

Facial expression recognition based on meta probability codes

Nacer Farajzadeh · Gang Pan · Zhaohui Wu

Received: 10 January 2012 / Accepted: 12 December 2012 / Published online: 1 January 2013
© Springer-Verlag London 2012

Abstract Automatic facial expression recognition has made considerable gains in the body of research available due to its vital role in human–computer interaction. So far, research on this problem or problems alike has proposed a wide variety of techniques and algorithms for both information representation and classification. Very recently, Farajzadeh et al. in *Int J Pattern Recognit Artif Intell* 25(8):1219–1241, (2011) proposed a novel information representation approach that uses machine-learning techniques to derive a set of new informative and descriptive features from the original features. The new features, so called meta probability codes (MPC), have shown a good performance in a wide range of domains. In this paper, we aim to study the performance of the MPC features for the recognition of facial expression via proposing an MPC-based framework. In the proposed framework any feature extractor and classifier can be incorporated using the meta-feature generation mechanism. In the experimental studies, we use four state-of-the-art information representation techniques; local binary pattern, Gabor-wavelet, Zernike moment and facial fiducial point, as the original feature extractors; and four multiclass classifiers, support vector machine, k -nearest neighbor, radial basis function neural network, and sparse representation-based classifier. The results of the extensive experiments conducted on three facial expression datasets, Cohn–Kanade, JAFFE, and

TFEID, show that the MPC features promote the performance of facial expression recognition inherently.

Keywords Facial expression · Information representation · Classification · Support vector machine · Radial basis function neural network · k -nearest neighbor · Sparse representation-based classifier · Local binary pattern · Gabor-wavelet · Zernike moment · facial fiducial point · Meta probability code

1 Introduction

In our daily life, we may change our faces many times based on the situation we are involved in. These changes in our faces, so called facial expressions, are a rich source of information that exposes our emotions. By perception and subtle changes in muscles of the face, we can understand a lot that we may alternatively need several words or phrases to express instead. In other words, facial expressions contain key features of our emotional reactions when facing or experiencing various occasions, and play a crucial role in our non-verbal communication. It is also very helpful for emotion-awareness in smart environments [2, 3].

Research on analysis of facial expressions was originated by Darwin [4] in the nineteenth century. Later, in 1974, Ekman and Friesen [5] defined six basic emotions that can be distinguished easily from human faces and are common among humans: anger, disgust, fear, happiness, sadness and surprise. Each of these emotions has its own unique characteristics and thus can be distinguished from the others without any difficulties.

We, as humans, are able to distinguish and interpret facial expressions easily. It is obvious that this ability or the other abilities alike are based on the way we observe

N. Farajzadeh (✉) · G. Pan (✉) · Z. Wu
College of Computer Science and Technology,
Zhejiang University, Hangzhou 310027, China
e-mail: nafa@zju.edu.cn

G. Pan
e-mail: gpan@zju.edu.cn

Z. Wu
e-mail: wzh@zju.edu.cn

(information representation) and the way we learn (classification) or interpret the observed activities over time. Yet, there is no clear explanation as to how we have established such a, perhaps highly parallel and indeed very accurate, vision system. Therefore, the research on automation of facial expression recognition (FER) or other related fields in machine vision literature has proposed a wide variety of techniques and algorithms for both information representation and classification towards an optimal vision system. A survey of recent advantages in the literature can be found in [6–9].

Generally speaking, it is difficult to group automatic facial expression recognition approaches based on the classification algorithms used to recognize different expressions. This is due to the fact that the role of classification has been considered as of secondary importance in machine vision, while the role of information representation is considered first [10]. The more intuitive reason, which supports the aforementioned statement, can be sought in the fact that the different classification approaches fit for different types of information representation; and, indeed, there is no classifier that can outperform other classifiers in any given problem due to the lack of a universal approach for information representation. Nevertheless, it is obvious that classification is one of the primary stages for any vision system, and the final performance heavily depends on the performance of the classifier utilized.

Very recently, Farajzadeh et al. [1] proposed a novel approach for information representation. Their approach derives a set of new features from the original features based on machine learning techniques. These features, which are called meta probability code (MPC), consist of the probabilities of the corresponding sample being a member of pair-wise classes. In other words, the MPC is a set of new features that represent an instance in terms of similarities between pairs of classes instead of the original attributes that have been measured or extracted. Therefore, the MPC features are more discriminative and informative than the original features, and have shown a good performance in both clustering and classification problems in a wide range of domains [1].

As the authors of [1] have mentioned, their approach is well suited for the problems with balanced datasets, i.e., the numbers of samples per classes are almost equal or good enough. In this paper, we aim to study how well the MPC features represent facial expressions and improve the recognition performance, where the datasets available for this problem are almost balanced.

To achieve the above-mentioned goal, we propose an MPC-based framework for the recognition of facial expressions in which any feature extractor and classifier can be incorporated in the proposed framework using the meta-feature generation mechanism.

In the experimental studies, we integrate different feature extractors and classification algorithms into the proposed framework yielding different single MPC-based FER systems. In order to compare the performances of the MPC-based FER systems with the other systems in a systematical and statistical [11] manner, the original features and the classifiers of interest are also integrated into several simple and common original-feature-based FER systems. To extract the original features, we use four state-of-the-art information representation approaches: local binary pattern [12], Gabor-wavelet [13], Zernike moment [14] and facial fiducial points, and, for the classification purpose, we use four multiclass classification algorithms; support vector machine [15], k -nearest neighbor [16], radial basis function neural network [17] and sparse representation-based classifier (SRC) [18].

The experimental results conducted on three publicly available datasets: Cohn–Kanade [19], JAFFE [20] and TFEID [21], show that, using a system based on the MPC features, the performance of facial expression recognition is significantly improved. The results also indicate that the MPC features markedly improve the generalization performance on across datasets evaluation, where one of the datasets of interests is used as the training set and the other two datasets are used as the testing sets.

This paper not only evaluates the performance of the newly proposed features MPC for the recognition of facial expression, but also conducts the experiments in a very promising manner to ensure the consistency and the reliability of the results. To the best of our knowledge, this is the first effort in the literature that the performance of different information representation techniques and classification algorithms for the recognition of facial expressions are compared systematically and statistically. It is obvious that the evaluation of different approaches for FER by directly quoting the reported results and comparing them is not consistent (e.g., [6–8]), as their exploited datasets, preprocessing techniques, comparison protocols, information representation techniques and classification algorithms are not the same. In other words, there are always two or three things out of five important things (datasets, preprocessing, protocols, classifiers and features) that are different in the available comparisons for the recognition of facial expressions. And more importantly, none of the studies have compared the performance of the approaches statistically to ensure the reliability of the results from the statistical view point.

The rest of this paper is organized as follows. In the next section, Sect. 2, a review of the previously proposed facial expression recognition approaches is presented; Sect. 3 describes our MPC-based facial expression recognition framework; Sect. 4 provides brief introductions of some of the commonly used information representation approaches

and classification algorithms for immediate reference; Sect. 5 gives the experimental results; and Sect. 6 is the conclusion.

2 Related work

Based on the research done in the literature, one can group automatic facial expression recognition approaches into roughly two categories according to the employed information representation methods [6]: holistic-based representation [22–30] and analytic-based representation [31–34]. However, we can also group the related works into another two main categories according to the characteristic of the observations [35]: image-based approaches, where the apex of an expression is captured and used as an observation; and video-based approaches, where the whole sequence of an expression or a subset of it is considered [36]. As this paper is involved with the image-based approaches, in the following, we provide a review of the previously proposed image-based approaches under the former categorization.

The *holistic*-based approaches extract features from the whole face or major parts of it. These approaches are also known as the template-based techniques, as they, for example, use a labeled graph or a 2D array of intensity values to represent facial images as a whole [6].

As one of the recent *holistic* approaches, Yu and Bhanu [22] used Gabor-wavelet (GW) to represent primitive features of a face. Then, they applied genetically inspired learning method for selecting the extracted features. The significant advantage of their method was the use of SVM to classify facial expressions. However, the accuracy of their approach (80.9 %) on JAFFE dataset was relatively low compared to the other approaches' achievement on the same dataset. Littlewort et al. [23] also used GW features to represent facial images and an SVM to classify them. But the technique they used for selecting a subset of extracted GW features was Adaboost. They achieved a classification accuracy of 93.3 % on Cohen–Kanade dataset.

By manually selecting the positions of eyes and using local binary patterns (LBP) as features, Feng et al. [24] proposed a new approach for human facial expression recognition. Using linear programming (LP), they produced 21 binary classifiers per each pairs in which each pair was composed of two expressions (e.g., fear-happiness, sadness-surprise, etc.). Then they generated a simple binary tournament tree (BT) using these binary classifiers in order to induce the final classifier. The successful classification rate of their approach on JAFFE dataset was 93.8 %. In another recent work, Shan et al. [25] also used LBP features to represent facial expressions. They

extensively evaluated the performance of LBP features with different classification algorithms. The best recognition rates that they reported on Cohn–Kanade dataset were obtained by using SVM; 92.6 % for six expressions and 88.9 % for seven expressions (including Neutral).

Xie and Lam [26] proposed a new model, spatially maximum occurrence model (SMOM), based on statistical characteristics of the training images for representing facial expressions. The technique they used to classify different expressions was the similarities among the images. To measure the similarity they applied elastic shape–texture matching (ESTM) technique. The successful recognition rates of their work were reported at 94.5 and 94.7 % on AR dataset and Yale dataset, respectively. In [27], Lajevardi and Hussain used Zernike moments (ZM) as the statistical features to represent facial expressions. They evaluated the performance of Zernike moments against noises and rotations. Applying LDA on Cohn–Kanade and JAFFE datasets, under different noises and rotations, the average successful recognition rates of their approach were 73.2 and 92.8 %, respectively. In the proposed approach by Yang et al. [28], facial images were divided into some patches according to the locations of action units. Then, Haar-like features were extracted from each patch and selected by Adaboost. They named these features as combined features (CF). Using minimum error-based optimization strategy (MEOS), the successful recognition rate of their approach on Cohn–Kanade dataset was 92.3 %.

