracted by Local Spatiotemporal Directional Features (LSTD). In order to extract more effective local features, 16 Regions of Interest (ROIs) are assigned based on the Facial Action Coding System (FACS). The experimental results on two micro-expression databases show the proposed method gain better performance. Moreover, the proposed method may further be used to extract other subtle motion information (such as lip-reading, the human pulse, and micro-gesture etc.) from video.

Keywords: Micro-expression Recognition, Sparse Representation, Dynamic Features, Local Binary Pattern, Subtle motion extraction.

1 Introduction

In our social life, deception is a reality. Its detection can be beneficial, not only to an individual but also to the whole society. Currently, the most widely used system is the polygraph which monitors uncontrolled changes in heart rate and electro-dermal response, as a result of the subject's arousal to deceit [12]. However, the polygraph is an overt system, which makes people realize that they are being monitored. As a result, some people may trick the machine by employing some anti-polygraph techniques, such as remaining calm and controlling their heart rate.

The recent studies [22][11][6][5] show that micro-expression can reveal an emotion that a person tries to conceal, especially in high-stake situations. Therefore, micro-expression might be treated as an effective cue for deception detection [5]. Compared with ordinary facial expressions, micro-expressions have three significant characteristics: short duration, low intensity and usually local movements (fragments of prototypical facial expressions).

Because of these characteristics, human beings are difficult to detect and recognize micro-expression. In order to improve the human's performance on recognizing micro-expression, Ekman [4] developed the Micro-Expression Training Tool (METT), which trains people to better recognize micro-expression. To better apply micro-expression as a cue to detect deception in practice, computer scientists try to train the computer to automatically recognize micro-expression.

Hitherto, there are just several papers on micro-expression recognition. Polikovsky *et al.* [16] used 3D-gradient descriptor for micro-expressions recognition. Wang *et al.* [19] treated a gray-scale video clip of micro-expression as a 3rd-order tensor and used Discriminant Tensor Subspace Analysis (DTSA) and Extreme Learning Machine (ELM) to recognize micro-expression. Pfister *et al.* [15] utilized a temporal interpolation model (TIM) [25] based on Laplacian matrix to normalize the frame numbers of micro-expression video clips. In addition, the LBP-TOP [23] was used to extract the motion and appearance features of micro-expressions and multiple kernel learning was used to classify the features.

We emphasize again the two important characteristics of micro-expression: short duration and low intensity, which makes that the micro-expression data are sparse in both temporal and spatial domains. The key problem is how to extract the sparse information. In this paper, the sparse information are extracted by Robust PCA (RPCA) [20]. RPCA is widely used for face recognition [18], video frame interpolation [2], brain imaging [8] and EEG signal processing [17] etc. RPCA leverage on the fact that the data are characterized by low-rank subspaces [3]. It decomposes the observed data matrix $\mathbf{D}$ into two parts:

$$\mathbf{D} = \mathbf{A} + \mathbf{E} \tag{1}$$

where $\mathbf{A}$ lies in a subspace of low rank and $\mathbf{E}$ is the error term. In many applications of RPCA, $\mathbf{A}$ is the deserved data and $\mathbf{E}$ is usually treated as noise and removed. In this paper, however, $\mathbf{E}$ includes the deserved subtle motion information of micro-expression. In the following, we use Local Spatiotemporal Directional Features (LSTD) [24] to extract local dynamic texture features of the subtle facial motion information from 16 Regions of Interest (ROIs) based on the Facial Action Coding System (FACS).

2 Extraction of motion information

In a micro-expression video clip, the subtle facial motion information is discriminant for recognizing micro-expression. Other information, such as identity information, is not useful for micro-expression recognition. This information accounts for the great proportion of whole information in a clip. Relatively, the

subtle facial motion information is *sparse*. We aim to extract this sparse information.

