Paper:

Dynamic Facial Expression Recognition Based on Optical Flow and Geometric Features

Da Ma[⋆1], **Kaoru Hirota**[⋆2], **Yaping Dai**[⋆3], **Zhiyang Jia**[⋆4]

[⋆1⋆2⋆3⋆4]School of Automation, Beijing Institute of Technology, No. 5, Zhongguancun South Street, Haidian District, Beijing, China
E-mail: zhiyang.jia@bit.edu.cn

Abstract. For the recognition of facial expressions in video images, we propose to use optical flow method to improve geometric features for dynamic expression recognition in this paper. The optical flow method can better capture facial motion and record dynamic features compared with traditional geometric methods. After feature extraction, support vector machines are used to classify the two types of features separately. Finally, a weighted fusion strategy is used for decision-level fusion of the two classifiers. The proposed method is verified on the extended Cohn-Kanade database and the overall recognition accuracy is 95%.

Keywords: Facial expression recognition; optical flow method; geometric feature; Delaunay triangulation; support vector machine

1. INTRODUCTION

With the rapid development of computer vision and machine learning technology, artificial intelligence is more and more appearing in life, and humans need machines to understand human emotions. Facial expressions are an important way for humans to express emotions. Obtaining human emotional states through facial expression recognition not only helps machine learn to understand some human perceptions and decisions, but also helps to establish intelligent and friendly human-computer interaction in the future.

Currently there are two mainstream facial expression recognition methods: static images-based and dynamic videos-based methods. Among these methods, the recognition based on static images mainly extracts features based on the information of a single picture when the expression occurs. The main methods can be divided into appearance-based methods and model-based methods[1]. For example, the appearance based methods mainly employ texture descriptors like Gabor filters [2], Histogram of Oriented Gradient(HOG) [3], Local Binary Patterns(LBP) [4], etc. to represent facial expressions.

The recognition method based on the static images has the characteristics of easy processing and fast speed, but lacks the movement information of the expression. So, the static images-based method can not reflect the change of the expression well. The recognition method based on the dynamic videos utilizes the image sequence for expression recognition, which considers the spatiotemporal information of the image sequence and records the dynamic characteristics of expression changes. So the dynamic videos-based method can describe the expression change process naturally and effectively [5]. Therefore, this paper takes the dynamic image sequences as the research object of facial expression recognition.

The difficulty of facial expression recognition based on videos is to extract the dynamic features of expressions from the image sequences. In response to this problem, two mainstream methods have been developed: one is appearance-based method and the other is geometry-based method. The appearance-based approach is mainly to extract the texture features of video images. For example: in [6], Zhao and Pietikainen considered the local binary mode of three orthogonal planes and proposed the LBP-TOP algorithm; Jiang et al. [7] proposed the local phase quantization algorithm of three orthogonal planes(LPQ-TOP); Chen et al. [8] proposed a gradient direction histogram algorithm for three orthogonal planes(HOG-TOP).

The geometry-based method extracts dynamic geometric features by calculating changes in the position of the face area. For example: in [9], Chen et al. proposed to use the position of the center point of the triangle network to calculate geometric features; in [10], Lucey applied active appearance model (AAM) to extract similar normalized shapes (SPTS) and canonical appearance (CAPP) from facial landmarks. In addition, Chew et al. [12] proposed a space-time representation to simulate the dynamics of facial expressions and established a sparse representation classifier to categorize facial expressions. Huang et al. [13] proposed a weighted component-based feature descriptor (CFD) for expression recognition in video sequences.

