Taylor & Francis
Taylor & Francis Group

IETE Technical Review

ISSN: 0256-4602 (Print) 0974-5971 (Online) Journal homepage: http://www.tandfonline.com/loi/titr20

A Review on Facial Expression Recognition: Feature Extraction and Classification

Xiaoming Zhao & Shiqing Zhang

To cite this article: Xiaoming Zhao & Shiqing Zhang (2016): A Review on Facial Expression Recognition: Feature Extraction and Classification, IETE Technical Review, DOI: 10.1080/02564602.2015.1117403

To link to this article: http://dx.doi.org/10.1080/02564602.2015.1117403

Published online: 13 Jan 2016.

IETE TECHNICAL REVIEW, 2016
http://dx.doi.org/10.1080/02564602.2015.1117403

Taylor & Francis
Taylor & Francis Group

A Review on Facial Expression Recognition: Feature Extraction and Classification

Xiaoming Zhao and Shiqing Zhang

Institute of Image Processing and Pattern Recognition, Taizhou University, Taizhou, Zhejiang, China

ABSTRACT
Facial expression recognition (FER) is currently a very active research topic in the fields of computer vision, pattern recognition, artificial intelligence, and has drawn extensive attentions owing to its potential applications to natural human–computer interaction (HCI), human emotion analysis, interactive video, image indexing and retrieval, etc. This paper is a survey of FER addressing the most two important aspects of designing an FER system. The first one is facial feature extraction for static images and dynamic image sequences. The second one is facial expression classification. Conclusions and future work are finally discussed in the last section of this survey.

KEYWORDS
Dynamic image sequences; facial expression; facial expression classification; facial feature extraction; human–computer interaction; static images

1. Introduction

Facial expression is an important mode of expressing and interpreting emotional states and mental states of human beings. In early psychology, Mehrabian [1] has found that only 7% of the whole information human expresses is conveyed through language, 38% through speech, and 55% through facial expression. Therefore, through facial expression a large amount of valuable information can be obtained so as to detect human beings' consciousness and mental activities. Facial expression recognition (FER) aims to develop an automatic, efficient, accurate system to distinguish facial expression of human beings so that human emotions can be understood through facial expression, such as happiness, sadness, anger, fear, surprise, disgust, etc. During the last two decades, automatic FER has attracted growing attentions in many fields such as computer vision, pattern recognition, and artificial intelligence, owing to its potential applications to natural human–computer interaction (HCI), human emotion analysis, interactive video, image indexing and retrieval, etc.

Generally, a basic FER system is comprised of two major steps: facial feature extraction, and facial expression classification. As a result, this paper only focuses on giving recent advances on these two steps, i.e. facial feature extraction, and facial expression classification on FER tasks. Although some previous work [2–4] on reviewing FER exist in the last decades, this paper aims to present a recent advance, especially from 2013 to 2015, on FER. Additionally, we performed FER experiment to present a comparative analysis of different classification methods. Finally, we identify several challenges in this area and put forward recommendations for future research.

The rest of this paper is organized as follows. In Section 2, facial feature extraction methods are introduced. In Section 3, facial expression classification methods are reviewed. In Section 4, performance comparison with experimental results are provided. Finally, conclusions and discussion are presented in Section 5.

2. Facial feature extraction

Facial feature extraction is to extract facial features from the input face images to effectively represent facial expression. According to different types of input images, facial feature extraction methods can be divided into two categories. One is these facial feature extraction methods for static images without variations. The other is these facial feature extraction methods for dynamic image sequences.

2.1. *Facial feature extraction methods for static images*

For static images, there are two types of facial feature extraction methods: geometric feature-based methods and appearance-based methods.

2.1.1. *Geometric feature-based methods*

It is known that a face is composed of eyebrows, eyes, brows, nose, mouth, chin, and so on. These organs' size, shape, direction, and position affect the generation of facial expression. Therefore, geometric features are able to depict the shape and locations of facial components such as mouth, nose, eyes and brows. The main purpose of geometric feature-based methods is to use the geometric relationships between facial feature points to extract facial features. However, extracting geometric features usually requests an accurate feature point detection technique. This is difficult to implement it in real-world complex background. In addition, geometric feature-based methods easily ignore the changes in skin texture such as wrinkles and furrows that are important for facial expression modelling.

There are three typical types of geometric feature-based extraction methods: active shape models (ASM), active appearance models (AAM), as well as scale-invariant feature transform (SIFT). The details for ASM, AAM, and SIFT are described below.

ASM: Active shape model (ASM) proposed by Cootes et al. [5] is a feature-matching method based on a statistical model. An ASM is comprised of a point-distribution model (PDM) learning the changes of valid shapes, and a number of flexible models capturing the grey levels around a number of landmark feature points. Figure 1 shows an example with the ASM feature extraction method in [6], defined by 58 facial landmark feature points. The ASM method includes two steps. First, shape models are built from the training samples with some annotated landmark feature points. Then, local texture models for each landmark feature point are also built. Second, according to the two building models, an iterative search procedure to deform the model example can be finished. Shbib and Zhou [7] used the geometric displacement among the projected ASM feature point coordinates and the mean shape of ASM as facial features for FER. In recent years, Anderson et al. [8] presented an enhanced version of ASM called active shape and statistical models (ASSM) for face recognition, which has potential applications for FER.

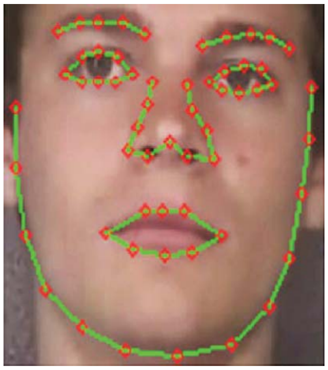
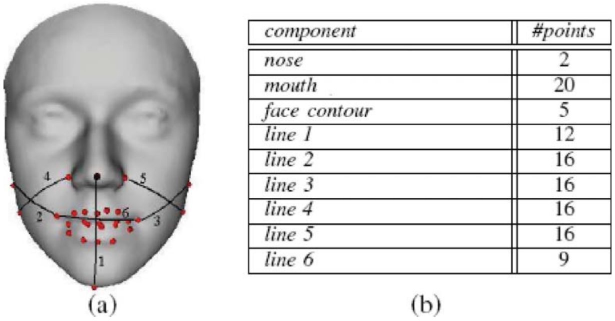


Figure 1: The used ASM feature extraction method in [6] based on 58 facial landmark feature points.

