Fusion of transformed shallow features for facial expression recognition

Fares Bougourzi, Karim Mokrani, Yassine Ruichek, Fadi Dornaika, Abdelkrim Ouafi, Abdelmalik Taleb-Ahmed

IET Image Processing 13/9 (2019) 1479-1489

Fusion of transformed shallow features for facial expression recognition

Fares Bougourzi[1] , Karim Mokrani[1], Yassine Ruichek[2], Fadi Dornaika[3,4], Abdelkrim Ouafi[5], Abdelmalik Taleb-Ahmed[6]

[1]LTII Laboratory, University of Bejaia, Bejaia, Algeria

[2]Le2i FRE2005, CNRS, Arts et Metiers, University of Bourgogne Franche-Comte, UTBM, F-90010 Belfort, France

[3]Department of Computer Science and Artificial Intelligence, University of the Basque Country, San Sebastian, Spain

[4]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

[5]LESIA Laboratory, University of Biskra, Algeria

[6]UPHF Laboratory, IEMN DOAE UMR, 8510 Valenciennes, France

E-mail: faresbougourzi@gmail.com

Abstract: Facial expression conveys important signs about the human affective state, cognitive activity, intention and personality. In fact, the automatic facial expression recognition systems are getting more interest year after year due to its wide range of applications in several interesting fields such as human computer/robot interaction, medical applications, animation and video gaming. In this study, the authors propose to combine between different descriptors features (histogram of oriented gradients, local phase quantisation and binarised statistical image features) after applying principal component analysis on each of them to recognise the six basic expressions and the neutral face from the static images. Their proposed fusion method has been tested on four popular databases which are: JAFFE, MMI, CASIA and CK+, using two different cross-validation schemes: subject independent and leave-one–subject-out. The obtained results show that their method outperforms both the raw features concatenation and state-of-the-art methods.

1 Introduction

The studies of how humans perceive and interpret facial expressions have attracted many disciplines such as neuroscience and psychology to pursuit human mechanisms. These studies have given a rise to several theories of how human encodes, represents and interprets the facial expressions. When the computer vision community first tried to define the problem of the machine analysis of facial expressions, it was only natural to resort to the psychology theories and adopt some of their theories, conventions and coding systems [1].

The studies of facial expression describe two main problems: the analysis of facial muscle actions and the recognition of prototypical facial expressions. In 1967 [2], Ekman developed the facial expression recogniser and the analysis of the facial expressions from photographs of the face muscular movement by electrical stimuli. This work of Ekman led to create the Facial Action Coding System (FACS 1978, [3]) which is based on the anatomical basis of facial action. For the second problem that seeks to recognise the prototypic facial expressions, it considers basic or non-basic emotions. Basic emotions refer to the affected model developed by Ekman and his colleagues, who argued that the production and interpretation of certain expressions are hardwired in our brain and recognised universally [4]. The emotions conveyed by these expressions are modelled with six classes: anger, disgust, fear, happiness, sadness and surprise.

In the last two decades, plenty of algorithms have been designed to detect the prototypic facial expressions from static images or images sequence. These algorithms can be classified, depending on the way features are extracted from the original data, into two pillars: handcrafted and learned. First, the handcrafted algorithms are obtained using a mathematical model which is designed with prior knowledge of certain characteristics which allow the method to overcome specific hurdles. For example, the local phase quantisation (LPQ) [5] descriptor proved its efficiency on blur data. Second, the learned features are widely used in the past few years, especially after the winning of the ImageNet challenge by using convolution neural network architecture called 'AlexNet' [6]. In contrast with the handcrafted methods, the deep-learning architectures can extract the relevant features directly from the data across their layers.

During the occurrence of the facial expressions, there are two kinds of features that appear: deformation and movement of the face components (e.g. eyes, eyebrows, mouth, nose etc.), and the facial appearance changes such as wrinkles, furrows and skin texture changes. The handcrafted feature extraction techniques are categorised into geometric, appearance, or hybrid approach, according to the kind of features that they aim to extract and represent to recognise the facial expression. First, the geometric techniques essentially depend on locating and tracking the facial landmarks; for the static images, the methods use the location of the landmark to measure meaningful distances and angles to recognise the facial expressions, whereas, for the image sequences, the motion of facial landmarks caused by the facial expression occurrence is extracted from frame to frame. Second, the appearance approach ciphers the changes on face texture by using a mathematical relation between the intensity of each pixel and its neighbour's intensities. Finally, the hybrid approach uses a combination of the previous approaches to the features extraction phase.

The facial expression recognition (FER) systems need to overcome many challenges to perfectly recognise facial expressions. The following challenges could badly influence the whole process. The first significant challenge affects the face detection process. Indeed, face detection can be very challenging and this is due to several factors such as the three-dimensional face pose, severe clutter, occlusion and variation of illumination [1, 7–9]. It is very important to cope these challenges concerning the face detection because they fail to detect the face or detect wrongly some of its regions will mislead the system, especially for the systems which depend on or aim to extract the features from specific regions of the face. Other factors affect the feature

extraction process such as illumination variation, the appearance of shadows on some regions of the face, the misalignment problems, the low quality and blurred images [1, 7–9]. The FER systems must effectively deal with these challenges because they directly influence the worthiness of the features. The objective in the decision phase is to distinguish between the facial expressions based on features that contain much irrelevant information and can be influenced by many factors. These factors include external ones such as the presence of glasses, facial hair, head hair and human demographic attributes (age, gender and ethnicity). Even the same person can produce facial expression in different ways for each time, for both intensity and shape [1, 7–9].

In this paper, we propose a fully automatic framework for recognising the basic facial expressions from static images. Our method effectively combines different feature types that are extracted by descriptors possessing different properties. The main contributions of this paper are:

- We propose an efficient way to transform descriptor features into new ones. The new features have low dimensionality with more discriminative power than the raw features.
- Our approach allows combining different feature types which overcome the weakness of the concatenation fusion method, where the different features are not considered equivalently.
- We conduct experiments on several facial expression databases. The obtained results show that our method outperforms state-of-the-art methods.
- We compare two evaluation schemes [the classic subject independent and leave-one–subject-out (LOSO)], to decide which one is better to evaluate and compare FER methods.

This paper is organised as follows: Section 2 summarises some of the previous works. The proposed approach is introduced and detailed in Section 3. In Section 4, we describe the experimental set-up. Section 5 presents the experimental results and comparison with state-of-the-art methods. Finally, we conclude this paper and give some future research directions in Section 6.

2 Related works

In the past decade, FER field has achieved a point of mature due to two main reasons: the first one is the availability of considerable databases, which are collected with plausible scenarios. The second factor is the abundance of algorithms that have achieved high performance on these databases. The developed algorithms aim to recognise the facial expression from two types of data either from images [10–13] or from dynamic image sequences [14–17]. The dynamic methods exploit both temporal and spatial information from a sequence of images; in contrast, the static methods use just the spatial information from a single image. In this section, we will focus on describing the works that were developed to recognise the basic facial expressions from the static images.

In the literature, there are many approaches to extract the facial expression features from the face. The appearance-based feature descriptors are among the most successful methods such as local binary pattern (LBP) [18], local mean binary pattern [19], local Gabor binary patterns [20], LPQ [21], local directional texture pattern (LDTP) [22] and Gabor wavelet [23]. Another feature extraction approach is to use the shape information by using the distances and angles between the facial landmarks. Some of the most successful methods are active appearance model [24], elastic bunch graph matching [17], Kanade–Lucas–Tomasi [25]. Kulkarni *et al.* [26] used eight distances between specific facial features such as inter-eyebrow distance and seven facial muscle contractions such as nose lines to train the neural network. There are many works that used a hybrid approach, which combine the two types of features [23, 27–29]. In addition, there are some techniques such as histogram of oriented gradients (HOG) [10, 30], which extract both types of features. The handcrafted methods are the closest ones to our approach. Usually, these existing approaches use raw descriptor features or concatenate more than one descriptor in order to form the final descriptor. On the other hand, in our proposed method, we transform the raw features to obtain a higher-level representation of each type of feature; in the last step, we concatenate these different high-level features in order to form the final descriptor.

In recent years, most of the dominant methods are deep-learning ones. The strength of this approach is the way of how the features are learnt from the data, which is inspired by the mechanism of the human brain [31]. Some of the recent works that have used deep learning are [11, 32–36].

In [32], Cai *et al.* proposed a new loss function called island loss to enhance the discriminative power of the deeply learnt features. Their IL-CNN architecture is constructed as follows: Conv1, Pool1, Conv2, Pool2, Conv3, Pool1, fully connected (FC), island loss layer and finally a softmax loss layer. Liu *et al.* [35] proposed a deep architecture that was inspired by the psychological theory which states that the expressions can be decomposed into multiple facial action units (AUs). Their AU-inspired deep network (AUDN) architecture is composed of three sequential modules; first, convolution and max-pooling layers to learn the micro-action-pattern (MAP) representation; second, feature grouping by combining correlated MAPs adaptively to simulate larger receptive fields, and finally, a group-wise sub-learning network to obtain higher-level representations. In [11], Mollahosseini *et al.* proposed their own architecture which consists of two convolutional layers each followed by max-pooling and then four inception layers. Meng *et al.* [36] designed a new architecture (termed IACNN) and a new loss function (expression-sensitive contrastive loss). They proposed to use two identical CNN architectures for extracting the expression and identity-related features in parallel to alleviate the inter-subject variations introduced by personal attributes. The used CNN's architecture consists of three convolutional layers, each of which is followed by a parametric rectified linear unit (PReLU) layer. The first two PReLU layers are followed by batch normalisation and max-pooling layers, whereas the third one is followed by two FC layers consisting of 1024 neurones. The final part is softmax and contrastive losses layers.