In this paper, in order to capture the global changes in facial expressions, we first extracted geometric features by constructing a triangulation of human face. With the consideration of the facial movements during the expression generation process, this paper uses the sparse optical flow method to track 68 landmarks of the human face in the neutral expression frame. Therefore, the dynamic features during the expression change process are obtained. After that, Support Vector Machines are used to classify the two types of features separately. Finally, a weighted

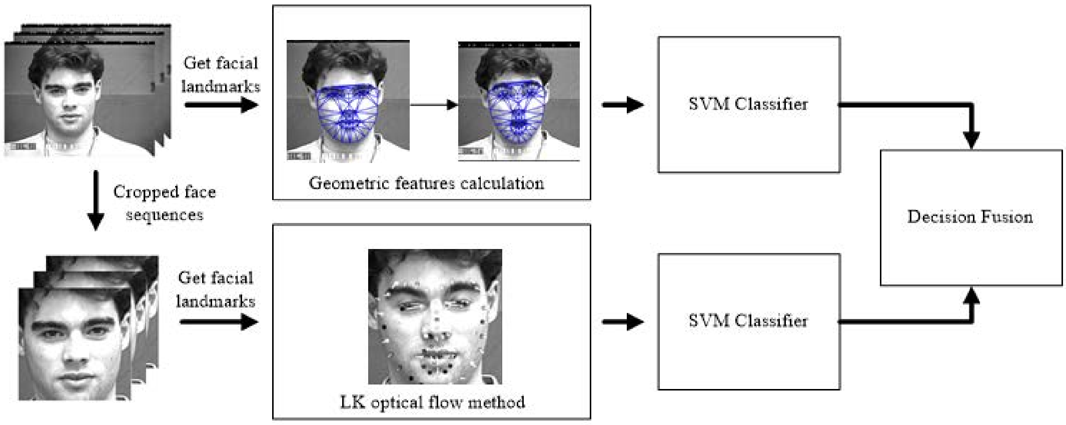

Fig. 1. The flow chart of our proposed method

fusion strategy is used for decision-level fusion of the two classifiers. The flow chart of our proposed method is shown in **Fig 1**.

The rest of the paper is organized as follows. Section 2 introduces our proposed method in detail. Experiments and results analysis are in Section 3. The paper is summarized in Section 4.

2. METHODOLOGY

This section introduces the details of our proposed method. Two features and classifiers are introduced.

2.1. Geometry Features

In order to obtain the geometric features of the facial expression changes, the facial landmarks are used to generate a facial triangle network. Suppose a given set of facial landmark points P = (x_1, y_1, x_2, y_2, ..., x_i, y_i), where x, y are the coordinates of facial landmarks, and i represents the number of facial landmarks. The facial landmark points and Delaunay triangulation algorithm [14] are used to divide the face into many triangle meshes, as shown in **Fig 2**. The displacement of facial landmarks includes changes of coordinates $\Delta x = x_E - x_N$, $\Delta y = y_E - y_N$, and distance $d = \sqrt{(\Delta x)^2 + (\Delta y)^2}$, between the corresponding facial landmarks. x_E, y_E are the coordinates of the facial landmarks of the expressive facial triangle network, x_N, y_N are the coordinate positions of the neutral facial triangle network. These displacements are characteristic of changes in facial rigidity. When the expression changes, the corresponding triangle in the facial triangle network changes as shown in **Fig 3**.

From **Fig 2**, we can see that the human face is covered by many triangles, and each triangle is a sub-region of the face. When a neutral surface becomes expressive, the triangles on the neutral face are deformed by the movement of facial muscles. In **Fig 3**, ΔABC is a triangle in the neutral expression frame, the coordinates of the three vertices are (x, y). The corresponding triangle in the peak expression frame is $\Delta A'B'C'$, and the coordinates of the corresponding vertex are (u, v). Each triangle includes three sides. We define the difference between the sides of two triangles as d = d_E - d_N, d_E is the length side $A'B'$,$A'C'$ or $B'C'$; d_N is the length of the side AB,AC orBC. From the coordinates of the vertices, the lengths of the edges d_E and d_N can be obtained. We can calculate the attributes of each pair of triangles on the face and connect them to a global vector to describe the changes in facial expressions.

