IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Emotion recognition from facial expressions using hybrid feature descriptors

ISSN 1751-9659
Received on 13th May 2017
Revised 24th November 2017
Accepted on 19th January 2018
E-First on 12th February 2018
doi: 10.1049/iet-ipr.2017.0499
www.ietdl.org

Tehmina Kalsum[1], Syed Muhammad Anwar[1] ✉, Muhammad Majid[2], Bilal Khan[3], Sahibzada Muhammad Ali[3]

[1]*Department of Software Engineering, University of Engineering and Technology Taxila, Pakistan*
[2]*Department of Computer Engineering, University of Engineering and Technology Taxila, Pakistan*
[3]*Department of Electrical Engineering, COMSATS Institute of Information Technology, Abbottabad, Pakistan*
✉ *E-mail: s.anwar@uettaxila.edu.pk*

Abstract: Here, a hybrid feature descriptor-based method is proposed to recognise human emotions from their facial expressions. A combination of spatial bag of features (SBoFs) with spatial scale-invariant feature transform (SBoF-SSIFT), and SBoFs with spatial speeded up robust transform are utilised to improve the ability to recognise facial expressions. For classification of emotions, *K*-nearest neighbour and support vector machines (SVMs) with linear, polynomial, and radial basis function kernels are applied. SBoFs descriptor generates a fixed length feature vector for all sample images irrespective of their size. Spatial SIFT and SURF features are independent of scaling, rotation, translation, projective transforms, and partly to illumination changes. A modified form of bag of features (BoFs) is employed by involving feature's spatial information for facial emotion recognition. The proposed method differs from conventional methods that are used for simple object categorisation without using spatial information. Experiments have been performed on extended Cohn–Kanade (CK+) and Japanese female facial expression (JAFFE) data sets. SBoF-SSIFT with SVM resulted in a recognition accuracy of 98.5% on CK+ and 98.3% on JAFFE data set. Images are resized through selective pre-processing, thereby retaining only the information of interest and reducing computation time.

1 Introduction

Automatic recognition of emotions from human's facial expressions has been a challenging and interesting task. In recent times, this has drawn more attention with an increase in surveillance, safety, and human–computer interactive applications [1]. It is the need of the hour to make computers intelligent so that they can detect user's emotion and thereby take appropriate actions. These intelligent systems can be used in a wide variety of application areas including robotics, human–computer interaction, and for entertainment purposes. This interaction would be more natural if computers become capable of recognising and responding to the human's non-verbal communication that is mostly carried out using facial expressions.

Emotions are those human reactions, which are specific to events that usually take place for a short duration of time. In our daily life, emotions play a vital role and provide useful information such as a person's state of mind. In interpersonal communication, human facial expressions play an essential role. It has been shown that spoken words comprising the verbal part of a message contribute merely 7% of the overall message [2]. The voice intonation, representing the vocal part, contributes 38% and the speaker's facial expressions contribute 55% to the overall effect of the spoken message. These numbers represent the importance of facial expressions in effectively delivering a message. The emotions that are commonly expressed in day-to-day human interaction irrespective of the cultural background are known as basic emotions [3]. According to a psychological theory, human emotions are classified into six basic emotions, i.e. fear, anger, sadness, disgust, surprise, and happiness [4].

The emotion recognition systems rely on image processing techniques for face detection and feature extraction. This is followed by the application of pattern recognition methods for classification of facial expressions. First, the facial part of a user is detected either from a static image or from a frame extracted from a video. The extracted facial region is further processed to determine significant facial features, which are helpful in identifying emotions. In various emotion recognition systems, a model or feature-based [5, 6] approach is employed. Physiological signals that include electroencephalography and electrocardiogram among others have been used to detect emotions, but generally give lower performance when compared with computer vision-based systems [7, 8].

In model-based approach, a face model is created for the purpose of classification, e.g. active appearance model (AAM) [9] and active shape model (ASM) [10]. The ASM technique has performed facial expression recognition with an accuracy of 92.1%. In these model-based approaches finding the right model that fits the faces of different types and sizes is difficult. On the other hand, in feature-based methods, features are extracted from the salient facial regions. These methods differ depending on whether features are extracted from local components or globally from the whole image. In component-based methods, subregions like eyes, lips, and nose etc., are extracted from the face [11]. The deformation of corresponding local features in training and test images is compared to detect emotions. Whereas in the case of global face segmentation methods, features are extracted from the whole segmented face [12]. A significant challenge in component-based approach is feature extraction and alignment for different face sizes and for one face with varying expressions.

In the proposed system, spatial scale-invariant feature transform (SIFT) and speeded up robust transform (SURF) descriptors are used for feature extraction that are illumination, scale, and rotation invariant from the segmented facial patches. An ordered spatial bag of feature (SBoF) [13] technique is used for the purpose of dictionary building. Also, parameters such as the facial region of interest (ROI) and feature vector length are optimised, which results in better recognition accuracies. The major contributions of this study are,

i. Spatial SIFT and SURF is proposed for the extraction of scale, rotation, and illumination invariant features within segmented

© The Institution of Engineering and Technology 2018

face part in order to prevent irrelevant feature point extraction caused by dense SIFT (DSIFT).

ii. A combination of spatial SIFT and SURF with SBoF is used for reduced number of comparisons during image training and testing.

iii. The recognition task is evaluated with *K*-nearest neighbour and SVM classifiers and an improvement in performance is achieved in detecting the basic emotions, i.e. sad, surprise, happy, fear, anger, disgust, and neutral.

