DISCRIMINATIVE VIDEO REPRESENTATION WITH TEMPORAL ORDER FOR MICRO-EXPRESSION RECOGNITION

Mingyue Niu[1,3], *Jianhua Tao*[1,2,3], *Ya Li*[1,3], *Jian Huang*[1,3], *Zheng Lian*[1,3]

[1]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Micro-expression recognition is a challenging task due to its low intensity and short duration and how to extract the subtle facial changes is a key issue in this field. Although there are many methods attempt to cope with this problem, they are difficult to encode the temporal order of all frames in the video clips. For these reasons, this paper employs rank pooling and $\ell_{2,1}$-norm to obtain the discriminative video representation with temporal order. In particular, we extract Local Two-Order Gradient Pattern (LTOGP) feature of each frame to describe the subtle information. Then, the video representation is generated by using rank pooling, which captures the temporal order among all frames. Furthermore, considering the sparsity of $\ell_{2,1}$-norm, we can select those discriminant features. Finally, micro-expression classification is accomplished using SVM. Experiments are conducted on two publicly available micro-expression databases i.e. CASME and CASME2. The results demonstrate that our method achieves better performance than the state-of-the-art algorithms.

Index Terms— Micro-expression recognition, LTOGP, temporal order, rank pooling, $\ell_{2,1}$-norm

1. INTRODUCTION

People sometimes hide their true emotions in the process of mutual communication [1], but some psychological studies [2, 3] indicate that micro-expression can reveal these true inner feelings. The reasons are that micro-expression is considered to be spontaneous and people can neither fake nor supress it [4]. Due to these characteristics of micro-expression, it has received more and more attention from many researchers and is applied to lie detection [5], clinical diagnosis [6] and security field [7].

Recognizing micro-expression through facial appearance is a challenging work since related works [5, 3] have shown that micro-expression occurs between 1/25 and 1/5 seconds and only appears low intensity in a few parts of the face. Therefore, how to extract the changes of facial subtle information in the video clips becomes a key issue for recognition. Although there are some methods [8, 9, 10, 11, 12] attempt to encode the subtle dynamic process, they only consider the temporal order among adjacent frames rather than all frames in a video clip. In addition, the face descriptors obtained in many works [11, 12, 13] contain redundant information such as identity and illumination.

In order to alleviate the above issues, this paper introduces rank pooling and feature selection method based on $\ell_{2,1}$-norm for micro-expression recognition. In particular, we firstly use LTOGP [14] to extract the feature of each frame. And then the video representation is generated using rank pooling method [15], which encodes the dynamic process of the face by capturing the temporal order among frame-level features. Furthermore, we take the advantage of the $\ell_{2,1}$-norm sparseness [16] to select the discriminative features related to micro-expression. Finally the support vector machine (SVM) is used for classification. To the best of our knowledge, it is the first time to explore rank pooling and $\ell_{2,1}$-norm for micro-expression recognition. We conduct experiments on Chinese Academy of Science Micro-expression Database (CASME) [17] and CASME2 [18], and the results demonstrate that our method obtains better performance.

The rest of this paper is organized as follows. In section 2, we review the works related to micro-expression recognition. In section 3, we provide a detailed description of the method in this paper. Our experimental results and discussion are presented in section 4, and section 5 concludes the paper.

2. RELATED WORKS

Video representation is an important issue for micro-expression recognition. Geometry-based and appearance-based features have been widely applied for face analysis. Specifically, geometry-based features are generated by calculating the shapes and locations of facial landmarks. However, such features are difficult to capture subtle facial movements (e.g. the eye wrinkles). Appearance-based features can extract not only detailed information by examining the texture changes of the faces, but also are robust to illumination [19]. Based on these considerations, many appearance-based feature extraction methods have been proposed.