In a very recent work, Zhi et al. [29] proposed a novel approach based on graph-preserving sparse non-negative matrix factorization (GSNMF). The GSNMF algorithm transforms a high dimension space into a locality-preserving subspace with sparse representation, where the sparse representation is obtained by minimizing the ℓ_1 -norm. Extensive experiments conducted on Cohn–Kanade and JAFFE datasets with clean and partially occluded facial images have shown the performance of GSNMF and its robustness to partial occlusions. In another recent work, Cotter [30] used sparse representation theory and a classification scheme based on this theory (SRC) on different type of features such as GW and raw pixels. His work also demonstrated promising performances on clean, noise corrupted, and partially occluded facial images of JAFFE.

In the *Analytic* or feature-based approaches, some facial points or contours of the key facial features such as eyes, eyebrows and mouth are used to model the face [6]. These feature points can either be used directly as the input feature vector, or can be used to measure the relative distances in between and construct a feature vector accordingly.

Kobayashi and Hara [31] manually measured 30 facial characteristic points (FCP) and they used them as inputs for

a back-propagation neural network (BPNN) to classify facial expressions. The classification rate of their approach reached the value of about 80.0 %. Ushida et al. [32] also used features introduced by Kobayashi and Hara and applied a bi-directional associative neural network (BANN) together with fuzzy logic (FL) to classify three expressions angry, happiness, and sadness. They achieved 79.0 % of successful classification accuracy. Sohail and Bhattacharya [33] used spatio-temporal representation of face with 11 facial characteristic points. The classification accuracy of their approach using a distance weighted k NN classifier was 90.8 % on JAFFE dataset. Using a semantic-based learning algorithm along with the analytical hierarchy process (AHP), Cheng et al. [34] introduced the hybridization of low-level features and high-level semantic

concept. The classification accuracy of their approach using weight-assigned semantic information supported k NN on JAFFE dataset for the recognition of five facial expressions (neutral, happiness, anger, sadness and surprise) was reported at 85.2 %.

A summary of the reviewed related works and the systems based on our proposed framework (see Sects. 3.2, 5.3 and 5.9) are provided in Table 1. In this table, the column *Comments* is to emphasize the differences between the experiments in terms of preprocessing and feature enhancement techniques, number of samples per subject, number of expressions, etc. Obviously, these differences can have a considerable effect on the final performance, and therefore, one should carefully take them into account when comparing the performance of different FER approaches.

Table 1 Summary of the reviewed FER systems

	Feature	Classifier	Dataset	Protocol	Accuracy (%)	Comments
1992 [31]	30 FCPs	BPNN	19 JS ^a	1/2	80.0	–
1993 [32]	30 FCPs	BANN + FL	19 JS	1/2	79.0	Three expressions were considered
2006 [22]	GW	SVM	JAFFE	10CV ^b	80.9	A genetically inspired process was used for feature selection
2006 [23]	GW	SVM	CK ^c	LOO ^d	93.3	Adaboost was used for feature selection
2007 [24]	LBP	LP+BT	JAFFE	10CV	93.8	The subsystem introduced in [37] was used for preprocessing
2007 [33]	11 FCPs	k NN	JAFFE	LOO	90.7	–
2007 [34]	14 FCPs + AHP	k NN	JAFFE	1/2	85.2	Five expressions were considered
2009 [25]	LBP	SVM	CK	10CV	88.9	The last three peak frames of each sequence were used
2009 [26]	SMOM	ESTM	AR	1/2	94.5	The face model proposed in [38] was adopted to produce a facial mask
2009 [26]	SMOM	ESTM	Yale	LOO	94.7	The face model proposed in [38] was adopted to produce a facial mask
2009 [27]	ZM	LDA	CK	1/2	73.2	Some amount of noises and several degrees of rotations were considered
2009 [27]	ZM	LDA	JAFFE	1/2	92.8	Some amount of noises and several degrees of rotations were considered
2010 [28]	CF	MEOS	CK	2/3	92.3	The last three peak frames of each sequence were used
2010 [30]	GW	SRC	JAFFE	LOO	93.9	PCA was applied to reduce the dimension of features
2011 [29]	GSNMF	k NN	CK	1/8 ^e	93.5	Six expressions were considered, and only a subset of 30 individuals was used
Our method	MPC	MPC-FER	CK	1/8	93.3	Six expressions were considered, and only a subset of 30 individuals was used
Our method	MPC	SVM	CK	1/8	93.1	Six expressions were considered, and only a subset of 30 individuals was used
Our method	MPC	RBFNN	CK	10CV	87.2	The last peak frame of each sequence was used

^a Japanese students

^b Tenfold cross validation

^c Cohn–Kanade

^d Leave-one-out

^e One out of the last eight peak frames of each sequence was used for training and the remaining frames were used for testing

3 MPC-based FER framework

The goal of automatic facial expression recognition is to use distinctive features of the expressions and label them with the respective emotions. In this section, we propose a facial expression recognition framework that uses MPC to represent facial expressions and an MPC-based classifier to recognize them. The next subsection introduces the MPC, followed by the proposed framework for automatic facial expression recognition.

3.1 Meta probability code (MPC)

One common and popular way to solve a multiclass classification problem via decomposition technique is one-against-one (OAO) strategy. In this strategy, given a training sample set \mathbf{X} and k class labels, $K = k(k - 1)/2$ binary classifiers are induced each of which separate a pair of classes:

$$h_{r,s}^B(\mathbf{x}_i) = \begin{cases} 1 & \text{if } \mathbf{x}_i \in C_r \\ 0 & \text{if } \mathbf{x}_i \in C_s \end{cases} \quad (1)$$

for $r = 1, \dots, k - 1$, and $s = r + 1, \dots, k$,

where the superscript B indicates that h^B is a binary classifier [1]. The class label of an unknown sample is assigned based on the majority voting scheme, where the votes are collected from h^B s results.

Borrowing the idea of stacked generalization [39], the authors of [1] have interpreted the outputs of h^B s as a set of new features for a further processing instead of using them for the voting scheme. Thus, a function that projects the original features to a new feature vector \mathbf{t} has been proposed [1]:

$$\mathbf{t} = f(\mathbf{x}), \quad (2)$$

$$f(\mathbf{x}) = \oplus h_{r,s}^B(\mathbf{x}), \quad f : \mathbb{R}^N \longrightarrow \mathbb{Q}^K,$$

where \oplus is the concatenation operator, and N and K are the dimensions of the original data space and the projected data space, respectively.

The new feature vector \mathbf{t} is called MPC if the outputs of h^B s are considered to be the class probabilities (real-valued outputs):

$$\text{MPC}(\mathbf{x}) = \oplus h_{r,s}^B(\mathbf{x}) \text{ where } h_{r,s}^B(\mathbf{x}) = p(r|\mathbf{x}). \quad (3)$$

We should note that the probability output of $h_{r,s}$ for a given sample \mathbf{x} (Eq. 3) is, in most cases, an estimation of the *posterior* probability. Therefore, following [1], we use the probability estimation method proposed by Wu [40] in order to produce the class probabilities in the proposed framework.

3.2 Proposed framework

The proposed MPC-based framework for automatic facial expression recognition consists of three steps. In the first step, a typical feature extraction approach is used to obtain the facial expression features. For this step, one can use any state-of-the-art feature extractor. The extracted features, referred to as the original features, are then sent to the next step in order to be projected into a new feature space, MPC.

In the second step, the MPC features are derived from the original features using the projection function in Eq. (3). The projection function in Eq. (3) is built based on the training set. That is, all the binary classifiers (h^B s), which constitute the projection function, are trained based on the original features extracted in the first step. In this step, one can also use any classification algorithm to induce h^B s and build the projection function.

The third step employs a multiclass classification algorithm to classify the MPC features. For this step, in particular, we use the classification algorithm introduced in [1], which is called MPC-clustering. We will refer to this particular system as MPC-FER. However, in order to evaluate the performance of the MPC features, only from the information representation point of view, we will also use several well-known and promising multiclass classification algorithms in the third step of the proposed framework.

In the third step of MPC-FER, given projected samples' set $\mathbf{T} = \{(\mathbf{t}_i, y_i), i = 1, \dots, l\}$, a clustering scheme is used to cluster \mathbf{T} ; $\mathbf{T} = \bigcup_{s=1}^S \ell_s$, where S is the number of produced clusters and the elements of ℓ_s are pairwise disjoint. Then using the samples in each and every produced clusters, $\ell_s = \{(\mathbf{t}_j, y_j), j \in \mathcal{N}_s\}$, where \mathcal{N}_s is the set of samples' indexes in cluster ℓ_s , S multiclass classifiers are induced such that $h_s^M(\mathbf{t}_j) = y_j$ (the superscript M is to indicate that h^M is a multiclass classifier) [1]. For an unknown sample, its original features are extracted first, and then applying the projection function in Eq. (3), its MPC features are produced; subsequently, the cluster that the sample belongs to is determined. Finally, the corresponding individual multiclass classifier is used to label the sample. The block diagram of the proposed framework is demonstrated in Fig. 1.

4 Facial expression representation and recognition

4.1 Expression representation

In this section, a brief introduction of four state-of-the-art information representation (extraction) approaches along with their properties used in the experiments is presented.

We select three holistic face representation approaches; local binary patterns, Gabor-wavelets, Zernike moments, and one analytic approach; facial fiducial points. In this study, the features extracted by these approaches are referred to as the original features (ORG).

4.1.1 Local binary pattern

The local binary pattern (LBP) is one of the most popular image descriptors due to its efficiency in descriptiveness and computational complexity. The LBP operator, introduced by Ojala et al. [12], assigns a label to every pixel of an image by thresholding the gray-level of a given pixel’s neighbors with the gray-level of the pixel itself, and considering the result as an integer number.

In order to capture more dominant features in some textures, the basic 3×3 LBP operator was extended to use different sizes of neighborhoods and radii by means of interpolation of the adjacent pixels. Another extension to the original LBP is to use only a subset of patterns out of 2^P total binary patterns that are more informative, called

uniform patterns. The uniform local binary pattern with P neighborhood pixels and R radius is indicated by $LBP_{P,R}^{u2}$.