Deep-learning architectures are facing two main drawbacks: (i) the requirement of a huge labelled dataset and (ii) the high computational cost of both training and testing phases and that requires powerful workstation with large RAM space and powerful GPU. These make finding the optimal hyperparameters a tedious task and lead to an expensive computational cost [37–39]. In contrast, handcrafted methods can find the right trade-off between accuracy and computational efficiency [37].

Face representation is another important component in FER. Generally, there are three common facial representations: the holistic representation extracts the features from all face regions, as used in [10, 18, 19, 40]. The second representation type is the part based that uses some parts from the face which are assumed to have a direct relationship with the recognition of facial expressions. Some works that have used this representation type are [12, 22, 41, 42]. The third one, boosting representation, also has proved its efficiency for the recognition of facial expressions. Shan *et al.* [13] and Bartlett *et al.* [43] used the AdaBoost algorithm [44] to learn the significant patches from Gabor and LBP features, respectively.

For the decision phase, the support vector machine (SVM) is the most used technique. In [13], Shan *et al.* used SVM with one-against-rest technique and grid search for the best hyperparameters for the different kernel functions [linear, polynomial and radial basis function (RBF)]. Ryu *et al.* [22] used SVM for multi-classification using one-against-one technique and searched about the best hyperparameters for the RBF kernel. In [10], the one-against-one strategy was used and RBF kernel with non-linearly separable parameters $C = 1000$ and $\gamma = 0.05$. Happy and Routray [12] used one-against-one SVMs for multi-classification purpose and they selected RBF kernel for its superior classification performance, after several experimental comparisons with linear and polynomial kernels. Another classification option that has been used for FER is two stages based classification. Turan and Lam [42] extracted first the LPQ features from the eyes and the mouth windows. In the first-stage classification, they fed the features to one-versus-all SVM to classify the facial expressions. If the difference between the highest output and the second-highest one is

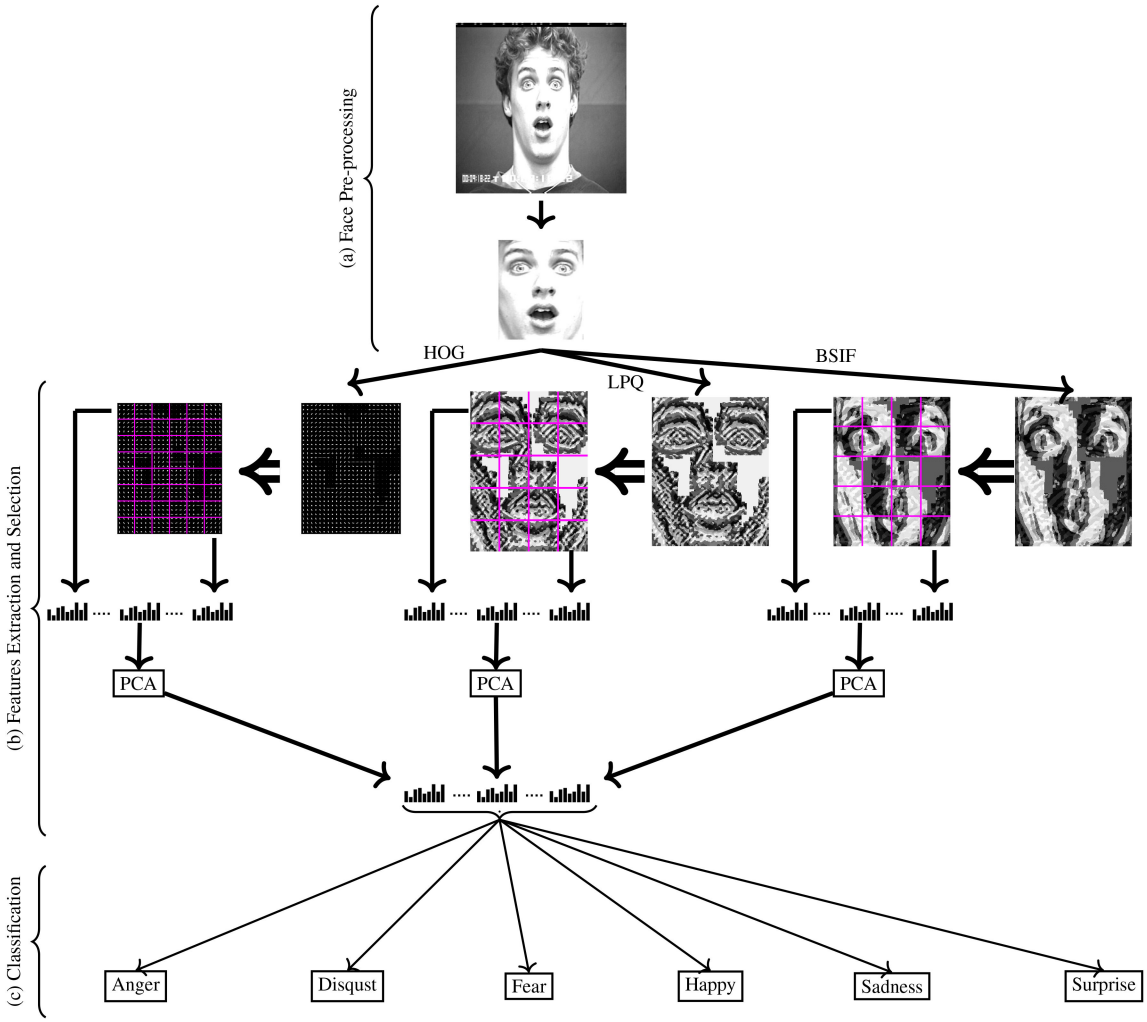


Fig. 1 *General structure of the proposed approach*
(a) Face pre-processing, ***(b)*** Features extraction and selection, ***(c)*** Classification

<0.1, the canonical correlation analysis is used to fuse the features of the two windows, then they fed these fused features to a second classification stage which is one-versus-one SVM. In [45], Xue *et al.* proposed a hierarchical approach which uses one-against-rest SVM and the grid search to pick the best hyperparameters. In the first stage, they fed SVM with LBP and displacement features extracted from the whole face to merge the confused expressions together into one class. In the second stage, the expressions in the merged class are separated by a second SVM which is learnt by the mouth and eyebrows locations and displacements. Some other works used different classifiers and investigated the best one. Wang and Yin [46] used different classifiers which are: quadratic discriminant classifier, linear discriminant analysis (LDA) and Naive Bayesian network classifier and in their experiment LDA classifier achieved the highest accuracy. Sebe *et al.* [47] also compared between different classifiers which are: Bayesian networks, SVMs and decision trees (*k*NN) and they found that the best classifier is *k*NN with $K = 3$.

3 Methodology

3.1 Our approach

In our approach, we propose a fully automatic system to recognise the basic facial expressions from the static images. Fig. 1 describes the overall structure of our approach; as most of FER systems, our system consists of three main steps, which are: (a) pre-processing, (b) feature extraction and selection and (c) expression recognition.

For the pre-processing phase, first, we detected 68 facial fiducial points using Dlib library [48]. The points of the eyes region were used to assign the centre of the eyes that we used to align and crop the face. Then, we resized all the faces into 240×192 and converted them into grey-scale space. In the feature extraction and selection phase, we extracted first the features from the face image using the three descriptors HOG, LPQ and binarised statistical image features (BSIFs). In more details, the selected HOG parameters are: 8 and 32 for cell size and block size without overlapping. The chosen LPQ parameters are: the size of the local window is 13 and the frequency estimation method is the Gaussian derivative quadrature filter pair. For BSIF, we chose $17 \times 17 \times 11$ filter. All the descriptors' parameters are chosen experimentally and they do not significantly deviate from the values used in other image analysis problems. To consider information from face regions, we compute histograms from 5×4 equal face blocks for both LPQ and BSIF and the 32×32 block size for HOG. For each descriptor, the overall histogram is the concatenation of all block histograms. Second, we propose to use the principal component analysis (PCA) method to transform the features into their eigenvectors to have the same feature vector length from different descriptors, and then select the most discriminative features which correspond to the highest variances. The final feature vector is the concatenation of the transformed descriptors features. In the last phase, we feed the resulted histogram to linear SVM using lib-linear library [49].

3.2 Descriptors

The feature extraction phase is considered as the most important step in machine learning tasks. We selected three of the most successful local descriptors in various computer vision problems.