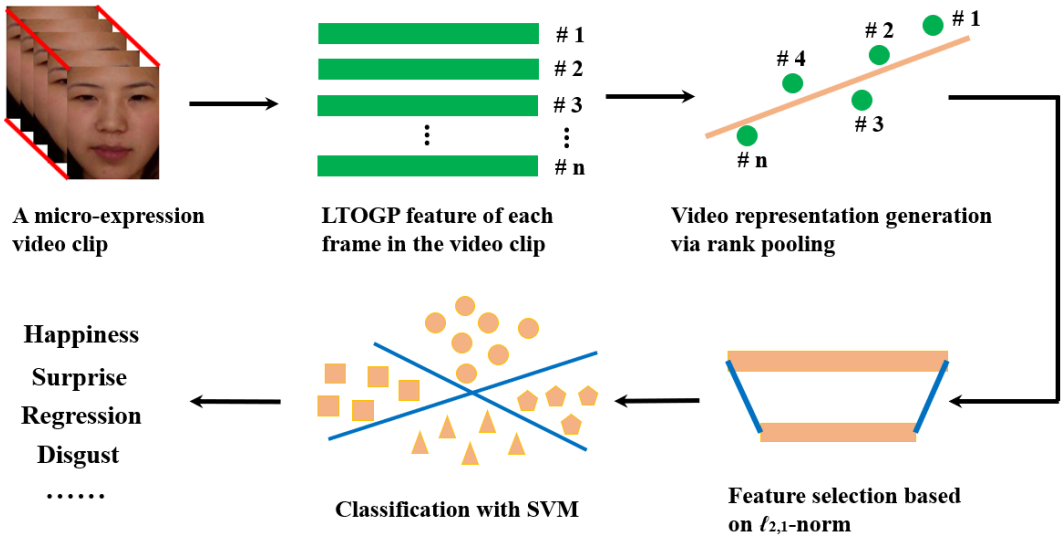

Fig. 1. Illustration of the proposed framework for micro-expression recognition.

Le Ngo et al. [20] used sparse sampling to get the details of each frame, but they only considered the changes of adjacent frames when generating the video representation. Thus, the method was difficult to describe the dynamic process in the video clips. Wang et al. [10] viewed the video clips as tensors and proposed Sparse Tensor Canonical Correlation Analysis (STCCA) for micro-expression classification. Since this method required normalizing the number of frames in the video clips, it interfered with the original data itself and affected the accuracy of recognition. Huang et al. [19] used the low rank representation method to erase the identity attribution of the faces, and then employed the discriminative spatiotemporal local binary pattern based on a revisited integral projection (DiSTLBP-RIP) to extract the dynamic features of the video clips for recognition. But the revisited integral projection used in that method is insensitive to temporal order among all frames. D.Kim [13] adopted convolutional neural networks (CNN) to obtain the visual description of each frame, and then video representation was generated by long short-term memory (LSTM) recurrent neural networks. However, those network parameters couldn't be optimized very well due to the limited data in micro-expression databases.

Different from the above works, this paper employs rank pooling method to retain the order of all frames in an entire video clip. In addition, to further improve the discriminability, we take the advantages of sparsity of $\ell_{2,1}$-norm for feature selection. The experimental results illustrate the effectiveness of our method.

3. PROPOSED METHOD

In order to explore the temporal information in micro-expression video clips, this paper introduces a new integrated framework. Firstly, we extract the feature of each frame in the video clips based on LTOGP. Secondly, the video representation is generated by adopting rank pooling to encode temporal order among all frames. Thirdly, given the sparsity of $\ell_{2,1}$-norm, we select discriminative features that are beneficial to micro-expression recognition. Finally, SVM is used for classification. The integrated framework of our method is shown in Fig. 1.

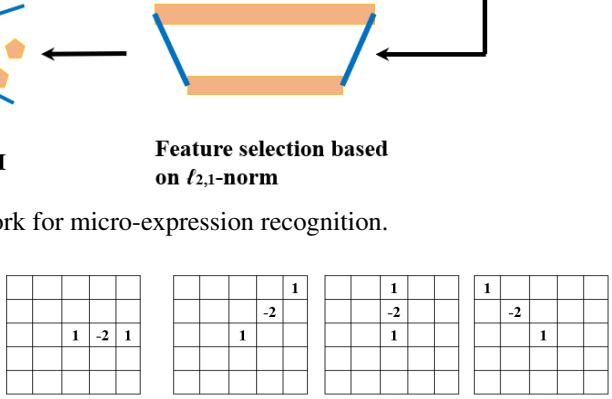

Fig. 2. Eight templates used in LTOGP. Note that the blank space in the templates is 0.