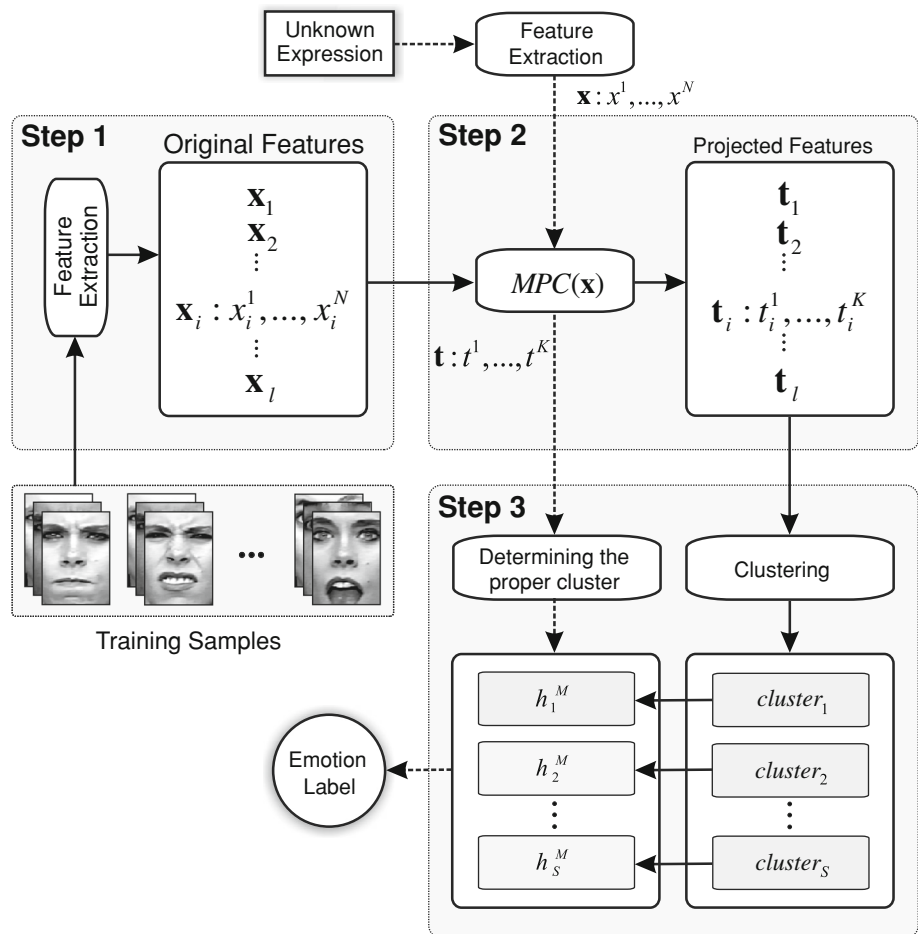
The purpose of the LBP operator is to extract and codify the local micro-patterns such as edges, corners, spots and flat areas of a given image [12]. The local micro-patterns are then used to describe the image statistically by the means of their distribution over the whole image.

In our experiments, in order to have a good trade-off between feature vector length and the recognition performance, we follow the settings used in [41]: a given face image is divided into 42 (6×7) non-overlapping regions, and $LBP_{8,2}^{u2}$ operator is applied on each region separately (Fig. 2). Concatenating the LBP histograms of the regions results in a feature vector of length 2,478 (59×42) [25].

4.1.2 Gabor-wavelet

The Gabor filter is a linear filter that was originally used for edge detection in images [13]. The similarity of Gabor filters in terms of frequency and orientation representations to those of human visual system have made them very appropriate technique for image description [42–44].

Fig. 1 Block diagram of the proposed MPC-based framework for automatic recognition of facial expressions. In the diagram, solid arrows and dashed arrows show the training and the testing steps, respectively



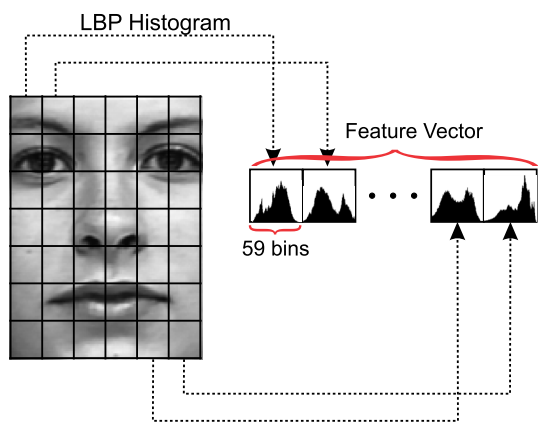


Fig. 2 A given facial expression image is divided into 42 regions and the LBP histograms of the regions are extracted and concatenated into a single feature vector

A set of Gabor-wavelet (GW) functions $\Psi = \{\psi_{1,1,1}, \dots, \psi_{P,Q,R}\}$ with P frequencies and Q orientations at R feature points is defined as follows:

$$\psi_{i,j,k}(x,y) = \frac{f_i^2}{2\pi} \exp\left\{-0.5f_i^2 \left[(x - c_{x_k})^2 + (y - c_{y_k})^2\right]\right\} \times \sin\left\{2\pi f_i \left[(x - c_{x_k}) \cos \theta_j + (y - c_{y_k}) \sin \theta_j\right]\right\} \tag{4}$$

where f_i is the frequency, θ_j is the orientation, c_{x_k} and c_{y_k} are the positions of the wavelet. To obtain the GW features of an image, the convolution of the image with the Gabor filters bank given in Eq. (4) is calculated.

In the experiments, a bank of Gabor filters with eight different orientations and five spatial frequencies is used to represent face images [45]. For an image of 110×150 pixels (in the experiment, every image is normalized to 110×150 pixels) the length of the feature vector is 660,000 ($40 \times 110 \times 150$), which is far greater than the original data for the image. To reduce the computational burden and to have a good generalization performance, the length of the feature vector is reduced to 42,560 via down-sampling Gabor filters by a factor of 16 [46].

4.1.3 Zernike moments

The orthogonal moments, also known as the statistical information representation approaches, have gained considerable attention in the literature due to their invariant properties¹ [14]. Among the well-known orthogonal moments, such as Legendre moments, Fourier–Mellin moments, and pseudo-Zernike moments, Zernike moments (ZM) have been frequently used as an image descriptor,

¹ If a given image is changed in terms of scale, position, rotation or a combination of them, its statistical features will remain unchanged.

and have shown a good performance in face and facial expression recognition problems [27, 47].

The ZMs are calculated in polar coordinates, and thus, to utilize them as a descriptor one needs to map a given image to a unit disc and set the center of the image as the center of the unit disc, i.e., $x^2 + y^2 \leq 1$. The complex ZM of order n and repetitions m subject to $n - |m| = \text{even}$ and $|m| \leq n$ is defined as follows:

$$ZM_{nm} = \frac{n+1}{\pi} \sum_x \sum_y F_{xy} V_{nm}(x,y), \tag{5}$$

where F_{xy} represents the current pixel and $V_{nm}(x,y)$ is the Zernike polynomial in polar coordinate as follows:

$$V_{nm}(r,\theta) = R_{nm}(r) \exp(jm\theta), \quad r = \sqrt{x^2 + y^2} \quad \text{and} \quad \theta = \tan^{-1}\left(\frac{y}{x}\right). \tag{6}$$

The real-valued orthogonal radial polynomial, R_{nm} , is defined as:

$$R_{nm}(r) = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s r^{n-2s} \frac{(n-s)!}{s! \left(\frac{n+|m|}{2} - s\right)! \left(\frac{n-|m|}{2} - s\right)!}. \tag{7}$$

As it has been mentioned in [47], finding the best order and repetitions for an invariant moment-based image descriptor is an NP-hard problem. Thus, a straightforward approach to form an optimal feature vector based on invariant moments has been suggested; the feature vector for ZM with lower bound k and upper bound N is defined as follows:

$$FV_{k,N}^{ZM} = \{ZM_{nm}\}, \tag{8}$$

where $n = k, \dots, N$, and $m = 0, 2, \dots, n$ when n is even, and $m = 1, 3, \dots, n$ when n is odd.

In our experiments, according to our empirical studies, we set the value of k and N equal to 2 and 15, respectively, which results in a feature vector of length 70.

4.1.4 Facial fiducial points

The facial fiducial point (FFP) or facial characteristic point is another approach for representing facial expressions. In this approach, after localizing a face in an image, the precise positions of the center of the eyes is determined. These points are then used to extract and normalize the face sub-image. Subsequently, other facial components including eyes, eyebrows, nose, and mouth, are localized in order to extract more fiducial points such as the tip of nose, lip corners, their upper and lower mid-points, mid-point, etc. Finally, all the extracted fiducial points are concatenated to form a feature vector. Two examples of different facial fiducial points are demonstrated in Fig. 3.

In the experiments, we use publicly available FFPs for the Cohn–Kanade dataset. There are 59 fiducial points in total for each image in Cohn–Kanade dataset, resulting in a feature vector of length 118 (2×59). For the JAFFE dataset, we use the fiducial points introduced in [48], where there are 34 fiducial points for each image, resulting in a feature vector of length 68 (2×34). For the TFEID, there are no publicly available fiducial points. Therefore, we do not consider these features for TFEID in our experiments.

4.2 Expression recognition

In this section, we provide very short introductions of some well-known and promising multiclass classification algorithms being employed and examined in this study as well as the detailed information regarding their parameter settings and training procedures. The classifiers of interest are: support vector machine, radial basis function neural network, k -nearest neighbor, and sparse representation-based classifier.

4.2.1 Support vector machine

The support vector machine (SVM) is a class of linear classification algorithms proposed by Vapnik [15], in which it aims to find a separating hyperplane with as wide a margin as possible between two different categories of data. Unfortunately, the linear optimization problem proposed in SVM algorithm is not enough for practical usage due to the linearly inseparable nature of the data in real-world applications. One possible approach to overcome this problem is to map the data to an alternative dimension space, which is higher (possibly infinite) than the original space, in the hope that the data will be linearly separable in that space. To employ this approach efficiently, a trick known as the kernel trick is utilized. This trick allows us to compute dot products between the vectors in a high

dimension space within the original space without ever having to compute the mapping explicitly. There are several popular kernel functions that can be employed in SVM algorithm, among which we use Gaussian function in our experiments.