3.2.1 Local phase quantisation: Since the development of LBP [50], a lot of LBP variants have been proposed to deal with

different computer vision tasks [51]. LPQ [5] is one of the most successful LBP variants for computer vision problems including the recognition of facial expression [21]. LPQ is based on quantising the Fourier transform phase in local neighbourhoods to overcome the sensitivity of LBP to image blurring. Ojansivu and Heikkilä [5] proposed the LPQ descriptor which is based on quantising the Fourier transform phase in local neighbourhoods N_x. For each pixel $x = (x_1, x_2)^{\mathrm{T}}$ from the input image f, the local frequency is computed using the short-term Fourier transform by

$$F(u, x) = \sum_{y \in N_x} f(y) w_{\mathrm{R}}(y - x)\, \mathrm{e}^{-\mathrm{j}2\pi u^{\mathrm{T}} y} \tag{1}$$

where $w_{\mathrm{R}}(x)$ is a window function defining the neighbourhood N_x. In the case of regular LPQ, w_{R} is an M by M rectangle given as $w_{\mathrm{R}}(x) = 1$ if $|x_1|$ and $|x_2| < M/2$ and 0 otherwise. LPQ considers four frequencies for u, which are $u_1 = [a, 0]^{\mathrm{T}}$, $u_2 = [0, a]^{\mathrm{T}}$, $u_3 = [a, a]^{\mathrm{T}}$ and $u_4 = [a, -a]^{\mathrm{T}}$, where a is selected as a sufficiently small scalar. So, each pixel position results in a vector

$$F(x) = [F(u_1, x), F(u_2, x), F(u_3, x), F(u_4, x)]\,. \tag{2}$$

The phase information in the Fourier coefficients is directly related to signs of the real and imaginary parts of each component in $F(x)$. LPQ descriptor uses $G(x) = [\mathrm{Re}\{F(x)\}, \mathrm{Im}\{F(x)\}]$, where Re and Im are the real and the imaginary parts of $F(x)$, respectively, to obtain a binary code by

$$q_i = \begin{cases} 1, & \text{if } g_i \geq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where g_j is the jth component of $G(x)$. Finally, the eight binary code is transformed into a number by: $f_{\mathrm{LPQ}(x)} = \sum_{j=1}^{8} (q_j 2^{(j-1)})$.

3.2.2 Binarised statistical image features: BSIF [52] is another LBP variant, which proved its efficiency in many computer vision tasks [51]. To the best of our knowledge, BSIF has not been used for FER yet. In contrast with LBP and LPQ, BSIF uses a fixed set of filters which are automatically learnt from a small set of natural images, instead of using handcrafted filters. The filters are learnt using independent components analyses for estimating the independent components.

The s_i filter response of each pixel (u, v) of the input image is obtained by

$$s_i = \sum_{u, v} W_i(u, v) X(u, v) = w_i^{\mathrm{T}} x, \tag{4}$$

where W_i is the learnt set of filters of size $l \times l \times k$ and X is the image patch of size $l \times l$ pixels that corresponds to the (u, v) pixel. For $i = 1, \ldots, k$, the s_i response is binarised as follows:

$$b_i = \begin{cases} 1, & \text{if } s_i > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

So b_i contains k binary digits, the BSIF code is obtained by $f_{\mathrm{BSIF}(x)} = \sum_{i=1}^{k} (b_i \times 2^{(k-1)})$. As a result, the BSIF feature is a histogram of $(0{:}2^{(k)} - 1)$ codes. The code value of a pixel is considered as a local descriptor of the image intensity pattern in the pixel's surroundings. Furthermore, histograms of pixels' code values allow characterising texture properties within image subregions.

3.2.3 Histogram of oriented gradients: In contrast to LPQ and BSIF which were designed to extract just the appearance features, HOG extracts the shape information. Indeed, the HOG descriptor was originally designed for human detection [53] than it has been extended to various computer vision tasks including FER as in [10, 30]. HOG counts the occurrences of gradient orientation in localised portions of the image. The process includes four steps: first, the horizontal and vertical gradients of the whole image are computed by using two centred derivative masks [1,0,–1] and [1,0,–1], then these gradients are used to compute the magnitude and orientation on each pixel by

$$M_{xy} = \sqrt{(G_x)^2 + (G_y)^2}, \quad \text{and} \tag{6}$$

$$\theta_{xy} = \arctan\left(\frac{G_y}{G_y}\right) \tag{7}$$

where G_x and G_y are the horizontal and vertical gradients of the pixel (x, y). Second, the image is divided into cells. The gradient orientations θ_{xy} of the cell pixels are used to vote into nine corresponding orientation bins equally spaced between 0° and 180°. The vote is weighed by the gradient magnitude M. The third step of HOG is to group the cells together into larger blocks, and the block histogram is the concatenation of its cell histograms. The overall image histogram is the concatenation of normalised block histograms. There are different methods for block normalisation, some of them are described in [53].

3.3 Principal component analysis

Dimensionality reduction techniques have been widely used in pattern recognition and computer vision tasks. One of the most used techniques is PCA, which considers $n \times d$ data matrix $\boldsymbol{X}$, where each of columns represents a data sample and each of rows represents the observations on each feature variable. The PCA linearly transforms the original d-dimensional space into a D-dimensional subspace, where $D \leq d$. The new feature vectors are defined by

$$y_i = W^{\mathrm{T}} x_i \tag{8}$$

where $x_i = \boldsymbol{X}_{(i,\, 1:d)}$ (x_i is the ith row of the matrix $\boldsymbol{X}$), and the columns of W are the eigenvectors e_i that are obtained by solving the eigenstructure decomposition

$$\lambda_i \boldsymbol{e}_i = \boldsymbol{Q} \boldsymbol{e}_i \tag{9}$$

where $\boldsymbol{Q}$ is the covariance matrix of the data matrix $\boldsymbol{X}$ and λ_i is the eigenvalue associated with the eigenvector $\boldsymbol{e}_i$ [54].

In our work, we calculated the eigenvectors of the covariance matrix associated with the transpose of the data matrix $\boldsymbol{X}$. The obtained $n \times n$ eigenvector matrix is used as the projected data matrix, $Y = (y_1^{\mathrm{T}}; y_2^{\mathrm{T}}; y_3^{\mathrm{T}}; \ldots; y_n^{\mathrm{T}})$.

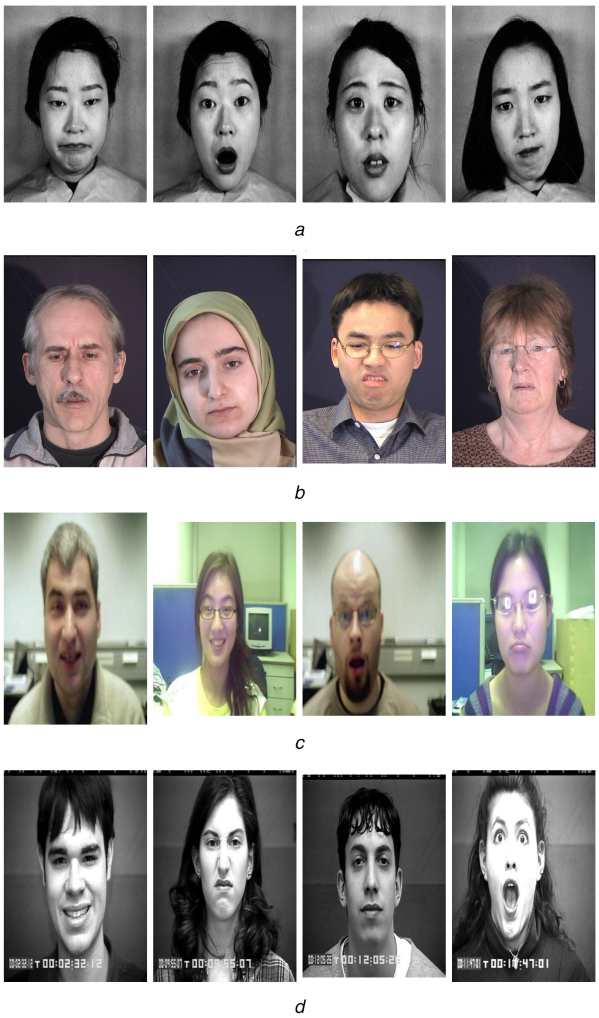
4 Experimental set-up

4.1 Evaluation protocols

In the literature of basic FER, the k-fold cross-validation protocol is widely used to evaluate the performance of the proposed methods. Cross-validation protocol consists of repeating the training–testing process K times, where, at each time, one fold is left for the test phase and the rest of the folds are used for training the model. The overall accuracy is the average of the accuracies obtained from testing all the folds. In fact, the evaluation of FER methods using cross-validation protocol has two main schemes which are subject dependent and subject independent. In the first scheme, the samples of each class of the whole database are randomly divided into K equal subsets, in order to create K-folds, each fold contains one subset from each class. However, in the subject-independent scheme, the subjects are randomly divided into K equal subsets, so each fold contains the samples of subjects corresponding to this fold. In addition, there is a special case of a subject-independent scheme, which is known as LOSO. In this scheme, the samples of just one subject are used for testing and the rest of the subjects samples are used for training the model. Thus, the number of the folds equals to the number of persons of the database.