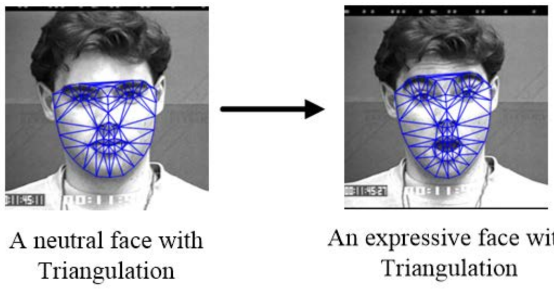

Fig. 2. Facial structure composed of triangle mesh

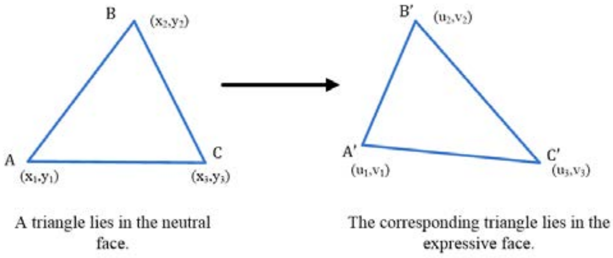

Fig. 3. A triangle deforms due to facial muscle movements.

2.2. Lucas-Kanade Optical Flow Method

The geometric features extracted by the triangulation only consider the start frame of expression and the peak frame of expression. In order to obtain the dynamic information of the facial expression change process, we track the position information of 68 face landmarks in neutral expression frames by Lucas-Kanada (LK) optical flow algorithm [15]. Finally, the position difference of the feature points when the neutral expression changes to the peak expression is obtained as a dynamic feature.

Optical flow is a motion vector formed between frames before and after the same pixel in an image. Optical flow calculation is actually based on two assumptions. One is the brightness constancy assumption, i.e., the movement process of the tracked target pixel should keep the brightness unchanged. The second is the spatial smoothness assumption, which means that the change of the optical flow field caused by the moving object should be continuous and smooth, i.e., the speed of the moving object is slow and smooth[16]. The change of facial expression in the dynamic sequence satisfies these two conditions. Therefore, the optical flow method can be used to calculate the sparse optical flow to realize the tracking of facial expression feature points.

The Lucas-Kanade (LK) optical flow method is based on local constraints, assuming that the motion vector remains constant in a small area with a certain point as the center, and the optical flow is estimated by weighted least squares of different points in the area[17].

Set $I = (x, y, t)$ is the gray value of the point $m = (x, y)^T$ at time t. The gray value after time interval Δt is $I(x+\Delta x, y+\Delta y, t+\Delta t)$. When Δt tends to 0, the formula can be derived as:

$$I(x+\Delta x, y+\Delta y, t+\Delta t) = I(x, y, t) \tag{1}$$

Expand the above formula by Taylor series to obtain

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \varepsilon = 0 \tag{2}$$

where ε represents the higher-order infinitesimal term, which can be ignored because dt tends to 0. Let $u = \frac{dx}{dt}$, $v = \frac{dy}{dt}$ represent the optical flow in the$x$ and y directions. Let $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$, $I_t = \frac{\partial I}{\partial t}$ be combined with (2) to obtain

$$I_x u + I_y v + I_t = 0 \tag{3}$$

The above formula is the optical flow constraint equation, expressed in vector form as

$$\Delta I \cdot v_m + I_t = 0 \tag{4}$$

among them, $\Delta I = (I_x, I_y)$ is the gradient of the image at point m and $v_m = (u, v)$ is the optical flow at point m. The LK optical flow method is only for sparse optical flow. So it is necessary to make the two images have the closest pixel distribution in a small range centered at point m. The local neighborhood minimization equation of m is shown as equation (5).

