Multimedia Tools and Applications (2023) 82:2663–2688
https://doi.org/10.1007/s11042-022-13327-8

Facial expression recognition based on hybrid geometry-appearance and dynamic-still feature fusion

Ruyu Yan[1] · Mingqiang Yang[1] · Qinghe Zheng[1] · Deqiang Wang[1] · Cheng Peng[1]

Received: 12 December 2020 / Revised: 3 April 2022 / Accepted: 31 May 2022 /
Published online: 1 July 2022

Abstract

Emotion recognition through facial expression is regarded as one of the most effective methods to directly reflect a person's inner emotional state for affective computing. However, a key issue of facial expression recognition (FER) is how to design and fuse features from videos rapidly and thus extract representative features to improve the recognition accuracy efficaciously. In this paper, we propose a novel expression recognition framework to mitigate this issue. Specifically, we first present a new descriptor, the improved Local Binary Pattern from Three Orthogonal Planes (I-LBP-TOP), which can extract both the static and dynamic features in changing expressions, and set Gabor's magnitude feature (GMF) as texture information. Meanwhile, the facial landmarks of the peak frame are proposed to represent geometric feature (GF) and the spatiotemporal geometric feature (ST-GF) is obtained by extending it to time dimension. Then we integrate multiple features of image sequences to overcome the limitation of using one single feature descriptor. A support vector machine (SVM) with multiple kernels is applied to train three base classifiers. Finally, to realize reliable expression classification, a decision-level feature fusion method based on a relative majority voting (MV) strategy is also employed. Intensive experiments are conducted on the CK+ and Oulu-CASIA databases, where the experimental results demonstrate that our proposed method achieves an improved performance compared with the existing state-of-the-art hand-crafted approaches.

Keywords Facial expression recognition · LBP-TOP · Gabor feature · Geometric feature · Feature fusion

1 Introduction

Facial expression, an indispensable component of human emotion expression systems, is usually regarded as a non-verbal language reflecting the state of human emotions. In [37], A. Mehrabian's research has showed that facial expression contains the most emotional information, and 55% of what the speaker wants to say comes from facial expression. As

✉ Mingqiang Yang
imageinstitute@outlook.com

Extended author information available on the last page of the article.

FER has made considerable progress in recent years, it provides a wide range of applications in scientific fields, such as human robot interaction (HRI) [8], safe driving [25], medical diagnosis [24], and so on. Although FER has been studied for decades and notable advances have been obtained in both software and hardware systems [18], the recognition of facial expression with high accuracy remains to be realized due to the impact of interpersonal variations, facial occlusion, and changes in facial pose. As illustrated in Fig. 1, psychologists Ekman and Friesen [10] have originally proposed that human beings have six prototypical emotions, namely anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU), each of which reflects a unique psychological activity with a particular expression.

Feature extraction obtains high-level semantic expression by calculating the texture, shape, spatial structure and other information of the original image, which plays an important role in FER [3]. According to the type of input data, there are two mainstream approaches in the research of FER currently: image-based and video-based approaches. The image-based approaches analyze and extract expression features through peak frames, while video-based approaches detect the temporal and motion information from image sequences with facial expressions. With regard to image-based approaches, they are normally divided into two categories: geometric feature-based and appearance feature-based approaches. For the former, the locations of many facial key points are extracted and subsequently combined into a feature vector to represent facial geometric information (e.g., angle, distance, and position) [14]. Besides, the appearance feature-based approaches model the appearance variations by applying descriptors on holistic or local regions to convolute and extract features [45].

Many previous studies are based on either a still frame or an image sequence, nevertheless, little work has been done to combine these two methods together. Actually, features extracted by these two methods are complementary. The peak frame of facial expression has strong discrimination while the temporal information is indispensable in special video classification tasks. Since a single feature is not comprehensive and rich enough to capture all dominant global information, it is necessary to fuse multiple complementary features to design a robust feature descriptor [29, 54]. In recent years, there is a trend to apply neural network model [42] in FER, which yields state-of-the-art results [32]. However, many follow-up problems come into being, such as the low availability of big data, poor generalization ability of models, excessive consumption of processing time and memory, and so on.

This paper presents an effective system using both static and dynamic features to enhance the recognition accuracy of FER in video. We utilize two kinds of static information when facial expression occurs: Gabor magnitude pictures with multiple scales and multiple directions are used to achieve extraction of texture features and 49 key points of facial expression are applied for geometric features. In addressing dynamic features, we improve the LBP-TOP method proposed by Zhao et al. [58], which demonstrated the contribution of XY plane is less than that of XT and YT planes in FER. To enhance the contribution rate of XY plane, we substitute the LBP histogram extracted from the image sequence in the XY plane for that from the peak frame. Compared with the original LBP-TOP method, the improved

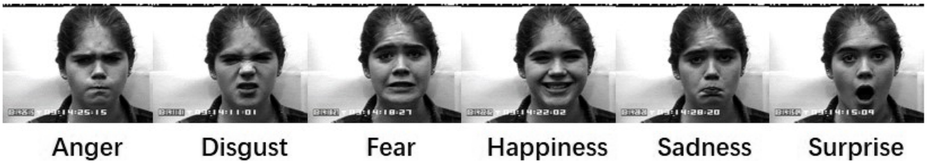


Fig. 1 Example of six prototypical expressions from the CK+ database

LBP-TOP method has computational simplicity and higher effectiveness for characterizing spatiotemporal texture features. The improved LBP-TOP is represented as I-LBP-TOP in this paper. Subsequently, the SVM in the shogun toolbox is used to obtain corresponding results of three classifiers [60], and MV strategy can determine the final classification result.

The main contributions of this paper are as follows:

- Firstly, we put forward an improved LBP-TOP descriptor (I-LBP-TOP), which remedy the shortcomings of original LBP-TOP descriptor in FER.
- Secondly, a new geometric feature (GF) is proposed and extended to time dimension (ST-GF).
- Thirdly, a framework that integrates the spatiotemporal motion features (I-LBP-TOP and ST-GF) with static features (GMF) is proposed, which takes into account geometry-appearance and dynamic-still information simultaneously.
- Finally, the experimental results prove that the recognition performance of the system is greatly improved due to the introduction of multi-feature fusion method at the decision-level.

The remainder of this paper is arranged as follows. Previous related work is briefly reviewed in Section 2. In Section 3, we introduce the main contribution of this paper. Section 4 discusses and analyzes the experimental results and Section 5 concludes the paper.

2 Related work

2.1 Static image based approaches

Over the past few decades, most of researches have been devoted to expression analysis based on still frames. For example, a general approach is Local Binary Pattern (LBP), which was first introduced by Ojala et al. [43] and implemented by Shan et al. [47] in the field of FER. As a simple descriptor with rotation and illumination invariance, LBP is widely used in expression recognition. This descriptor has many variants, including Completed LBP (CLBP) [16], support LBP (sLBP) [40] and scale selective LBP (SSLBP) [15]. Gabor wavelets have also been proven to be a powerful tool with an optimal localization in both the spatial as well as the frequency domain [26]. Gabor magnitude features are commonly used for modeling face changes, while there are also several Gabor phase based approaches like HGPP [55] and LGXP [51], which show competitive performance for facial feature extraction. What is more, the Histogram of Orientated Gradients (HOG) features are constructed by calculating and counting intensity gradient distribution of the local image region to represent shape and appearance information of facial image [9]. For geometric features, Liliana et al. utilized landmarks in a facial component to analyze facial expression [30]. Furthermore, in [6] and [27], the displacement and the coordinates difference of facial landmarks between a neutral face and an emotional face were calculated to characterize facial rigid changes, respectively. However, there are not enough neutral expressions in some databases or in the real system, which makes it impossible to calculate geometric features by far.

2.2 Dynamic image sequence based approaches

As a matter of fact, natural facial expression activity is a dynamic process, and its changing process can be disassembled into three stages: the onset, the apex and the offset. Therefore, video-based approaches in FER have become an active topic in recent studies. Both volume local binary pattern (VLBP) and LBP-TOP are extensions of LBP descriptor in time dimension, combining motion and appearance textures [57]. LBP-TOP, as a simplified descriptor of VLBP, has shown its promising performance for FER system [23]. However, the histograms extracted from XY plane are not as significant as those from XT and YT plane. Authors of [2] have proposed the LGBP-TOP descriptor, in which LBP-TOP was used on each Gabor magnitude sequence to further enhance the feature extraction. One drawback of this method is that the computational cost can be very high when a facial expression sequence is represented as 40 Gabor magnitude sequences. By using the optical flow approach, Guojiang et al. analyzed the dynamic information of facial expressions and extracted the characteristic flow which could reflect the facial expression changes effectively [17].

2.3 Multi-feature fusion based approaches

Latest studies suggest that the fusion of multi-feature can yield better results than only performing a single feature in emotion recognition system. The method reported in [1], proved that simple combinations of both static and dynamic approaches can break through their respective limitations. Moreover, Zhao et al. proposed a novel framework for facial expression analysis concatenating dynamic and static information in video sequences [59]. However, it is difficult to generalize universal features across different persons only from the extracted spatiotemporal texture information. Fan et al. combined PHOG-TOP and dense optical flow according to the weighting strategy to extract both the spatial and dynamic motion information of facial expressions [11]. In [12], Feng et al. focused on two-stream-CNN with LBP-TOP to capture spatial and temporal streams.