To generalize a two-class SVM to a multiclass SVM, we use three strategies in our experiments: one-against-one (OAO), one-against-all (OAA), and a single machine approach² SVM proposed in [49]. In the experiments, we use a publicly available implementation of SVM, *libsvm* [50], where the optimal parameter selection is done based on the grid optimization strategy [51].

4.2.2 Radial basis function neural network

The radial basis function neural network (RBFNN) is a type of non-linear classifier which is well suited for regression and complex (non-linear) pattern classification problems [17]. The basic architecture for a RBFNN is a 3-layer network: the first layer, input layer; the hidden layer, RBF units; and the third layer, output layer. The unique characteristic of the RBFNN is that the units in the hidden layer are assumed to be the centers of the possible clusters (also known as the prototypes) in a given dataset. Therefore, to build the RBFNN, we need to know the number of units forming the hidden layer in advance. To this end, we use k -means clustering scheme to find the existing clusters in the training set and assign them to the hidden layer units. The proper number of clusters is found by cross validation on the training set. The radius of the units (clusters' widths) are all set to a single value as half the average distance between the set of centers, and the weights are tuned by means of gradient descent algorithm [52].

4.2.3 k -nearest neighbor

The k -nearest neighbor classifier (k NN) is the most straightforward classifier in machine learning [16]. In this classifier, the generalization task is postponed until the classification of a sample is required. That is, there is no effort to gain prior assumptions about the distribution of the training samples, and due to this, the learning algorithm in k NN is called lazy or instance-based algorithm. In its simplest form (1NN), once an unseen sample is presented, its label is assigned based on the nearest training sample's label; and in its general form (k NN), the majority label of k -nearest training samples is assigned to the unknown sample. In our experiments, the number of nearest neighbors, k , is set to 10.

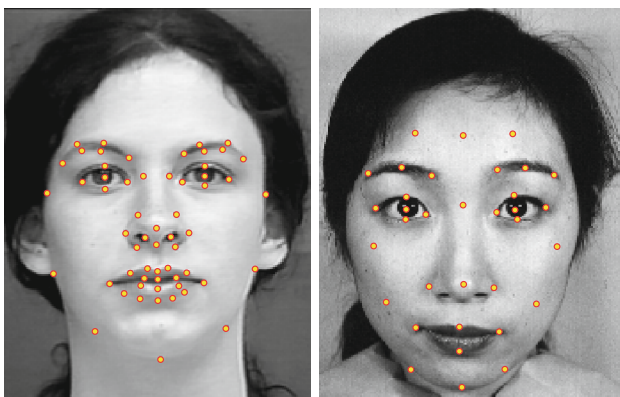


Fig. 3 Two different examples of facial fiducial points: *left* Cohn–Kanade dataset, *right* JAFFE dataset

² In the single machine approach, a binary classifier is generalized by adapting its internal operations to a single multiclass optimization problem.

4.2.4 Sparse representation-based classifier

Sparse representation is a recently developed theory for signal processing in the compressed sensing. It has been shown that the sparse representation can be very efficiently used for acquiring, representing, and compressing high-dimensional signals [53], as the signals such as images have naturally sparse representations. With the help of sparse representation, it is possible to exactly reconstruct sparse signals from a small number of linear measurements [54].

To employ the sparse representation theory in the classification context, where there are a number of training samples available, examples of different classes are considered as the measurements. That is, forming a dictionary of training vectors from all available samples, a relationship between the test vector and the training vectors must be found to classify the test vector [30]. This idea has been proposed by Wright et al. [18] and has been applied successfully for the recognition of face and facial expressions [18, 30]. The main idea in [18] is to represent test sample i as efficiently as possible merely using a linear combination of the training samples from class i [30]. The solution for this combination problem is formed by using a small number of training vectors from the large training set, and thus, it is sparse and can be achieved by solving the following equation:

$$\min \|\alpha\|_1 \quad \text{subject to} \quad \mathbf{A}\alpha = y, \quad (9)$$

where \mathbf{A} is the matrix of all the training samples, α denotes the weights on each of the dictionary vectors, and y is the test vector.

In our experiment, following [30], we use an implementation of Basis Pursuit from the SparseLab software package [55] to find the solution to Eq. (9), which results in SRC.

5 Experiments

5.1 Datasets

In this study, experimental studies are carried out on three facial expression image datasets: Cohn–Kanade, JAFFE and TFEID. In the following, the descriptions of these datasets are presented.

5.1.1 Cohn–Kanade

The Cohn–Kanade facial expression dataset consists of 100 adult subjects aged from 18 to 30 and of which 69 % were female, 81 % European-American, 13 %

African-American, and 6 % other groups [19]. The subjects were asked to perform six emotions starting from a neutral emotion and ending with the target emotion. Some of the subjects were asked to perform one of the emotions twice. The image sequences of each performance were captured and digitized into 640×490 pixel arrays.

For our experiments, we carefully label the emotion of each sequence and chose only the peak frame of the selected sequence as the target emotion. The images for Neutral emotion are collected from the first frame of 97 different sequences. For those subjects that have more than one performance for a given emotion, we only consider selecting one sequence, resulting 407 images: 36 anger, 40 disgust, 33 fear, 84 happiness, 97 neutral, 42 sadness, and 75 surprise.

5.1.2 JAFFE

The Japanese Female Facial Expression (JAFFE) image dataset consists of 10 Japanese female facial expression images [20]. Every subject in this dataset has 2–4 images for each expression, 213 images in total of size 256×256 pixels: 30 anger, 29 disgust, 32 fear, 31 happiness, 30 neutral, 31 sadness, and 30 surprise. In our experiments, we use all the images in this dataset.

5.1.3 TFEID

The Taiwanese Facial Expression Image Dataset (TFEID) consists of seven different facial expressions captured from 40 Taiwanese models (50 % male) [21]. There is only one image of each expression available for each subject, totally 268 images of size 480×600 pixels: 34 anger, 40 disgust, 40 fear, 40 happiness, 39 neutral, 39 sadness, and 36 surprise. All the images of this dataset are used in our experiments.

5.2 Preprocessing

All the facial images in the datasets of interest are normalized to a fixed distance between the center of the eyes and are cropped to sub-images of size 110×150 pixels [25]. The eye coordinates for Cohn–Kanade and JAFFE datasets are from the available facial fiducial points, and for TFEID, they are manually labeled. The cropped facial sub-images are then rotated to place the center of the eyes in line. We should note that no further preprocessing procedures such as the subsystem used in the CSU Face Identification Evaluation System [37], the face model proposed in [38], illumination correction, histogram equalization, etc., are applied. Figure 4 shows some pre-processed sample images of the facial expressions used in the experiments.

5.3 Systems of interest

In the experiments, 11 different systems are evaluated from which five systems, SVM_{ORG}^{OAO} , SVM_{ORG}^{OAA} , kNN_{ORG} , $RBFNN_{ORG}$, and SRC_{ORG} are based on the original features. The other five systems, SVM_{MPC}^{OAO} , SVM_{MPC}^{OAA} , kNN_{MPC} , $RBFNN_{MPC}$, and SRC_{MPC} are based on the MPC features. The last one is MPC-FER; for this system we use binary SVMs and multiclass SVMs to induce h^B s and h^M s, respectively. For the clustering step of the MPC-FER, self organizing map (SOM) [56] is employed (see [1] for more information).

5.4 Protocol

To assess the performance of the facial expression recognition systems on a given dataset, we perform tenfold cross-validation strategy. That is, we divide the samples of a given dataset into 10 disjoint and equally sized subsets, and we use nine subsets as the training sets and one subset as the testing set. This procedure is repeated 10 times using one subset exactly once as the testing set. The obtained results from 10 runs are averaged at the end.



Fig. 4 Samples of facial expressions images. The *first two rows* are samples of Cohn–Kanade, the *second two rows* are from JAFFE, and the *last two rows* are samples of TFEID

6 Results

The results are broken down by the dataset category and are presented in Tables 2, 3 and 4. The demonstrated results also include the average performances and the improvements in the performances achieved by the MPC features on the same classifiers. Note that the demonstrated averages in the tables are the overall performances of the classifiers, not the averages of the columns. The best rates are in bold face.

From the results, it can be seen that although the FFP features used in Cohn–Kanade dataset have been carefully labeled, and as a result a good performance was expected for them, the LBP features achieved better performance than FFPs. The average classification rates for LBP and FFP features are 81.2 and 75.6 %, respectively³. In contrast, it can be seen that the average classification accuracies of LBP and FFP features are almost the same (80.1 % for LBPs and 79.4 % for FFPs) on JAFFE dataset. The reason can be found in the fact that the number of fiducial points (59 points) used in the Cohn–Kanade dataset may not be as appropriate as we hoped for. Also, their positions⁴ in different expressions may overlap or be very close, which can cause the classifiers to suffer from over-fitting. We, therefore, can conclude that choosing the right number of the fiducial points and their positions are the key points for FFP information representation approach.

However the length of feature vector in LBP is almost one-eighteenth of the GW features, the LBP features outperform GW features in Cohn–Kanade and TFEID datasets, and it is slightly lower than GW features on JAFFE dataset (1.3 % on average). Therefore, choosing LBP features has advantages and is effective for both facial expression representation and recognition. From the results, we can also conclude that the orthogonal moments, ZM in particular, compared to GW, have a good potential to be used for facial expression representation, as the feature vector length in ZM is very low and its performance is somewhat comparable to that of GW features.

As it has been mentioned, the MPC features consists of meta features that are enriched by class-wise similarity, while the original features are simply extracted from the instances and no further processing has been done to enhance them. Therefore, as it was expected and the results clearly demonstrate, the MPC features outperform the original features in most of the cases from which the

³ Since the classification rates obtained by SRC using ZM and FFP features are very low, we do not include all of its results in the averaging processes throughout this section in order to avoid unfair comparisons.