$$\sum_{x\in\Omega} W^2(x)[\Delta I(x,t) \cdot v + I_t(x,t)]^2 \tag{5}$$

In the equation (5): Ω is a small area centered on the point m, $W(x)$ is the weight, and the size is proportional to the distance from the m point.Solving equation (5), we can get:

$$A = [\Delta I(x_1), \Delta I(x_2), \ldots, \Delta I(x_n)]^T \tag{6}$$

$$W = diag[W(x_1), W(x_2), \ldots, W(x_n)] \tag{7}$$

$$b = -[I_t(x_1), I_t(x_2), \ldots, I_t(x_n)]^T \tag{8}$$

the solution of the final equation is:

$$v = (A^T W^2 A)^{-1} A^T W^2 b \tag{9}$$

In order to reduce the calculation error caused by the sudden change of expression, the layered Gaussian pyramid LK optical flow algorithm is used. When the image is down-sampled and decomposed to a certain layer, the large motion of the image between adjacent frames is reduced to a small motion that satisfies the optical flow constraints. The initial motion speed can be corrected to obtain an accurate optical flow estimate[18]. As shown in **Fig 4**, I^0 is the original image, and sequentially generates n layers of pyramid image sequences, namely $I^0, I^1, I^2, \ldots, I^{n-1}$, where the initial optical flow of the $L(0 \le L \le n-1)$ layer is $g^L = [g_x^L \quad g_y^L]^T$. The initial optical flow of the top layer L_m is $g^L = [0 \quad 0]^T$, d^L is the optical flow increase of the L layer, and the optical flow calculation formula between two adjacent layers is:

$$g^{L_m} = [0 \quad 0]^T \tag{10}$$

$$g^{L-1} = 2(g^L + d^L) \tag{11}$$

the final optical flow calculation result is:

$$d = g^0 + d^0 = \sum_{L=0}^{L_m} 2^L d^L \tag{12}$$

The combination of the optical flow increment and the initial optical flow is used as the initial value of the next layer, and iterative calculation is performed to obtain the LK optical flow of the original image. Pyramid LK optical flow tracking feature point results are shown in **Fig 5**.

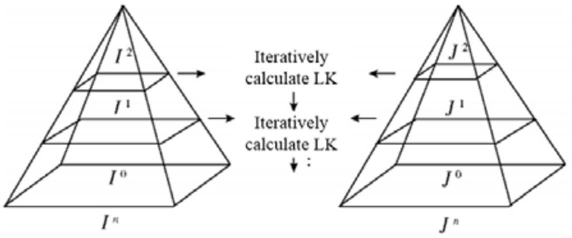

Fig. 4. Pyramid LK optical flow method

Fig. 5. Pyramid LK optical flow method tracking results

2.3. Support Vector Machine Classifier

Next, we need to build a classifier to train and classify the extracted features. Support vector machine (SVM) is a pattern classification method based on statistical learning theory[19]. The basic SVM is oriented to two classification problems. So it can not directly deal with multi-category issues. When targeting 7 types of basic expression classification, it is necessary to design a SVM classifier that supports multiple types. The main idea is to extract two types in sequence from multiple categories and combine them, and convert multiple types of problems into multiple two types of problems for training and learning. The discriminant function of the two-class SVM classifier is shown in equation (13).

$$f(x) = \sum_{i=1}^{n} a_i y_i K(x, x_i) + b\star \tag{13}$$

where n is the number of samples, a_i is the Lagrange factor, and y_i and $b\star$ can determine the optimal classification hyperplane. $K(x, x_i)$ determines the category of x according to the symbol of $f(x)$.

Use the standard SVM calculation process to construct multiple decision boundaries in an orderly manner to achieve multi-classification of samples, usually implemented as 'one-versus-rest' and 'one-versus-one'. 'One-versus-rest' SVM establishes m decision boundaries for m categories, and each decision boundary determines the ownership of one category to all other categories. The advantage of this method is that only m classifiers need to be trained. The number of training classifiers is small, and its classification speed is relatively fast. However, the disadvantage is that the training of each classifier uses all samples as training samples. In this way, when solving the quadratic programming problem, the training speed will slow down sharply as the number of training samples increases. Besides, when new categories are added, all models need to be retrained.

'One-versus-one' SVM is a voting method, the calculation process is to establish decision boundaries for any 2 of the m categories, i. e., there are $\frac{m(m-1)}{2}$ decision boundaries in total. The category of the sample is selected according to the category with the highest score among the discriminant results of all decision boundaries. The advantage of this approach is that do not need to retrain all SVMs when increasing the samples. We only need to retrain and increase the sample-related classifiers. Therefore, when training a single model, the relative speed is faster. However, the number of binary classifiers to be constructed and tested by this method grows into a quadratic function with respect to m. So the total training time and testing time are relatively slow.