In addition, multiple features can be fused at feature-level or decision-level. For instance, Hu et al. [22] employed Center-Symmetric Local Signal Magnitude Pattern (CS-LSMP) descriptor on multiple features for obtaining fused features. Rathee et al. fused Gabor, HOG and DWT features, using Multiview Distance Metric Learning (MDML), which employed complementary features of images to extract details while eliminating redundant information [44]. In [7], HOG-TOP, acoustic and geometric features were combined, and multiple kernel SVM was used for classification at the feature-level. On the contrary, in [46], both audio and visual information were fused at the decision-level with a decision rule to identify emotions. And in [13], towards the extracted SIFT and LBP descriptors, Gao et al. used the improved DS-evidence theory for decision-level information fusion to improve the robustness of face recognition in complex conditions. In general, system performs excellently when combining multiple complementary features.

3 Proposed model

This section presents the detailed methodology of our proposed framework, including three types of feature descriptors (I-LBP-TOP, GMF, ST-GF) and a classification method for multi-feature fusion at the decision-level. As illustrated in Fig. 2, the process of the proposed FER system includes following steps: (1) for the preprocessed image sequence, extracting LBP histograms on XT and YT planes and combining with LBP histograms on XY plane of peak image to generate I-LBP-TOP descriptor; (2) employing Gabor operator to extract GMFs from preprocessed peak image; (3) utilizing facial key points for the sampled image sequence to obtain ST-GF; (4) for the above three features, three SVM base classifiers are

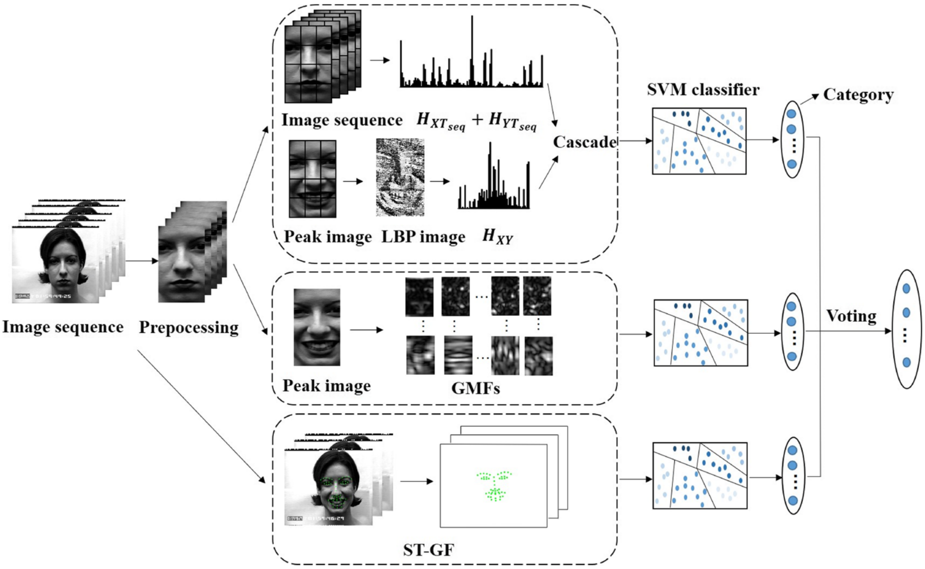


Fig. 2 Overview of our proposed framework for FER

trained, and the final classification result of an unknown sample is determined by MV strategy of the base classifiers. The detailed algorithms of the framework are in the following sections.

3.1 Texture and Spatial-Temporal Motion LBP Descriptor

The LBP is a descriptor for extracting local texture features, which calculates the pixels in an image successively. For each pixel in the image to be processed, the neighboring pixels are thresholded to generate a binary code that is usually converted to a decimal number to represent the LBP value of the central pixel. It reflects the texture information of local neighborhood (see Fig. 3 for an illustration). In this way, each pixel in the picture is redistributed according to the values of its neighboring pixels to obtain the LBP feature.

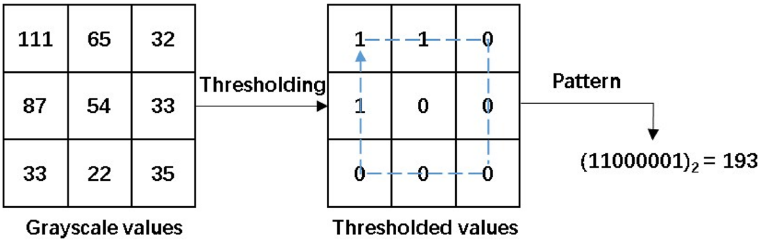


Fig. 3 The Original LBP descriptor

Denoting by g_p the pth neighboring pixel of the central pixel g_c, by P the total number of all involved neighbors and by R the radius of the neighborhood, the calculation method of the original LBP descriptor is given in (1).

$$LBP_{P,R} = \sum_{p=0}^{P-1} s_1(g_p, g_c)2^p, \tag{1}$$

where the function s_1 can be formulated as follows:

$$s_1(x, y) = \begin{cases} 1 & \text{if } x - y \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

The LBP maps involve the local information, while their statistical histograms are utilized as feature vectors to take global information into consideration. For an image of size $M \times N$, the histogram after LBP encoding can be defined as:

$$H(i) = \sum_{m=1}^{M} \sum_{n=1}^{N} s_2(LBP_{P,R}(m, n), i), \ \forall i \in [1, I], \tag{3}$$

where the function s_2 is defined as:

$$s_2(x, y) = \begin{cases} 1 & \text{if} x - y = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where i is the number of patterns after LBP mapping and I is the maximal LBP pattern value. Note taht 59-dimensional histogram of uniform pattern is adopted in this paper.

The co-occurrences on XT and YT planes of LBP-TOP are applied concerning both the spatial and temporal domain information. It is known that a single frame is a two-dimensional plane and a video sequence is a three-dimensional volume. Accordingly, a video sequence can be considered as a stack of XY planes in the temporal dimension T, and similarly for XT and YT planes but in the axis Y and X, respectively. Figure 4 shows a 33-frame images sequence and its corresponding example images from three planes. Figure 4a

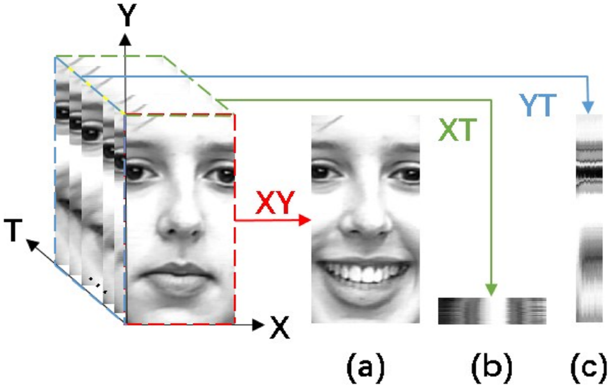


Fig. 4 The example images in XY, XT and YT planes. (a) Image in the XY plane (128×256) in t = 8. (b) Image in the XT plane (128×33) in y = 128. (c) Image in the YT plane (33×256) in x = 64

shows the frontal face image of frame 8 in the XY plane, while Fig. 4b and c describe the top and side views of the volume, both of which give the visual impression of motion changing in temporal space. The LBP histograms obtained from image sequence on XY, XT and YT planes are represented by $H_{XY_{seq}}$, $H_{XT_{seq}}$ and $H_{YT_{seq}}$, respectively. Then a single histogram is derived by the concatenation of these statistical histograms corresponding to separate planes. Figure 5 demonstrates the procedure of the LBP-TOP feature extraction for a block. The final histogram can be formulated as follows:

$$\begin{aligned} &H = (H^1, H^2, H^3), \\ &H^j(i) = \sum_{x,y,t} s_2(LBP_{P,R}(x, y, t), i), \ \forall i \in [1, I], \end{aligned} \tag{5}$$

where $LBP_{P,R}(x, y, t)$ represents the LBP code of the central pixel in the jth plane, where $j = 1, 2, 3$ denotes the XY, XT, YT planes respectively. In the block-based approach, the histogram of each block volume needs to be cascaded to obtain the final feature vector.

However, in FER, these three planes of LBP-TOP contribute differently for feature expression, and not all components are of equal importance. Compared with XT and YT planes, the features extracted from the XY plane contribute less. The $H_{XY_{seq}}$ is achieved from neutral face to emotional face of dynamic sequence, while the neutral face does not contain the corresponding expression information, which may weaken the feature expression ability of $H_{XY_{seq}}$. In contrast, $H_{XT_{seq}}$ and $H_{YT_{seq}}$ explain more about the movement of facial muscles.

Similar to [59], $H_{XY_{seq}}$ is abandoned, and the joint $H_{XT_{seq}}$ and $H_{YT_{seq}}$ construct spatial-temporal motion LBP features. However, unlike in [59], we utilize the LBP histogram of the peak frame instead of Gabor multiorientation fusion histogram to enhance spatial texture information. On the one hand, this is because compared with Gabor multiorientation fusion histogram, the LBP histogram has lower calculation expense and can make up for the deficiency of the original LBP-TOP descriptor in XY plane. On the other hand, with the uniform pattern coding, the dimension of LBP histogram is the same as that of the LBP-TOP histogram in single plane, both of which are 59 dimensions. In this paper, we combine the LBP texture feature of the peak frame (H_{XY}) with the spatiotemporal motion feature of image sequence, which is called I-LBP-TOP descriptor. The detail of I-LBP-TOP algorithm is presented in Algorithm 1.