4.2 Experimental data

In our experiments, we used four popular available facial expression databases which are: JAFFE [23], MMI [55, 56], CASIA [57] and CK+ [58]. The JAFFE [23] dataset consists of 213 images of the six basic facial expressions plus neutral face. These images were obtained from ten Japanese females, each subject has from two to four samples for each expression. The images are grey scale and their sizes are 256×256 px^2. Fig. 2*a* shows some image samples. This database has just a few numbers of subjects (ten), and that is a very challenging task to any FER system to reach high efficiency.



a

b

c

d

Fig. 2 *Databases samples*
(a) JAFFE, ***(b)*** MMI, ***(c)*** CASIA, ***(d)*** CK+

The MMI facial expression database [55, 56] is collected from students and research staff members of both sexes aged 19–62 years. For the six basic emotions, there are 207 clips, which are obtained from 30 subjects. Each video starts with the neutral expression, then passing through a peak phase in the middle and ends with the neutral face. For our experiments, we collected the three peak frames from the middle of each sequence. Thus, we obtained 621 images in total, all the images are the frontal or near-frontal view of the participant's faces and digitised into 720×576 px^2 with 24 bit colour values. The MMI database contains a lot of challenges for facial expression systems including illumination, gender, ageing, ethnicities, insufficient number of subjects and many of subjects wear accessories (e.g. glasses and scarfs) and have facial hair (moustache). Fig. 2*b* shows some image samples from the MMI database.

The Oulu-CASIA [57] NIR visible (VIS) facial expression database consists of six expressions (surprise, happiness, sadness, anger, fear and disgust) from 80 subjects between 23 and 58 years old, 73.8% of the subjects are males, and most of them are Finnish and Chinese people. In our experiments, we choose the VIS (VIS lighting) image sequences with the normal illumination conditions. The first frame of each sequence is the neutral face and the last one is the peak expression. We collected the three last frames from each sequence. We obtained in total 1440 images, the images are colour images and their resolution is 320×240 px^2. Despite that this database consists of a considerable number of subjects, it still has many challenging aspects such as human demographic attributes (gender, age and ethnicity), external factors (glasses and facial hair). The most different aspect from the other databases is the low quality of pictures which makes the facial appearance changes (such as wrinkles and furrows) not clear as shown in Fig. 2*c*.

The CK+ database [58, 59] is one of the most comprehensive face databases which has been widely used by the research community. It consists of 593 image sequences from 123 subjects. The image sequences vary in duration from 10 to 60 frames; each sequence begins with the neutral face and ends with the peak expression. In our experiments, we selected 309 sequences from 106 subjects. The only selection criterion is that the sequences are labelled as one of the six basic emotions. We took the three last frames from each sequence and the first one as a neutral face. We have obtained 927 images for the six basic emotions and 1236 for the six basic expressions plus neutral. The images are frontal views and their resolution is either 640×490 or 640×480 px^2 arrays with 8 bit grey scale or 24 bit colour value. Fig. 2*d* shows some image samples. Tables 1 and 2 summarise some statistics and properties about the used databases, which we will be used to analyse the experimental results.

5 Experimental results

Our experimental phase is divided into three parts. In the first part, we use ten-fold subject-independent cross-validation protocol to evaluate our proposed approach, and then we compare the obtained

Table 1 Databases statistics

Database	Angry	Surprise	Disgust	Happy	Fear	Sadness	Neutral
JAFFE	30	31	29	31	32	30	30
MMI	96	123	96	126	84	96	—
CASIA	240	240	240	240	240	240	—
CK+	138	252	174	207	72	84	—
CK + 7	138	252	174	207	72	84	309

Table 2 Databases properties

Database	Number of subjects	Number of sessions	Original images	Selection criteria	Number of images
JAFFE	10	—	static	—	213
MMI	30	207	video	three middle frames	621
CASIA	80	480	sequence	three last frames	1440
CK + 6	109	309	sequence	three last frames	927
CK + 7	109	309	sequence	first and three last frames	1236

Table 3 Our experimental results using different descriptors and the fusion between them for the subject-independent protocol

Technique	JAFFE	MMI	CASIA	CK + 6	CK + 7
HOG	60.09	65.62	67.92	94.45	90.85
LPQ	57.28	55.99	69.7361	92.2654	88.81
BSIF	54.46	61.90	71.25	88.17	87.32
descriptors fusion	60.09	63.03	73.47	92.56	91.26

results with the state-of-the-art ones. In the second part, we evaluate our approach using leave-one-subject-out scheme and compare the results with the state-of-the-art results. Finally, we compare the classic subject-independent and the LOSO cross-validation schemes.

5.1 Subject independent

The subject-independent experimental scheme has widely investigated in the past years because it is more plausible for the real applications, which need to recognise the facial expressions from new persons. For each database, we divided the database into ten folds with the condition that all samples of one subject appear just in one fold. In addition, we repeated the whole process ten times and the accuracy is the mean accuracy. We used ten-fold subject independent to compare between the descriptors (LPQ, BSIF and HOG) and the fusion between them (the concatenation of their features). Table 3 contains the obtained results on JAFFE, MMI, CASIA, CK + 6 and CK + 7 databases.

From the results, we find that the fusion by concatenating the features, obtained from the three descriptors (HOG, LPQ and BSIF), did not always achieve the highest accuracy compared with the use of a single descriptor. In fact, each descriptor provides a different feature vector length (HOG: 1080, LPQ: 5120, BSIF: 40,960 and the fusing: 47,160); consequently, the classifier did not consider the descriptors equivalently.

To deal with this issue, we use PCA not only to have the same feature size from each descriptor, but also to obtain more discriminative features from the features of each descriptor. Furthermore, PCA allows reducing the dimensionality of the whole system. For each database, we searched for the optimal number of PCA features (the eigenvectors) that have the most discriminative power to distinguish between the basic facial expressions. The optimal sizes for fusion descriptors are 35, 50, 150, 130 and 210 from each descriptor features transformation for JAFFE, MMI, CASIA, CK + 6 and CK + 7, respectively, as shown in Fig. 3. We emphasise that in Fig. 3 and the following ones, the size of the fusion feature is three times the value depicted on the *x*-axis; the fusion is the concatenation of three features, each having a size *x*. The variation of the optimal PCA-features number from one database to another is due to the fact that the databases are captured in different conditions. Furthermore, the number of classes (six basic expressions and six basic expressions plus neutral) and the number of subjects play a crucial role in the optimal size of PCA-fusion features. In more details, for the CK+ database, we find that the optimal PCA-fusion features number to recognise the seven expressions (210) are considerably high compared with the one for six expressions recognition (130), and that is very plausible due to the added class (neutral face), which has many common features with all of the other expressions. The number of subjects influence is clear, where we observe that the databases which consist of bigger subjects number (first column of Table 2) have longer optimal PCA features and vice versa.

With varying the number of PCA features, first, we compared between the use of single descriptor and fusion of all descriptors (Fig. 3), then we compared the fusion between two descriptors and the fusion of all descriptors (Fig. 4). From the experiments on all databases, we observe the following statements:

(i) For few PCA feaures, the recognition accuracy is low and with the augmentation of PCA features the accuracy increases to reach a peak-accuracy interval. Hence, the number of PCA features within this interval has a considerable amount of discriminative features that can highly distinguish between the different facial expressions. For example, for the CASIA database, we can consider [100,190] as the optimal interval (Fig. 3).

(ii) When the number of PCA features is within the peak-accuracy interval, the accuracy of using PCA features of each descriptor (Fig. 3) outperforms the use of its raw features (results of Table 3).

(iii) Most of the time the use of PCA-fusion between all descriptors outperforms the use of one descriptor PCA features (Fig. 3) or a PCA features a fusion of two descriptors (Fig. 4).

(iv) The peak-accuracy interval is larger when the number of subjects is sufficient as we can see for CASIA, CK + 6, and CK + 7 in Figs. 3*c–e* and 4*c–e*. In contrast, when the number of subjects is small, this interval is very small as we can see for JAFFE and MMI databases in Figs. 3*a*, *b* and 4*a*, *b*. Indeed, we observe that the accuracy increases with the number of PCA features until the optimal value, and then it fast decreases.

(v) As shown in Fig. 3, there is no descriptor from the chosen ones (HOG, LPQ and BSIF) that always performs better than the others. For example, for CK + 7 database, BSIF outperforms the other descriptors Fig. 3*c*, in contrast, for the CK + 6 the HOG descriptor is better than the others Fig. 3*d*.

(vi) For the combination of two PCA feaures (Fig. 4), there is no combination that always gives better accuracy than the other two combinations. For example, on CASIA database (Fig. 4*c*), the combination HOG-LPQ is the best combination within the interval [50,90], whereas in the interval [100,200], the PCA features of LPQ-BSIF is the best fusion.

Now, we will compare our fusion method with raw features fusion, and then with state-of-the-art methods. We observe from Table 4, comparing our proposed method with the use of raw features concatenation, that the optimal numbers of PCA features give better accuracies by 13, 10, 6.5, 3.5 and 4.5% for JAFFE, MMI, CASIA, CK + 6 and CK + 7, respectively. The comparison to the results obtained by state-of-the-art approaches is somehow difficult to achieve due to the lack of the knowledge about the used data (i.e. the followed procedure and number of the selection frames, the selected subjects and their number), and the type of declared accuracy. We compared our results with the systems that used the same data, evaluation protocol, and declared the overall accuracy. It should be mentioned that for the MMI database, most of the existing works are different in terms of the number of subjects and sequences, and the selection criterion of frames. As shown in Table 5, our result is better than [32, 60] which used similar experiment on the MMI database. On the other hand, for the CASIA database, due to the lack of static experimental works, we compared just with one work of static images [32] and many works for image sequence. Our proposed method obtained better recognition accuracy than [32] and competitive performance with the dynamic algorithms as shown in Table 6. The CK+ database is the most used database in the literature for both the first version [59] and the extended one [58].