⁴ The number of fiducial points (34) and their positions in JAFFE dataset, as shown in Fig. 3, are different from those of Cohn–Kanade dataset.

Table 2 Classification accuracy (%) of various FER systems on Cohn–Kanade dataset

	Trained on ORG features					Trained on MPC features					MPC-FER
	SVM ^{OAo}	SVM ^{OAA}	kNN	RBFNN	SRC	SVM ^{OAo}	SVM ^{OAA}	kNN	RBFNN	SRC	
LBP											
Anger	61.11	69.44	19.44	58.33	50.00	63.89	72.22	69.44	75.00	55.56	72.22
Disgust	82.50	82.50	52.50	70.00	72.50	80.00	82.50	82.50	90.00	52.50	85.00
Fear	57.58	63.64	18.18	30.30	51.52	63.64	72.73	54.55	60.61	30.30	75.76
Happiness	94.05	94.05	79.76	95.24	91.67	95.24	91.67	92.86	94.05	55.95	94.05
Neutral	90.72	92.78	68.04	82.47	87.63	89.69	88.66	87.63	89.69	24.74	88.66
Sadness	71.43	71.43	07.14	66.67	64.29	69.05	73.81	71.43	80.95	42.86	73.81
Surprise	94.67	94.67	81.33	86.67	92.00	96.00	94.67	97.33	96.00	54.67	94.67
Average	84.03	85.75	56.76	76.66	79.12	84.52	85.50	84.03	87.22	44.47	86.49
IMPC ^a						+0.49	-0.25	+27.27	+10.56	-34.65	
GW											
Anger	50.00	50.00	13.89	66.67	55.56	47.22	52.78	58.33	41.67	27.78	50.00
Disgust	60.00	60.00	15.00	40.00	47.50	62.50	65.00	57.50	65.00	42.50	62.50
Fear	48.48	48.48	0.00	45.45	78.79	51.52	45.45	30.30	27.27	30.30	48.48
Happiness	85.71	85.71	76.19	70.24	94.05	88.10	86.90	92.86	86.90	40.48	83.33
Neutral	78.35	78.57	56.12	70.41	47.42	80.61	82.47	82.65	78.57	40.82	81.63
Sadness	52.38	54.76	14.29	58.54	69.05	54.76	54.76	45.24	42.86	40.48	61.90
Surprise	96.00	96.00	84.00	81.33	94.64	94.67	93.33	96.00	98.67	56.00	97.33
Average	73.71	74.02	48.77	65.85	71.25	75.00	75.18	74.51	71.57	41.67	75.49
IMPC						+1.29	+1.16	+25.74	+5.72	-29.58	
ZM											
Anger	52.78	50.00	25.00	44.44	22.22	41.67	47.22	47.22	41.67	47.42	47.22
Disgust	70.00	62.50	55.00	52.50	07.50	62.50	62.50	62.50	65.00	55.00	62.50
Fear	45.45	39.39	06.06	27.27	18.18	48.48	54.55	39.39	36.36	36.36	51.52
Happiness	85.71	85.71	83.33	88.10	19.05	90.48	86.90	85.71	95.24	57.14	90.48
Neutral	81.44	79.38	87.63	83.51	18.56	80.41	79.38	81.44	83.51	26.80	83.51
Sadness	42.86	40.48	30.95	07.14	19.05	52.38	50.00	52.38	52.38	42.86	57.14
Surprise	89.33	93.33	86.67	92.00	29.33	93.33	90.67	94.67	97.33	66.67	93.33
Average	73.22	71.74	65.36	67.08	19.90	74.20	73.46	73.46	75.92	47.42	76.17
IMPC						+0.98	+1.72	+8.10	+8.84	+27.52	
FFP											
Anger	63.89	55.56	05.56	55.56	00.00	66.67	61.11	58.33	66.67	38.89	63.89
Disgust	62.50	80.00	02.50	42.50	15.00	77.50	77.50	80.00	82.50	47.50	78.57
Fear	54.55	54.55	00.00	60.61	06.06	57.58	48.48	51.52	60.61	36.36	64.52
Happiness	90.48	90.48	33.33	54.76	07.14	95.24	94.05	92.86	91.67	48.81	94.05
Neutral	89.69	96.91	37.11	59.79	09.28	89.69	87.63	91.75	88.66	28.87	92.78
Sadness	66.67	76.19	00.00	54.76	30.95	76.19	76.19	73.81	76.19	57.14	80.95
Surprise	96.00	98.67	70.67	84.00	26.67	100	98.67	98.67	97.33	49.33	98.67
Average	80.84	85.01	29.48	60.69	13.76	85.50	83.29	84.03	84.77	43.00	86.73
IMPC						+4.66	-1.72	+54.55	+24.08	+29.24	

^a Improvement/decline by the MPC features

improvements in kNN and RBFNN-based systems are more noticeable; on average kNN_{MPC} and RBFNN_{MPC}, respectively, improved 28.9 and 14.3 % classification accuracy in Cohn–Kanade dataset, 23.8 and 8.8 % in JAFFE dataset, and 15.9 and 5.6 % in TFEID dataset.

However the performance of SRC using LBP and GW features is comparable to the other classifiers, its performance drops dramatically when using ZM and FFP features. This is due to a very few number of features in these two techniques (70 and 59 features for ZM and FFP,

Table 3 Classification accuracy (%) of various FER systems on JAFFE dataset

	Trained on ORG features					Trained on MPC features					MPC-FER
	SVM ^{OAo}	SVM ^{OAA}	kNN	RBFNN	SRC	SVM ^{OAo}	SVM ^{OAA}	kNN	RBFNN	SRC	
LBP											
Anger	83.33	80.00	43.33	73.33	83.33	83.33	73.33	76.67	80.00	23.13	80.00
Disgust	79.31	82.76	65.52	68.97	79.31	86.21	86.21	89.66	68.97	13.79	86.21
Fear	75.00	78.13	71.88	71.88	87.50	71.88	78.13	78.13	78.13	28.13	78.13
Happiness	87.10	96.77	74.19	90.32	90.32	100	100	100	100	03.23	96.77
Neutral	66.67	90.00	63.33	76.67	93.33	86.67	83.33	83.33	86.67	10.00	83.33
Sadness	83.87	67.74	45.16	77.42	61.29	64.52	77.42	54.84	74.19	32.26	80.65
Surprise	96.67	96.67	40.00	90.00	90.00	93.33	90.00	90.00	93.33	00.00	100
Average	81.69	84.51	57.75	78.40	83.57	83.57	84.04	81.69	83.10	15.96	86.38
IMPC						+1.88	-0.47	+23.94	+4.70	-67.61	
GW											
Anger	90.00	90.00	33.33	79.31	80.00	90.00	86.67	90.00	90.00	50.00	86.67
Disgust	79.31	79.31	37.93	66.67	72.41	79.31	86.21	82.76	79.31	41.38	89.66
Fear	96.88	87.50	65.63	59.38	87.50	93.75	96.88	84.38	87.50	62.50	81.25
Happiness	83.87	90.32	64.52	74.19	61.29	83.87	87.10	87.10	90.32	51.61	100
Neutral	93.33	80.00	76.67	70.00	90.00	96.67	93.33	90.00	96.67	50.00	86.67
Sadness	77.42	74.19	38.71	70.97	61.29	74.19	77.42	70.97	70.97	45.16	74.19
Surprise	80.00	96.55	60.00	86.67	73.33	93.33	90.00	90.00	93.33	36.67	96.67
Average	85.92	85.38	53.99	72.30	75.12	87.32	88.26	84.98	86.85	48.36	87.79
IMPC						+1.4	+2.88	+30.99	+14.55	-26.76	
ZM											
Anger	80.00	73.33	76.67	80.00	16.67	83.33	80.00	83.33	80.00	60.00	86.67
Disgust	72.41	82.76	62.07	62.07	06.90	79.31	79.31	82.76	79.31	55.17	89.66
Fear	78.13	96.88	78.13	71.88	00.00	78.13	84.38	81.25	78.13	31.25	75.00
Happiness	90.32	87.10	67.74	80.65	25.81	87.10	90.32	90.32	87.10	29.03	93.55
Neutral	93.33	96.67	60.00	70.00	16.67	90.00	96.67	96.67	93.33	33.33	83.33
Sadness	58.06	61.29	38.71	64.52	09.68	61.29	54.84	45.16	70.97	32.26	61.29
Surprise	83.33	90.00	70.00	83.33	13.33	90.00	90.00	93.33	83.33	53.33	96.67
Average	79.34	84.04	64.79	73.24	12.68	81.22	82.16	81.69	81.69	41.78	83.57
IMPC						+1.88	-1.88	+16.9	+8.45	+29.10	
FFP											
Anger	76.67	73.33	66.67	76.67	13.33	73.33	83.33	76.67	73.33	40.00	76.67
Disgust	65.52	62.07	65.52	68.97	06.90	75.86	72.41	75.86	68.97	41.38	89.66
Fear	84.38	87.50	68.75	87.50	21.88	84.38	87.50	84.38	78.13	31.25	84.38
Happiness	87.10	83.87	58.06	67.74	03.23	80.65	83.87	90.32	90.32	19.35	87.10
Neutral	93.33	90.00	53.33	63.33	30.00	90.00	93.33	93.33	93.33	30.00	93.33
Sadness	77.42	77.42	48.39	74.19	22.58	80.65	77.42	80.65	83.87	38.71	67.74
Surprise	80.00	86.67	66.67	86.67	20.00	93.33	83.33	90.00	90.00	63.33	96.67
Average	80.75	80.28	61.03	75.12	16.90	82.63	83.10	84.51	82.63	37.56	84.98
IMPC						+1.88	+2.82	+23.48	+7.51	+20.66	

respectively). This problem has been pointed out in [18], where the dimensionality of feature space has been indicated as one of the critical points of SRC, i.e., the number of features should be sufficiently large for a given classification problem. The same reason is true for the SRC_{MPC}

systems, as the number of MPC features is very low (21 features).