AAM: Active appearance model (AAM) was developed by Cootes et al. [9] in 2001. AAM essentially extends ASM by capturing the shape and texture information jointly. In detail, AAM first builds a statistical model based on training data for statistical analysis, and then employ this statistical model to implement fitting calculation for testing data. Different with ASM, AAM not only takes advantage of the global shape and texture information, but also conducts statistical analysis on local texture information so as to find out the relationships between shape and texture information. Cheon and Kim [10] presented an FER method by using differential-AAM and manifold learning. First, the difference of AAM parameters between the input images and the reference images (such as neutral expression images) is calculated to extract the differential-AAM features (DAFs). Second, manifold learning methods are used to embed the DAFs on the smooth and continuous feature space. Finally, the input facial expression is identified. Recently, several advanced versions of AAM have been also developed, such as histogram of oriented gradient (HOG)-based AAM [11], dense-based AAM [12], regression-based AAM [13]. It is an interesting task to investigate the performance of these recently-developed AAM variants on FER.

SIFT: Scale-invariant feature transform (SIFT) is a local image descriptor for image-based matching proposed by David Lowe [14,15]. The SIFT features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or three-dimensional (3D) projection. Figure 2 gives an example of the SIFT feature extraction method used in Berretti et al.[16], in which they took facial landmarks located in important morphological regions of the face as key points, and then the SIFT feature extractor was implemented on these located key points in order to obtain the SIFT descriptor. Recently, Soyel and Demirel [17] gave a discrimination of scale-invariant feature



component	*#points*
nose	2
mouth	20
face contour	5
line 1	12
line 2	16
line 3	16
line 4	16
line 5	16
line 6	9

Figure 2: The used SIFT feature extraction method in [16]. (a) A sample image, and the lines along which the 85 additional landmarks were located. (b) Number of landmarks for every face component they belong to.

transform (D-SIFT) method, which can effectively make decisions on the overall appearance features. Li et al. [18] presented a new scale-invariant feature transform called GA-SIFT for multispectral image using geometric algebra (GA). At first, based on the theory of the GA, a novel representation of multispectral images with spectral and spatial information was presented. Second, finding the scale space of a multispectral image was given. Third, similar to SIFT, GA-based difference of Gaussian images were obtained. Finally, the feature points can be detected and described based on the theory of GA.

2.1.2. Appearance-based methods

Appearance-based methods aim to use the whole-face or specific regions in a face image to reflect the underlying information in a face image, especially the subtle changes of the face, such as wrinkles and furrows. So far, there are mainly two representative appearance-based feature extraction methods, i.e. local binary patterns (LBP) [19] and Gabor wavelet representation

LBP: Local binary pattern (LBP) [19] is an effective texture description operator, which can be used to measure and extract the adjacent texture information in an image. The advantage of using the LBP operator is that the LBP operator has a good rotation invariance and grey invariance, and overcomes the problems of disequilibrium displacement, rotation and illumination in an image. Moreover, the LBP operator has a relatively simple calculation. Figure 3 shows an example of the LBP feature extraction for FER, as described in [20]. The used LBP feature extraction method in [20] contains three crucial steps. At first, a facial image was divided into various non-overlapping blocks. Second, LBP histograms were worked out for each block. Third, the block LBP histograms were concatenated into a single vector represented by the LBP code. In our previous works [21,22], we investigated the performance of the LBP operator with dimensionality reduction methods such as local fisher discriminant analysis on FER tasks. In recent years, some variants of the LBP operator can be found in the literature [23]. Till now the typical LBP variants contain volume local binary patterns (VLBP) [24], LBP on three orthogonal planes (LBP-TOP) [24], local directional patterns (LDP) [25], local transitional patterns (LTP) [26], and so on. Recently, Li et al. [27] has proposed the polytypic multi-block local binary patterns (P-MLBP) for fully automatic 3D FER. The P-MLBP involves both the feature-based irregular divisions to accurately represent the facial expression, and integrate the depth and texture information of 3D models to enhance facial features.

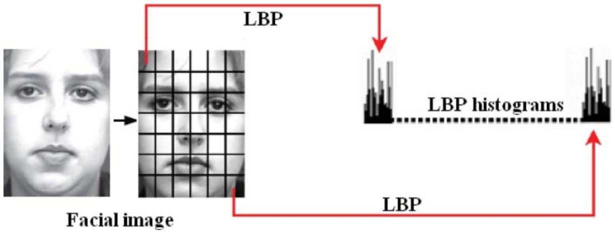


Figure 3: The used LBP feature extraction method in [20].

Gabor: Gabor wavelet representation [28] is a classical method to extract facial expression features. In detail, an image is filtered by a set of filters, and the filtered results can reflect the relationship (gradient, texture correlation, etc.) between local pixels. Gabor wavelet representation method has been widely used for facial expression feature extraction. It is able to detect multi-scale, multi-direction changes of texture, and has a little impact on illumination changes. Figure 4 presents an example of Gabor wavelet representation used in [28], in which the total 18 Gabor kernels at three scales and six orientations were employed. Liu et al. [29] proposed an FER method based on Gabor wavelet features and kernel principal component analysis (KPCA). In this scheme, they used a local Gabor filter to replace the traditional Gabor filter, resulting in the fact it can speed up the computation speed. Gu et al. [30] performed FER by using the radial encoding of local Gabor features and classifier synthesis. In this study, the input images were first subjected to local, multi-scale Gabor-filter operations, and then the resulting Gabor decompositions were used to be encoded with radial grids. Recently, Owusu et al. [31] presented a neural-AdaBoost-based FER system in which Gabor feature extraction techniques were used to extract a large number of facial features representing various facial deformation patterns.

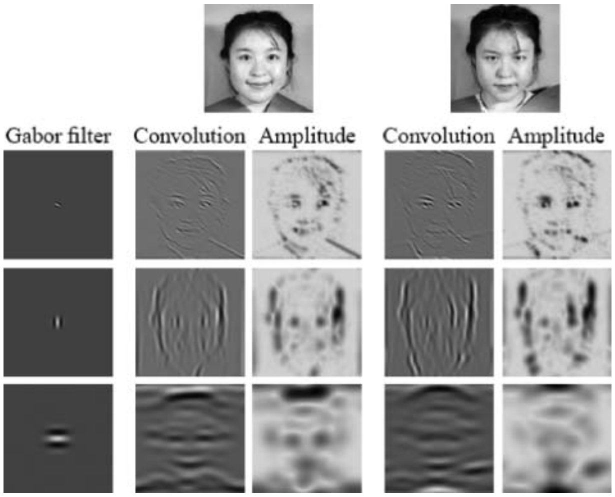


Figure 4: The Gabor wavelet representation method used in [28].

2.2. Facial feature extraction methods for dynamic image sequences

Dynamic image sequences reflect the continuous process of facial expression movements. Facial expression features for dynamic image sequences are mainly represented by deformation and facial muscle movements. At present, two popular feature extraction methods for dynamic image sequences are given as follows: optical flow, and feature point tracking.