Given a gray video clip $\mathcal{V} \in \mathbb{R}^{h \times w \times f}$ with h pixels height, w pixels width, and f frames. Each of its frame is vectorized as a column of matrix $\mathbf{D}$ with $h \times w$ rows and f columns. $\mathbf{D}$ consists of two parts: $\mathbf{D} = \mathbf{A} + \mathbf{E}$. Each column of $\mathbf{A}$ is the same with each other as possible. $\mathbf{E}$ includes the derived sparse subtle motion information. This may be formulated as follows:

$$\min_{\mathbf{A},\mathbf{E}} rank(\mathbf{A}) + \|\mathbf{E}\|_0 \qquad \text{subject to} \qquad \mathbf{D} = \mathbf{A} + \mathbf{E} \tag{2}$$

where $rank(\cdot)$ denotes the rank of matrix and $\|\cdot\|_0$ denotes ℓ_0-norm, which counts the number of nonzero entries. This is a non-convex problem. RPCA [20] converted Eq.(2) into the following convex optimization problem:

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 \qquad \text{subject to} \qquad \mathbf{D} = \mathbf{A} + \mathbf{E} \tag{3}$$

where $\|\cdot\|_*$ denotes the nuclear norm, which is the sum of its singular values, and $\|\cdot\|_1$ denotes ℓ_1-norm, which is the sum of the absolute values of matrix entries. λ is a positive weighting parameter.

Eq.(3) involves minimizing a combination of both the ℓ_1-norm and the nuclear norm. In [20], the iterative thresholding technique is used. However, the iterative thresholding scheme converges extremely slowly. It can not be used to deal with the large-scale micro-expression video clips. Lin *et al.* [10] applied the method of augmented Lagrange multipliers (ALM) to solve Eq.(3) and improved the efficiency by more than five times of those in [20]. ALM is introduced for solving the following constrained optimization problem:

$$\min f(X) \qquad \text{subject to} \qquad h(X) = 0 \tag{4}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$. The augmented Lagrangian function can be defined as follows:

$$L(X, Y, \mu) = f(X) + \langle Y, h(X)\rangle + \frac{\mu}{2}\|h(X)\|_F^2 \tag{5}$$

Let X be $(\mathbf{A}, \mathbf{E})$, $f(X)$ be $\|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1$, and $h(X)$ be $\mathbf{D} - \mathbf{A} - \mathbf{E}$. Eq. (5) is re-written:

$$L(\mathbf{A}, \mathbf{E}, Y, \mu) = \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 + \langle Y, \mathbf{D} - \mathbf{A} - \mathbf{E}\rangle + \frac{\mu}{2}\|\mathbf{D} - \mathbf{A} - \mathbf{E}\|_F^2 \tag{6}$$

In [10], Lin *et al.* proposed two algorithms: exact ALM and inexact ALM. A slight improvement over the exact ALM leads to the inexact ALM, which converges practically as fast as the exact ALM, but the required number of partial SVDs is significantly less. Here, we chose inexact ALM to extract the subtle facial motion information.

Fig. 1 shows several frames (Figs.1(a)-1(e)) of an micro-expression video clip and its corresponding subtle motion frames (Figs.1(f)-1(j)) of $\mathbf{E}$. It is difficult