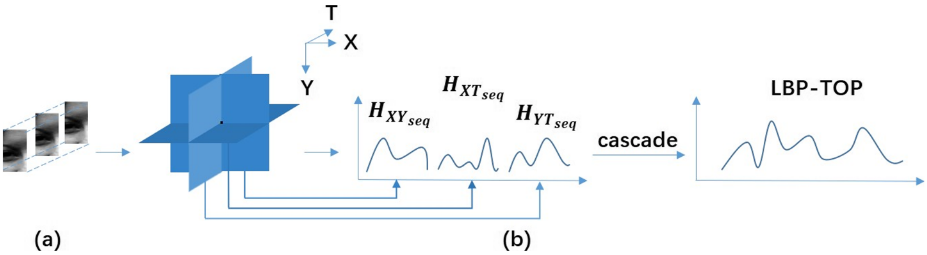


Fig. 5 LBP-TOP histogram of a block volume. (a) Block volume. (b) LBP histogram from three planes

Algorithm 1 Computation of the I-LBP-TOP.

```
Require: Video sequence $V$, which contains $N$ frames with the same width $W$ and height
    $H$, the radius of circle in XY, XT, YT planes $R_X$, $R_Y$, $R_T$.
Ensure: The histogram of improved LBP-TOP.
if $t = N$ then
    for $x = R_X + 1$ to $W - R_X$ do
        for $y = R_Y + 1$ to $H - R_Y$ do
            Compute the LBP value using (1);
            Compute the histogram using (3), i.e., $H_{XY}$;
        end for
    end for
else
    for $t = R_T + 1$ to $N - R_T$ do
        for $x = R_X + 1$ to $W - R_X$ do
            for $y = R_Y + 1$ to $H - R_Y$ do
                Compute the LBP value using (1);
                Compute the histogram using (5) in XT and YT planes, i.e., $H_{XT_{seq}}$ and $H_{YT_{seq}}$;
            end for
        end for
    end for
end if
Normalize the $H_{XY}$, $H_{XT_{seq}}$ and $H_{YT_{seq}}$ respectively.
Concatenate the three histograms into a long histogram.
```

3.2 Gabor magnitude descriptor

Gabor transform can analyze the gray changes of images in multi-resolution and multi-orientation effectively and thus is capable of solving the problem of different expressions with different scales. Meanwhile, it has good properties for information extraction in local spatial and frequency domain. Based on the above advantages, Gabor feature has been successfully applied in the field of FER. For point $z = (x, y)$, the 2-D Gabor filter commonly used is defined as follows:

$$G_{(\mathrm{u,v})}(z) = \frac{||k_{u,v}||^2}{\sigma^2} \exp\left(-\frac{||k_{u,v}||^2||z||^2}{2\sigma^2}\right) \times \left[\exp(ik_{u,v}z) - \exp(-\frac{\sigma^2}{2})\right] \tag{6}$$

where $k_{u,v} = k_v e^{i\phi_u}$, $k_v = k_{\max}/\lambda^v$, and $\phi_u = \pi u/8$, $k_{\max}$ is the maximum frequency, λ is the spacing factor between filters in the frequency domain, u, v corresponds to the direction and scale of Gabor filter, respectively, and $|| \cdot ||$ is the norm descriptor.

The Gabor representation of an image is the convolution of image $I(z)$ with Gabor filters, where five scales and eight orientations are used:

$$F_{u,v}(z) = I * G_{u,v}(z), \tag{7}$$

where $u \in \{0, 1, \cdots, 7\}$, $v \in \{0, 1, \cdots, 4\}$ and $*$ stands for the convolution operator. It should be noted that the coefficients of the Gabor wavelet are complex, hence the response $F_{u,v}$ of a Gabor filter is complex. Unlike the Gabor phase information of the transform which is time-varying, the magnitude of Gabor response is relatively smooth and stable. As a consequence, we exploit the magnitude of the response $F_{u,v}$ to yield the Gabor feature. In

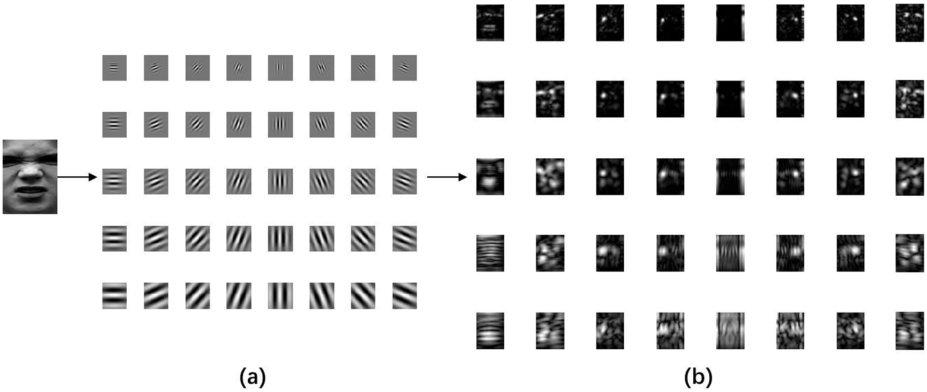


Fig. 6 The example of Gabor wavelet transformation. (a) Real part of Gabor kernel. (b) 40 Gabor magnitude pictures

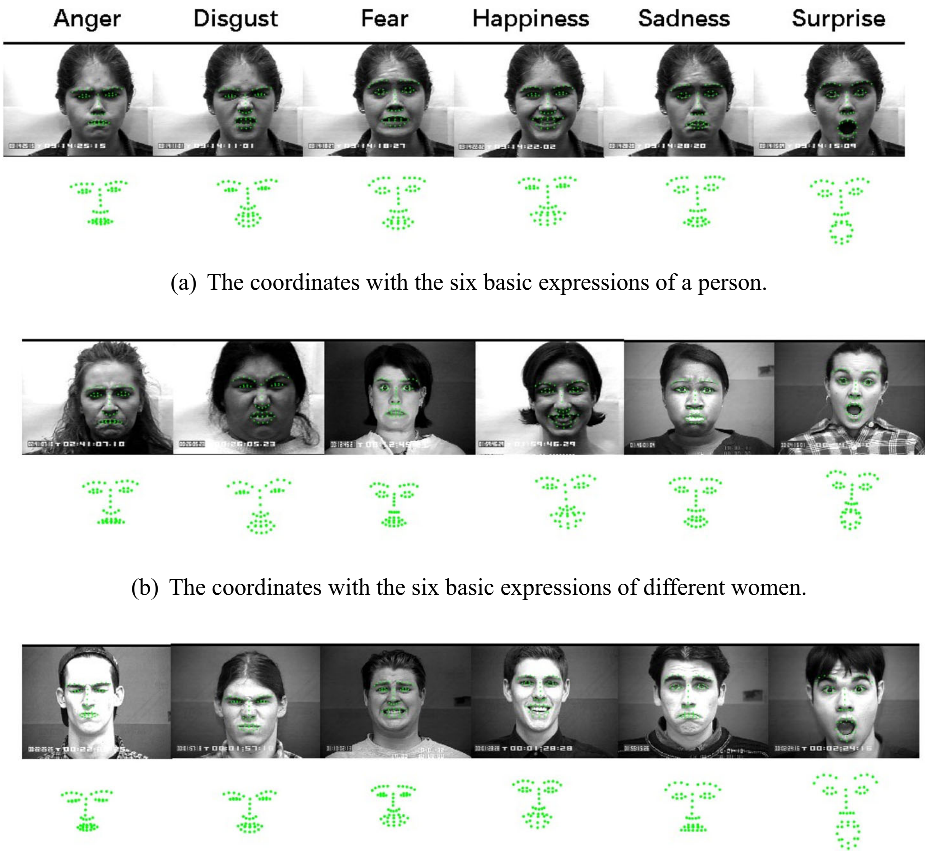
this paper, we generate 40 Gabor magnitude maps for each individual face image by Gabor wavelet transformation. An example of the Gabor wavelet transformation with five scales and eight orientations is given in Fig. 6.

3.3 Geometric descriptor

Facial expression not only encompasses the objective appearance and shape information, but also involves the unexpected identity characteristics of different subjects. The advantage of facial key points features is that they are not influenced by person identity such as face shape, gender, age, race and illuminance in the input video. Accordingly, we employ the facial key points as a geometric feature in this paper. The major task is to locate the key points from the face, including the corners of the eyebrows, eyes, mouth and nose in an image. It is encouraging that there have been many mature algorithms for facial key point detection [19] and face alignment [41], which are widely used in expression recognition, face tracking and face recognition. We utilize Supervised Descent Method (SDM) algorithm to locate the 49 key points of a face [52]. The coordinates of the facial image with emotion are shown in Fig. 7. As depicted in Fig. 7, no matter the same person, different women or men, the coordinates of the same expression are very similar according to columns. Furthermore, the coordinates of different expressions do have a great difference according to rows, especially the mouth, eyes and eyebrows, which proves that the key points of the face can remove the common underlying structure for the face images and extract the shape attributes of expressions effectively. Therefore, we have adequate grounds to take the coordinates of these 49 facial key points as geometric feature. Supposing that (x_i, y_i) represents the coordinates of the i-th facial landmark, a set of facial landmarks can be expressed as (8):

$$V_e = [x_1,\ y_1,\ x_2,\ y_2,\ \cdots,\ x_n,\ y_n], \quad n = 49, \tag{8}$$

where V_e is the geometric vector of emotion e. Then, the coordinates of X and Y axis are standardized with the mean value of 0 and the variance of 1 respectively which transforms them to dimensionless pure values. As a consequence, a total of 98-length coordinates are constructed to represent the geometric feature. The extracted 98-length vector is learned through the multi-kernel SVM and achieves high recognition accuracy.