2.2.1. Optical flow

Negahdaripour [32] redefines the optical flow method as geometry and radiation changes of dynamic images. The basic principle of the optical flow method is that each pixel in an image is assigned to a velocity vector. These velocity vectors form a motion field for an image. In a motion moment, the image point corresponds to the actual object point. The optical flow method has an obvious advantage, that is, the optical flow not only carries the motion information of the target, but also has the rich information about the 3D structure of the target. In the field of FER, the optical flow method is widely used to extract facial expression features from dynamic image sequences since it highlights facial deformation and reflects the motion trend of image sequences. Figure 5 shows an example of the optical flow feature extraction method used in [33], which was performed on two facial expression sequences. Lien [33] analysed holistic face motion by means of wavelet-based multi-resolution dense optical flow, and then figured out PCA-based-eigenflows both in horizontal and vertical directions for a compacter representation of the resulting flow fields.

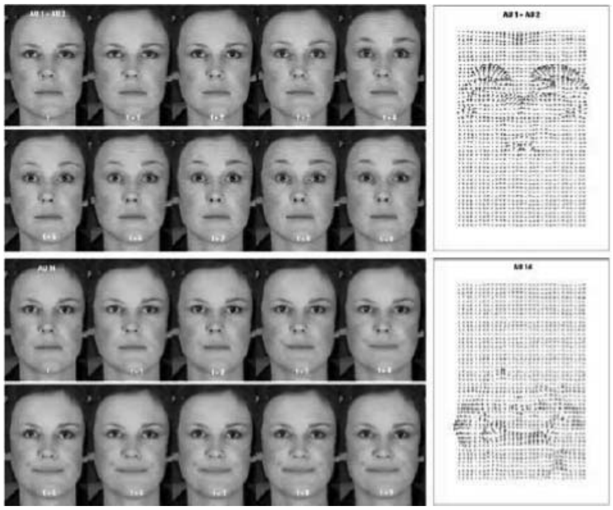


Figure 5: The optical flow method used in [33] in which on the left-hand side two sample facial expression sequences are presented and on the right-hand side the corresponding optical flow images are given.

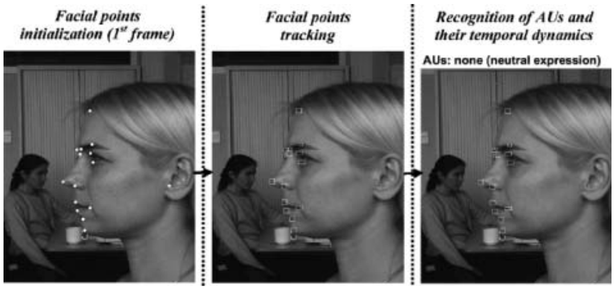

Facial points initialization (1st frame)
Facial points tracking
Recognition of AUs and their temporal dynamics
AUs: none (neutral expression)

Figure 6: The feature point tracking method used in [36].

Yacoob et al. [34] used the optical flow field and the gradient field between successive frames to represent the temporal and spatial variations of images. Then, according to the changes of the motion vectors of features, the facial muscle movements were calculated to classify different expression. Sánchez et al. [35] compared systematically two optical flow-based methods for FER. One was featural and aimed to select a reduced set of highly discriminant facial points. The other was holistic and used much more points that were uniformly distributed on the central face region.

2.2.2. Feature point tracking

The feature point tracking methods often select some feature points with large changes in the corners of eyes and mouth. Then, following these points will be able to get facial feature displacement or deformation information.

Figure 6 presents an example of the feature point tracking method used in Pantic et al.[36], in which 15 feature points were selected based on facial action coding system (FACS), and then the particle filter was used to track the movements of feature points in image sequences. Tie et al. [37] proposed a method which could automatically extract 26 reference points in a facial model from video sequences, and tracked the reference points through a number of particle filters. Fang et al. [38] used the salient facial point tracking method to extract salient information from video sequences but did not rely on any subjective preprocessing or additional user-supplied information to select frames with peak expression.

3. Facial expression classification

Facial expression classification aims to design an appropriate classification mechanism to identify facial expression. The representative classification methods for FER include hidden Markov model (HMM), artificial neural network (ANN), Bayesian network (BN), *K*-nearest neighbour (KNN), support vector machines (SVM), sparse representation-based classification (SRC), and so on.

3.1. Hidden Markov model

HMM is a Markov process with hidden unknown parameters, and can be used to describe the random signal information of a statistical model. HMM consists of two interrelated processes. One is the underlying and unobservable Markov chain with a certain number of states. The other is a set of probability density distribution corresponding to each state. An HMM can be defined by the following triplet:

$$\lambda = (A,\ B,\ \pi), \tag{1}$$

where A is the state transition probability matrix, B is the observation probability distribution, and π is the initial state distribution. In a discrete density HMM, B represents a matrix of probability entries, and in a continuous density HMM, B is denoted by the parameters of the probability distribution function of observations, such as the Gaussian distribution or a mixture of Gaussians. The widely-used general representation of the model probability density function (pdf) is defined as the following finite mixture form:

$$b_i(O) = \sum_{k=1}^{M} c_{ik} N(O, \mu_{ik}, U_{ik}), 1 \le i \le N, \tag{2}$$

where M denotes different observation symbols, N represents the number of states in the HMM model, c_{ik} is the mixture parameter for the kth mixture for state i, $N(O, \mu_{ik}, U_{ik})$ is the Gaussian pdf with the average vector μ_{ik} and the covariance matrix U_{ik}.

Aleksic and Katsaggelos [39] presented an automatic multistream HMM-based FER system, in which the multistream HMM method was used for introducing facial expression and facial animation parameters (FAPs) group dependent stream reliability weights. Sun and Akansu [40] proposed a regional hidden Markov model (RHMM) method for automatic FER in video sequences. They used RHMMs to describe facial action units for the states of facial regions: eyebrows, eyes and mouth registered in a video.

3.2. Artificial neural network

ANN is a flexible mathematical structure which is able to distinguish complex nonlinear relationships between input data and output data. In recent years, there are mainly two types of ANN used for FER, i.e. the multilayer perceptron (MLP) network, and the radial basis function neural network (RBFNN). Since these two types of ANN, i.e. RBFNN and MLP, are very similar, the basic idea of RBFNN is just given below.

The well-known RBFNN for classification is a three-layer feedforward network containing one input layer, one hidden layer as well as one output layer. Figure 7 gives an example of the RBFNN framework. For input data, every input neuron of the input layer associates with a component of an input vector x. The hidden layer aims to cluster input data and derive features. The hidden layer contains n neurons and one bias neuron. The widely-used Gaussian radial basis function (RBF) for the hidden layer is defined as

$$y_i = \begin{cases} \exp\left(-\dfrac{\|x - p_i\|}{2\sigma_i^2} \right), & i = 1, 2, \ldots, n \\ 1, & i = 0 \end{cases}, \tag{3}$$

in which p_i and σ_i separately denotes the centre and the width of the neuron, and the symbol $\|\ \|$ represents the Euclidean distance. The weight vector between the input layer and the ith hidden layer neuron corresponds to the centre p_i. The closer x is to p_i, the higher the value of the Gaussian function can give.