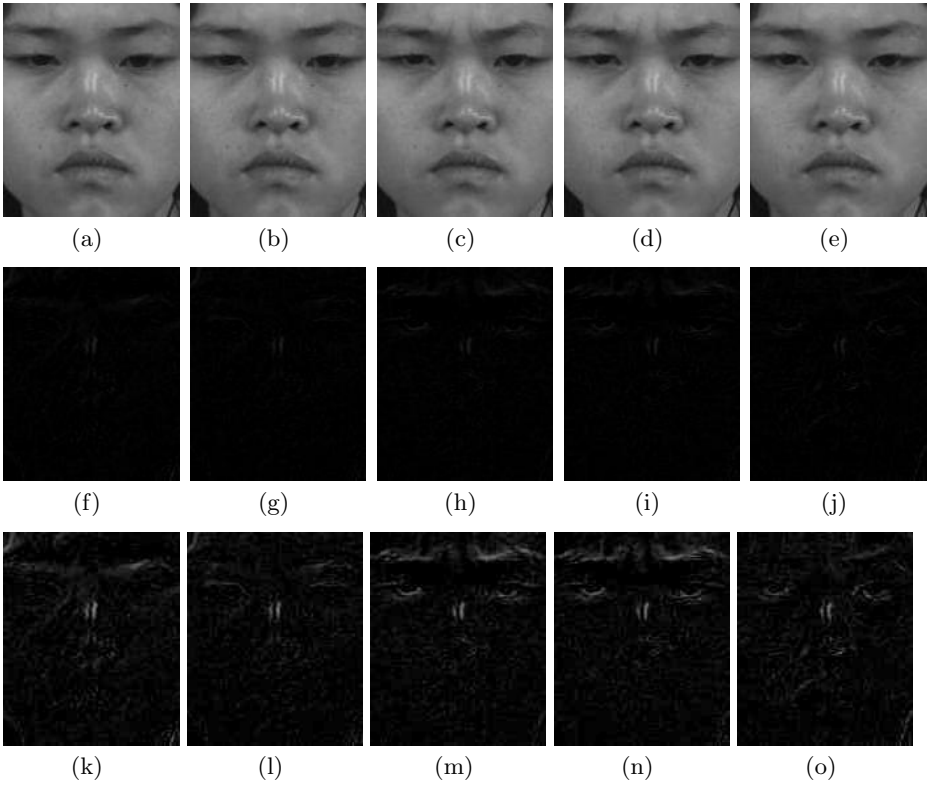

Fig. 1. An examples of extracting the subtle facial motion information. (a)-(e)indicate the original micro-expression frames sequence; (f)-(j)indicate the extracted subtle facial motion information; (k)-(o) indicates the enhanced display for (f)-(j) by multiplying each pixels with 4 (Considering some monitors may not well display the subtle facial motion information since the maximum gray value of (f)-(j) is less than 40)

for people to detect the subtle facial movement from Figs.1(a)-1(e). However, it is easy to detect the movement of eyebrows from Figs.1(f)-1(j). Moreover, there is very little identity information in Figs.1(f)-1(j). This may further improve subsequent classification accuracy.

It should be noticed that the frame numbers of micro-expression video clips are not the same. This leads to different sparse results for micro-expression with different frame numbers when setting the same λ. In order to address the problem, the frame numbers of all video clips are normalized by using linear interpolation.

3 Local Spatiotemporal Directional Features

In this section, we use Local Spatiotemporal Directional Features (LSTD) [24] to extract dynamic texture features of the subtle facial motion information. LSTD is built on the basis of the local binary pattern (LBP) [13].

In order to extract spatiotemporal texture, the local binary pattern from three orthogonal planes (LBP-TOP) [23] was proposed. With this approach the ordinary LBP for static images was extended to spatiotemporal domain. The LBP-TOP were developed for facial expression recognition [23] and micro-expression recognition [15].

Comparing with LBP-TOP, LSTD can extract more detailed spatiotemporal directional changes. In XY plane, there are mainly X and Y directions. Likewise, in XT and YT planes, there are X (Y) and T directions. The motion information changes in each direction further contributes to micro-expression recognition.

Fig. 2 shows a subtle motion video got in previous section and its images from the three planes. For each plane, we code it from two directions. Given a pixel, we can obtain a 3×3 neighboring area around it in each plane. For each row (column), a binary code is produced by thresholding its neighborhood with the value of the center pixel in this row (column). Fig. 2 shows the calculation for the X direction and Y direction in XY plane. Similarly, we can obtain the X direction and T direction in XT plane and Y direction and T direction in YT plane.

For each central pixel, we can get eight neighboring points and in total nine points in the calculation. The sampling distance of each direction can be changed. Fig. 3 shows that the sampling radius in X direction is three, the radius in Y direction is two, and the radius in T direction is four. So we can set R_x, R_y, and R_t with different values to represent the sampling radii in three directions. Then we could obtain $P_0, P_1, \ldots, P_8$ corresponding to, e.g. $I(x_c - R_x, y_c - R_y, t_c)$, $I(x_c, y_c - R_y, t_c)$, $I(x_c + Rx, y_c - Ry, tc)$, $I(x_c - Rx, y_c, t_c)$, $I(x_c, y_c, t_c)$, $I(x_c + R_x, y_c, t_c)$, $I(x_c - R_x, y_c + R_y, t_c)$, $I(x_c, y_c + R_y, t_c)$, $I(x_c + R_x, y_c + R_y, t_c)$ in XY plane, where R_x and R_y are sampling radii in X direction and Y direction, respectively. Similarly we can have R_t as sampling radius in T direction of XT and YT planes. From this area the local spatial-temporal feature is calculated for central pixel $P_4 = I(x_c, y_c, t_c)$.