We tested our method for 6 and 7 expressions on the CK+ database to compare with different algorithms that have achieved the highest accuracies on that database. Tables 7 and 8 contain the comparison for 6 and 7 expressions, respectively. Our method outperforms all the algorithms, despite that the works [11, 61] selected just the last frame representing the most exaggerated expression from each sequence. When we selected just the last frame from each sequence, the optimal PCA-fusing vector gives 96.47% (against 93.20% in [11]) and 96.27% (against 95.70% in [61]) for the CK + 6 and CK + 7 databases, respectively.

In addition, we compared our results with [13] (see Tables 7 and 8) which used the first version of the CK database, where some works showed that the experiment on CK+ is much harder than CK

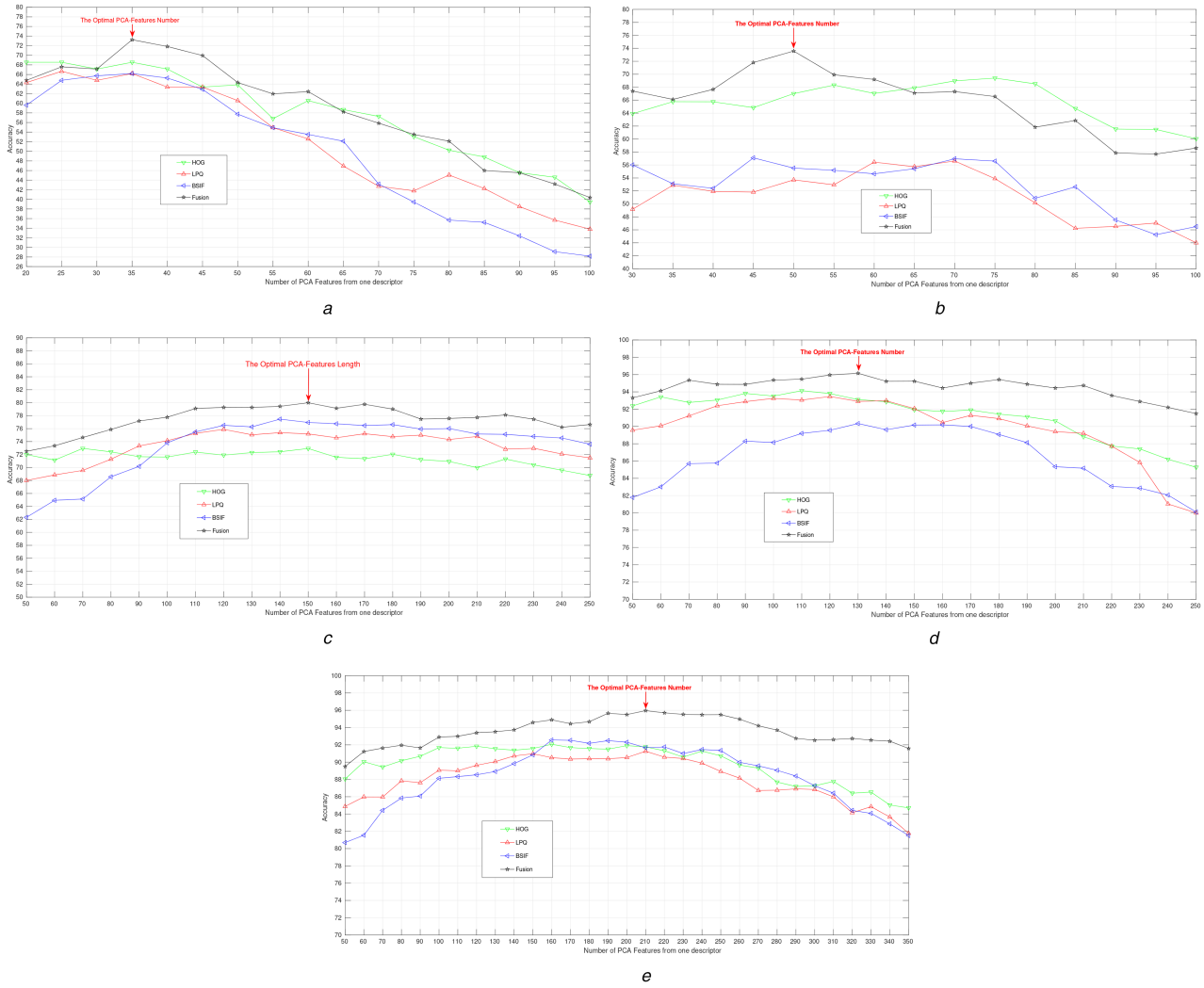


Fig. 3 *Comparison between the descriptors and the fusion between them according to PCA-features number*
(a) JAFFE, ***(b)*** MMI, ***(c)*** CASIA, ***(d)*** CK + 6 expressions, ***(e)*** CK + 7 expressions

database such as [64]. Finally for the JAFFE database, Table 9 shows that our method performs better than [22].

5.2 Leave-one–subject-out

In addition to the classic subject-independent scheme, we produced experiments using LOSO scheme on MMI, CASIA and CK + 6 expressions and 7 expressions databases. The number of the optimal PCA features are 50, 150, 130 and 240 from each descriptor features transformation, for MMI, CASIA, CK + 6 and CK + 7, respectively, as shown in Fig. 5. To obtain a better picture of our method behaviour on the level of recognising the individual expressions, we compared our fusion method with the raw features fusion and the use of the features from the single descriptors (HOG, LPQ and BSIF), Fig. 5 on JAFFE, MMI, CASIA, CK + 6 and CK + 7. We observe that the accuracy of PCA-fusion for recognition individual expressions is often better than all the accuracies that obtained using single descriptors, and that is due to the PCA-fusion power of combining properties from each descriptor. We observe also from Fig. 5 that the PCA-fusion exceeds the raw features combination in the level of recognising the individual expressions.

On the other hand, we note that the expressions (fear, disgust), (fear, sadness) and (disgust, angry) are the most confused expressions for JAFFE, MMI and CASIA, respectively. While for CK + 6 and CK + 7, the PCA-fusion recognition accuracies of the individual expressions are high, and that probably due to the plausible number of subjects in the CK+ database.

Another observation from the individual expressions recognition is that there is a descriptor that often performs better than the others in the recognition of one expression. For example, in the JAFFE database Fig. 5*a* HOG gave the better accuracy for angry, LPQ performed better for disgust, happy and sadness, BSIF is better for fear, while for neutral expression both BSIF and HOG provided almost the same accuracy.

LOSO scheme has not been widely used in the literature. To the best of our knowledge, there are only two works [22, 60] that have used this scheme in their experiments. The comparison with our method is summarised in Table 10. Unfortunately, in the works [22, 60], there are no experiments on CASIA and CK + 6 databases. From Table 10, we observe that our method's performance is better than the two works on MMI and CK + 7 databases.

In addition to the previous databases, we tested our method on the SoFace database [65]. The SoFace database consists of 2662 original images which have a variety of challenges including and not limited to different illuminations and subjects wearing glasses. The faces are labelled with four facial expressions/classes, which are neutral, happy, sad/angry/disgusted and surprised/fearful. Our fusion method achieved an accuracy of 73.1% using the LOSO scheme. Unfortunately, to the best of our knowledge, there is no published result on this database for FER to compare our result with.

5.3 Discussion

Finally, we compare the classic subject-independent scheme and LOSO. From Fig. 6, we note that the accuracies obtained using LOSO are frequently higher than classic subject-independent scheme because LOSO has the advantage of using more data in the training phase. To calculate the accuracy for the subject-independent scheme, there is a need to repeat ten times the process by dividing the subjects randomly into ten equal subsets which