The output layer is comprised of m neurons corresponding to the possible categories of the problems. Every output layer neuron is fully connected to the hidden layer and computes a linear weighted sum of the outputs of the hidden neurons by using the following form:

$$z_j = \sum_{i=0}^{n} y_i w_{ij}, \quad j = 1, 2, \ldots, m, \tag{4}$$

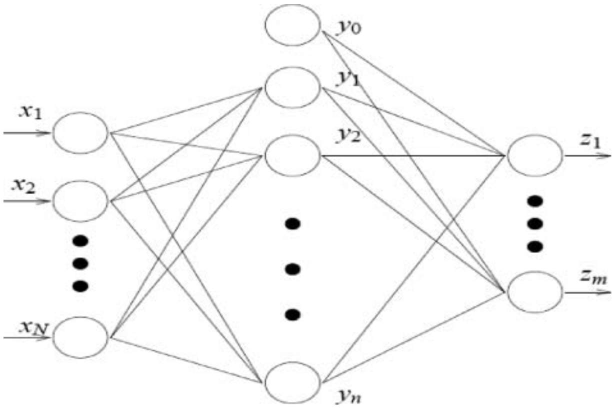


Figure 7: The RBFNN framework.

where w_{ij} represents the weight between the ith hidden layer neuron and the jth output layer neuron.

Ma and Khorasani [41] developed a new technique for FER, which employed the 2D discrete cosine transform (DCT) over the entire face image as a feature detector and a constructive one-hidden-layer feedforward neural network as a facial expression classifier. The obtained best recognition rates came up to 100% and 93.75% (without rejection), for the training and generalizing images, respectively. De Silva et al. [42] presented a modified version of RBFs called cloud basis functions (CBFs) for holistic recognition of six universal facial expression from static images. The CBF neural network method gave the best recognition accuracy of 96.1%. Recently, Kaburlasos et al. [43] described a fundamentally novel extension, namely, flrFAM, of the fuzzy ARTMAP (FAM) neural classifier as a fuzzy neural network for FER.

3.3. Bayesian network

BN is a graphical network based on the probabilistic reasoning. The so-called probabilistic reasoning uses some variable information to obtain the other probability information. BN based on the probabilistic reasoning is developed to solve the uncertainty and incompleteness problem. A BN classifier represents the dependencies among feature data and sample labels by using a directed acyclic graph. This graph is the basic structure of the BN. Generally, BN classifiers can be learned by using a fixed structure – the well-known example is the naive-Bayes classifier.

Given a BN classifier with parameter set θ, the optimizing classification rule based on the maximum likelihood (ML) idea to classify an observed feature vector $x \in R^n$ with n dimension, to one of $|C|$ class labels, $c \in \{1, 2, \cdots, |C|\}$, is denoted by:

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(x \mid c; \theta). \tag{5}$$

There are two types of decision rules when designing BN classifiers. The first one is to select the structure of the network, which is used to determine the dependencies among all the variables in the graph. The second one is to determine the distribution of feature data.

Cohen et al. [44] employed different BN classifiers for identifying facial expression from video, focusing on changes in distribution assumptions, and feature dependency structures. In particular, they used Naive BN classifiers and changed the distribution from Gaussian to Cauchy, and employed Gaussian tree-augmented naive Bayesian (TAN) classifiers to learn the dependencies among different facial motion features. Zhao et al. [45] present a unified probabilistic framework based on a novel Bayesian belief network (BBN) for 3D facial expression and action unit (AU) recognition. The proposed BBN performs Bayesian inference based on statistical feature models (SFMs) and Gibbs–Boltzmann distribution and featured a hybrid approach in fusing both geometric and appearance features along with morphological ones.

3.4. K-nearest neighbour

KNN is a type of instance-based learning classification algorithm. The principle of the KNN method is that in the feature space one sample has k closest samples, and its label is assigned to the class most common among its KNNs by using a majority vote of its neighbours. Without prior knowledge, the KNN classification algorithm frequently employs the Euclidean distance as the distance metric. Given two vector $x = (x_1, x_2, \cdots, x_m)$ and $y = (y_1, y_2, \cdots, y_m)$, their Euclidean distance is represented as

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \tag{6}$$

Sebe et al. [46] used the geometric features to obtain the best classification accuracy of 93% on the Cohn–Kanade database with the KNN method. Gu et al. [30] presented an FER method by using radial encoding of local Gabor features and classifier synthesis based on the KNN method with $k = 1$.

3.5. Support vector machine

SVM are developed based on the structural risk minimization principle, which has been shown to be superior to the traditional empirical risk minimization principle used by conventional neural networks. The principle of SVM is to transform the input vectors to a higher dimensional space by a nonlinear transform, and then an optimal hyperplane which separates the data can be found.

Given the training data set $(x_1, y_1), \cdots, (x_l, y_l), y_i \in \{-1, 1\}$, to seek the optimal hyperplane, a nonlinear transform, $Z = \Phi(x)$, is utilized to make training data be linearly dividable. A weight w and offset b is decided by the following criteria:

$$\begin{cases} w^T z_i + b \geq 1, & y_i = 1 \\ w^T z_i + b \leq -1, & y_i = -1 \end{cases}. \tag{7}$$

The above procedure is finished by using the following optimizing problem:

$$\begin{aligned}&\min_{w,b}\Phi(w)=\frac{1}{2}(w^Tw)\\&s.t.\ \ y_i(w^Tz_i+b)\geq 1,\quad i=1,2,\cdots,n\,.\end{aligned}\tag{8}$$

By means of using the Lagrange method, the decision function can be written as

$$f=\operatorname{sgn}\left[\sum_{i=0}^{l}\lambda_i y_i(z^Tz_i)+b\right].\tag{9}$$

From the kernel idea, a non-negative symmetrical function $K(u,v)$ uniquely denotes a Hilbert space H:

$$K(u,v)=\sum_i \alpha\varphi_i(u)\varphi_i(v),\tag{10}$$

where K is the kernel function in the space H. This represents an internal product in the Hilbert space H:

$${z_i}^Tz=\Phi(x_i)^T\Phi(x)=K(x_i,x).\tag{11}$$

Then the decision function could be rewritten as

$$f=\operatorname{sgn}\left[\sum_{i=1}^{l}\lambda_i y_i K(x_i,x)+b\right].\tag{12}$$

There are four typical kernel functions for the used SVM model, such as the linear kernel, the polynomial kernel, the RBF kernel, and the sigmoid kernel, which are described below.