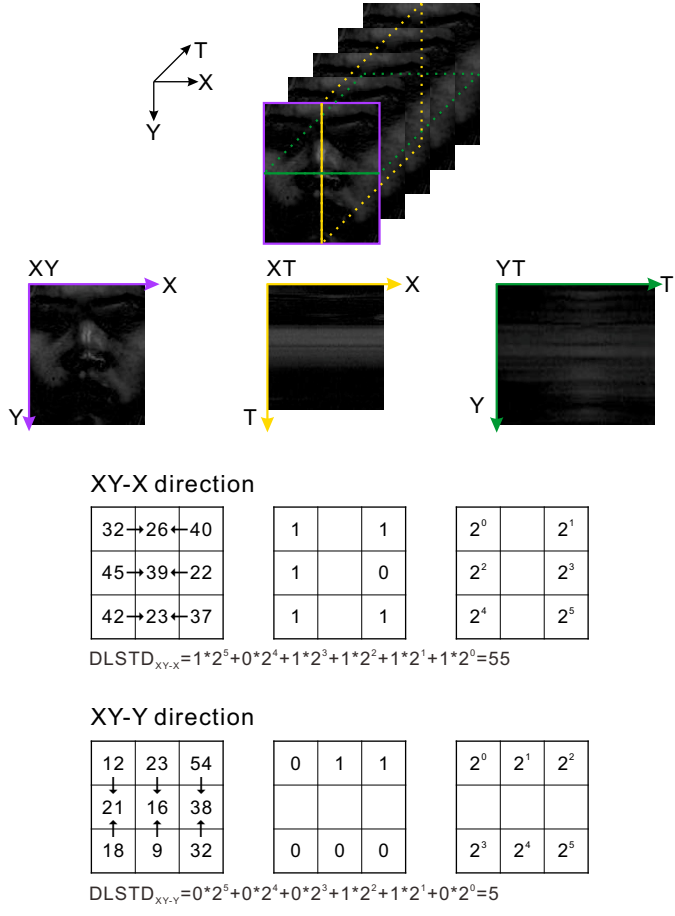

Fig. 2. Illustration of Local Spatiotemporal Directional Features.

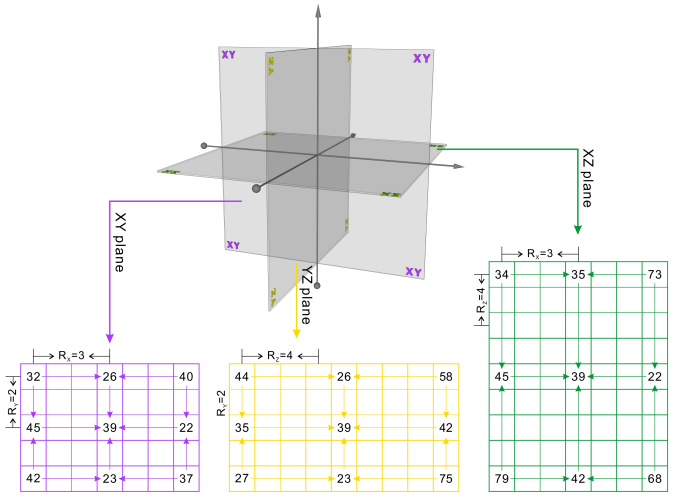

Fig. 3. Sampling in three planes with radius three in X direction, radius two in Y direction, and radius four in T direction.

The obtained neighboring area can be formulated as a 3×3 matrix

$$\mathbf{F} = \begin{bmatrix} P_0 & P_1 & P_2 \\ P_3 & P_4 & P_5 \\ P_6 & P_7 & P_8 \end{bmatrix} \tag{7}$$

We left-multiply $\mathbf{F}$ by $\mathbf{W}_l$:

$$\mathbf{W}_l = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \tag{8}$$

and obtain

$$\mathbf{W}_l\mathbf{F} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} P_0 & P_1 & P_2 \\ P_3 & P_4 & P_5 \\ P_6 & P_7 & P_8 \end{bmatrix} = \begin{bmatrix} P_0 - P_3 & P_1 - P_4 & P_2 - P_5 \\ P_6 - P_3 & P_7 - P_4 & P_8 - P_5 \end{bmatrix}. \tag{9}$$

Here, we denote every entries of the obtained 2×3 matrix as

$$\begin{bmatrix} b_0 & b_1 & b_2 \\ b_3 & b_4 & b_5 \end{bmatrix}. \tag{10}$$

If the neighboring area matrix $\mathbf{F}$ comes from the XY plane, the code of LSTD in the $XY - Y$ direction is calculated by:

$$LSTD = \sum_{n=0}^{5} sign(b_n)2^n \tag{11}$$

where $sign(b)$ is 1 if $b \geq 0$ and 0 otherwise. Similarly, we can obtain the codes of LSTD in the $XT-T$ and $YT-Y$ directions.