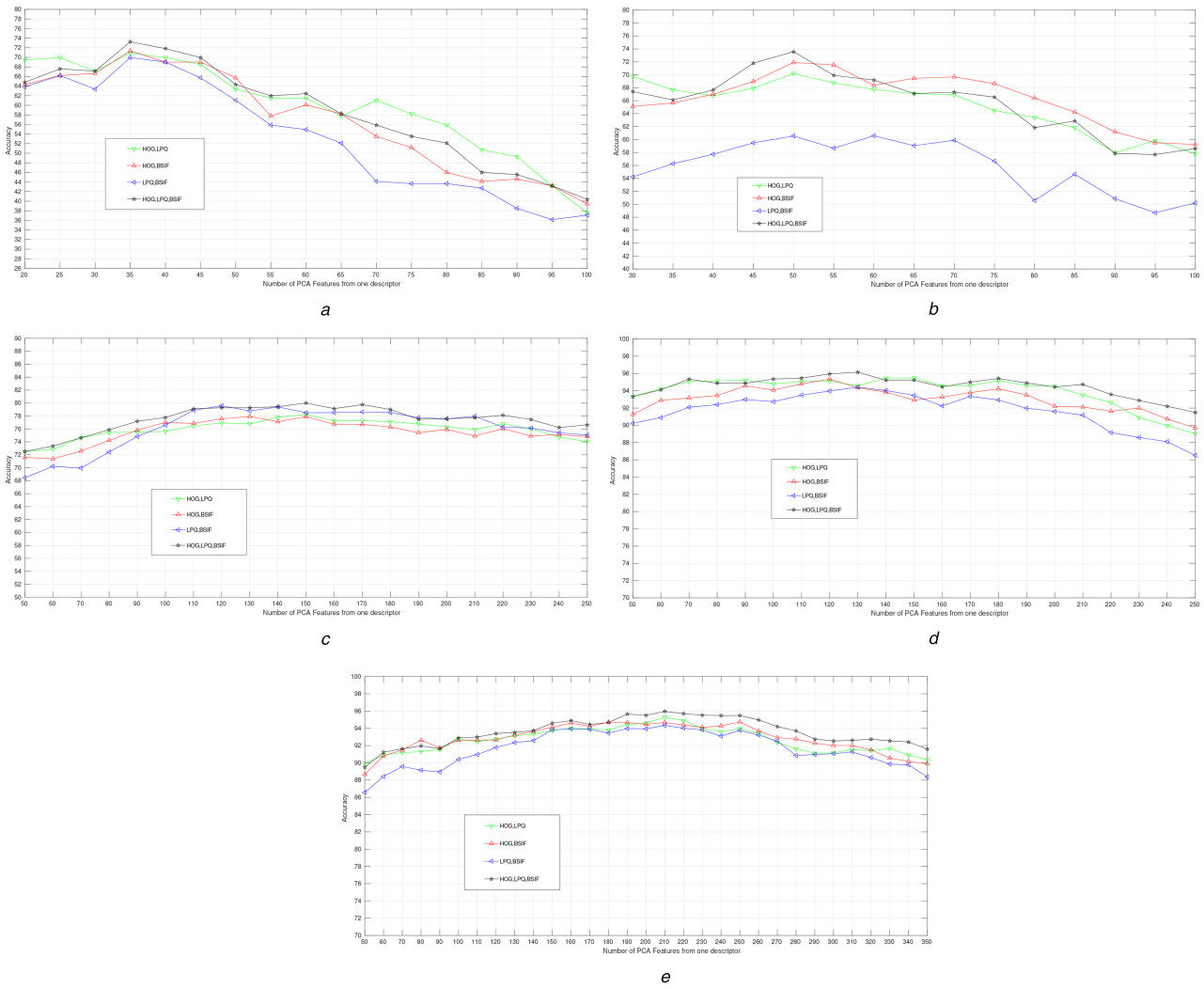


Fig. 4 *Comparison between the fusion between two descriptors and the fusion between all of them according to PCA-features number* ***(a)*** JAFFE, ***(b)*** MMI, ***(c)*** CASIA, ***(d)*** CK + 6 expressions, ***(e)*** CK + 7 expressions

Table 4 Comparison between the raw-fusion and PCA-fusion using the subject-independent scheme

Database	JAFFE	MMI	CASIA	CK + 6	CK + 7
raw-fusion	60.09	63.03	73.47	92.56	91.26
PCA-fusion	73.24	73.57	79.99	95.98	95.96

Table 5 Comparison to state-of-the-arts methods on MMI database using subject-independent cross-validation

Article	Method	Accuracy
2017 [60]	IACNN	69.48
2017 [32]	IL-CNN	70.67
our method	PCA-fusion	73.57

Table 6 Comparison to state-of-the-art methods on CASIA database using subject-independent cross-validation

Article	Method	Feature type	Accuracy
2015 [14]	ADTAGN (joint)	dynamic	81.46
	DTAGN (weighted sum)	dynamic	80.62
	DTGN	dynamic	74.17
	DTAN	dynamic	74.38
2016 [15]	UMM Dis-ExpLet	dynamic	79.0
	UMM ExpLet	dynamic	76.90
2014 [16]	STM-ExpLet	dynamic	74.59
2017 [32]	IL-CNN	static	77.29
our method	PCA-fusion	static	79.99

give ten accuracies, then the overall accuracy is the average of these accuracies. In contrast, the LOSO scheme has known folds. In addition, repeating a subject-independent experiment gives different accuracies, may be the difference between two accuracies is neglected, but sometimes that difference is significant. When we studied the recognition of the individual expressions, we noted that repeating a subject-independent experiment gives different recognition accuracies for the individual expressions, so no unique result can be concluded from the subject-independent scheme. We conclude that this is better to use the LOSO scheme because it gives a unique accuracy for each experiment and allows studying the recognition of the individual expressions.

6 Conclusion

In this paper, we proposed to use PCA-fusion between three descriptors (HOG, LPQ and BSIF) for recognising the basic facial expressions (anger, disgust, fear, happiness, sadness and surprise plus the neutral face) from the static images. Our experiments are produced on JAFFE, MMI, CASIA, CK+, and CK + 7 databases using two different cross-validation schemes. The obtained results show that our fusing method outperformed the use of the

Table 7 Comparison to state-of-the-art methods on CK + 6 database using subject-independent cross-validation

Article	Method	Accuracy
2017 [33]	DLP-CNN	95,78
2015 [40]	lp-norm MKL multiclass-SVM	95.50
2009 [13]	Boosted-LBP	95.10
	LBP uniform	92.60
2016 [11]	deep NN architecture	93.20
2013 [45]	two-stage classification of (LBP + shape)	89.20
our method	PCA-fusion	95.97

Table 8 Comparison to state-of-the-art methods on CK + 7 database using subject-independent cross-validation

Article	Method	Accuracy
2017 [61]	Boosting-POOF	95.70
2017 [36]	IACNN	95.37
2017 [32]	IL-CNN	94.35
2017 [62]	triplet-wise-based of GSF	94.09
2015 [35]	AU-inspired deep networks (AUDN GSL = 2)	93.70
2015 [40]	lp-norm MKL multiclass-SVM	93.60
2013 [63]	AU-aware deep networks (AUDNs) (OR)	91.44
	AUDN (AURF)	92.22
	AUDN	92.05
2009 [13]	Boosted-LBP	91.40
	LBP uniform	88.90
our method	PCA-fusion	95.96

Table 9 Comparison to state-of-the-arts methods on JAFFE database using subject-independent cross-validation

Article	Method	Accuracy
2017 [22]	LDTP	64.79
our method	PCA-fusion	73.23

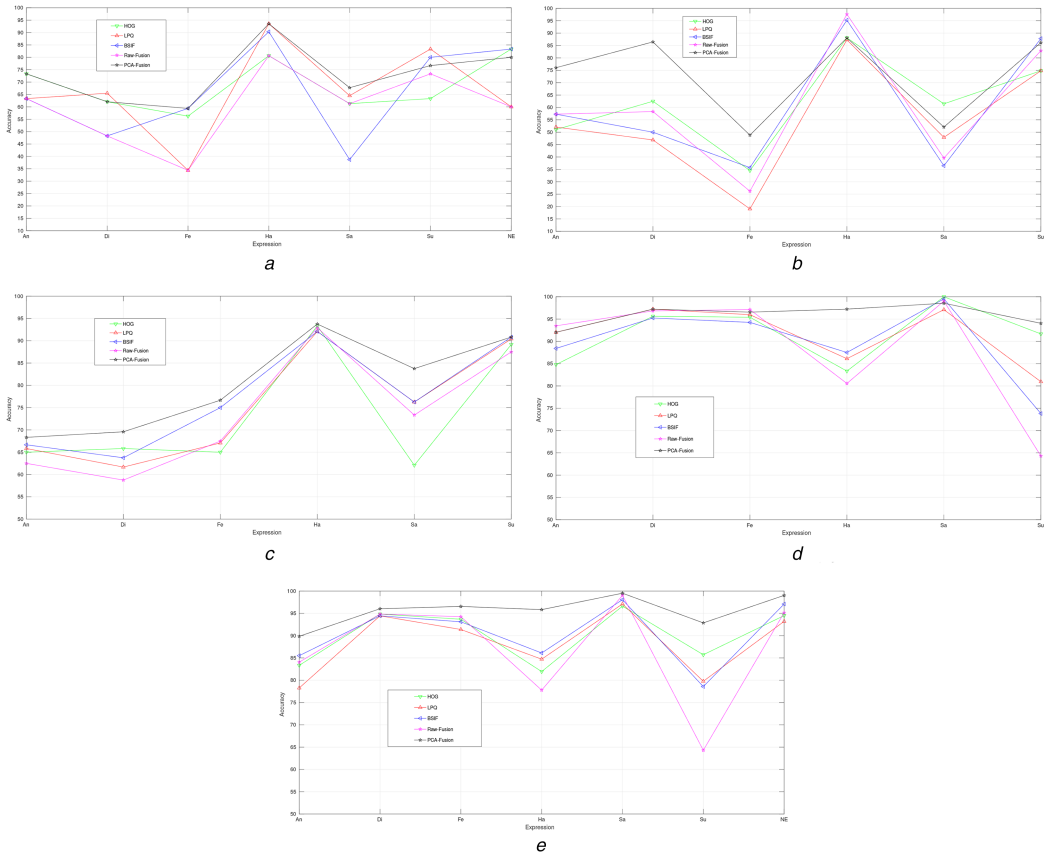


Fig. 5 *Comparison between the descriptors and the fusion between them on the recognition of the individual expressions*
(a) JAFFE database, ***(b)*** MMI database, ***(c)*** CASIA database, ***(d)*** CK + 6 database, ***(e)*** CK + 7 database