The linear kernel function is given as

$$K(x_i,x_j)={x_i}^Tx_j.\tag{13}$$

The polynomial kernel function is given as

$$K(x_i,x_j)=(\gamma {x_i}^Tx_j+\text{coefficient})^{\text{degree}}.\tag{14}$$

The RBF kernel function is given as

$$K(x_i,x_j)=\exp(-\gamma\,|\,x_i-x_j\,|^2.\tag{15}$$

The sigmoid kernel function is given as

$$K(x_i,x_j)=\tanh(\gamma {x_i}^Tx_j+\text{coefficient}).\tag{16}$$

Yurtkan and Demirel [47] used the SVM classifier to develop a feature selection system for improved FER utilizing 3D geometrical facial feature point positions. Ghimire and Lee [48] presented a method of geometric feature-based FER in facial image sequences and reported an accuracy of 97.35% using the SVM classifier on the Cohn–Kanade database.

3.6. Sparse representation-based classification

SRC [49] is developed based on compressed sensing (CS) [50]. The principle of the SRC method is based on an assumption that the whole set of training samples are used to constitute a dictionary, and then the classification problem is regarded as one of discriminatively seeking a sparse representation of the test sample as a linear combination of training samples by solving the l_1 -norm optimization problem.

Formally, for the training samples of a single class, this assumption could be formulated as

$$\begin{aligned}y_{k,\text{test}}&=\alpha_{k,1}y_{k,1}+\alpha_{k,2}y_{k,2}+\cdots+\alpha_{k,n_k}y_{k,n_k}+\varepsilon_k\\&=\sum_{i=1}^{n_k}\alpha_{k,i}y_{k,i}+\varepsilon_k,\end{aligned}\tag{17}$$

where $y_{k,\text{test}}$ represents the test sample of the kth class, $y_{k,i}$ denotes the ith training sample of the kth class, $\alpha_{k,i}$ represents the weight corresponding weight and ε_k denotes the approximation error.

For the training samples from all c object classes, the aforementioned Equation (17) can be reformulated as

$$\begin{aligned}y_{k,\text{test}}=&\ \alpha_{1,1}y_{1,1}+\cdots+\alpha_{k,1}y_{k,1}+\cdots\\&+\alpha_{k,n_k}y_{k,n_k}+\cdots+\alpha_{c,n_c}y_{c,n_c}+\varepsilon.\end{aligned}\tag{18}$$

In a matrix form, that is,

$$y_{k,\text{test}}=\mathbf{A}\boldsymbol{\alpha}+\varepsilon,\tag{19}$$

where

$$\begin{cases}\mathbf{A}=[y_{1,1}\,|\cdots|\,y_{1,n_1}\,|\cdots|\,y_{k,1}\,|\cdots|\,y_{k,n_k}\,|\cdots|\,y_{c,1}\,|\cdots|\,y_{c,n_c}]\\ \boldsymbol{\alpha}=[\alpha_{1,1}\cdots\alpha_{1,n_1}\cdots\alpha_{k,1}\cdots\alpha_{k,n_k}\cdots\alpha_{c,1}\cdots\alpha_{c,n_c}]'\end{cases}.$$

To achieve the weight vector $\boldsymbol{\alpha}$, the following l_1 -norm minimization problem needs to be solved:

$$\min_{\alpha}\|\boldsymbol{\alpha}\|_1,\quad\text{subject to}\quad\|y_{k,\text{test}}-\mathbf{A}\boldsymbol{\alpha}\|_2\leq\varepsilon.\tag{20}$$

This is a convex optimization problem and could be solved by the quadratic programming method. When a sparse solution of $\boldsymbol{\alpha}$ is presented, the classification procedure of SRC is summarized below:

Step 1: Solve the l_1-norm minimization Equation (20) problem.

Step 2: For every class i, work out the residuals between the reconstructed sample $y_{\text{recons}}(i) = \sum_{j=1}^{n_i} \alpha_{i,j} y_{i,j}$ and the given test sample by $r(y_{\text{test}}, i) = \|y_{k,\text{test}} - y_{\text{recons}}(i)\|_2$.

Step 3: The class label of the given test sample is decided by using the rule: identify $(y_{\text{test}}) = \arg\min_i r(y_{\text{test}}, i)$.

In our previous works [51,52], the performance of SRC was investigated when classifying clean or occluded facial expression images. They found that SRC had better performance and greater robustness when compared with the nearest neighbour (NN), the nearest subspace (NS), and SVM. Mohammadi et al. [53] presented a PCA-based dictionary building for sparse representation and classification of universal facial expression. In detail, expressive facials images of each subject were first subtracted from a neutral facial image of the same subject. Then the PCA method was applied to these difference images to model the variations within each class of facial expression. The learned principal components were employed as the atoms of the dictionary. Finally, for classification a given test image was sparsely represented as a linear combination of the principal components of six basic facial expression. Ouyang et al. [54] recently developed an accurate and robust FER by fusing multiple sparse representation-based classifiers, i.e. combining HOG+SRC and LBP+SRC.

4. Performance comparison

To verify the performance of different classifiers on FER, we performed FER experiments on two popular databases, i.e. the JAFFE database [55] and the Cohn−Kanade database [56]. These two databases contains seven facial expression, i.e. anger, joy, sadness, neutral, surprise, disgust and fear.

The JAFFE database has 213 images of female facial expression. Each image has a resolution of 256 × 256 pixels. Some sample images are shown in Figure 8. The Cohn−Kanade database contains 100 university students. Each image has a resolution of 640×490 pixels. Figure 9 presents some sample images from the Cohn−Kanade database. As done in [21,57], on the Cohn−Kanade database we selected 320 image sequences from 96 subjects, with 1 to 6 emotions per subject. For every sequence, the neutral face and one peak frames were employed for prototypic expression recognition, giving in total 470 images (32 anger, 100 joy, 55 sadness, 75 surprise, 47 fear, 45 disgust and 116 neutral).

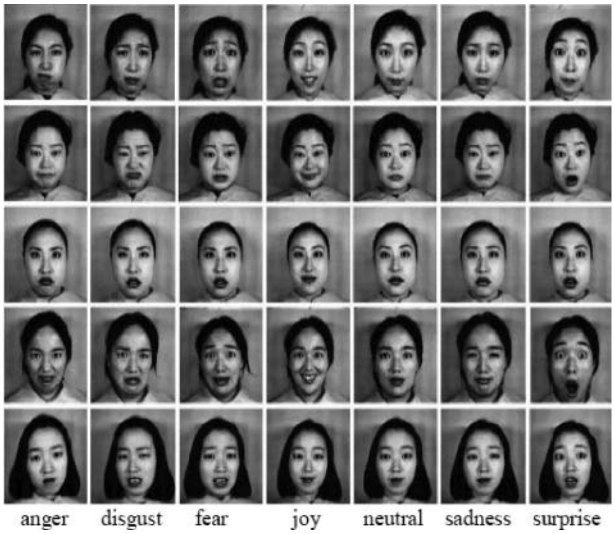


Figure 8: Sample images from the JAFFE database.