Only from the classification point of view can we see that SVMs with OAA strategy are performing better than SVMs with OAO strategy; on average, SVM_{ORG}^{OAA} and

Table 4 Classification accuracy (%) of various FER systems on TFEID dataset

	Trained on ORG features					Trained on MPC features					MPC-FER
	SVM ^{OA0}	SVM ^{OAA}	kNN	RBFNN	SRC	SVM ^{OA0}	SVM ^{OAA}	kNN	RBFNN	SRC	
LBP											
Anger	79.41	88.24	47.06	73.53	82.35	85.29	85.29	82.35	85.29	35.29	94.12
Disgust	90.00	87.50	75.00	95.00	90.00	95.00	97.50	95.00	97.50	07.50	90.00
Fear	82.50	85.00	35.00	80.00	82.50	77.50	80.00	82.50	82.50	07.50	85.00
Happiness	97.50	100	97.50	100	100	100	100	97.50	97.50	10.00	100
Neutral	97.44	100	87.18	84.62	100	97.44	97.44	97.44	97.44	30.77	97.44
Sadness	76.92	76.92	41.03	58.97	74.36	79.49	84.62	76.92	76.92	15.38	82.05
Surprise	97.22	97.22	91.67	100	97.22	100	100	100	100	55.56	100
Average	88.81	90.67	67.91	84.70	89.55	90.67	92.16	90.30	91.04	22.39	92.54
IMPC						+1.86	+1.49	+22.39	+6.34	-67.16	
GW											
Anger	76.47	70.59	58.82	70.59	70.59	76.47	79.41	73.53	76.47	41.18	85.29
Disgust	82.50	87.50	75.00	70.00	77.50	77.50	85.00	82.50	85.00	50.00	82.50
Fear	60.00	72.50	40.00	57.50	70.00	65.00	70.00	62.50	67.50	22.50	65.00
Happiness	87.50	87.50	95.00	95.00	77.50	95.00	97.50	97.50	95.00	52.50	90.00
Neutral	76.92	79.49	66.67	61.54	87.18	84.62	71.79	71.79	76.92	12.82	92.31
Sadness	66.67	64.10	35.90	71.79	61.54	58.97	76.92	58.97	61.54	23.08	71.79
Surprise	97.22	100	91.67	94.44	97.22	88.89	100	94.44	94.44	44.44	100
Average	77.99	80.22	66.04	74.25	77.24	77.99	82.84	77.24	79.48	35.07	83.58
IMPC						0	+2.62	+11.2	+5.23	-42.17	
ZM											
Anger	85.29	76.47	55.88	70.59	17.65	76.47	79.41	73.53	79.41	14.71	76.47
Disgust	87.50	85.00	70.00	87.50	12.50	90.00	90.00	90.00	90.00	30.00	87.50
Fear	65.00	72.50	32.50	69.23	25.00	72.50	67.50	67.50	65.00	25.00	77.50
Happiness	95.00	95.00	97.50	100	12.50	97.50	97.50	97.50	97.50	45.00	90.00
Neutral	76.92	74.36	84.62	74.36	17.95	84.62	82.05	76.92	82.05	38.46	79.49
Sadness	56.41	64.10	17.95	35.00	17.95	53.85	58.97	51.28	58.97	28.21	71.79
Surprise	94.44	97.22	100	100	05.56	100	100	100	99.99	55.56	97.22
Average	79.85	80.60	65.30	76.49	15.67	82.09	82.09	79.48	81.72	33.96	82.84
IMPC						+2.24	+1.49	+14.18	+5.23	+18.29	

SVM^{OAA}_{MPC} improve 2.1 and 1.0 % the performances of SVM^{OA0}_{ORG} and SVM^{OA0}_{MPC}, respectively,⁵. Moreover, considering the MPC-FER as a complex classifier in which it uses clustering and classification together, one can see that this method improves the performances of SVM^{OAA}_{MPC} and SVM^{OAA}_{MPC} about 2.2 and 1.2 %, respectively.

As Table 2 shows, the best overall performance on Cohn–Kanade dataset (87.2 %) belongs to RBFNN_{MPC}, where the original features used to produce MPCs are LBPs, indicated by RBFNN_{MPC}(LBP). The SVM^{OAA}_{MPC}(GW), with an average accuracy of 88.3 % on JAFFE dataset, is the dominating system (Table 3), and as Table 4 demonstrates, MPC-FER(LBP) is the best system among the other systems with an

⁵ In the averaging process only the results of the systems using LBP features are considered.

average accuracy of 92.5 % on TFEID dataset. The confusion matrices of the best performing systems on Cohn–Kanade, JAFFE and TFEID datasets are provided in Tables 5, 6 and 7, respectively.

From the confusion matrices, we can observe that the two most confused expressions over all the datasets are Neutral and Sadness. By contrast, Surprise is the least confused expression, and the highest recognition rate, considering only the best performing systems, also belongs to Surprise with an average accuracy of 95.3 %.

6.1 Statistical comparison of the FER systems

In order to statistically compare the performance of the facial expression recognition systems, we follow the two-step procedure recommended by Demšar [11]. The first

step is to accept or reject the null-hypothesis. The null-hypothesis indicates that the performances of the systems of interest are the same and there are no significant differences between their performances. If the null-hypothesis is rejected, we proceed to our comparison with a post hoc test to analyze the results in more detail.

In the first step, which is called the Friedman test, average ranks, $R_j = \frac{1}{N} \sum_i r_i^j$, are calculated for every system, where r_i^j is j th system's rank on i th dataset. In case of tie, the average rank is assigned to r_i^j . Table 8 shows a summary of the classification accuracies of the FER systems along with the assigned ranks.

Once the average ranks are assigned, the Friedman statistic is computed as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right], \quad j = 1, \dots, k, \quad (10)$$

where k and N indicate the number of classifiers and datasets, respectively. In our experiments, the value of k and N are both equal to 11. Substituting in Eq. (10), we

obtain the Friedman statistic with a value of 91.01. It has been shown that when k and N are not large enough, the Friedman statistic is not appropriate and it is undesirably conservative, thus, the following correction has been proposed [57]:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2}. \quad (11)$$

The F_F statistic is distributed according to the F-distribution with $(k-1)$ and $(k-1) \times (N-1)$ degrees of freedom. Substituting the value of χ_F^2 in Eq. (11), we obtain $F_F = 47.94$. The critical value for $F(10,100)$ with a significance level of $\alpha = 0.05$ is 1.93. Therefore, we can quite safely reject the null-hypothesis ($F_F > 1.93$), which is to say that the performances of the systems of interest are not the same.

The next step is to study the differences between the performances in detail. To this end, a step-down procedure, introduced by Holm [58], is applied. In this test, the hypotheses (systems) are sorted in an ascending manner according to their significance value, p_i , and are then

Table 5 Confusion matrix of RBFNN_{MPC}(LBP) on Cohn–Kanade dataset

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	27	7			1	1	
Disgust		36			4		
Fear	1	2	20		1	5	4
Happiness				79	3		2
Neutral	2		2	1	87	4	1
Sadness	1	1			6	34	
Surprise				3			72

Table 6 Confusion matrix of SVM_{MPC}^{OAA} (GW) on JAFFE dataset

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	26	2				2	
Disgust	2	25				2	
Fear			31			1	
Happiness				27	3	1	
Neutral					28	1	1
Sadness	1	1	2	1	2	24	
Surprise			2	1			27

Table 7 Confusion matrix of MPC-FER (LBP) on TFEID dataset

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	32			1		1	
Disgust	1	36	1		1	1	
Fear		1	34	2	1	2	
Happiness				40			
Neutral	1				38		
Sadness		2	2		3	32	
Surprise							36

Table 8 Summary of the FER systems’ performances. The ranks are indicated in the parentheses

	Trained on ORG features					Trained on MPC features					MPC-FER
	SVM ^{OA0}	SVM ^{OAA}	kNN	RBFNN	SRC	SVM ^{OA0}	SVM ^{OAA}	kNN	RBFNN	SRC	
CK											
LBP	84.0 (6.5)	85.7 (3)	56.8 (10)	76.7 (9)	79.1(8)	84.5 (5)	85.5 (4)	84.0 (6.5)	87.2(1)	44.4 (11)	86.5 (2)
GW	73.7 (6)	74.0 (5)	48.8 (10)	65.8 (9)	71.2(8)	75.0 (3)	75.2 (2)	74.5 (4)	71.6(7)	41.6 (11)	75.5 (1)
ZM	73.2 (6)	71.7 (7)	65.4 (9)	67.1 (8)	19.9(11)	74.2 (3)	73.5 (4.5)	73.5 (4.5)	75.9(2)	47.4 (10)	76.1 (1)
FFP	80.8 (7)	85.0 (3)	29.5 (10)	60.7 (8)	13.7(11)	85.5 (2)	83.3 (6)	84.0 (5)	84.8(4)	43.0 (9)	86.7 (1)
JAFFE											
LBP	81.7 (7.5)	84.5 (2)	57.7 (10)	78.4 (9)	83.6(4.5)	83.6 (4.5)	84.0 (3)	81.7 (7.5)	83.1(6)	15.9 (11)	86.4 (1)
GW	85.9 (5)	85.4 (6)	54.0 (10)	72.3 (9)	75.1(8)	87.3 (3)	88.3 (1)	85.0 (7)	86.8(4)	48.3 (11)	87.8 (2)
ZM	79.3 (7)	84.0 (1)	64.8 (9)	73.2 (8)	12.6(11)	81.2 (6)	82.1 (3)	81.7 (4.5)	81.7(4.5)	41.7 (10)	83.6 (2)
FFP	80.75 (6)	80.3 (7)	61.0 (9)	75.1 (8)	15.9(11)	82.6 (4.5)	83.1 (3)	84.5 (2)	82.6(4.5)	37.5 (10)	85.0 (1)
TFEID											
LBP	88.8 (8)	90.7 (4.5)	67.9 (10)	84.7 (9)	89.5(7)	90.7 (4.5)	92.2 (2)	90.3 (6)	91.0(3)	22.3 (11)	92.5 (1)
GW	78.0 (5.5)	80.2 (3)	66.0 (10)	74.2 (9)	77.2(7.5)	78.0 (5.5)	82.8 (2)	77.2 (7.5)	79.4(4)	35.0 (11)	83.6 (1)
ZM	79.8 (6)	80.6 (5)	65.0 (9)	76.4 (8)	15.6(11)	82.1 (2.5)	82.1 (2.5)	79.4 (7)	81.7(4)	33.9 (10)	82.8 (1)
Average rank (<i>R</i>)	6.409	4.227	9.636	8.545	8.909	3.955	3.000	5.591	4.000	10.455	1.273

sequentially tested by comparing p_i with the adjusted α , i.e., $\alpha/(k - i)$. If p_i is below adjusted α , we reject the corresponding hypothesis and proceed to examine the next hypothesis. Once a certain null-hypothesis cannot be rejected, we hold the remaining hypothesis.