(a) The coordinates with the six basic expressions of a person.

(b) The coordinates with the six basic expressions of different women.

(c) The coordinates with the six basic expressions of different men.

Fig. 7 The coordinates of facial expression images

As we all know, facial expression can be considered as a dynamic process, in which the same facial key point will produce relative displacement between frames. Therefore, the change of neutral face to expressive face in time dimension can be described by the trajectory of facial key points. To obtain the trajectory data and relative position information of the facial key points simultaneously, we consider extending the above GF descriptor to time dimension. Specifically, the original image sequence is normalized to a fixed length: L. Then the GFs of the L frames are concatenated into a one-dimensional vector, which is represented by V. Consequently, the ST-GF descriptor can be calculated as follows:

$$V = \left[V_{\mathrm{e}}^{1}, V_{e}^{2}, \cdots, V_{e}^{L}\right] \tag{9}$$

3.4 Classification

Classification is the process of training features to get the optimal mapping model between features and tags, and realizing the correct prediction of unknown samples. SVM is considered as one of the most effective and robust classifier for FER due to its following properties:

(1) a good balance between model complexity and generalization error and (2) a capability to deal with high dimensional data. We denote by $\{(x_i, y_i), i = 1, ..., L\}$, $x_i \in R^n$, and $y_i \in \{-1, 1\}$ a set of training data with labels. A new test data x is classified by

$$f(x) = \text{sign}\left(\sum_{i=1}^{L} \alpha_i y_i K(x, x_i) + b\right), \tag{10}$$

where α_i are Lagrange multipliers of a dual optimization problem, describing the separated hyper-plane, b is a bias, and $K(\cdot, \cdot)$ is a kernel function. For linear separable data, SVM finds a hyperplane to maximizes the margin with respect to support vectors. For nonlinear data, the processing method of SVM is to map the data into a higher dimensional space by selecting appropriate kernel function. Among various kernel functions, the most frequently used are polynomial and radial basis function (RBF) kernels. However, different feature vectors have different dimensions and are of different importance for recognition. It is difficult to ensure that the parameters and kernel functions are suitable for all feature vectors when performing SVM. On this condition, multiple kernel learning (MKL) is also a good choice, which employs a convex combination of multiple kernels to substitute for the single kernel:

$$\begin{aligned} K(x, x') &= \sum_{m=1}^{M} d_m K_m(x, x'), \\ s.t. \quad d_m &\geq 0, \quad \sum_{m=1}^{M} d_m = 1, \end{aligned} \tag{11}$$

where M is the total number of kernel functions and K_m represents basic kernel function.

Since SVM is a typical binary classifier, for multi-class problem, the one-vs-one and one-vs-rest approaches are simple but effective technique. In our study, we adopt one-vs-one method to deal with the six-class problem. In this strategy, SVM is trained between any two types of samples and 15 binary SVM classifiers need to be designed.

3.5 Multi-feature fusion at decision-level

According to the different stages of feature fusion, it is mainly divided into feature-level fusion and decision-level fusion. Feature-level fusion is performed before classification, which concatenates multiple features directly or according to weight ratio to form high-dimensional feature vectors. On the contrary, decision-level fusion is carried out after classification, and the final category of sample is determined by MV strategy of ensemble learning. We utilize feature-level fusion method to deal with features with small differences, such as I-LBP-TOP descriptor, while for multiple types of features, decision-level fusion method is adopted.

As proved by Bonab et al. [4], the MV strategy of ensemble learning combines multiple classifiers to obtain at least equal to the average performance of all individual components. Based on the above theory, We train three SVM base classifiers for three different types of features (I-LBP-TOP, GMF, ST-GF), and the predicted results are combined through MV. In the testing phase, for each sample x, the method of calculating the MV is as follows:

$$H(x) = \mathfrak{c}_{\kappa}, \quad \kappa = \arg\max_{j} \left\{\sum_{t=1}^{T} {h_t}^j(x)\right\}, \tag{12}$$

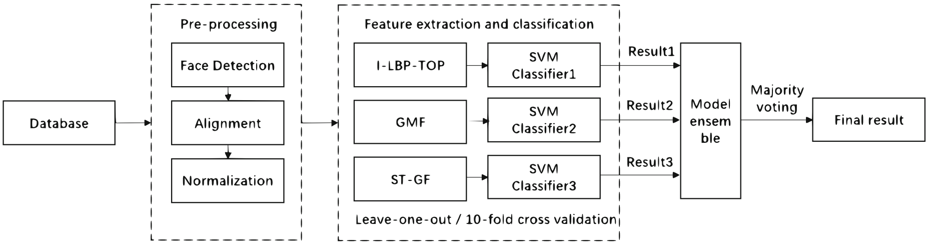


Fig. 8 Flow chart of our proposed FER system

where T is the number of base classifiers, ${h_t}^j(x) \in \{0, 1\}$ denotes the class tag, if h_t predicts sample x as class $\mathfrak{c}$, the value is 1, otherwise it is 0. When an unknown sample is classified, the category with the largest number of votes is the final classification result.

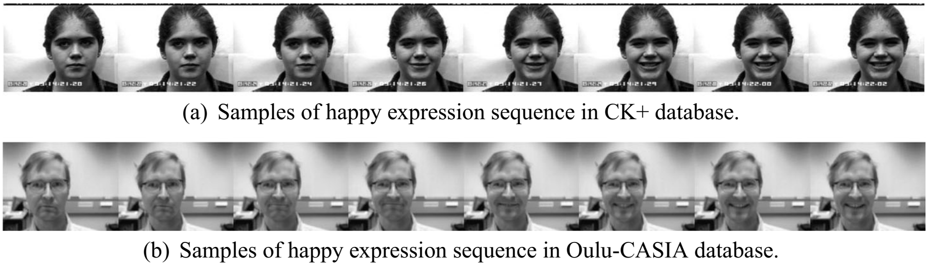
4 Experimental results

To evaluate the performance of our proposed model, we perform experiments on the CK+ and Oulu-CASIA facial expression databases. The details of the experiments and results are shown below. They are based on Windows OS with CPU Intel Core (TM) i5-1035G1 and 16GB of RAM. The feature extraction phase is based on the MATLAB platform of version R2019a. In addition, the design of classifier uses Shogun toolbox, which is based on Python platform of version 3.6. In order to test the predictive performance of our model, we take the leave-one-out cross validation. In this method, all the expression of each face can be used as a test set, which is very commonly used in expression recognition. Besides, the 10-fold cross validation is also used on CK+ database to compare with the existing methods. As shown in Fig. 8, our proposed FER system comprises the following stages: pre-processing, feature extraction, classification and feature fusion.

4.1 Facial expression databases and preprocessing

The extended Cohn-Kanade (CK+) database is an effective and general database to verify expression recognition system [35]. The expression sequence of this database starts from neutral expression and gradually changes to the peak of expression. It contains 593 video sequences from 123 individuals between the age of 18 and 30, while only 327 facial expression sequences are labeled with seven universal emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise). Most of the papers have abandoned the contempt expression because of its small amount of data (only 18 expression sequences). In order to facilitate the comparative experiments, we select 309 facial expression sequences with six basic expressions, excluding contempt.

The Oulu-CASIA database is another widely used video based database, which contains six basic expressions from 80 subjects of 23 to 58 years old (a mix of male/female and glasses/without glasses) [56]. Facial expressions are captured by a VIS camera under three different light conditions: normal, weak and dark. Similar to the CK+ database, all expression sequences are also changing from neutral to peak of emotion. We evaluate our model with 480 sequences (80 subjects by six expressions) in the normal illumination condition. The face sequence samples of CK+ and Oulu-CASIA databases are illustrated in Fig. 9.


(a) Samples of happy expression sequence in CK+ database.

(b) Samples of happy expression sequence in Oulu-CASIA database.

Fig. 9 Samples of expression sequence

It should be noted that in the process of collecting expression data, the acquired facial image can be tilted inevitably due to the change of facial muscles or head deflection. Hence, it is necessary to align any in-plane rotation so that the eyes are on the same horizontal line to optimize the recognition performance. After getting enough aligned samples, we note that even if the images come from benchmark facial expression databases, there is too much background information independent of expression. Accordingly, face detection and normalization are applied to remove the irrelevant information and improve the quality of subsequent feature extraction and recognition for FER [20, 21]. Support that the distance between the pupils of two eyes is d, the height and width of the cropped image are $2.25 \times d$ and $1.2 \times d$ based on the middle position of the pupils of two eyes, as shown in Fig. 10. These factor values are determined empirically [34]. Finally, in CK+ database, for the LBP feature, the cropped face is normalized to 256×128, while for the Gabor feature, it is normalized to 112×96 to reduce dimensions. Since the original image pixels of the Oulu-CASIA database are low, only 320×240, the cropped face is normalized to 112×64 and 64×48 respectively for the above two features.