In our study, we built a multi-class linear support vector machine classifier for each feature for classification training with LIBSVM[20], and its multi-classifier is implemented according to the 'one-versus-one' method.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. The Database

We conducted experiments on the extended Cohn-Kanade(CK+) database[10]. The CK+ database is a database commonly used in facial expression recognition at present, which is expanded on the basis of Cohn-Kanade Dataset. The CK+ database includes 123 subjects and 593 image sequences. The last frame of each image sequence has a label of action units, and among these 593 image sequences, there are 327 sequences with emotion labels. The emotion tags of the data set are divided into 7 categories: anger, happy, sad, contempt, fear, surprise and disgust.

This article mainly extracts two kinds of features. One is the geometric feature, and the other is the dynamic feature. The geometric feature is obtained by calculating the triangle network changes. The dynamic feature is obtained by using the LK optical flow method to track the facial landmark points in the expression change process. To extract these features we need to preprocess the CK+ database. When extracting geometric features, first select the first picture of each image sequence as the neutral expression frame, and select the last picture of each image sequence as the peak expression frame. Gray-scale the selected image, and then the face range of each image is detected and cropped. Then extract 68 facial landmarks points from the cropped face range. Finally, 68 facial landmarks are used to generate a triangulation network containing 105 triangles, and its deformation can be calculated to obtain a geometric feature of length 315.

When using the LK optical flow method to extract dynamic features, the entire image sequence is grayed out first, and the face range is detected and cropped. Then select the first picture as the initial expression frame, and extract 68 facial landmarks points of its face. After that, the LK optical flow method is used to track the landmark points, and the position changes of the landmark points in the process from the neutral expression to the peak expression are obtained. Finally a dynamic feature with a length of 136 is obtained.

3.2. Experiment and Analysis

We randomly selected 65 (approximately 20%) of the 327 video sequences with emotion tags in the data set as the validation set, and the remaining 262 (approximately 80%) as the training set. Firstly, the geometric features are extracted by means of triangulation, and the obtained geometric feature dimension is 315. Then we use the linear kernel SVM classifier to classify and train it. The confusion matrix obtained in the experiment is shown in **Table 1**(where 'AN' stands for 'anger', 'CO' stands for 'contempt', 'DI' stands for 'disgust', 'FE' stands for 'fear', 'HA' stands for 'happy', 'SA' stands for 'sad', 'SU' stands for 'surprise'). From the data in the table, it can be concluded that the overall accuracy obtained using geometric features is 89.28%. Besides, it can be seen from **Table 1** that the use of triangulation to extract geometric features for expression classification can achieve better classification results for five types of expressions. However, it can also be found that the classification accuracy of 'angry' and 'contempt' is not good. This result is caused by the corners of the mouth moving down when people are feeling angry, disgusted, or sad. And while frowning when people are feeling contempt, fear, or sadness. All of this cause some facial features to be similar, and eventually a classification error occurs.

Table 1. Confusion matrix of geometric feature.

	AN	CO	DI	FE	HA	SA	SU
AN	0.78	0.03	0.08	0.03	0	0.08	0
CO	0	0.75	0	0.03	0	0.17	0.05
DI	0.08	0	0.92	0	0	0	0
FE	0	0	0	0.94	0	0.06	0
HA	0	0	0	0	1	0	0
SA	0	0	0	0.17	0	0.83	0
SU	0	0	0	0	0.03	0.03	0.94

Next, we used the LK optical flow method to track the landmark points of the face, and obtained the dynamic features in the expression change process. After that, the linear multi-class SVM classifier is used for classification training. The confusion matrix obtained by the experiment is shown in **Table 2**. From the table, it can be obtained that the overall accuracy of tracking the marked points using the optical flow method is 90.24%. It can be seen from **Table 2** that when using the LK optical flow method to extract dynamic features for expression classification, better classification results can be obtained for 6 types of expressions. But it can also be seen that 'contempt' is misinterpreted as 'fear'. This result is caused by changes in the corners of the mouth and the enlarged eyes when people are feeling contempt and fear. Which cause some facial features to be similar, and eventually a classification error occurs.