We right-multiply $\mathbf{F}$ by $\mathbf{W}_r$:

$$\mathbf{W}_r = \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{bmatrix}, \tag{12}$$

we have

$$\mathbf{F}\mathbf{W}_r = \begin{bmatrix} P_0 & P_1 & P_2 \\ P_3 & P_4 & P_5 \\ P_6 & P_7 & P_8 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} P_0 - P_1 & P_2 - P_1 \\ P_3 - P_4 & P_5 - P_4 \\ P_6 - P_7 & P_8 - P_7 \end{bmatrix}. \tag{13}$$

Here, we denote every entries of the obtained 3×2 matrix as

$$\begin{bmatrix} b_0 & b_1 \\ b_2 & b_3 \\ b_4 & b_5 \end{bmatrix}. \tag{14}$$

If the neighboring area matrix $\mathbf{F}$ comes from the XY plane, the code of LSTD in the $XY-X$ direction is calculated by Eq. (11). Similarly, we obtain the codes of LSTD in the $XT-X$ and $YT-T$ directions. Through the procedures above, six directions have been encoded. Since the coefficients of these directions are correlated, the coefficients should be decorrelated before quantization [14].

Assuming Gaussian distribution, independence can be achieved using a whitening transform:

$$\mathbf{g} = \mathbf{V}^T \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & b_4 & b_5 \end{bmatrix}^T \tag{15}$$

where $\mathbf{V}$ is an orthonormal matrix derived from the Singular Value Decomposition (SVD) of the covariance matrix of the transform coefficient vector $\mathbf{b}$ and $\mathbf{V}$ can be solved in advance. For details, please refer to [14]. The code of Decorrelated LSTD (DLSTD) is calculated by:

$$DLSTD = \sum_{n=0}^{5} sign(g_n) 2^n \tag{16}$$

where g_n are the elements of $\mathbf{g}$.

4 Action Unit and Region of Interest

For each micro-expression video clip $\mathcal{V} \in \mathbb{R}^{h \times w \times f}$ with h pixels height, w pixels width, and f frames, we used RPCA to extract its subtle motion information $\mathbf{E}$. The matrix $\mathbf{E}$ is reshaped as a 3 dimensional array $\mathcal{E} \in \mathbb{R}^{h \times w \times f}$. LSTD is used on $\mathcal{E}$ to extract texture feature. Similar to LBP, LSTD is also based on local texture feature. With LSTD, the object has to be divided into patches. For each patch, six directional histograms are calculated by LSTD. In this paper, we divided a face image into 16 Regions of Interest (ROIs) based on the Facial Action Coding System (FACS) [7]. Each ROI is a patch.

FACS is an objective method for quantifying facial movement based on a combination of 57 elementary components. These elementary components, known as action units (AUs) and action descriptors (ADs), can be seen as the *phonemes* of facial expressions: words are temporal combinations of phonemes. Similar to facial expressions, micro-expressions are spatial combinations of AUs. Each AU depicts a local facial movement. We selected a frontal neutral facial image as the template face and divided the template face into 16 ROIs. These ROIs do not exactly correspond to the AUs. Since there are overlaps between some AUs, the ROIs were modified to be more independent between each other.

The template face is used not only to draw ROIs but also to avoid the large variations in the spatial appearance of micro-expressions. All faces were normalized to a template face. First, the template face is marked with 68 landmarks ψ_M by using the Active Shape Model (ASM) [1]. Second, the first frame of a sample micro-expression clip was marked 68 landmarks ψ_{f1}, and estimated the 2D geometric transformation of the template face as $\psi_M = T\psi_{f1}$, where T is the transformation matrix. Third, the remaining frames are registered to the template face by applying the transformation T. Because there is little head movement in the video clip, the transformation T can be used to all the frames in the same video clip. The sizes of each frame of samples are normalized to 163×134 pixels.

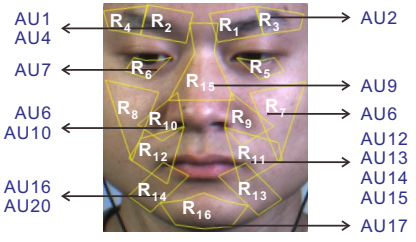

Fig. 4. The template face and 16 ROIs.