4.2 The effect of block numbers and overlapping ratio on I-LBP-TOP descriptor

First of all, we perform experiments on CK+ database to determine the appropriate number of blocks and whether to use overlapping blocks for I-LBP-TOP descriptor. We can learn from the previous researches that too few blocks could make the extracted features insufficient to get terrible accuracy, while a large number of blocks may also lead to the problem of too high feature dimensions, increasing time complexity and decreasing accuracy. What

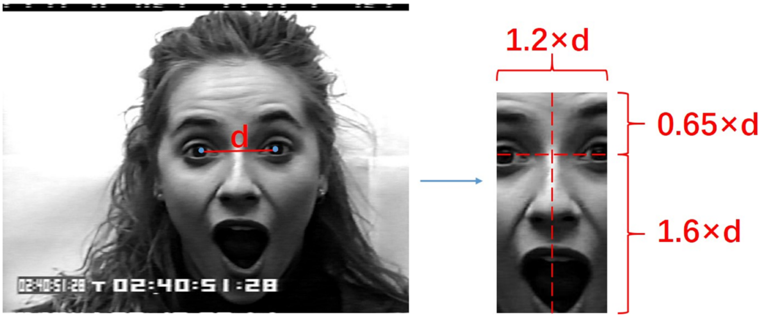


Fig. 10 The original image and the image normalized to 256×128 after cropping on CK+ database

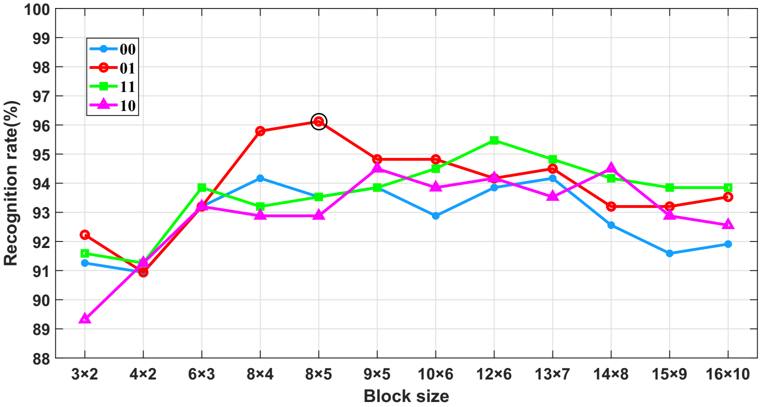


Fig. 11 Recognition rate of I-LBP-TOP descriptor according to the block size and whether performing overlapping ratio of 80% on CK+ database, where the black circle means the best result

is more, the overlapping ratio of 80% of original image has been shown to obtain the best result. Figure 11 shows how the accuracy of expression recognition varies with the number of blocks. Assume that the overlapping ratio of 80% of original image is represented by '1', and non-overlapping block is represented by '0'. Then '00' indicates that both the H_{XY} and $H_{(XT+YT)_{seq}}$ features are not overlapped; '01' means: the H_{XY} feature performs without overlapping, while $H_{(XT+YT)_{seq}}$ feature performs 80% overlap; '11' represents both the H_{XY} and $H_{(XT+YT)_{seq}}$ features with an overlapping ratio of 80%; and '10' indicates that the blocking method with an overlapping ratio of 80% is adopted by H_{XY}, whereas the $H_{(XT+YT)_{seq}}$ feature does not.

It can be found that the best result is obtained with 8×5 blocks, and a small or a large value of blocks will degrade the recognition performance. Consequently, 8×5 blocks are selected for all facial images. Besides, for spatial-temporal LBP histogram, we adopt the overlapping ratio of 80% of original block, whereas for LBP histogram of the peak frame, the non-overlapping blocks are used to obtain more position information.

4.3 Comparison of I-LBP-TOP with original LBP-TOP and other components

Secondly, we separate the LBP-TOP descriptor in three planes and investigate the performance of LBP histogram of peak frame (i.e., H_{XY}), LBP histogram of three individual planes (i.e., $H_{XY_{seq}}$, $H_{XT_{seq}}$, $H_{YT_{seq}}$), pairwise combination of different plane components (i.e., $H_{XY}+H_{XT_{seq}}$, $H_{XY}+H_{YT_{seq}}$, $H_{(XY+XT)_{seq}}$, $H_{(XY+YT)_{seq}}$, $H_{(XT+YT)_{seq}}$), I-LBP-TOP histogram and original LBP-TOP histogram on CK+ database. All the LBP descriptors are coded with uniform patterns. Based on the 8×5 blocks division method, the obtained feature vector is $59 \times 8 \times 5 = 2360$ dimensions, which is reduced to 308 dimensions by using principal component analysis (PCA) [61].

We utilize ablation study to analyze the contribution of different components, as shown in Table 1. The highest performance is obtained by combining the LBP histogram from peak frame with the spatiotemporal LBP histogram, which is 3.56% higher than original LBP-TOP descriptor. Compared with the LBP histogram of XY plane from image sequence whose recognition rate is only 84.14%, the accuracy of LBP histogram from peak image can reach 92.88%. We can conclude that the features extracted from XY plane of image

Table 1 Ablation study of LBP histogram from either XY, XT, YT plane or their combination on CK+ database

$H_{YT_{seq}}$				✓		✓	✓		✓	✓	✓
$H_{XT_{seq}}$			✓		✓		✓	✓		✓	✓
$H_{XY_{seq}}$		✓						✓	✓		✓
H_{XY}	✓				✓	✓				✓	
Accuracy (%)	92.88	84.14	90.94	92.56	93.53	95.47	93.20	90.61	92.88	96.12	92.56

sequence are affected by a series of weak expressions (i.e., neutral expressions and changing expressions) and cannot extract representative texture features. In addition, we can figure it out that the LBP descriptor in XT and YT planes can effectively extract the spatiotemporal information. The effect of YT plane is better than the other two planes, indicating that the shape information in the vertical direction plays more important role than that in the horizontal direction. The experimental results validate that our proposed I-LBP-TOP descriptor has the ability to extract more effective spatiotemporal texture features.

In order to test the adaptability of LBP histogram components in different classifiers, experiments are performed on K-Nearest Neighbor (KNN), Random Forest (RF), Artificial Neural Networks (ANN) and SVM classifiers, as shown in Fig. 12. It can be seen from the above figure that the change of recognition rate of each LBP histogram component has basically the same tendency regardless of the classifiers involved. For instance, the H_{XY} performs much better than $H_{XY_{seq}}$ and the recognition rate of I-LBP-TOP descriptor is always higher than that of original LBP-TOP descriptor in all kinds of classifiers. Furthermore, we also find that SVM is more suitable for our LBP histogram features compared with other classifiers.

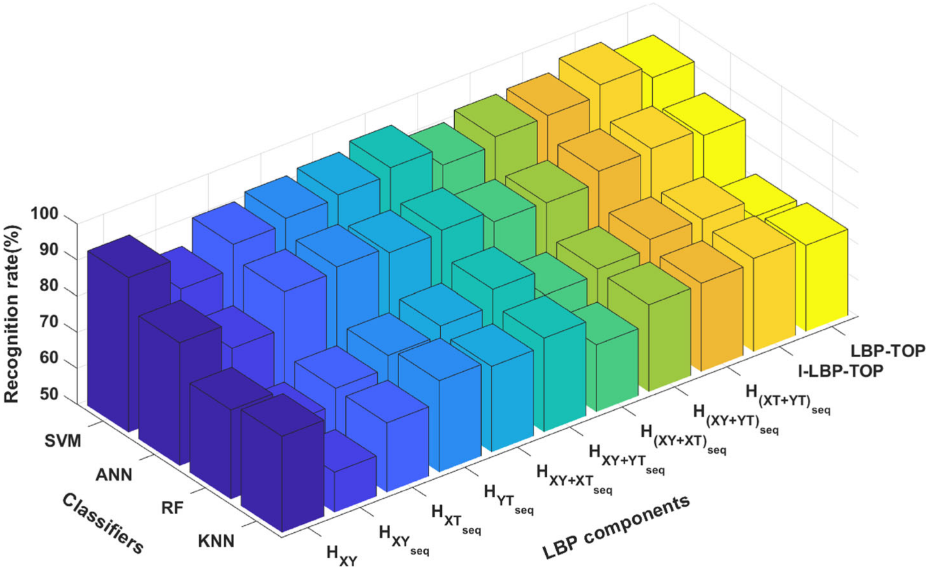


Fig. 12 The recognition rate of LBP histogram from either XY, XT, YT plane or their combination on KNN, RF, ANN and SVM classifiers on CK+ database

Table 2 Comparison results of GF and ST-GF descriptors

Descriptor	Database	AN	DI	FE	HA	SA	SU	Average (%)
GF	CK+	88.90	96.60	84.00	98.60	71.40	100.0	93.53
ST-GF	CK+	93.30	96.60	96.00	97.10	82.10	100.0	95.79
GF	Oulu-CASIA	60.00	62.50	68.80	86.30	68.80	88.80	72.50
ST-GF	Oulu-CASIA	66.30	61.30	67.50	91.30	70.00	92.50	74.79

In addition, we compare the computational speed of original LBP-TOP and I-LBP-TOP descriptors on CK+ database. The computation time varies with the length of expression sequence and the number of blocks. Under the same condition of expression sequence (19) and block size (8 × 5), the computation time of original LBP-TOP is 3.96s, while that of I-LBP-TOP is 3.27s. Moreover, when the length of expression sequence is set to 40, the computation time of original LBP-TOP and I-LBP-TOP is 8.38s and 7.21s, respectively. With the increase of sequence length and block number, the time superiority of I-LBP-TOP descriptor is more obvious.