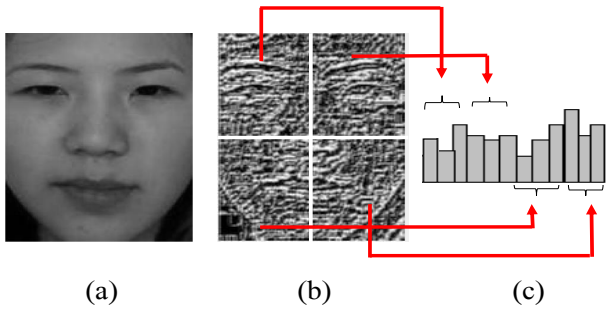

Fig. 3. Frame-level feature generation process. (a), (b) and (c) are the original image, LTOGP filter result and frame-level feature, respectively.

3.1. Frame-Level Feature Extraction Using LTOGP

Since the micro-expression is subtle facial changes, it is important for subsequent video representation to extract the detailed information of each frame in the video clips. For these reasons, we employ LTOGP [14] to extract the frame-level feature, which describes subtle information of texture in the local neighborhood.

In particular, we firstly use the templates in Fig. 2 to get the two-order gradient in different directions. Secondly, 8 bits are obtained and converted into a decimal number by Eq. (1) and Eq. (2), respectively. Finally, the frame-level feature is generated by concatenating grayscale statistical histogram of each block. This process is illustrated in Fig. 3.

$$b(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{1}$$

$$\mathbf{LTOGP} = \sum_{i=0}^{7} b_{i+1} \cdot 2^i \tag{2}$$

3.2. Video Representation with Temporal Order

Obtaining the video representation is a key stage in micro-expression recognition. In this paper, we employ rank pooling to encode the temporal order among all frames to describe the dynamic changes of the face in the video clips.

$$\begin{aligned} \arg\min_{\mathbf{u}} \quad & \frac{1}{2}\|\mathbf{u}\|^2 + C \sum_{\forall i,j\ \mathbf{v}_{t_i} \succ \mathbf{v}_{t_j}} \varepsilon_{ij} \\ \text{s.t.} \quad & \mathbf{u}^{\mathrm{T}} \cdot (\mathbf{v}_{t_i} - \mathbf{v}_{t_j}) \geq 1 - \varepsilon_{ij} \\ & \varepsilon_{ij} \geq 0 \end{aligned} \tag{3}$$

Specifically, the objective function is shown in Eq. (3), where $\mathbf{v}_{t_i}$ is the frame-level feature at time t_i. $\mathbf{v}_{t_i} \succ \mathbf{v}_{t_j}$ defines a patial order relationship, which means that if $t_i < t_j$ then $\mathbf{u}^{\mathrm{T}} \cdot (\mathbf{v}_{t_i} - \mathbf{v}_{t_j}) \geq 1 - \varepsilon_{ij}$. C is the penalty factor. ε_{ij} is a non-negative constant. The resulting vector $\mathbf{u}$ is the corresponding representation of a micro-expression video clip.

Note that Eq. (3) is not the objective fucntion in SVM. It encodes the temporal order among all frames in a video clip by the partial order relationship. In other words, we capture the facial subtle changes in the micro-expression video clips.

3.3. Discriminative Feature Selection Based on $\ell_{2,1}$-Norm

In micro-expression video clips, individual identity, illumination and other information account for a large proportion [19]. Therefore, it is necessary to select the discriminative features related to micro-expression from the above video representation to improve the performance of recognition.

Considering the advantages of $\ell_{2,1}$-norm sparsity, this paper obtains the discriminative features through Eq. (4), where $\mathbf{W}$ is the weight matrix and the $\ell_{2,1}$-norm definition of any matrix $\mathbf{A}$ is shown in Eq. (5). γ is a positive constant. In this paper, let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n] \in \mathbb{R}^{d \times n}$ be data matrix, each column of which is a representation of a micro-expression video clip. And $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times c}$ is the label matrix, which contains c classes and uses one-hot encoding.

$$\min_{\mathbf{W}} \gamma \|\mathbf{U}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\|_{2,1} + \|\mathbf{W}\|_{2,1} \tag{4}$$