For facial feature representation, the LBP [19] features were extracted due to its computation simplicity and promising performance. According to the normalized value of the eye distance, a resized image of 110 × 150 pixels was cropped from original images. Similar to the settings in [21,57], we employed the 59-bin operatorLBP$^{u2}_{P,R}$, and divided the cropped images of 110 × 150 pixels into 18 × 21 pixels regions, yielding a feature vector length of 2478 (59 × 42) represented by the LBP histograms. A 10-fold cross-validation scheme is implemented in seven-class FER experiments, and the average recognition results are reported.

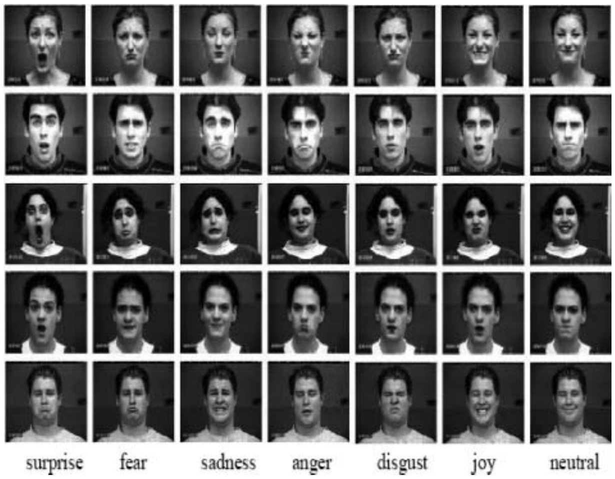


Figure 9: Sample images from the Cohn−Kanade database.

Table 1: **Performance comparison of different classification methods with the LBP features on the JAFFE database.**

Methods	Accuracy (%)
HMM	78.64
Naive-Bayes	70.57
ANN	68.09
KNN	80.95
SVM	79.88
SRC	84.76

Table 2: **Performance comparison of different classification methods with the LBP features on the Cohn−Kanade database.**

Methods	Accuracy (%)
HMM	94.76
Naive-Bayes	93.81
ANN	93.45
KNN	96.22
SVM	95.24
SRC	97.14

For HMM, we employed a seven-state discrete HMM model with the left−right structure, in which each state corresponded to one facial expression. For ANN, RBFNN with a three-layer feedforward network containing one input layer, one hidden layer as well as one output layer, is used for its computational simplicity. The number of hidden layer neurons in RBFNN is set to be the number of training samples. The goal of training error is 0.0001. For the BN, we used the naive-Bayes classifier. For KNN, we set K to be 1 for its satisfying performance. For SVM, we used the LIBSVM package, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, to perform the SVM algorithm with the linear kernel function, one-against-one for multi-class problems. The experiment platform is Intel CPU 2.10 GHz, 1G RAM memory, MATLAB 2012a.

Tables 1 and 2 separately present the recognition results of six different classification methods, including HMM, naive-Bayes, ANN, KNN, SVM, SRC on the JAFFE database and the Cohn−Kanade database, As shown in Tables 1 and 2, it can be seen that SRC obtains the best performance on FER tasks, followed by KNN, SVM, HMM, naive-Bayes, ANN. This confirms the validity and high performance of SRC for FER. To further show the detailed accuracy of each expression, Tables 3 and 4 show the confusion matrix of recognition results of SRC on the JAFFE database and the Cohn−Kanade database, respectively.

Table 3: **Confusion matrix of recognition results of SRC with LBP features on the JAFFE database.**

	Anger (%)	Joy (%)	Sadness (%)	Surprise (%)	Disgust (%)	Fear (%)	Neutral (%)
Anger	93.33	0	6.67	0	0	0	0
Joy	0	100	0	0	0	0	0
Sad	3.22	3.22	74.19	3.22	3.22	6.48	6.45
Surprise	0	3.45	3.45	82.76	0	10.34	0
Disgust	10.35	0	6.89	0	82.76	0	0
Fear	0	0	12.52	3.12	9.37	71.87	3.12
Neutral	3.45	0	0	6.89	0	0	89.66

Table 4: **Confusion matrix of recognition results of SRC with LBP features on the Cohn−Kanade database.**

	Anger (%)	Joy (%)	Sadness (%)	Surprise (%)	Disgust (%)	Fear (%)	Neutral (%)
Anger	90	0	0	0	0	0	10
Joy	0	100	0	0	0	0	0
Sad	3.33	0	90	0	0	0	6.67
Surprise	0	0	0	100	0	0	0
Disgust	0	0	0	0	100	0	0
Fear	0	0	0	0	0	100	0
Neutral	0	0	0	0	0	0	100

5. Conclusions and discussion

Over the last decade, increasing attentions have been directed to the study of FER. This paper presented a review of FER addressing facial feature extraction and facial expression classification, which are the most two important steps in an FER system. For facial feature extraction methods, geometric feature-based methods and appearance-based methods, used for static images, were first reviewed. Then, these facial feature extraction methods for dynamic image sequences, including optical flow and feature point tracking, were also investigated. For facial expression classification, six typical classification methods including HMM, ANN, BN, KNN, SVM and SRC, were also surveyed. In addition, we performed FER experiments on the JAFFE database and the Cohn−Kanade database, and presented a comparative study of different classification methods based on the extracted LBP features. Experiment results show that SRC outperforms the other used methods such as KNN, SVM, HMM, naive-Bayes, and ANN.

Although extensive efforts have been devoted to FER and many recent successes have been achieved, as mentioned above, many questions are still open. In our opinions, the following several points should be considered in future.

(1) How do human beings correctly identify facial expression?

So far, psychological and medical researches on human perception and cognition have lasted for a long time, but it is still ambiguous how human beings identify facial expression. Which kinds of parameters can be employed by humans and how are they processed?

(2) How do we identify facial expression in real-world sceneries?

Due to subtle facial deformations, frequent head movements, and ambiguous and uncertain facial motion measurements, identifying spontaneous facial expression in real-world sceneries is far more difficult than the acted FER widely studied to date. Considering the fact that a spontaneous facial expression can be characterized by rigid head movements and non-rigid facial muscular movements, recent work [58] in modelling of spontaneous head motion recognition and action unit recognition in spontaneous facial expression was an exciting development. They developed a unified probabilistic facial action model [58] to simultaneously and coherently represent rigid and non-rigid facial motions, as well as their spatiotemporal dependencies. How to elaborate the unified probabilistic facial action model is needed to be in such work, which is as yet a research question. In addition, many previous works involved to the privacy problems since the existing available standard datasets were pose-based datasets. To solve the privacy problems, recent work [59] was an important direction, in which a depth camera-based FER system using multilayer scheme was developed.

(3) How do we automatically learn more effective facial features for facial expression recognition?

It is worth pointing out that the abovementioned hand-designed feature extraction methods usually rely on manual operations with labelled data. In other words, these methods are supervised. In addition, these hand-designed features such as LBP and Gabor wavelets representation are able to capture low-level information of facial images, except for high-level representation of facial images. In recent years, deep learning [60–63], as a recently-emerged machine learning theory, is based on the hierarchical architecture of information processing in the primate visual perception system, and has shown how hierarchies of features can be directly learned from original data in an unsupervised manner. How to use deep learning techniques to automatically learn more effective facial features is an important direction for FER.