To calculate the significance value p for each and every system, z statistic is computed as follows:

$$z_j = (R_0 - R_j) \sqrt{\frac{k(k + 1)}{6N}}, \quad j = 0, \dots, k. \tag{12}$$

where R_0 is the average rank of the system that we are interested in the comparison of its performance with the other systems. Then, the value of p is found from the normal distribution table based on z . Note that the value of p is multiplied by two, as a two-tailed test is used.

In order to have meaningful p values, we only consider comparing the seven best ranked systems among the 11 systems, which are as follows: SVM^{OA0}_{ORG}, SVM^{OAA}_{ORG}, SVM^{OA0}_{MPC}, SVM^{OAA}_{MPC}, RBFNN_{MPC}, kNN_{MPC} and MPC-FER (see Table 8). We select the SVM^{OA0}_{ORG} as the controller system with an average rank of $R_0 = 6.409$. Table 9 shows the ordered hypothesis according to their p values. This table also includes corresponding z statistics and adjusted α s.

According to Table 9, the first null-hypothesis is rejected as its p value (0.0003) is below the adjusted α (0.0083). This is to say that the MPC-FER outperforms the other systems and the difference between its performance and the other remaining systems is statistically significant. The remaining hypotheses are retained, as the p value of the next hypothesis, SVM^{OAA}_{MPC}, is greater than the adjusted α (0.0100).

6.2 Visualizing the effectiveness of MPC

As it pointed out in [59], compactness and separability of regions in the input feature space are two basic assumptions for a given pattern recognition problem. So, the more compact and separate the patterns, the better classification performance it will be. In this section, to further the empirical results, we aim to visually demonstrate the effectiveness of MPC in terms of compactness and separability. To this end, a two-dimensional SOM network is used to map features’ spaces to 2D spaces, so that we can plot 2D maps of the MPC features and the original features in order to visually study their effectiveness. The size of the SOM network is chosen to be 200×200 , and the Euclidean distance is used as a distance measure. Figure 5 shows the resulting clusters as 2D maps generated by SOM on Cohn–Kanade dataset. The demonstrated results are drawn from one run of tenfold cross validation.

Table 9 Ordered hypotheses in an ascending manner according to their p values ($k = 7$ and $\alpha = 0.05$)

<i>i</i>	System	<i>z</i>	<i>p</i>	Adjusted α [i.e., $\alpha/(k - i)$]
1	MPC-FER	3.6320	0.0003	0.0083
2	SVM ^{OAA} _{MPC}	2.4106	0.0159	0.0100
3	SVM ^{OA0} _{MPC}	1.7356	0.0826	0.0125
4	RBFNN _{MPC}	1.7035	0.0885	0.0166
5	SVM ^{OAA} _{ORG}	1.5428	0.1229	0.0250
6	kNN _{MPC}	0.5785	0.5629	0.0500

Figure 5 clearly demonstrates that the MPC features are clustered very well and their compactness and separability are markedly better than the corresponding original features. It indicates that using pair-wise class similarity as a feature will result in more homogenous features for a group of the same facial expressions. As a result, the classifiers, which are trained based on the MPC features, and the clustered MPC features in particular, will have a better performance, as the empirical results have already shown.

6.3 Generalization performance on across datasets

In this section our goal is to evaluate the generalization performance of the facial expression recognition systems in a more challenging manner, one in which is more likely to appear in real-world applications. This evaluation is an across dataset evaluation where the training and the testing sets are not from the same dataset [23]. More precisely, we use one dataset among the introduced datasets as a training set and the other datasets as the testing sets. To this end, the Cohn–Kanade dataset is chosen as the training dataset and all the selected samples (totally 407 samples) from this dataset are used to train the systems. For the evaluation purpose, we use the JAFFE and TFEID datasets as the testing sets. For these datasets, we also use all the samples that we used in the previous experiments. In this experiment, we skip evaluating the RBFNN and k NN-based systems. The results are presented in Table 10.

Table 10 shows that the MPC-based systems are again performing better than the ORG-based systems on the across datasets evaluation; considering the LBP features as the best original features, MPC-FER improves the best recognition rates by 6.7 and 3.0 % on JAFFE and TFEID datasets, respectively. We, therefore, can conclude that MPCs contain more informative features than the original features that help the classifiers to be trained with a better

generalization for unseen samples from different datasets. However, the generalization performances of the systems using original GW features are better than the systems using GW-based MPC features. This is because of the huge number of the features in GW that may result in high-variance h^B s [60]. As a consequence, the produced MPC features may not contain generalized pair-wise class similarities, and the classifiers trained on these features may not have an acceptable performance on different datasets accordingly. We can also observe that the performances of the systems using original LBP features are better than the other ORG-based systems. This clearly indicates that the LBP, compared to GW and ZM, is the dominating information representation approach.

As can be seen from Table 10, the results on TFEID dataset are better than those of JAFFE. This is due to the fact that the samples demonstrating expressions in TFEID are more authentic than the samples of JAFFE as some of the subjects in JAFFE did not perform the requested emotions correctly or perspicuously enough [24].

6.4 Comparison with other methods

In order to fairly compare the performance of the proposed framework with one of the most recent works introduced in [29], GSNMF, we follow the experimental setup used in [29] and report the results (as *our method*) in Table 1 to ease the comparison. The experimental setup is as follows [29]; a subset of 30 individuals of Cohn–Kanade dataset is selected and only six expressions (excluding neutral) are considered. Then, the training set is composed using one of the last eight peak frames of each sequence and the remaining frames are used to compose the testing set.

To avoid any effects of one single run, we repeat the aforementioned procedure 10 times in which 30 individuals are randomly selected at each run and the results are

Fig. 5 2D maps of generated clusters on Cohn–Kanade dataset using SOM. In the demonstrated maps, anger, disgust, fear, happiness, neutral, sadness and surprise are indicated by red, gray, yellow, orange, green, blue and pink, respectively

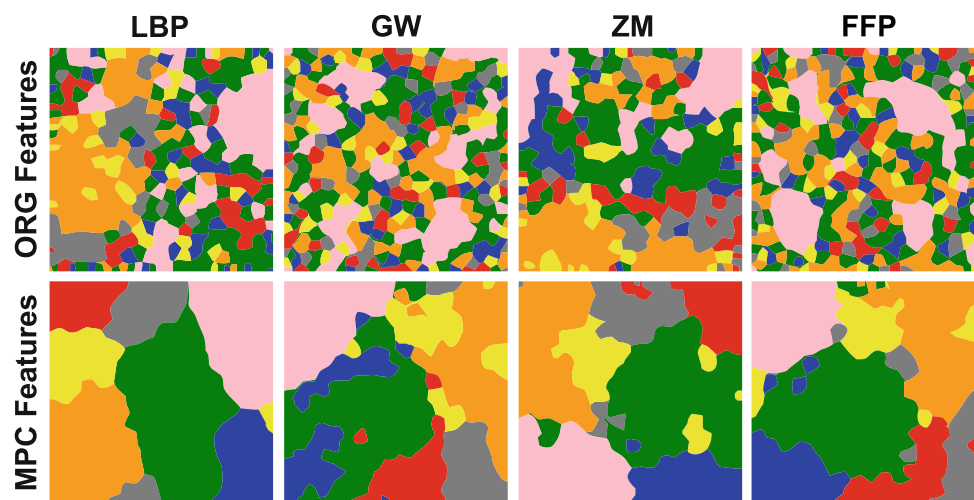


Table 10 Generalization performances of the systems on across datasets; trained on Cohn–Kanade dataset and tested on JAFFE and TFEID datasets

The improvements/declines in the performances achieved by the MPC features on the same classifiers are indicated in the parentheses

	SVM ^{OAo} _{ORG}	SVM ^{OAA} _{ORG}	SVM ^{OAo} _{MPC}	SVM ^{OAA} _{MPC}	MPC-FER
JAFFE					
LBP	38.50	34.27	47.42 (+08.92)	45.54 (+11.27)	45.19
GW	21.13	25.82	20.66 (−00.47)	23.00 (−02.82)	20.66
ZM	35.68	32.39	44.13 (+08.45)	42.72 (+10.36)	40.29
TFEID					
LBP	48.13	47.39	55.60 (+07.47)	54.10 (+06.71)	51.13
GW	24.25	30.22	20.52 (−03.73)	25.37 (−04.85)	24.12
ZM	44.40	43.28	48.51 (+04.11)	46.27 (+02.99)	45.76

averaged at the end. And to derive the MPC features, we use the LBP features as the original features.

As the results show in Table 1, the performance of two systems based on the proposed framework, namely MPC-FER and SVM^{OAA}_{MPC}, are comparable with the performance of GSNMF, where their performances are, respectively, only 0.2 and 0.4 % lower than GSNMF.

7 Conclusions and future work

The purpose of the MPC approach is to derive a set of new discriminative and informative features from the original features by the means of pair-wise class similarities. In this paper, we studied and assessed the effectiveness of the MPC features for the representation and recognition of facial expressions via an MPC-based framework. In the experiments, we introduced 11 different systems from which five were based on the original features. The other five were based on the derived MPC features, and the last one was MPC-FER. The original features used in the experiments were LBP, GW, ZM and FFP, and the classification algorithms included SVM^{OAo}, SVM^{OAA}, *k*NN, RBFNN and SRC. Based on the extensive experiments conducted on three publicly available datasets, Cohn–Kanade, JAFFE, and TFEID, we draw our conclusions as follows:

- It was observed that among the original features of interest, LBP features were the dominating features for the representation of facial expressions. It was also observed that the MPC features, derived from the LBP features, outperformed the other MPC features.
- The results indicated that the MPC features improved the classification accuracy in most of the cases, among which the improvements in *k*NN and RBFNN-based systems were remarkable.
- We statistically showed that the MPC features improved the performance of automatic facial expression recognition significantly. It was also shown that

the MPC features markedly improved the generalization performance on across dataset evaluation.