Fig. 4 shows the template face, the 16 ROIs and their corresponding AUs. Fox example, ROI R_1 (or R_2) corresponds to AU1 and AU4 which represent the movements of inner eyebrows. Table 1 lists the 16 ROIs, the corresponding AUs and the facial movements. The ROIs are drawn to exclude some noises such as the nose tip and the eye ball movement.

5 Experiments

5.1 SMIC

Spontaneous MIcro-expression Corpus(SMIC) [9] consists of 164 samples from 16 participants. These micro-expressions were elicited by emotional video episodes

Table 1. ROIs, the corresponding AUs and the facial movements.

ROIs	AUs	Facial Movements
R_1, R_2	AU1, AU4	inner eyebrows
R_3, R_4	AU2	outer eyebrows
R_5, R_6	AU7	lower eyelid
R_7, R_8	AU6	cheeks
R_9, R_{10}	AU6, AU10	side of the nose
R_{11}, R_{12}	AU12, AU13, AU14, AU15	mouth corner
R_{13}, R_{14}	AU16, AU20	side of the chin
R_{15}	AU9	nose root
R_{16}	AU17	chin

in a lab situation, with the resolution of 640×480 pixels. SMIC includes 3 data sets: HS (recorded by a high speed camera), VIS (recorded by a normal visual camera), and NIR (recorded by a near-infrared camera). In the experiments, we use HS data set, which consists of 164 samples. These samples are labeled into 3 classes: positive (51 samples), negative (70 samples) and surprise (43 samples). Positive indicates happiness, while negative may include disgust, fear, sadness and anger. Fig. 5 illustrates an example from SMIC.

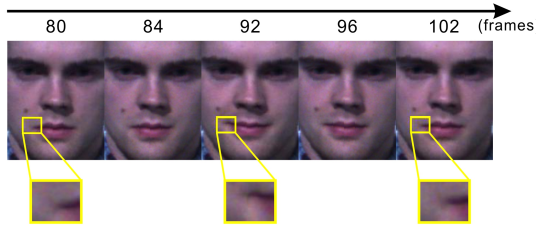

Fig. 5. A demonstration of the frame sequence in a micro-expression from SMIC. The apex frame presents at about 92 frames. The AUs for this micro-expression is 12. The three rectangles below the images show the lip corners (AU 12) in *zoom in* mode.

The frame numbers of all samples are normalized to 60 by using linear interpolation. The size of each frame is normalized to 163×134 pixels. Thus, each sample was normalized to a three dimensional array with the size of $163 \times 134 \times 60$. We used the leave-one-subject-out cross-validation in the experiments.

We conduct LBP-TOP, LSTD, and DLSTD on the micro-expression videos and the corresponding sparse subtle motion videos. When we use Eq.(3) to obtain the sparse parts $\mathbf{E}$, the parameter λ is set as 0.0017. For LBP-TOP, the number of neighboring points in the XY, XT and YT planes were all set as 4. The radius in axes X and Y were assigned various values from 1 to 4. To avoid too much combinations of parameters, we made $R_x = R_y$ and denoted as R_{xy}. The

radii in axes T was assigned various values from 2 to 4. The results are listed in Table 2.

Table 2. Micro-expression recognition accuracies (%) of LBP-TOP, LSTD, and DLSTD on SMIC. +V means the algorithm is conducted on the original micro-expression videos. +S means the algorithm is conducted on the sparse subtle motion videos.

R_{xy}	R_t	LBP-TOP+V	LBP-TOP+S	LSTD+V	LSTD+S	DLSTD+V	DLSTD+S
1	2	**67.0732**	54.8780	64.0244	52.4390	60.3659	55.4878
	3	65.2439	53.0488	**71.3415**	53.0488	60.3659	57.3171
	4	63.4146	56.7073	**65.8537**	57.3171	60.3659	62.1951
2	2	60.9756	**61.5854**	54.8780	58.5366	59.7561	60.3659
	3	57.3171	59.7561	59.7561	59.7561	61.5854	**64.6341**
	4	61.5854	61.5854	57.3171	57.3171	60.9756	**67.6829**
3	2	62.1951	59.7561	60.3659	63.4146	59.1463	**64.6341**
	3	60.9756	61.5854	62.8049	62.8049	59.7561	**66.4634**
	4	60.9756	64.0244	60.9756	62.8049	59.7561	**67.6829**
4	2	57.9268	**65.2439**	61.5854	64.0244	58.5366	63.4146
	3	62.1951	**67.6829**	66.4634	65.2439	59.7561	67.0732
	4	59.7561	64.0244	64.0244	68.2927	59.7561	**68.2927**