(4) How may we integrate facial expression analysis with other modalities?

Emotion transfers the psychological information of human beings, since emotion is conveyed by various physiological changes, such as changes in heart-beating rate, sweating degree, blood pressure, etc. Emotion is also expressed by affective speech, facial expression, body gesture and so on. To promote emotion recognition performance, integrating multiple affective modalities, such as speech, facial, physiological and lexical information, is a very active subject [64–67] in recent years. Nevertheless, how to effectively integrating heterogeneous modalities of emotion expression to further improve multimodal emotion recognition performance is still an open question.

Funding

This work is supported by National Natural Science Foundation of China [grant number 61203257], [grant number 61272261] and Zhejiang Provincial Natural Science Foundation of China [grant number LY16F020011], [grant number LQ15F020001].

References

[1] A. Mehrabian, “Communication without words,” *Psychol. Today*, Vol. 2, pp. 53–5, 1968.

[2] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, “Static and dynamic 3D facial expression recognition: A comprehensive survey,” *Image Vis. Comput.*, Vol. 30, no. 10, pp. 683–97, Oct. 2012.

[3] Y. Tian, T. Kanade, and J. Cohn, “Facial expression analysis,” in *Handbook of face recognition*. Springer, 2005, pp. 247–75.

[4] C.-D. Caleanu, “Face expression recognition: A brief overview of the last decade,” in *IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, 2013, pp. 157–61.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models-their training and application,” *Comput. Vis. Image Underst.*, Vol. 61, no. 1, pp. 38–59, Jan. 1995.

[6] Y. Chang, C. Hu, R. Feris, and M. Turk, “Manifold based analysis of facial expression,” *Image Vis. Comput.*, Vol. 24, no. 6, pp. 605–14, Jun. 2006.

[7] R. Shbib, and S. Zhou, “Facial expression analysis using active shape model,” *Int. J. Signal Process. Image Process. Pattern Recognit*, Vol. 8, no. 1, pp. 9–22, 2015.

[8] L. A. Cament, F. J. Galdames, K. W. Bowyer, and C. A. Perez, “Face recognition under pose variation with local Gabor features enhanced by active shape and statistical models,” *Pattern Recognit.*, Vol. 48, no. 11, pp. 3371–84, Nov. 2015.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 23, no. 6, pp. 681–5, Jun. 2011.

[10] Y. Cheon, and D. Kim, “Natural facial expression recognition using differential-AAM and manifold learning,” *Pattern Recognit.*, Vol. 42, no. 7, pp. 1340–50, Jul. 2009.

[11] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. Zafeiriou, “Hog active appearance models,” in *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, 2014, pp. 224–8.

[12] R. Anderson, B. Stenger, and R. Cipolla, “Using bounded diameter minimum spanning trees to build dense active appearance models,” *Int. J. Comput. Vis.*, Vol. 110, no. 1, pp. 48−57, Oct. 2014.

[13] Y. Chen, C. Hua, and R. Bai, “Regression-based active appearance model initialization for facial feature tracking with missing frames,” *Pattern Recognit. Lett.*, Vol. 38, pp. 113−9, Mar. 2014.

[14] D. G. Lowe, “Object recognition from local scale-invariant features,” in *the Seventh IEEE International Conference on Computer vision*, Kerkyra, 1999, pp. 1150−7.

[15] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, Vol. 60, no. 2, pp. 91−110, Nov. 2004.

[16] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and D. Mohamed, “A set of selected SIFT features for 3D facial expression recognition,” in *20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 4125−8.

[17] H. Soyel, and H. Demirel, “Facial expression recognition based on discriminative scale invariant feature transform,” *Electron. Lett.*, Vol. 46, no. 5, pp. 343−5, Mar. 2010.

[18] Y. Li, W. Liu, X. Li, Q. Huang, and X. Li, “GA-SIFT: A new scale invariant feature transform for multispectral image using geometric algebra,” *Inform. Sci.*, Vol. 281, pp. 559−72, Oct. 2014.

[19] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognit.*, Vol. 29, no. 1, pp. 51−9, Jan. 1996.

[20] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,” *Image Vis. Comput.*, Vol. 27, no. 6, pp. 803−16, May 2009.

[21] S. Zhang, X. Zhao, and B. Lei, “Facial Expression Recognition Based on Local Binary Patterns and Local Fisher Discriminant Analysis,” *WSEAS Trans. Signal Process.*, Vol. 8, no. 1, pp. 21−31, 2012.

[22] X. Zhao, and S. Zhang, “Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding,” *EURASIP J. Adv. Signal Process.*, Vol. 2012, no. 1, pp. 20, Dec. 2012.

[23] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, “Local binary patterns and its application to facial image analysis: a survey,” *IEEE Trans. Syst. Man, Cybernet. Part C: Appl. Rev.*, Vol. 41, no. 6, pp. 765−81, Nov. 2011.

[24] G. Zhao, and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, no. 6, pp. 915−28, Jun. 2007.

[25] T. Jabid, M. H. Kabir, and O. Chae, “Robust facial expression recognition based on local directional pattern,” *ETRI J.*, Vol. 32, no. 5, pp. 784−94, Oct. 2010.

[26] T. Ahsan, T. Jabid, and U.-P. Chong, “Facial expression recognition using local transitional pattern on Gabor filtered facial images,” *IETE Tech. Rev.*, Vol. 30, no. 1, pp. 47−52, Jan.−Feb. 2013.

[27] X. Li, Q. Ruan, Y. Jin, G. An, and R. Zhao, “Fully automatic 3D facial expression recognition using polytypic multi-block local binary patterns,” *Signal Process.*, Vol. 108, pp. 297−308, Mar. 2015.

[28] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron,” in *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998*. Proceedings, Nara 1998, pp. 454−9.

[29] S.-s. Liu, and Y.-t. Tian, “Facial expression recognition method based on Gabor wavelet features and fractional power polynomial kernel PCA,” *Adv. Neural Netw.-ISNN 2010*. Vol. 6064, no. Part 2, pp. 144−51, 2010.

[30] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, “Facial expression recognition using radial encoding of local Gabor features and classifier synthesis,” *Pattern Recognit.*, Vol. 45, no. 1, pp. 80−91, Jan. 2012.

[31] E. Owusu, Y. Zhan, and Q. R. Mao, “A neural-AdaBoost based facial expression recognition system,” *Expert Syst. Appl.*, Vol. 41, no. 7, pp. 3383−90, Jun. 2014.

[32] S. Negahdaripour, “Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, no. 9, pp. 961−79, Sep. 1998.

[33] J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, “Subtly different facial expression recognition and expression intensity estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 853−9.