- Finally, from the classification point of view, we observed that the cluster-based classifier and the SVM with OAA strategy preformed better than *k*NN, RBFNN, SVM with OAo strategy and SRC.

In this study, we used several basic information representation approaches as the original features and derived the MPC features based on them accordingly. However, it is of interest to see how well the performance of facial expression recognition can be improved when some of the enhanced information representation approaches such as; boosted-GW [23], boosted-LBP [25], GMFA [61], boosted-WM [62], etc., are used to derive the MPC features. For example, Littlewort et al. [23] used Adaboost to select GW features, and they improved the recognition rate of their system by 5.3 %. Shan et al. [25] showed that the boosted-LBP features, compared to the LBP features, improved the classification accuracy by about 2.5 %. In [61], Wang and Guo used a Gabor-based marginal Fisher analysis (GMFA) approach to enhance the GWs, and they improved the classification accuracy of GW + LDA + *k*NN system by 1.4 and 3.6 % on ORL and FERET datasets, respectively. In another work [62], the authors used wavelet moment (WM) invariants to represent facial expressions and AdaBoost to select effective features. Their results indicated that the performance of the FER system on JAFFE dataset using boosted-WM improved by 4.9 and 1.2 % compared to the systems using GW features and ZM features, respectively. Hence, as our future work, we are motivated to study and examine the effect of some of the enhanced features as the original features on the performance of the MPC-based FER system (step 1). We will also study the effect of different combinations of the classifiers (2nd and 3rd steps) on the performance of the MPC-based FER systems.

Acknowledgements We would like to thank Prof. Guodong Guo for providing us the facial fiducial points of JAFFE dataset. This work is partly supported by NSF of China (No. 61070067) and National 973 Program (2013CB329504).

References

1. Farajzadeh N, Pan G, Wu Z, Yao M (2011) Multiclass classification based on meta probability codes. *Int J Pattern Recognit Artif Intell* 25(8):1219–1241
2. Pan G, Xu Y, Wu Z, Li S, Yang LT, Lin M, Liu Z (2011) TaskShadow: toward seamless task migration across smart environments. *IEEE Intell Syst* 26(3):50–57
3. Sun J, Wu Z, Pan G (2009) Context-aware smart car: from model to prototype. *J Zhejiang Univ Sci A* 10(7):1049–1059
4. Darwin C (1872) *The expression of the emotions in man and animals*. John Murray, London
5. Ekman P, Friesen WV (1974) *Unmasking the face*. Prentice Hall, New Jersey
6. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. *IEEE Trans Pattern Anal Mach Intell* 22(12):1424–1445
7. Fasel B, Luetttin J (2003) Automatic facial expression analysis: a survey. *Pattern Recognit* 36(1):259–275
8. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans Pattern Anal Mach Intell* 31(1):39–58
9. Bettadapura V (2012) *Face expression recognition and analysis: the state of the art*. Technical, College of Computing, Georgia Institute of Technology
10. Sebe N, Cohen I, Grag A, Huang T (2005) *Machine learning in computer vision*. Springer, Berlin
11. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
12. Ojala T, Pietikinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distribution. *Pattern Recognit* 29(1):51–59
13. Lee TS (1996) Image representation using 2D gabor wavelets. *IEEE Trans Pattern Anal Mach Intell* 18(10):1–13
14. Teh CH, Chin RT (1988) On image analysis by the methods of moments. *IEEE Trans Pattern Anal Mach Intell* 10(4):496–513
15. Vapnik VN (1998) *Statistical learning theory*. Wiley, New York
16. Shakhnarovich G, Darrell T, Indyk P (2005) *Nearest-neighbor methods in learning and vision: theory and practice*. MIT Press, Boston
17. Fausett LV (1994) *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall, Upper Saddle River
18. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. *IEEE Trans Pattern Anal Mach Intell* 31(2):210–227
19. Kanade T, Cohn J, Tian Y (2000) Comprehensive database for facial expression analysis. In: *Proceedings of the IEEE international conference on automatic face and gesture recognition*
20. Lyons MJ, Budynek J, Akamatsu S (1999) Automatic classification of single facial images. *IEEE Trans Pattern Anal Mach Intell* 21(12):1357–1362
21. Chen LF, Yen YS (2011) TFEID: Taiwanese facial expression image database, November
22. Yu J, Bhanu B (2006) Evolutionary feature synthesis for facial expression recognition. *Pattern Recognit Lett* 27(11):1289–1298
23. Littlewort G et al (2006) Dynamics of facial expression extracted automatically from video. *Image Vis Comput* 24(6):615–625
24. Feng X, Pietikainen M, Hadid A (2007) Facial expression recognition with local binary patterns. *Pattern Recognit Image Anal* 17(4):592–598
25. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis Comput* 27(6):803–816
26. Xie X, Lam KM (2009) Facial expression recognition based on shape and texture. *Pattern Recognit* 42(5):1003–1011
27. Lajevardi SM, Hussain ZM (2009) Zernike moments for facial expression recognition. In: *Proceedings of International conference on communication, computer and power*, pp 15–18
28. Yang P, Liu Q, Metaxas D (2010) Exploring facial expressions with compositional features. In: *Proceedings of IEEE conference on computer vision and pattern recognition*, pp 2638–2644
29. Zhi R, Flierl M, Ruan Q, Kleijn W (2011) Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Trans Syst Man Cybern Part B Cybern* 41(1):38–52
30. Cotter SF (2010) Sparse representation for accurate classification of corrupted and occluded facial expressions. In: *Proceedings of the international conference on acoustics speech and signal processing*, pp 838–841
31. Kobayashi H, Hara F (1992) Recognition of mixed facial expressions and their strength by neural network. In: *Proceedings of IEEE international workshop on robot and human communication*, pp 381–386
32. Ushida H, Takagi T, Yamaguchi T (1993) Recognition of facial expressions using the conceptual fuzzy sets. In: *Proceedings of the 2nd international conference on fuzzy systems*, pp 594–599
33. Sohail ASM, Bhattacharya P (2007) Classification of facial expressions using *k*-nearest neighbor classifier. In: *Advances in computer vision and computer graphics. Lecture notes in computer science*. 4418:555–566
34. Cheng S, Chen M, Chang H, Chao T (2007) Semantic-based facial expression recognition using analytical hierarchy process. *Expert Syst Appl* 33(1):86–95
35. Yang P, Liu Q, Metaxas D (2011) Dynamic soft encoded patterns for facial event analysis. *Comput Vis Image Underst* 115(3):456–465
36. Buenaposada JM, Muñiz E, Baumela L (2008) Recognising facial expressions in video sequences. *Pattern Anal Appl* 11(2-3):101–116
37. Bolme D, Teixeira M, Beveridge J, Draper B (2003) The CSU face identification evaluation system user's guide: its purpose, feature and structure. In: *Proceedings of the 3rd international conference on computer vision systems*, pp 304–313
38. Wong KW, Lam KM, Siu WC (2001) An efficient algorithm for human face detection and facial feature extraction under different conditions. *Pattern Recognit* 34(10):1993–2004
39. Wolpert DH (1992) Stacked generalization. *Neural Netw* 5:241–259
40. Wu T-F, Lin C-J, Weng RC (2003) Probability estimates for multiclass classification by pairwise coupling. *J Mach Learn Res* 5:975–1005
41. Ahonen T, Hadid A, Pietik M (2004) Face recognition with local binary patterns. In: *Proceedings of the 8th European conference on computer vision*, pp 469–481
42. Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J Opt Soc Am A* 2(7):1160–1169
43. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with Gabor wavelets. In: *Proceedings of the 3rd IEEE international conference on automatic face and gesture recognition*, pp 200–205
44. Shen L, Bai L (2006) A review on gabor wavelets for face recognition. *Pattern Anal Appl* 9(2-3):273–292
45. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 568–573
46. Donato G, Bartlett MS, Hager J, Ekman P, Sejnowski T (1999) Classifying facial actions. *IEEE Trans Pattern Anal Mach Intell* 21(10):974–989

47. Farajzadeh N, Faez K, Pan G (2010) Study on the performance of moments as invariant descriptors for practical face recognition systems. *IET Comput Vis* 4(4):272–285
48. Guo G, Dyer CR (2005) Learning from examples in the small sample case: face expression recognition. *IEEE Trans Syst Man Cybern Part B Cybern* 35(3):477–488
49. Crammer K, Singer Y (2001) On the algorithmic implementation of multiclass svms. *J Mach Learn Res* 2:265–292
50. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines
51. Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Technical Report, Department of Computer Science, National Taiwan University
52. Snyman JA (2005) Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms. Springer, Berlin
53. Donoho DL (2006) Compressed sensing. *IEEE Trans Inf Theory* 52(4):1289–1306
54. Candes EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inf Theory* 52(2):489–509
55. Tsaig Y, Donoho D, Stodden V (2007) Sparselab, May
56. Kohonen T (1997) Self organizing maps, 2nd edn. Springer, Berlin
57. Iman RL, Davenport JM (1980) Approximations of the critical region of the friedman statistic. *Commun Stat A9*:571–595
58. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6(2):65–70
59. Sommer D, Golz M (2002) Multiple training of vector-based neural networks to detect density centers in input space. In: Proceedings of the European symposium on intelligent technologies, hybrid systems and their implementation on smart adaptive systems, pp 135–137
60. Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. *Neural Comput* 4:1–58
61. Wang C, Guo C (2010) Face recognition based on gabor-enhanced manifold learning and SVM. In: Advances in neural networks. Lecture notes in computer science. 6064:184–191
62. Zhi R, Ruan Q (2009) Robust facial expression recognition using selected wavelet moments invariants. In: Proceedings of the global congress on intelligent system, pp 508–512