Table 2. Confusion matrix of LK optical flow.

	AN	CO	DI	FE	HA	SA	SU
AN	1	0	0	0	0	0	0
CO	0	1	0	0	0	0	0
DI	0.09	0	0.82	0	0.09	0	0
FE	0	0.25	0	0.75	0	0	0
HA	0	0	0	0	1	0	0
SA	0	0	0	0	0	1	0
SU	0	0	0	0.14	0	0	0.86

Finally, we used the weighted fusion strategy to fuse the decision results of the classification results of the two classifiers. The results of classification using two features are compared with the results after fusion, as shown in **Fig 6**. From the **Fig 6** we can get the overall recognition accuracy is 95%. From the result, the recognition accuracy after fusion is higher than the classification accuracy obtained by training the two features separately.

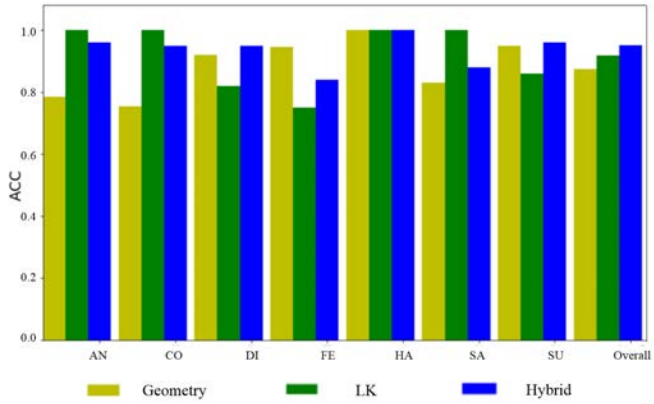

Fig. 6. The performance obtained by using three different features.

In addition, we can see from **Fig 6** that the classification accuracy after weighted fusion has been improved. At the same time, the classification effect of the two expressions of 'angry' and 'contempt' using geometric feature methods has been greatly improved. And the classification effect of the expression of 'fear' using the optical flow method has also been greatly improved. It can be seen that the dynamic features obtained by the optical flow method can fully record the process of facial expression changes and supplement the geometric features generated by the triangulation network. After the two features are fused, the two features can supplement the parts that are not well recognized. Ultimately improve the overall recognition accuracy. However, the 'fear' emotion is still not well recognized and classified though the recognition effect has been improved. This result may be because the rigid deformation of the face during the change of the 'fear' expression is similar to other expressions. Only geometric features and optical flow methods cannot capture the dynamic features of the 'fear' expression. We also compare the method proposed in this paper with several other advanced methods, the comparison results are shown in **Table 3**.

Table 3. A Comparison of our method with the other methods on the CK+ database.

	Ours	[10]	[6]	[12]	[13]	[8]
Angry	0.96	0.75	0.76	0.86	0.98	0.98
Contempt	0.95	0.84	0.93	0.56	0.91	0.78
Disgust	0.95	0.95	0.89	1.00	0.93	0.97
Fear	0.84	0.65	0.80	0.92	0.65	0.76
Happy	1.00	1.00	0.99	1.00	0.98	1.00
Sad	0.88	0.68	0.79	0.86	0.86	0.75
Surprise	0.96	0.96	0.93	0.98	0.99	0.99
Overall	0.950	0.884	0.89	0.949	0.932	0.936

As shown in **Table 3**, the SPTS+CAPP method is used in [10], the LBP-TOP method is uesd in [6], the Sparse Temporal Representation (STR) is used in [12], the Component-based Feature Descriptor (CFD) is used in [13], the hybrid features using HOG-TOP and Geometric features is used in[8]. From the **Table 3**, we can see that our method can achieve better performance (0.95) than SPTS+CAPP (0.884) and LBP-TOP (0.89). Our method achieves a competitive accuracy compared with STR (0.949), CFD (0.932) and HOG-TOP+Geometry (0.936).