Table 10 Comparison to state-of-the-art methods using LOSO scheme

Database	LDTP [22]	F-Bases [60]	Our method
MMI	67.86	57.56	74.72
CK + 7	94.2	94.81	96.52

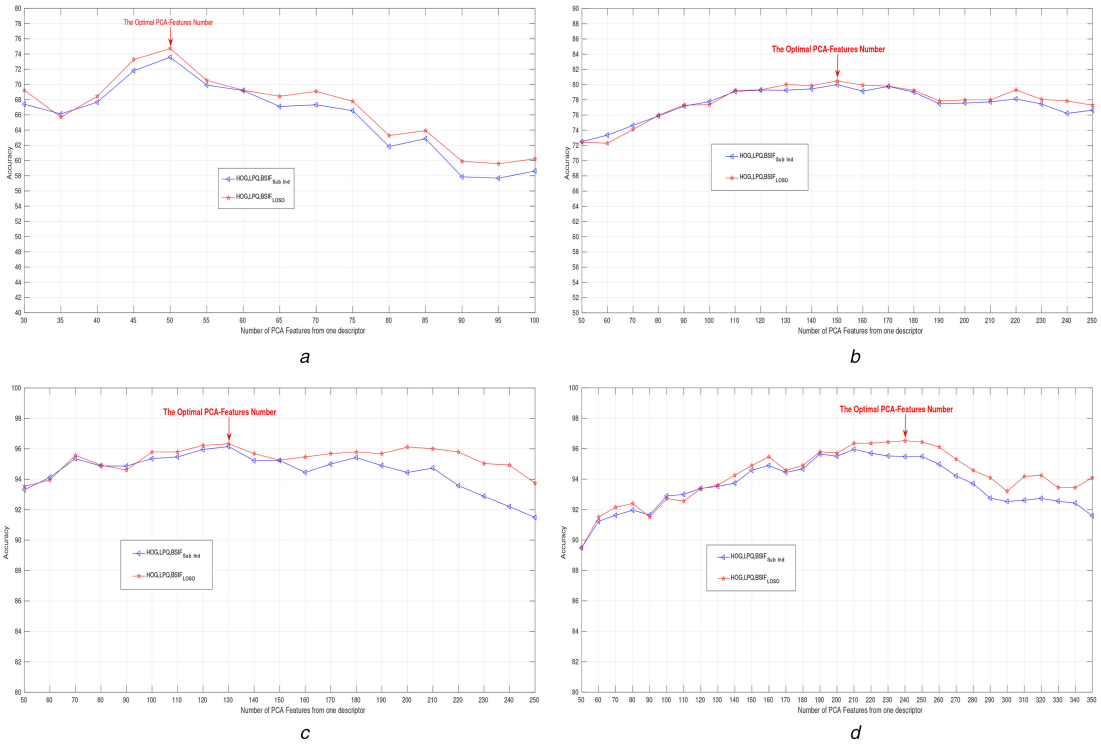


Fig. 6 *Comparison between classic subject-independent and LOSO schemes according to PCA features*
(a) MMI database, ***(b)*** CASIA database, ***(c)*** CK + 6 database, ***(d)*** CK + 7 database

traditional fusing and the existing works of different state-of-the-art approaches.

Our proposed approach is based on making different descriptors equivalent to the classifier by using PCA to transform the descriptors features into their eigenvectors. The power of our method is not only due to the use of PCA that allows our approach to obtain efficient discriminative features from the raw ones, but also the fusion between different features types made our approach stronger; appearance features using (HOG, LPQ and BSIF), shape features using (HOG) and learned features using (BSIF). The combination of these specific descriptors plays a crucial role by complimenting each other to overcome the different challenges that are facing the recognition of facial expressions, which are different from one database to another. In this paper, we also compared between two different cross-validation schemes (ten-fold cross-validation and LOSO), and we concluded that the use of the LOSO scheme is preferred due to many advantages.

As future work, we propose to use supervised techniques to fuse different features. We also plan to use more descriptors that can provide different feature types to deal with more FER challenges. Finally, we fit our method to recognise the facial expressions from videos.

7 References

[1] Martinez, B., Valstar, M.F.: 'Advances, challenges, and opportunities in automatic facial expression recognition', in Kawulok, K., Celebi, E., Smolka, B. (Eds.): '15th Ibero-American Conference on AI', San José, Costa Rica, November 23–25, 2016, pp. 63–100

[2] Ekman, P., Friesen, W.V.: 'Non-verbal leakage and clues to deception', *Psychiatry*, 1969, **32**, (1), pp. 88–106

[3] Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997

[4] Ekman, P., Campos, J.J., Davidson, R.J., *et al.*: 'Emotions inside out (130 years after Darwin's 'The expression of the emotions in man and animals')', *Ann. N.Y. Acad. Sci.*, 2003, **1000**, pp. 205–221

[5] Ojansivu, V., Heikkilä, J.: 'Blur insensitive texture classification using local phase quantization', in Elmoataz, A., Lezoray, O., Nouboud, F., *et al.* (Eds.): *'Image and signal processing', Lecture Notes in Computer Science* (Springer, Berlin, Heidelberg, 2008), pp. 236–243. Available at https://link.springer.com/chapter/10.1007/978-3-540-69905-7_27, accessed April 2018

[6] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems, Lake Tahoe, USA, 2012, pp. 1097–1105

[7] Tian, Y.L., Kanade, T., Cohn, J.F.: 'Facial expression analysis', in Li, S.Z., Jain, A.K. (Eds.): '*Handbook of face recognition*' (Springer, New York, NY, 2005), pp. 247–275

[8] Cohn, J.F., De la Torre, F.: 'Automated face analysis for affective', in Calvo, R., D'Mello, S., Gratch, J., *et al.* (eds.): '*The Oxford handbook of affective computing*' (Oxford University Press, England, 2014), p. 131

[9] Sariyanidi, E., Gunes, H., Cavallaro, A.: 'Automatic analysis of facial affect: a survey of registration, representation, and recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (6), pp. 1113–1133

[10] Carcagnì, P., Coco, M., Leo, M., *et al.*: 'Facial expression recognition and histograms of oriented gradients: a comprehensive study', *SpringerPlus*, 2015, **4**, (1), p. 645

[11] Mollahosseini, A., Chan, D., Mahoor, M.H.: 'Going deeper in facial expression recognition using deep neural networks'. 2016 IEEE Winter Conf. Applications of Computer Vision (WACV), Lake Placid, USA, 2016, pp. 1–10

[12] Happy, S.L., Routray, A.: 'Automatic facial expression recognition using features of salient facial patches', *IEEE Trans. Affective Comput.*, 2015, **6**, (1), pp. 1–12

[13] Shan, C., Gong, S., McOwan, P.W.: 'Facial expression recognition based on local binary patterns: a comprehensive study', *Image Vis. Comput.*, 2009, **27**, (6), pp. 803–816

[14] Jung, H., Lee, S., Yim, J., *et al.*: 'Joint fine-tuning in deep neural networks for facial expression recognition'. 2015 IEEE Int. Conf. Computer Vision (ICCV), Las Condes, Chile, 2015, pp. 2983–2991

[15] Liu, M., Shan, S., Wang, R., *et al.*: 'Learning Expressionlets via universal manifold model for dynamic facial expression recognition', *IEEE Trans. Image Process.*, 2016, **25**, (12), pp. 5920–5932

[16] Liu, M., Shan, S., Wang, R., *et al.*: 'Learning Expressionlets on spatiotemporal manifold for dynamic facial expression recognition'. 2014 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Columbus, USA, 2014, pp. 1749–1756

[17] Ghimire, D., Lee, J.: 'Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines', *Sensors*, 2013, **13**, (6), pp. 7714–7734. Available at http://www.mdpi.com/1424-8220/13/6/7714, accessed April 2018

[18] Shan, C., Gong, S., McOwan, P.W.: 'Robust facial expression recognition using local binary patterns'. IEEE Int. Conf. Image Processing 2005, Genova, Italy, 2005, vol. **2**, p. II–370–3

[19] Goyani, M.M., Patel, N.: 'Recognition of facial expressions using local mean binary pattern', *Electron. Lett. Comput. Vis. Image Anal.*, 2017, **16**, (1), pp. 54–67. Available at http://www.raco.cat/index.php/ELCVIA/article/view/327325, accessed April 2018

[20] Moore, S., Bowden, R.: 'Local binary patterns for multi-view facial expression recognition', *Comput. Vis. Image Underst.*, 2011, **115**, (4), pp. 541–558. Available at http://www.sciencedirect.com/science/article/pii/S1077314210002511, accessed April 2018

[21] Wang, Z., Ying, Z.: 'Facial expression recognition based on local phase quantization and sparse representation'. 2012 Eighth Int. Conf. Natural Computation, Chongqing, China, 2012, pp. 222–225

[22] Ryu, B., Rivera, A.R., Kim, J., *et al.*: 'Local directional ternary pattern for facial expression recognition', *IEEE Trans. Image Process.*, 2017, **26**, (12), pp. 6006–6018