[34] Y. Yacoob, and L. S. Davis, “Recognizing human facial expressions from long image sequences using optical flow,” *IEEE Trans Pattern Anal. Mach. Intell.*, Vol. 18, no. 6, pp. 636−42, Jun. 1996.

[35] A. Sánchez, J. V. Ruiz, A. B. Moreno, A. S. Montemayor, J. Hernández, and J. J. Pantrigo, “Differential optical flow applied to automatic facial expression recognition,” *Neurocomputing*, Vol. 74, no. 8, pp. 1272−82, Mar. 2011.

[36] M. Pantic, and I. Patras, “Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences,” *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, Vol. 36, no. 2, pp. 433−49, Apr. 2006.

[37] Y. Tie, and L. Guan, “A deformable 3D facial expression model for dynamic human emotional state recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 23, no. 1, pp. 142−57, Jan. 2013.

[38] H. Fang, N. Mac Parthaláin, A. J. Aubrey, G. K. Tam, R. Borgo, P. L. Rosin, P. W. Grant, D. Marshall, and M. Chen, “Facial expression recognition in dynamic sequences: An integrated approach,” *Pattern Recognit.*, Vol. 47, no. 3, pp. 1271−81, Mar. 2014.

[39] P. S. Aleksic, and A. K. Katsaggelos, “Automatic facial expression recognition using facial animation parameters and multistream HMMs,” *IEEE Trans Inf. Forensics Secur.*, Vol. 1, no. 1, pp. 3−11, Mar. 2006.

[40] Y. Sun, and A. Akansu, “Facial expression recognition with regional hidden Markov models,” *Electron. Lett.*, Vol. 50, no. 9, pp. 671−3, Apr. 2014.

[41] L. Ma, and K. Khorasani, “Facial expression recognition using constructive feedforward neural networks,” *IEEE Tran. Syst. Man Cybern. Part B: Cybern.*, Vol. 34, no. 3, pp. 1588−95, Jun. 2004.

[42] C. R. De Silva, S. Ranganath, and L. C. De Silva, "Cloud basis function neural network: a modified RBF network architecture for holistic facial expression recognition," *Pattern Recognit.*, Vol. 41, no. 4, pp. 1241–53, Apr. 2008.

[43] V. G. Kaburlasos, S. E. Papadakis, and G. A. Papakostas, "Lattice Computing Extension of the FAM Neural Classifier for Human Facial Expression Recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, Vol. 24, no. 10, pp. 1526–38, Oct. 2013.

[44] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Comput. Vis. Image Underst.*, Vol. 91, no. 1, pp. 160–87, Jul.–Aug. 2003.

[45] X. Zhao, E. Dellandréa, J. Zou, and L. Chen, "A unified probabilistic framework for automatic 3D facial expression analysis based on a Bayesian belief inference and statistical feature models," *Image Vis. Comput.*, pp. 231–45, Mar. 2013.

[46] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image Vis. Comput.*, Vol. 25, no. 12, pp. 1856–63, Dec. 2007.

[47] K. Yurtkan, and H. Demirel, "Feature selection for improved 3D facial expression recognition," *Pattern Recognit. Lett.*, Vol. 38, pp. 26–33, Mar. 2014.

[48] D. Ghimire, and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, Vol. 13, no. 6, pp. 7714–34, Jun. 2013.

[49] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 31, no. 2, pp. 210–27, Feb. 2009.

[50] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, Vol. 52, no. 4, pp. 1289–306, Apr. 2006.

[51] S. Zhang, X. Zhao, and B. Lei, "Robust facial expression recognition via compressive sensing," *Sensors*, Vol. 12, no. 3, pp. 3747–61, Mar. 2012.

[52] S. Zhang, X. Zhao, and B. Lei, "Facial expression recognition using sparse representation," *Wseas Trans. Syst.*, Vol. 11, no. 8, pp. 440–52, 2012.

[53] M. Mohammadi, E. Fatemizadeh, and M. Mahoor, "PCA-Based dictionary building for accurate facial expression recognition via sparse representation," *J. Vis. Commun. Image Represent.*, Vol. 25, no. 5, pp. 1082–92, Jul. 2014.

[54] Y. Ouyang, N. Sang, and R. Huang, "Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers," *Neurocomputing*, Vol. 149, pp. 71–8, Feb. 2015.

[55] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 21, no. 12, pp. 1357–362, Dec. 1999.

[56] T. Kanade, Y. Tian, and J. Cohn, "Comprehensive database for facial expression analysis," in International Conference on Face and Gesture Recognition, Grenoble, France, 2000, pp. 46–53.

[57] C. Shan, S. Gong, and P. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image Vis. Comput.*, Vol. 27, no. 6, pp. 803–16, May 2009.

[58] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 32, no. 2, pp. 258–73, Feb. 2010.

[59] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan, and S. Lee, "Depth camera-based facial expression recognition system using multilayer scheme," *IETE Tech. Rev.*, Vol. 31, no. 4, pp. 277–86, Aug. 2014.

[60] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, Vol. 313, no. 5786, pp. 504–07, Jul. 2006.

[61] D. Yu, and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Process. Mag.*, Vol. 28, no. 1, pp. 145–54, Jan. 2011.

[62] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, Vol. 521, pp. 436–44, May 2015.

[63] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Tech. Rev.*, Vol. 32, no. 5, pp. 347–55, Sep.–Oct. 2015. doi:10.1080/02564602.2015.1017542.

[64] S. Zhang, X. Wang, G. Zhang, and X. Zhao, "Multimodal emotion recognition integrating affective speech with facial expression," *WSEAS Trans. Signal Process.*, no. 10, pp. 526–37, 2014.

[65] J. Kim, and M. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Trans. Affective Comput.*, Vol. 6, no. 4, pp. 371–84, Oct.–Dec. 2015. doi:10.1109/TAFFC.2015.2411273.

[66] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognit. Lett.*, Vol. 66, pp. 22–30, Nov. 2015. doi:10.1016/j.patrec.2014.11.007.

[67] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal Bayesian fusion for affect sensing: combining video, audio, and lexical modalities," *IEEE Trans. Cybernet.*, Vol. 45, no. 9, pp. 1927–41 Sep. 2015. doi:10.1109/TCYB.2014.2362101.

Authors

Xiaoming Zhao received the BS degree in mathematics from Zhejiang Normal University in 1990 and the MS degree in software engineering from Beihang University in 2006. He is currently a professor of department of computer science, Taizhou University, China. His research interests include image processing, machine learning and pattern recognition.

E-mail: tzxyzxm@163.com

Shiqing Zhang received the BS degree in electronics and information engineering from Hunan University of Commerce in 2003, the MS degree in electronics and communication engineering from Hangzhou Dianzi University in 2008, and the PhD degree at school of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. Currently, he works as an assistant professor of department of physics and electronics engineering, Taizhou University, China. His research interests include image processing, affective computing and pattern recognition.

E-mail: tzczsq@163.com