4.4 Evaluation of the proposed geometric descriptors

In this subsection, we analyze the validation and superiority of our proposed geometric descriptors. As mentioned in Subsection 3.3, the dimensions of GF and ST-GF descriptors are 98 and $98 \times L$, respectively. In the experiment, L is set to 6 in CK + database and 9 in Oulu-CASIA database, which is the minimum length of the image sequence. Table 2 shows the comparison results of GF and ST-GF descriptors on two databases (CK + and Oulu-CASIA). We can see that compared with the GF, the overall recognition rate of ST-GF is increased, especially for angry and sadness. ST-GF can enhance GF and better capture the subtle changes of those two expressions.

Furthermore, our proposed ST-GF is compared with other geometric feature extraction algorithms on CK+ database, as illustrated in Table 3. The method [36] extracted geometric features according to facial key regions. For eyebrows and lips, the coordinate differences of key points among frames are calculated as displacement information, while for the eye regions, the projection ratio of horizontal distance to vertical distance is utilized. In [33], Euclidean distances of facial landmarks are put into graph-based network to obtain DAUGN-G expression recognition model. And [50] employed Riemannian sparse coding and dictionary learning to code shape trajectories of 2D facial landmarks. From the comparison results, it is witnessed that our proposed ST-GF achieves a superior performance, which outperforms the geometric features in recent years.

Table 3 Comparison results of different geometric features on CK+ database

Reference	Method	Verification method	Class	Accuracy (%)
[36] (2018)	Ratio+displacement information(SVM)	10-fold	6	93.50
[33] (2020)	DAUGN-G	10-fold	6	93.55
[50] (2020)	Extrinsic SCDL(Bi-LSTM)	10-fold	7	93.75
Ours	ST-GF(SVM)	Leave-one-out	6	95.79
		10-fold		95.48

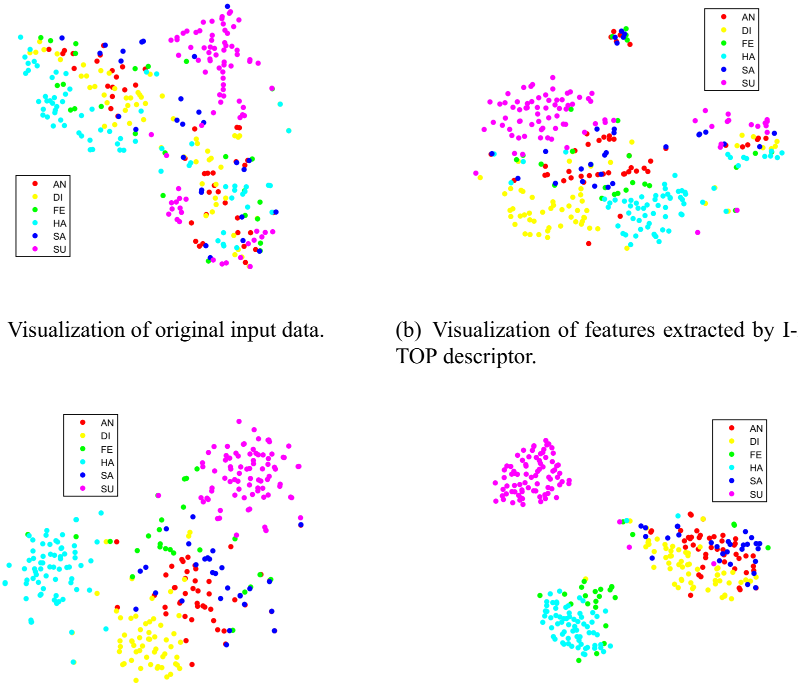
4.5 Feature visualization

To further demonstrate the separability of the proposed features, we utilize T-Distributed Stochastic Neighbor Embedding (t-SNE) [28] algorithm to visualize the feature vector extracted by I-LBP-TOP, Gabor and spatiotemporal geometric descriptor, respectively. The t-SNE algorithm is a useful visualization technique to convert high-dimensional data into two space. Figure 13 shows the visualization results on CK+ database.

Figure 13(a) shows the random distribution of original input data, with various types of samples mixed together. A total of 6 clusters is having, representing each facial expression. It may be observed from Fig. 13(b)to 13(d) that our feature extraction method can effectively distinguish six kinds of expressions to a certain extent. Particularly, the expressions of surprise, happiness and disgust can be well clustered together in the same category.

4.6 The influence of feature fusion on expression recognition rate

Subsequently, we explore the effectiveness of combining I-LBP-TOP feature, GMF and ST-GF at the decision-level. The above three features utilize SVM classifiers with their appropriate kernel functions and save the prediction results separately. For each single sample, the three classifiers adopt the principle that a few obeys the majority to determine the final category. If the prediction result of every classifier is different, a classification is



(a) Visualization of original input data.

(b) Visualization of features extracted by I-LBP-TOP descriptor.

(c) Visualization of features extracted by Gabor descriptor.

(d) Visualization of features extracted by spatiotemporal geometric descriptor.

Fig. 13 The visualization results on CK+ database

Table 4 Confusion matrix of I-LBP-TOP feature on CK+ database

Expression	AN	DI	FE	HA	SA	SU
AN	97.80	0	0	0	2.20	0
DI	1.70	96.60	0	1.70	0	0
FE	0	0	84.00	8.00	8.00	0
HA	0	0	0	100.0	0	0
SA	17.90	0	0	0	82.10	0
SU	0	0	0	0	0	100.00
Average						96.12

Table 5 Confusion matrix of GMF on CK+ database

Expression	AN	DI	FE	HA	SA	SU
AN	88.90	2.20	2.20	0	6.70	0
DI	0	100.00	0	0	0	0
FE	4.00	0	72.00	12.00	4.00	8.00
HA	0	0	2.90	97.10	0	0
SA	10.70	0	0	0	89.30	0
SU	0	0	0	0	0	100.00
Average						95.15

Table 6 Confusion matrix of ST-GF on CK+ database

Expression	AN	DI	FE	HA	SA	SU
AN	93.30	2.20	2.20	0	2.20	0
DI	1.70	96.60	0	0	1.70	0
FE	0	0	96.00	0	0	4.00
HA	0	0	2.90	97.10	0	0
SA	14.30	3.60	0	0	82.10	0
SU	0	0	0	0	0	100.00
Average						95.79

Table 7 Confusion matrix of hybrid feature at decision-level on CK+ database

Expression	AN	DI	FE	HA	SA	SU
AN	100.00	0	0	0	0	0
DI	0	100.00	0	0	0	0
FE	0	0	100.00	0	0	00
HA	0	0	0	100.00	0	0
SA	7.10	0	0	0	92.90	0
SU	0	0	0	0	0	100.00
Average						99.35

randomly selected as the final result. Tables 4, 5, 6 and 7 shows the classification accuracy obtained by applying I-LBP-TOP feature, GMF, ST-GF and hybrid feature on CK+ database, respectively. The horizontal axis represents a predicted class among six emotions, and the vertical axis represents the target class which is the correct label.

From the results in the table, we observe that fear and sadness are the most confused expressions. On the one hand, because of the relatively small sample numbers of these two expressions, it is difficult to extract a group of features that reveal the internal rules. On the other hand, the slight dynamic variations of facial critical areas for both fear and sadness create more difficulties to distinguish them clearly. Moreover, anger and sadness tend to be identified with each other incorrectly, which may be due to their similar mouth motion. As shown in Fig. 14, it is difficult to accurately distinguish some expressions even for human. However, happiness and surprise can be easily recognized with an accuracy of nearly 100%, which is attributed to their relatively large muscle deformations and drastic changes in appearance.

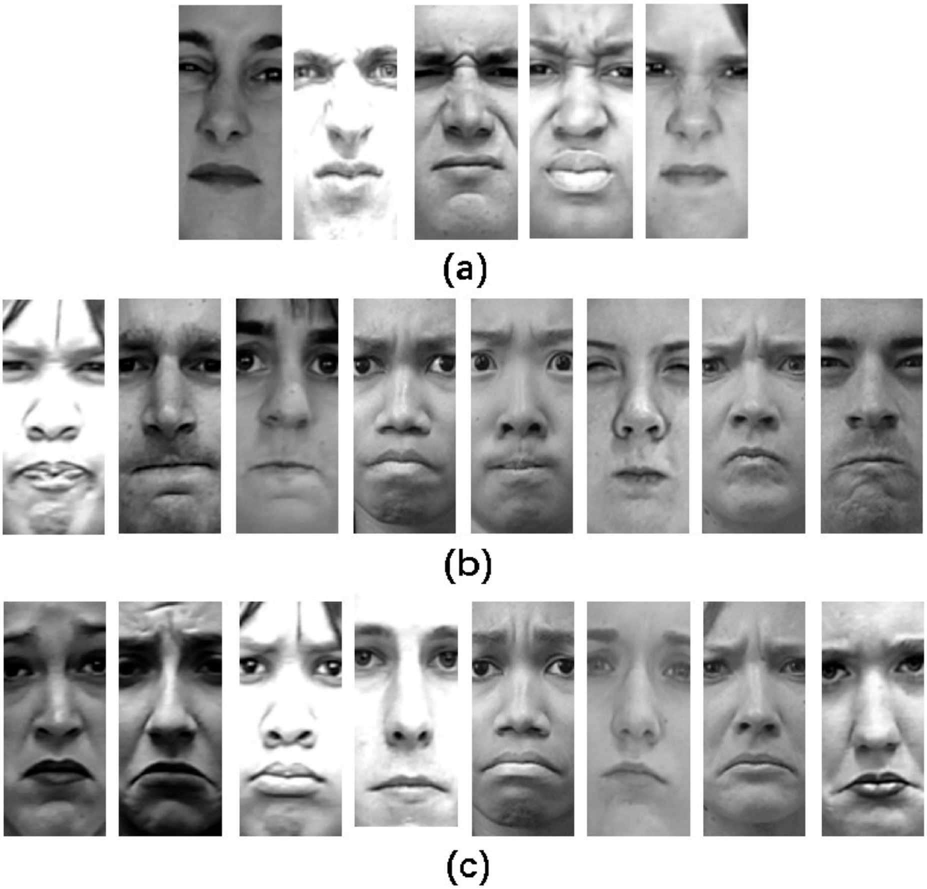


Fig. 14 Failed cases in CK+ database. (a) Expressions are mislabeled as Disgust from Anger. (b) Expressions are mislabeled as Sad from Anger. (c) Expressions are mislabeled as Anger from Sad