In most cases, the algorithms on the sparse subtle motion videos overperform those on original videos. However, the performance of algorithms on the micro-expression videos are better than those on sparse videos when $R_{xy} = 1$. In other cases. This result may stem from the fact that there exists flickering light (because of the frame rate of the camera is higher than the frequency of the alternative current) in video samples in SMIC. From the table, we can also see that the DLSTD on the sparse subtle motion videos has the best performance in many cases. The reason is that DLSTD decorrelates the coefficients of codes. The combination of the decorrelated codes gets better performance.

5.2 CASME 2

The CASME 2 [21] database includes 246 spontaneous facial micro-expressions recorded by a 200 fps camera. These samples were selected from more than 2,500 facial expressions. Compared with SMIC, the database is improved in increased sample size, fixed illumination, and higher resolution (both temporal and spatial). The selected micro-expressions in this database either had a total duration less than 500 ms or an onset duration (time from onset frame to apex frame) less than 250 ms. These samples are coded with the onset and offset frames, as well as tagged with AUs and emotions. Fig. 6 is an example. There are 5 classes of the micro-expressions in this database: happiness (32 samples), surprise (25 samples), disgust (60 samples), repression (27 samples) and tense (102 samples).

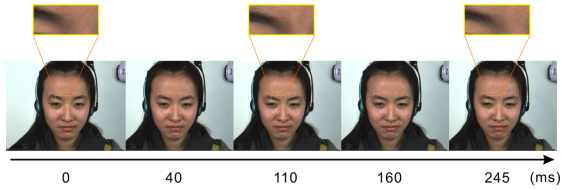

Fig. 6. A demonstration of the frame sequence in a micro-expression from CASME 2. The apex frame presents at about 110 ms. The AUs for this micro-expression is 4+9 (with AU 17 kept almost unchanged), which indicates disgust. The three rectangles above the images show the right inner brow (AU 4) in *zoom in* mode.

In these samples, the frame number of the shortest sample is 24 and that of the longest sample is 146. The frame numbers of all samples are normalized to 150 by using linear interpolation. The size of each frame is normalized to 163×134 pixels. So, each sample was normalized to a three dimensional array with the size of $163 \times 134 \times 150$. The parameter λ is set as 0.00095. Other experimental sets are the same with SMIC. The results are listed in Table 3.

Table 3. Micro-expression recognition accuracies (%) of LBP-TOP, LSTD, and DLSTD on CASME 2. +V means the algorithm is conducted on the original micro-expression videos. +S means the algorithm is conducted on the sparse subtle motion videos.

R_{xy}	R_t	LBP-TOP+V	LBP-TOP+S	LSTD+V	LSTD+S	DLSTD+V	DLSTD+S
1	2	49.5935	55.2846	50.8130	58.5366	54.4715	**62.1951**
	3	53.2520	**60.9756**	51.2195	57.7236	56.0976	**60.9756**
	4	54.4715	60.5691	52.4390	**60.9756**	56.0976	59.7561
2	2	55.6911	59.3496	53.2520	59.7561	58.5366	**63.0081**
	3	57.3171	**60.9756**	56.9106	56.5041	59.7561	60.1626
	4	58.1301	55.6911	56.5041	56.0976	60.5691	**63.0081**
3	2	53.2520	59.3496	52.4390	60.9756	61.7886	**63.0081**
	3	52.4390	60.5691	53.2520	60.9756	60.9756	**63.4146**
	4	52.8455	58.9431	53.6585	58.5366	**60.9756**	60.5691
4	2	52.8455	57.7236	55.2846	**63.8211**	58.5366	62.1951
	3	53.2520	61.3821	53.2520	**63.8211**	57.7236	63.4146
	4	54.8780	59.7561	53.6585	**65.4472**	58.5366	65.0407

From the table, we can see that the algorithms on the the sparse subtle motion videos obtain the best performance in every cases but $R_{xy} = 3$ and $R_t = 4$. In CASME 2, the illumination is fixed. So the 'sparse' processing can have a better effect. Moreover, this processing removes the identify information, which is considered noise for micro-expression recognition. From the table, we can also see that (D)LSTD is better than LBP-TOP. This also shows the code of six directions is better than the code of three planes.