In rest of the paper, Section 2 presents a review of the related work for facial expression recognition. The details of the proposed framework are presented in Section 3. The experimental results are discussed in Section 4, followed by conclusion in Section 5.

2 Related work

In recent studies, feature-based approach is preferred for the purpose of emotion classification since fitting a model to the varying face shapes and sizes is a challenging task. The algorithms that use global facial features without segmentation and incorporation of local spatial information are simple and faster but the recognition accuracy decrease with a change in the object pose and illumination [14]. Whereas, algorithms that are based on local features are more appropriate for the purpose of human emotion recognition due to their robustness to illumination and pose variations. It has been shown that local feature descriptors for the selected regions of interest perform well for image processing applications such as object recognition, image matching, and object categorisation [15]. The main focus of recent research has been on making these feature descriptors more robust to object transformations.

SIFT descriptor [16] is found to be invariant to transformations such as image scaling, rotation, translation, projective transform, and partially to illumination changes. However, in order to use SIFT for facial expression recognition, exponential comparisons are required resulting in high computation time. The computation time required for SIFT feature matching can be reduced by using bag of feature (BoF) [17]. The BoF model is inspired by the bag of words (BoW) model [18], which is used for document categorisation. The originally proposed BoF [17] generates an orderless collection of local features without considering feature's spatial information required for facial expression recognition. In [19], a manifold supervised learning algorithm is proposed that is based upon kernel and local linear embedding for the purpose of expression recognition. The facial features are extracted by using local binary pattern technique and further classified by using SVM. An accuracy of 79.8, 79.8, and 81.0% is achieved for Japanese female facial expression (JAFFE) data set with polynomial, linear, and RBF kernel, respectively. Also, an average accuracy rate of >90% is achieved for the CK data set. An efficient method for facial salient point detection from a video sequence having frames with varying expression is presented in [20]. A total of 26 points are selected for facilitation of expression recognition with the help of shift invariant feature detectors. Also for tracking purpose, a differential evolution Markov chain particle filter is applied. A detailed analysis is done with kernel correlation approach to maximise the similarity measure among the candidate and target points. Experiments are performed on three publicly available data sets achieving an accuracy of 93%.

The performance of an emotion recognition system is evaluated based on the accuracy of extracted features. An efficient feature extraction method is proposed using subtle variations in spatial and temporal domain [21]. The Viola–Jones algorithm is used for face detection and a three-dimensional Harris detector is used for feature modelling. The information is further represented in the form of histogram using a block-based method and SVM classifier is used for emotion classification. In [22], spontaneous facial expressions are recognised on the basis of histogram of gradients (HoG) features. Face detection and extraction is done using the continuously adaptive mean shift (CAMshift) algorithm achieving an accuracy of 95%. The face detected by CAMshift algorithm contains a lot of side region information such as neck, hair, and some ear parts. The significant facial feature points are marked manually for feature extraction. This manual feature region marking and extraction for a data set comprising a large number of images would be very time-consuming. In [23], a detailed review regarding the suitability of HoG feature descriptor for facial expression recognition is presented and also proposed a facial expression recognition system for frontal images of 65×59 dimension. Experiments are performed using 10-fold cross-validation with SVM classifier on CK+ data set.

The available methods suffer from optimal feature selection, specifically when transformation invariant feature descriptors are used. DSIFT can take into account most of the relevant features but end up with a relatively large feature vector. This translates into more computational burden that needs to be reduced for real-time and mobile applications. At the same time, it is also important to capture spatial information in local facial areas for achieving higher classification accuracy in the emotion recognition task. In this paper, results are presented from extensive experiments for facial emotion recognition task using feature-based approach with global face segmentation. The segmented face area is further preprocessed to retain only the significant information. The spatial BoF method is employed for an ordered feature selection with spatial information. A detailed analysis is performed by using a hybrid of SBoF with spatial SIFT and SURF as feature descriptors. Different classifiers are trained and the performance is evaluated to select the best parameters.

3 Proposed framework

The proposed emotion recognition framework comprises a number of steps shown in Fig. 1. Preprocessing, feature extraction, codebook construction, and classification are the major steps involved in the proposed framework. The details of each step are as follows.