[23] Lyons, M., Akamatsu, S., Kamachi, M., *et al.*: 'Coding facial expressions with Gabor wavelets'. Proc. Third IEEE Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 200–205

[24] Choi, H.C., Oh, S.Y.: 'Real-time facial expression recognition using active appearance model and multilayer perceptron'. 2006 SICE-ICASE Int. Joint Conf., Busan, South Korea, 2006, pp. 5924–5927

[25] Kotsia, I., Pitas, I.: 'Facial expression recognition in image sequences using geometric deformation features and support vector machines', *IEEE Trans. Image Process.*, 2007, **16**, (1), pp. 172–187

[26] Kulkarni, S.S., Reddy, N.P., Hariharan, S.: 'Facial expression (mood) recognition from facial images using committee neural networks', *Biomed. Eng. Online*, 2009, **8**, p. 16. Available at https://doi.org/10.1186/1475-925X-8-16, accessed April 2018

[27] Youssif, A.A.A., Asker, W.A.A.: 'Automatic facial expression recognition system based on geometric and appearance features', *Comput. Inf. Sci.*, 2011, **4**, (2), p. 115. Available at http://www.ccsenet.org/journal/index.php/cis/article/view/8627, accessed April 2018

[28] Tariq, U., Lin, K.H., Li, Z., *et al.*: 'Emotion recognition from an ensemble of features'. Face and Gesture 2011, Santa Barbara, USA, 2011, pp. 872–877

[29] Tariq, U., Lin, K.H., Li, Z., *et al.*: 'Recognizing emotions from an ensemble of features', *IEEE Trans. Syst. Man Cybern. B, Cybern.*, 2012, **42**, (4), pp. 1017–1026

[30] Lekdioui, K., Messoussi, R., Ruichek, Y., *et al.*: 'Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier', *Signal Process. Image Commun.*, 2017, **58**, pp. 300–312. Available at http://www.sciencedirect.com/science/article/pii/S0923596517301406, accessed April 2018

[31] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *Nature*, 2015, **521**, (7553), pp. 436–444. Available at https://www.nature.com/articles/nature14539, accessed April 2018

[32] Cai, J., Meng, Z., Khan, A.S., *et al.*: 'Island loss for learning discriminative features in facial expression recognition', arXiv preprint arXiv:171003144, 2017

[33] Li, S., Deng, W., Du, J.: 'Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild'. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017, pp. 2584–2593

[34] Burkert, P., Trier, F., Afzal, M.Z., *et al.*: 'Dexpression: deep convolutional neural network for expression recognition', arXiv preprint arXiv:150905371, 2015

[35] Liu, M., Li, S., Shan, S., *et al.*: 'Au-inspired deep networks for facial expression feature learning', *Neurocomputing*, 2015, **159**, pp. 126–136

[36] Meng, Z., Liu, P., Cai, J., *et al.*: 'Identity-aware convolutional neural network for facial expression recognition'. 2017 12th IEEE Int. Conf. Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 2017, pp. 558–565

[37] Nanni, L., Ghidoni, S., Brahnam, S.: 'Handcrafted vs. non-handcrafted features for computer vision classification', *Pattern Recognit.*, 2017, **71**, pp. 158–172

[38] Sze, V., Chen, Y.H., Yang, T.J., *et al.*: 'Efficient processing of deep neural networks: a tutorial and survey', *Proc. IEEE*, 2017, **105**, (12), pp. 2295–2329

[39] Ren, C.X., Lei, Z., Dai, D.Q., *et al.*: 'Enhanced local gradient order features and discriminant analysis for face recognition', *IEEE Trans. Cybern.*, 2016, **46**, (11), pp. 2656–2669

[40] Zhang, X., Mahoor, M.H., Mavadati, S.M.: 'Facial expression recognition using l_p-norm MKL multiclass-SVM', *Mach. Vis. Appl.*, 2015, **26**, (4), pp. 467–483

[41] Turan, C., Lam, K.M., He, X.: 'Facial expression recognition with emotion-based feature fusion'. 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA), Hong Kong, 2015, pp. 1–6

[42] Turan, C., Lam, K.M.: 'Region-based feature fusion for facial-expression recognition'. 2014 IEEE Int. Conf. Image Processing (ICIP), Paris, France, 2014, pp. 5966–5970

[43] Bartlett, M.S., Littlewort, G., Frank, M., *et al.*: 'Recognizing facial expression: machine learning and application to spontaneous behavior'. 2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'05), San Diego, USA, 2005, vol. **2**, pp. 568–573

[44] Freund, Y., Schapire, R.E.: 'A decision-theoretic generalization of on-line learning and an application to boosting', *J. Comput. Syst. Sci.*, 1997, **55**, (1), pp. 119–139. Available at http://www.sciencedirect.com/science/article/pii/S002200009791504X, accessed April 2018

[45] Xue, M., Liu, W., Li, L.: 'Person-independent facial expression recognition via hierarchical classification'. 2013 IEEE Eighth Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing, Piscataway, NJ, 2013, pp. 449–454

[46] Wang, J., Yin, L.: 'Static topographic modeling for facial expression recognition and analysis', *Comput. Vis. Image Underst.*, 2007, **108**, (1), pp. 19–34. Available at http://www.sciencedirect.com/science/article/pii/S107731420600227X, accessed April 2018

[47] Sebe, N., Lew, M.S., Sun, Y., *et al.*: 'Authentic facial expression analysis', *Image Vis. Comput.*, 2007, **25**, (12), pp. 1856–1863. Available at http://www.sciencedirect.com/science/article/pii/S0262885606002903, accessed April 2018

[48] King, D.E.: 'Dlib-ml: a machine learning toolkit', *J. Mach. Learn. Res.*, 2009, **10**, (Jul), pp. 1755–1758. Available at http://www.jmlr.org/papers/v10/king09a.html, accessed April 2018

[49] Fan, R.E., Chang, K.W., Hsieh, C.J., *et al.*: 'LIBLINEAR: a library for large linear classification', *J. Mach. Learn. Res.*, 2008, **9**, (Aug), pp. 1871–1874. Available at http://www.jmlr.org/papers/v9/fan08a.html, accessed April 2018

[50] Ojala, T., Pietikainen, M., Maenpaa, T.: 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (7), pp. 971–987

[51] Hadid, A., Ylioinas, J., Bengherabi, M., *et al.*: 'Gender and texture classification: a comparative analysis using 13 variants of local binary patterns', *Pattern Recognit. Lett.*, 2015, **68**, pp. 231–238. Available at http://www.sciencedirect.com/science/article/pii/S0167865515001348, accessed April 2018

[52] Kannala, J., Rahtu, E.: 'BSIF: binarized statistical image features'. Proc. 21st Int. Conf. Pattern Recognition (ICPR2012), Tsukuba, Japan, 2012, pp. 1363–1366

[53] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. 2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'05), San Diego, USA, 2005, vol. **1**, pp. 886–893

[54] Martinez, A.M., Kak, A.C.: 'PCA versus LDA', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (2), pp. 228–233

[55] Pantic, M., Valstar, M., Rademaker, R., *et al.*: 'Web-based database for facial expression analysis'. 2005 IEEE Int. Conf. Multimedia and Expo, Amsterdam, Netherlands, 2005, p. 5

[56] Valstar, M., Pantic, M.: 'Induced disgust, happiness and surprise: an addition to the mmi facial expression database'. Proc. Third Int. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, 2010, p. 65

[57] Zhao, G., Huang, X., Taini, M., *et al.*: 'Facial expression recognition from near-infrared videos', *Image Vis. Comput.*, 2011, **29**, (9), pp. 607–619. Available at http://www.sciencedirect.com/science/article/pii/S0262885611000515, accessed April 2018

[58] Lucey, P., Cohn, J.F., Kanade, T., *et al.*: 'The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression'. 2010 IEEE Computer Society Conf. Computer Vision and Pattern Recognition – Workshops, San Francisco, USA, 2010, pp. 94–101

[59] Kanade, T., Cohn, J.F., Tian, Y.: 'Comprehensive database for facial expression analysis'. Proc. Fourth IEEE Int. Conf. Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 2000, pp. 46–53

[60] Sariyanidi, E., Gunes, H., Cavallaro, A.: 'Learning bases of activity for facial expression recognition', *IEEE Trans. Image Process.*, 2017, **26**, (4), pp. 1965–1978

[61] Liu, Z., Li, S., Deng, W.: 'Boosting-POOF: boosting part based one vs. one feature for facial expression recognition in the wild'. 2017 12th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 2017, pp. 967–972

[62] Xie, W., Shen, L., Yang, M., *et al.*: 'Active AU based patch weighting for facial expression recognition', *Sensors*, 2017, **17**, (2), p. 275

[63] Liu, M., Li, S., Shan, S., *et al.*: 'Au-aware deep networks for facial expression recognition'. 2013 10th IEEE Int. Conf. Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 2013, pp. 1–6

[64] Rivera, A.R., Castillo, J.R., Chae, O.O.: 'Local directional number pattern for face analysis: face and expression recognition', *IEEE Trans. Image Process.*, 2013, **22**, (5), pp. 1740–1752

[65] Afifi, M., Abdelhamed, A.: 'AFIF4: deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces', arXiv preprint arXiv:170604277, 2017