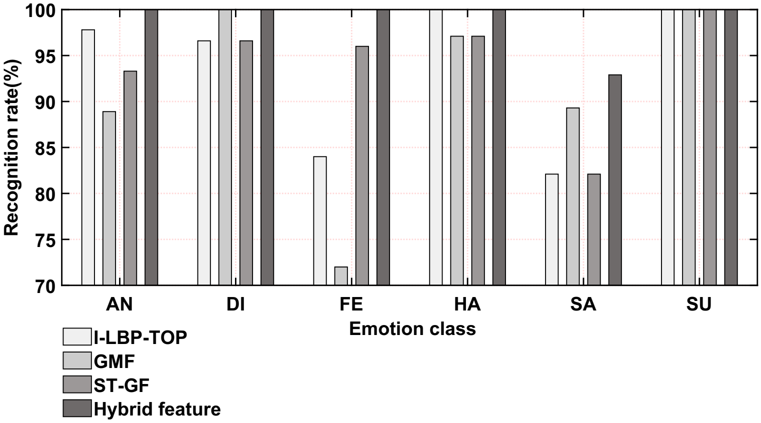


Fig. 15 Comparison of recognition rate between single feature and hybrid feature

According to the results illustrated in Fig. 15, we note that there is always a descriptor that performs better than the others when identifying a certain type of expression. For instance, the I-LBP-TOP descriptor is better at identifying anger; GMFs have greater advantages in recognizing disgust and sadness; and ST-GFs have the best recognition rate for fear. When merging at decision-level, the recognition rate of each expression can be improved owing to the different dominant expressions of these three descriptors.

Table 8 presents the comparison results of different methods on CK+ database. Fan et al. [11], Zhao et al. [59], Chen et al. [7], Bougourzi et al. [5], and Shanthi et al. [48] all used multiple hand-crafted feature fusion methods to evaluate the performance of model. Even compared with the recent convolutional neural network framework proposed by Yang et al. [53] and Kim et al. [27], our proposed traditional hand-crafted features perform better with the same 10-fold cross validation. Experimental results illustrate that fused features at decision-level are further enhanced based on the individual feature, achieving a final

Table 8 Comparison results of our proposed method and other seven methods on CK+ database

Reference	Method	Verification method	Class	Accuracy(%)
[11] (2015)	PHOG-TOP+dense flow (handcrafted)	Leave-one-out	7	83.70
[59] (2017)	LBP-XT+LBP-YT+Gabor (handcrafted)	Leave-one-out	6	95.80
[7] (2018)	HOG-TOP+Geometric (handcrafted)	Leave-one-out	7	95.70
[5] (2019)	PCA-fusion (handcrafted)	Leave-one-out	6	95.97
[48] (2020)	LBP+LNEP (handcrafted)	10-fold	6	97.86
[53] (2017)	WMDNN (ML-based)	10-fold	6	97.02
[27] (2019)	Hierarchical DNN (ML-based)	10-fold	6	96.46
[38] (2019)	FAN (ML-based)	10-fold	7	99.70
Ours	Feature fusion (handcrafted)	Leave-one-out	6	99.35
	Feature fusion (handcrafted)	10-fold		98.71

Table 9 Comparison results of our proposed method and other four methods on Oulu-CASIA database

Research	Method	Accuracy (%)
[31] (2014)	STM-ExpLet (handcrafted)	74.59
[49] (2015)	Exemplar-HMM (handcrafted)	75.62
[59] (2017)	LBP-XT+LBP-YT+Gabor (handcrafted)	74.37
[5] (2019)	PCA-fusion (handcrafted)	79.99
[39] (2019)	dynamic MTL (ML-based)	89.60
Ours	Improved LBP-TOP (handcrafted)	75.63
	Gabor (handcrafted)	69.58
	Geometric (handcrafted)	74.79
	Feature fusion (handcrafted)	80.63

recognition rate of 99.35% on CK+ database. Our whole approach outperforms other well-known algorithms, which reveals its effectiveness and advancement in processing dynamic expression sequences.

In addition, to further demonstrate the reasonability of our proposed method, the Oulu-CASIA database is also used to provide quantitative comparisons with several methods in other papers. The comparison results are shown in Table 9. The performance of our proposed method on Oulu-CASIA database is inferior to CK+ database. In particular, the Gabor feature of peak frame performs poorly. The main reason for the poor accuracy is that the expression changes of some peak frames are relatively small and the resolution of original image is low. However, the proposed method still outperforms the other four methods based on hand-crafted features. According to Tables 8 and 9, although the best result is the machine learning based(ML-based) method, it can not be ignored that it also at the expense of a large amount of calculation and high hardware cost.

4.7 Evaluation of computational time

In the experiment, we examine the computational cost of our proposed method and LBP-TOP on CK+ and Oulu-CASIA databases. Table 10 shows the average computational time of feature extraction (Matlab platform) and classification (Python platform). Although the feature extraction time on CK+ database is longer than that of Oulu-CASIA database because of its high resolution, the classification time is faster. The feature extraction time

Table 10 Computational cost of our proposed method and LBP-TOP on CK+ and Oulu-CASIA databases

Descriptor	Feature extraction time (s)		Classification time (ms)	
	CK+	Oulu-CASIA	CK+	Oulu-CASIA
I-LBP-TOP	3.070	0.960	0.51	0.98
LBP-TOP	3.830	1.040	0.69	1.00
GMF	0.079	0.047	0.70	0.95
GF	0.074	0.053	0.44	0.79
ST-GF	0.465	0.496	0.31	0.54
Hybrid feature	3.614	1.503	1.61	2.48

Table 11 Comparing the influence of SVM Kernel function on individual descriptor

Kernel	I-LBP-TOP (%)	GMF (%)	ST-GF (%)
Polynomial	96.12	95.15	95.47
RBF	26.86	26.86	94.17
MKL	93.53	94.50	95.79

of I-LBP-TOP, LBP-TOP and ST-GF comes from each image sequence, so it is more time-consuming than single frame. Compared to LBP-TOP, our proposed I-LBP-TOP descriptor shorten the runtime effectively. Moreover, the proposed geometric descriptor takes the minimum amount time in both feature extraction and classification, showing the potential for real-time implementation. Significantly, GMF is calculated by using *construct_Gabor_filters_PhD.m* and *filter_image_with_Gabor_bank_PhD.m* function in PhD toolbox. Even if per image is amplified to 40 Gabor amplitude images, the calculation speed is still very fast in feature extraction stage. Based on the experimental results, although computational cost of fused method is close to the sum of multiple single features, its remarkable performance for FER cannot be ignored.

4.8 A selection of SVM kernel function

Ultimately, we evaluate the appropriate SVM kernel function for each individual descriptor, as shown in Table 11. Since our data is linearly inseparable, we only verify the effects of the polynomial kernel function, RBF kernel function, and MKL on classification accuracy, where MKL stands for multiple kernel learning and is a linear combination of polynomial and RBF kernel function.

Obviously, the choice of SVM kernel function plays a critical role in the performance of classification. We find that the use of a RBF kernel function is better than a polynomial kernel function in the case of fewer feature dimensions. On the contrary, the polynomial kernel function is more effective when the feature dimensions are large. RBF kernel function works better for ST-GF whose dimensions (from 98×6 reduced to 282) are less than sample numbers (309). However, when the original dimensions of I-LBP-TOP and Gabor features are much larger than 309, the polynomial kernel function performs better. In addition, the representation ability of MKL is not always optimal. If the performance of single kernel function is not poor, the effect of their combination may be enhanced. For instance, the polynomial kernel function and RBF kernel function have the accuracy of 95.47% and 94.17% on ST-GF, respectively, accordingly the MKL obtains the best result with the accuracy of 95.79%.

5 Conclusion

In this paper, we present an efficient facial expression recognition framework combing I-LBP-TOP, GMF and ST-GF at the decision-level for more accurate and competitive classification. The I-LBP-TOP descriptor can extract not only dynamic texture features, but also static texture features to characterize facial appearance changes. The adopted GMF can obtain the orientation and scale information effectively. In addition, the proposed method that directly utilizes the facial key points as geometric feature achieves simple calculation and high recognition accuracy. In the fusion strategy at decision-level, each descriptor has the same weight and the advantages are fully exerted to boost the performance of recognition.

Experiments performed on the CK+ and Oulu-CASIA facial expression databases confirm that the superiority of the proposed approach over other existing methods. Our proposed hybrid feature has reached the improved performance on 6 basic expressions with an average recognition rate of 99.35% on CK+ database and 80.63% on Oulu-CASIA database. Nonetheless, there is still room for further improvement in the accuracy of our algorithm in identifying fear and sadness. In the future, we will consider using data augmentation methods to increase the sample size of fear and sadness to solve the problem of sample imbalance, and develop more powerful structures to simulate the movement of facial critical areas from videos. Further, robust features need to be designed to resist head pose variation, occlusion, illumination effect in real-time environment.