4. CONCLUSIONS

Facial expression recognition in videos has gradually attracted attention in the field of artificial intelligence, and has broad development prospects. In this article, we combine the optical flow method with geometric features for facial expression recognition in video. We use the triangulation network to extract the overall geometric features during the facial emotion change process, and at the same time use the LK optical flow method to track the key points of the face and obtain the dynamic features during the expression change process. Support vector machines were used to train the two features for emotion classification. Finally, a weighted fusion strategy was used to perform decision-level fusion on the two classifiers. Experiments conducted on the CK+ data set show that our proposed method can achieve better performance compared with other methods. In future work, we will try to add texture feature extraction and fusion on the basis of the existing framework, and test the results.

Acknowledgement. This work was supported by the Beijing Municipal Natural Science Foundation under Grant No. 3192028;

References:

[1] Zhang, Wei, et al, Multimodal learning for facial expression recognition, Pattern Recognition 48.10(2015):3191-3202.

[2] H. G. Feichtinger, T. Strohmer, Gabor analysis and algorithms:Theory and applications, Springer, 1998.

[3] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in Proc. IEEE Conf. Comput. Vision Pattern Recog, 2005, pp. 886-893.

[4] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution Gray-Scale and Rotation Invariant Texture and Classification with Local Binary Patterns, Pattern Analysis & Machine Intelligence, IEEE Transactions on, vol.24, pp.971-987, 2002.

[5] Jung H, Lee S, Yim J, et al, Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition, IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2983-2991.

[6] G. Zhao, M. Pietikainen, Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions, IEEE Transactions on Pattern Analysis & Machine Intelligence 29(2007):915-928.

[7] B. Jiang, M. Valstar, B. Martinez, et al, A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modelling, Cybernetics, IEEE Transactions on, vol. 44, pp. 161-174, 2014.

[8] J. Chen, Z. Chen, Z. Chi, et al, Dynamic texture and geometry features for facial expression recognition in video, in Proc.IEEE Int. Conf. Image Process.,2015, pp. 4967-4971.

[9] Chen, Junkai, et al, Facial Expression Recognition in Video with Multiple Feature Fusion, IEEE Transactions on Affective Computing (2018):1-1.

[10] P. Lucey, J. F. Cohn, T. Kanade, et al, The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 94-101.

[11] I. Kotsia, I. Pitas, Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 16(2007):172.

[12] S. W. Chew, R. Rana, P. Lucey, et al, Sparse Temporal Representations for Facial Expression Recognition, in Advances in Image and Video Technology, 2012, pp. 311-322.

[13] X. Huang, G. Zhao, M. Pietikinen, et al, Expression Recognition in Videos Using a Weighted Component-Based Feature Descriptor, in Proceedings of the 17th Scandinavian conference on Image analysis, 2011, pp. 569-578.

[14] Uddin, Md. Zia, T. S. Kim, et al, An Optical Flow Feature-Based Robust Facial Expression Recognition with HMM from Video, International journal of innovative computing, information & control: IJICIC 9.4(2013):1409-1421.

[15] B. D. Lucas, Generalized Image Matching by the Method of Differences, Carnegie Mellon University, 1985.

[16] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2 Morgan Kaufmann Publishers Inc. 1981.

[17] Z. Tang, Z. Xu, Dynamic Generation and Modification for Triangulation of Two-dimensional Points Set, Journal of Computer Aided Design & Computer Graphics (1990).

[18] B. Wang, Z. Weng, et al, Real-time Tracking Method for Eye Feature Points Based on Lucas-Kanada Optical Flow Algorithm, Computer Engineering (2015).

[19] A. J. Laszlo,M. G. Jeffrey, et al, Continuous AU intensity estimation using localized sparse facial feature space, IEEE International Conference and Workshops on Automatic Face and Gesture Recognition,2013:1-7.

[20] C. -C. Chang, C. -J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, 2011.