6 Conclusions

In this paper, the subtle facial movement information of micro-expression is extracted by RPCA. RPCA decomposes the observed data into two parts: a low rank part and a sparse part. In general, the low rank part is the deserved data and the sparse part is considered as the error or noise and is removed. In this paper, however, we remove the low rank part consisting of the identity information and reserve the sparse part consisting of the the subtle facial movement information. Following, LSTD is used extracted the texture feature from 16 ROIs based on FACS. The results on two micro-expression databases show the proposed method get better performance.

In the further work, we will apply the proposed method on the subtle motion (such as lip-reading, the human pulse, and micro-gesture etc.) detection in surveillance.

Acknowledgments

This work was supported by grants from 973 Program (2011CB302201), the National Natural Science Foundation of China (61379095, 61375009, 61175023), China Postdoctoral Science Foundation funded project, the Open Projects Program of National Laboratory of Pattern Recognition (201306295) and the open project program (93K172013K04) of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University. G. Z. was supported by the Academy of Finland and Infotech Oulu.

References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding 61(1), 38–59 (1995)
2. Dao, M., Suo, Y., Chin, S., Tran, T.: Video frame interpolation via weighted robust principal component analysis. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 1404–1408. IEEE (2013)
3. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1(3), 211–218 (1936)
4. Ekman, P.: Microexpression training tool (METT). San Francisco: University of California (2002)
5. Ekman, P.: Lie catching and microexpressions. The philosophy of deception pp. 118–133 (2009)
6. Ekman, P., Friesen, W.: Nonverbal leakage and clues to deception. Tech. rep., DTIC Document (1969)
7. Ekman, P., Friesen, W.V.: Facial action coding system: A technique for the measurement of facial movement, vol. 12. CA: Consulting Psychologists Press (1978)
8. Georgieva, P., De la Torre, F.: Robust principal component analysis for brain imaging. In: Artificial Neural Networks and Machine Learning–ICANN 2013, pp. 288–295. Springer (2013)

9. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: inducement, collection and baseline. In: IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (2013)
10. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. In: Neural Information Processing Systems (NIPS) (2011)
11. Matsumoto, D., Hwang, H.: Evidence for training the ability to read microexpressions of emotion. Motivation and Emotion 35(2), 181–191 (2011)
12. Michael, N., Dilsizian, M., Metaxas, D., Burgoon, J.K.: Motion profiles for deception detection using visual cues. In: Computer Vision–ECCV 2010, pp. 462–475. Springer (2010)
13. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
14. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Image and Signal Processing, pp. 236–243. Springer (2008)
15. Pfister, T., Li, X., Zhao, G., Pietikainen, M.: Recognising spontaneous facial micro-expressions. In: 12th IEEE International Conference on Computer Vision. pp. 1449–1456. IEEE (2011)
16. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In: 3rd International Conference on Crime Detection and Prevention. pp. 1–6. IET (2009)
17. Shi, L.C., Duan, R.N., Lu, B.L.: A robust principal component analysis algorithm for eeg-based vigilance estimation. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. pp. 6623–6626. IEEE (2013)
18. Wang, L., Cheng, H.: Robust principal component analysis for sparse face recognition. In: Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on. pp. 171–176. IEEE (2013)
19. Wang, S.J., Chen, H.L., Yan, W.J., Chen, Y.H., Fu, X.: Face recognition and micro-expression based on discriminant tensor subspace analysis plus extreme learning machine. Neural Processing Letters 39(1), 25–43 (2014)
20. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: Advances in neural information processing systems. pp. 2080–2088 (2009)
21. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLoS ONE 9(1), e86041 (2014)
22. Yan, W.J., Wu, Q., Liang, J., Chen, Y.H., Fu, X.: How fast are the leaked facial expressions: The duration of micro-expressions. Journal of Nonverbal Behavior pp. 1–14 (2013)
23. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 915–928 (2007)
24. Zhao, G., Pietikäinen, M.: Visual speaker identification with spatiotemporal directional features. In: Image Analysis and Recognition, pp. 1–10. Springer (2013)
25. Zhou, Z., Zhao, G., Pietikainen, M.: Towards a practical lipreading system. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 137–144. IEEE (2011)