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} a_{ij}^2}, \quad \forall \mathbf{A} \in \mathbb{R}^{m \times n} \tag{5}$$

We apply the method proposed in [16] to optimize the Eq. (4) and still denote the result as $\mathbf{W}$ for the sake of brevity. Let $\mathbf{w}_i$ denote the i-th row of W, $i = 1, ..., d$. Then they are sorted as $\|\mathbf{w}_{i_1}\|_2 \geq ... \geq \|\mathbf{w}_{i_d}\|_2$. If the number of selected feature is N, then the (i_1-th, ..., i_N-th) rows of $\mathbf{U}$ are considered to be discriminative features related to micro-expression.

4. EXPERIMENTS

In this section, firstly, we introduce the databases used in experiments briefly, and then describe the settings of parameters, finally experimental results and analysis are presented.

4.1. Database Description and Parameter Settings

To evaluate the proposed method, we conduct experiments on two public available micro-expression databases i.e. CASME [17] and CASME2 [18]. CASME contains 195 spontaneous micro-expression video clips. These clips record 20 individual micro-expression using a 60fps camera. In order to ensure the fairness of the comparison, like other advanced methods [17, 19], we select 171 video clips that include disgust (44 samples), surprise (20 samples), repression (38 samples) and tense (69 samples). For CASME2, it records 247 micro-expression video clips from 35 subjects using a 200fps camera. This database contains 5 classes of the micro-expression: happiness (32 samples), surprise (25 samples), disgust (64 samples), repression (27 samples) and others (99 samples). What's more, we use Leave-One-Subject-Out (LOSO) cross validation (i.e. one subject's video clips are used as the test data and the others are used as the training data) to evaluate the proposed method. It should be noted that we resize all the frames to 128×128 pixels. All the experiments are conducted on a PC with MATLAB 2017b.

In our method, there are four parameters i.e. the number of blocks divided for each frame to generate LTOGP, C in Eq. (3), γ in Eq. (4) and the number of features selected N. Note that we let $C = 1$ and $\gamma = 1$ in all experiments. For the other two parameters, Fig. 4 (a) and (b) show their impacts on the recognition accuracy on CASME and CASME2, respectively. As we can see from them, the best performance is achieved

when the number of blocks is 8×8 and the number of features selected (N) is 325 on CASME, while they are 8×8 and 450 on CASME2. It is necessary to illustrate that the dimension of the video representation without feature selection is $8 \times 8 \times 256 = 16384$. In addition, we use LIBSVM [21] with the linear kernel for classification.

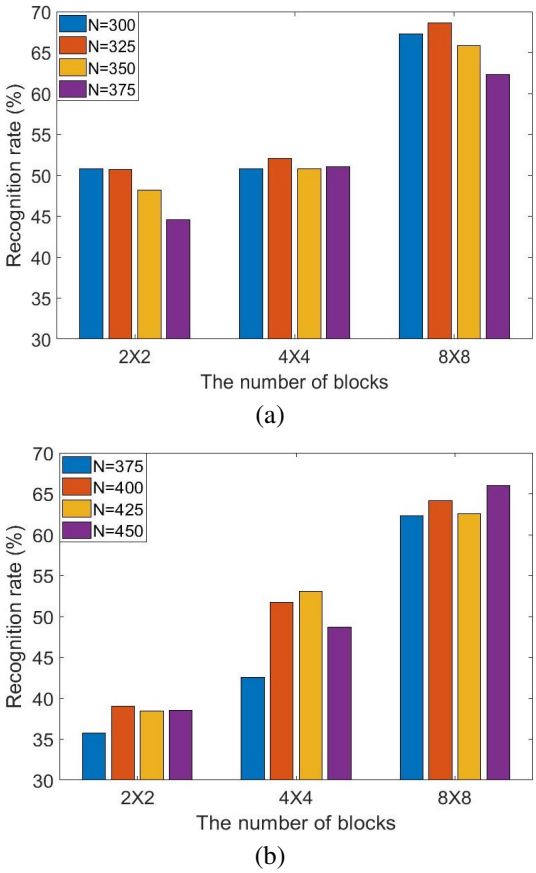

Fig. 4. Influence of the number of blocks and features selected (N) on recognition rate on CASME (a) and CASME2 (b). Note that the four different colors represent four different numbers of features selected.