Acknowledgements This research was supported by National Key R&D Program of China (2018YFC0831503), National Natural Science Foundation of China (61571275), Shenzhen Science and Technology Research and Development Funds (JCYJ20170818104011781).

Declarations

Conflict of Interests The authors declare that have no conflict of interest.

References

1. Acevedo D, Negri P, Buemi ME, Mejail M (2016) Facial expression recognition based on static and dynamic approaches. In: 2016 23rd International conference on pattern recognition (ICPR). IEEE, pp 4124–4129
2. Almaev TR, Valstar MF (2013) Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, pp 356–361
3. Awad AI, Hassaballah M (2016) Image Feature Detectors and Descriptors. Springer International Publishing
4. Bonab H, Can F (2019) Less is more: a comprehensive framework for the number of components of ensemble classifiers. IEEE Trans Neural Netw Learn Syst 30(9):2735–2745
5. Bougourzi F, Mokrani K, Ruichek Y, Dornaika F, Ouafi A, Taleb-Ahmed A (2019) Fusion of transformed shallow features for facial expression recognition. IET Image Process 13(9):1479–1489
6. Chen J, Chen Z, Chi Z, Fu H (2015) Dynamic texture and geometry features for facial expression recognition in video. In: IEEE international conference on image processing (ICIP), pp 4967–4971. IEEE
7. Chen J, Chen Z, Chi Z, Fu H (2018) Facial expression recognition in video with multiple feature fusion. IEEE Trans Affect Comput 9(1):38–50
8. Chen L, Wu M, Zhou M, She J, Dong F, Hirota K (2018) Information-driven multirobot behavior adaptation to emotional intention in human–robot interaction. IEEE Trans Cogn Dev Syst 10(3):647–658
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, vol 1, pp 886–893
10. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124
11. Fan X, Tjahjadi T (2015) A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. Pattern Recogn 48(11):3407–3416
12. Feng D, Ren F (2018) Dynamic facial expression recognition based on two-stream-cnn with lbp-top. In: 2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS). IEEE, pp 355–359
13. Gao T, Lei X-M, Hu W (2017) Face recognition based on sift and lbp algorithm for decision level information fusion. In: 2017 13th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD). IEEE, pp 2242–2246

14. Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors 13(6):7714–7734
15. Guo Z, Wang X, Zhou J, You J (2015) Robust texture image representation by scale selective local binary patterns. IEEE Trans Image Process 25(2):687–699
16. Guo Z, Zhang L, Zhang D (2010) A completed modeling of local binary pattern operator for texture classification. IEEE Trans Image Process 19(6):1657–1663
17. Guojiang W, Guoliang Y (2017) A modified optical flow algorithm and its application in facial expression recognition. In: 2017 3rd IEEE international conference on computer and communications (ICCC). IEEE, pp 1601–1605
18. Hassaballah M, Aly S (2015) Face recognition: challenges, achievements and future directions. IET Comput Vis 9:614–626
19. Hassaballah M, Bekhet S, Rashed AAM, Zhang G (2019) Facial features detection and localization. In: Recent advances in computer vision
20. Hassaballah M, Murakami K, Ido S (2011) An automatic eye detection method for gray intensity facial images. Int J Comput Sci Issues 8(4):272–282
21. Hassaballah M, Murakami K, Ido S (2013) Face detection evaluation: a new approach based on the golden ratio Φ, Signal Image & Video Processing
22. Hu M, Yang C, Zheng Y, Wang X, He L, Ren F (2019) Facial expression recognition based on fusion features of center-symmetric local signal magnitude pattern. IEEE Access 7:118435–118445
23. Huang X, Zhao G, Zheng W, Pietikäinen M (2012) Towards a dynamic expression recognition system under facial occlusion. Pattern Recogn Lett 33(16):2181–2191
24. Jan A, Meng H, Gaus YFBA, Zhang F (2017) Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Trans Cogn Dev Syst 10(3):668–680
25. Jeong M, Ko BC (2018) Driverą́s facial expression recognition in real-time for safe driving. Sensors 18(12):4270
26. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58(6):1233–1258
27. Kim J-H, Kim B-G, Roy PP, Jeong D-M (2019) Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. IEEE Access 7:41273–41285
28. Laurens VDM, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(2605):2579–2605
29. Li Y, Zou B, Deng S, Zhou G (2020) Using feature fusion strategies in continuous authentication on smartphones. IEEE Internet Comput 24:49–56
30. Liliana DY, Widyanto MR, Basaruddin T (2018) Geometric facial components feature extraction for facial expression recognition. In: 2018 International conference on advanced computer science and information systems (ICACSIS). IEEE, pp 391–396
31. Liu M, Shan S, Wang R, Chen X (2014) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1749–1756
32. Liu Y, Zhang X, Lin Y, Wang H (2019) Facial expression recognition via deep action units graph network based on psychological mechanism. IEEE Trans on Cogn Dev Syst
33. Liu Y, Zhang X, Lin Y, Wang H (2020) Facial expression recognition via deep action units graph network based on psychological mechanism. IEEE Trans Cogn Dev Syst 12(2):311–322
34. Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn 61:610–628
35. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, pp 94–101
36. Majumder A, Behera L, Subramanian VK (2018) Automatic facial expression recognition system using deep network based data fusion. IEEE Trans Cybern 48(1):103–114
37. Mehrabian A (2008) Communication without words. Commun theory:193–200
38. Meng D, Peng X, Wang K, Qiao Y (2019) Frame attention networks for facial expression recognition in videos. In: 2019 IEEE International Conference on Image Processing (ICIP)
39. Ming Z, Xia J, Luqman MM, Burie J-C, Zhao K (2019) Dynamic multi-task learning for face recognition with facial expression. arXiv:1911.03281
40. Nguyen VD, Nguyen DD, Nguyen TT, Dinh VQ, Jeon JW (2013) Support local pattern and its application to disparity improvement and texture classification. IEEE Trans Circuits Syst Vid Technol 24(2):263–276
41. Ning X, Duan P, Li W, Zhang S (2020) Real-time 3d face alignment using an encoder-decoder network with an efficient deconvolution layer. IEEE Signal Process Lett 27:1944–1948

42. Ning X, Xu S, Li W, Nie S (2020) Fegan: flexible and efficient face editing with pre-trained generator. IEEE Access 8:65340–65350
43. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
44. Rathee N, Vaish A, Gupta S (2017) Emotion detection through fusion of complementary facial features. In: 2017 7th International conference on communication systems and network technologies (CSNT). IEEE, pp 163–166
45. Sadeghi H, Raie A-A, Mohammadi M-R (2013) Facial expression recognition using geometric normalization and appearance representation
46. Sahoo S, Routray A (2016) Emotion recognition from audio-visual data using rule based decision level fusion. In: 2016 IEEE studentsąŕtechnology symposium (TechSym). IEEE, pp 7–12
47. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Compu 27(6):803–816
48. Shanthi P, Nickolas S (2020) An efficient automatic facial expression recognition using local neighborhood feature fusion. Multimed Tools Appl:1–26
49. Sikka K, Dhall A, Bartlett M (2015) Exemplar hidden markov models for classification of facial expressions in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 18–25
50. Tanfous AB, Drira H, Amor BB (2020) Sparse coding of shape trajectories for facial expression and action recognition. IEEE Trans Pattern Anal Mach Intell 42(10):2594–2607
51. Xie S, Shan S, Chen X, Chen J (2010) Fusing local patterns of gabor magnitude and phase for face recognition. IEEE Trans Image Process 19(5):1349–1361
52. Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 532–539
53. Yang B, Cao J, Ni R, Zhang Y (2017) Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access 6:4630–4640
54. Zhang Q, Li H, Sun Z, Tan T (2018) Deep feature fusion for iris and periocular biometrics on mobile devices. IEEE Trans Inf Forensics Secur 13:2897–2912
55. Zhang B, Shan S, Chen X, Gao W (2006) Histogram of gabor phase patterns (hgpp): a novel object representation approach for face recognition. IEEE Trans Image Process 16(1):57–68
56. Zhao G, Huang X, Taini M, Li SZ, PietikäInen M. (2011) Facial expression recognition from near-infrared videos. Image Vis Comput 29(9):607–619
57. Zhao G, Pietikäinen M (2006) Dynamic texture recognition using volume local binary patterns. In: Dynamical vision. Springer, pp 165–177
58. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
59. Zhao L, Wang Z, Zhang G (2017) Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram. Math Probl Eng, vol 2017
60. Zheng Q, Tian X, Yang M, Su H (2019) The email author identification system based on support vector machine (svm) and analytic hierarchy process (ahp). IAENG Int J Comput Sci 46(2):178–191
61. Zheng Q, Tian X, Yang M, Wu Y, Su H (2019) Pac-bayesian framework based drop-path method for 2d discriminative convolutional network pruning. Multidim Syst Sign Process:1–35

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Ruyu Yan[1] **· Mingqiang Yang**[1] **· Qinghe Zheng**[1] **· Deqiang Wang**[1] **· Cheng Peng**[1]

✉ Deqiang Wang
wdq_sdu@sdu.edu.cn

Ruyu Yan
yan17860779713@163.com

Qinghe Zheng
15005414319@163.com

Cheng Peng
chengpeng8169@163.com

[1] School of Information Science and Engineering, Shandong University, Qingdao, Shandong 266237, China