4.2. Results and Discussion

Based on the above settings, we perform experiments on CASME and CASME2, respectively. The results are shown in Table 1 and Table 2.

As illustrated in those two tables, our method with feature selection is 7.57% and 2.97% higher than without feature selection on CASME and CASME2, respectively. The reasons lie in that the first term in Eq. (4) is robust to noise and the second term is effective to select discriminative features using the sparsity of $\ell_{2,1}$-norm [16].

In addition, compared to other algorithms, our method with feature selection achieves better performance on both CASME and CASME2. Mutilinear principal component analysis (MPCA) [17] extracts the principal component directly from the video clips but it is difficult to obatin the details in the frames. Instead, we adopt LTOGP to get the subtle information in each frame. For DiSTLBP-RIP [19] and the sparse sampling [20], the details of each frame are extracted, but they only capture the temporal order among adjacent frames. While our method is able to get the dynamic changes among all frames, thus we obtain better accuracy. It is worth mentioning that our method is superior to the approach with neural network [13]. This is can be explained that there are limited data in the micro-expression databases, which degrade the performance due to over-fitting.

Table 1. Performance comparison with the state-of-the-art methods on CASME. FS means Feature Selection.

Methods	Recognition accuracy (%)
MPCA[17]	41.01
DiSTLBP-RIP[19]	64.33
Our method without FS	61.07
Our method with FS	**68.64**

Table 2. Performance comparison with the state-of-the-art methods on CASME2. FS means Feature Selection.

Methods	Recognition accuracy (%)
Sparse sampling[20]	49.00
DiSTLBP-RIP[19]	64.78
CNN+LSTM[13]	60.98
Our method without FS	63.03
Our method with FS	**66.00**

5. CONCLUSION

In this paper, a discriminative video representation with temporal order is proposed for micro-expression recognition. Specifically, rank pooling is introduced to encode the subtle changes of frame-level feature, which describes the detailed information with LTOGP in the face. Moreover, to enhance the discriminability, we select the relevant features through $\ell_{2,1}$-norm. Experimental results on CASME and CASME2 indicate that our method is superior than the state-of-the-art approaches. In the future, we will combine neural networks with this work to get better performance.

6. ACKNOWLEDGEMENT

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002804) and the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61603390).

7. REFERENCES

[1] S. Weinberger, "Intent to deceive?," *Nature*, vol. 465, no. 7297, pp. 412, 2010.

[2] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

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

[4] E. A Haggard and K. S Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of research in psychotherapy*, pp. 154–165. Springer, 1966.

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

[6] M. Frank, M. Herbasz, K. Sinuk, A. Keller, A. Kurylo, and C. Nolan, "I see how you feel: training laypeople and professionals to recognize fleeting emotions, 2009," in *annual meeting of the International Communication Association. http://www. allacademic. com/meta/p15018_index. html*.

[7] P. Seidenstat and F. X Splane, *Protecting airline passengers in the age of terrorism*, ABC-CLIO, 2009.

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

[9] 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.

[10] S. Wang, W. Yan, T. Sun, G. Zhao, and X. Fu, "Sparse tensor canonical correlation analysis for micro-expression recognition," *Neurocomputing*, vol. 214, pp. 218–232, 2016.

[11] Y. Oh, A. C. Le Ngo, R. C-W Phari, J. See, and H. Ling, "Intrinsic two-dimensional local structures for micro-expression recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1851–1855.

[12] A. C. Le Ngo, Y. Oh, R. C-W Phan, and J. See, "Eulerian emotion magnification for subtle expression recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1243–1247.

[13] D. 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.

[14] M. Niu, Y. Li, J. Tao, and S. Wang, "Micro-expression recognition based on local two-order gradient pattern," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[15] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2017.

[16] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[17] W. Yan, Q. Wu, Y. Liu, S. Wang, and X. Fu, "Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces," in *Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on*. IEEE, 2013, pp. 1–7.

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

[19] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikainen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, 2017.

[20] A. C. Le Ngo, J. See, and R. C-W Phan, "Sparsity in dynamics of spontaneous subtle emotions: analysis and application," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 396–411, 2017.

[21] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.