3.1 Preprocessing

First, the facial part of images is detected by using the Viola–Jones face detection algorithm [24]. Some portion of the detected face is further cropped from the left, right, and top side. Let (P_1, P_2) be the top left and (P_3, P_4) be the top right corner of detected face. The image is cropped from left, right, and top calculating updated pixel values, represented as (P_{n1}, P_{n2}) and (P_{n3}, P_{n4}), which are calculated as

$$P_{n1} = P_1 + \lfloor \alpha X \rfloor, \tag{1}$$

$$P_{n2} = P_2 + \lfloor \beta Y \rfloor, \tag{2}$$

$$P_{n3} = P_3 + \lfloor \alpha X \rfloor, \tag{3}$$

$$P_{n4} = P_4 + \lfloor \beta Y \rfloor, \tag{4}$$

where X and Y are the image dimensions in horizontal and vertical directions, respectively. α and β are the cropping scale factor in horizontal and vertical image direction, respectively. The cropping operation gives a core central facial part where the major changes can be seen when an emotion is expressed. The cropping of the bottom side of facial images negatively effected the system performance specifically when detecting the surprise emotion.

The resulting cropped image is shown in Fig. 2 and only contains the core central facial area. The optimised values for cropping scale factors α and β are selected after extensive experimentation. The scaled values are rounded off to the nearest floor value, so that an integer value for pixel shift is calculated.

The cropped images in the JAFFE and CK+ data sets are of different dimensions. All images are resized to 96×128 using bilinear interpolation so that an equal number of features are extracted from each image. In addition, discrete wavelet transform (DWT) is applied to these scaled images and the resulting information in low–low (LL) frequency subband is retained. The advantage of applying DWT and retaining the LL frequency subband is a reduction in the number of pixels that needs to be

© The Institution of Engineering and Technology 2018

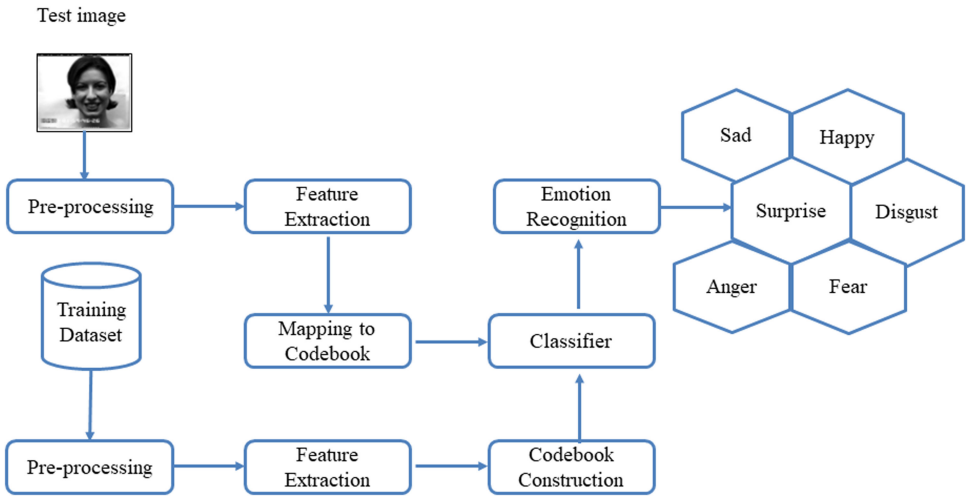

Fig. 1 *Proposed framework to detect basic emotions, i.e. sad, surprise, happy, fear, anger, and disgust*

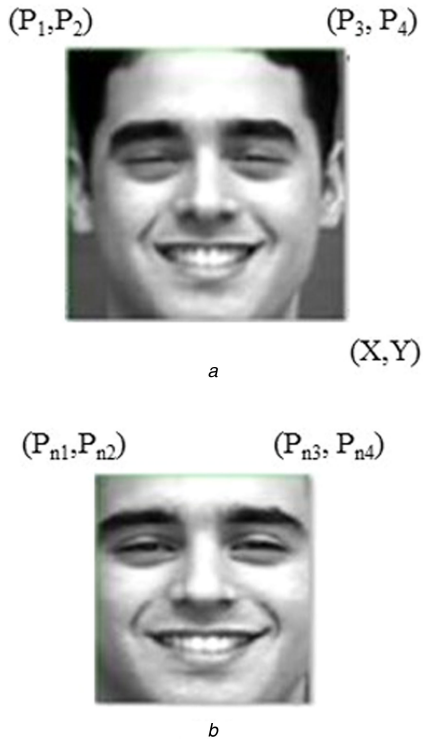

Fig. 2 *Significant facial area selection for emotion recognition*
(a) Face detected by the Viola–Jones algorithm, ***(b)*** Face after side cropping to exclude insignificant facial regions

processed. It is also observed that these down-sampled images still include some side facial areas that remain unchanged when different emotions are expressed. This can cause feature overlapping among different emotions in the matching phase of classification. The images contained in CK+ data set contain many variations in face type, user ethnicity, and illumination as shown in Fig. 3*a*. Feature overlapping for such a challenging scenario can cause significant performance degradation. The pixels in these areas are neglected for key point extraction. The shaded areas shown in Figs. 3*b* and *c* do not contribute in codebook construction and the resulting feature vector. The effect of the selecting this area is also observed by varying its height based on which an optimal area is chosen. This helped in further reducing the facial area and hence the number of pixels that are processed for key point extraction. This results in an increased accuracy by considering only the highly differentiating core central facial features.

3.2 Feature extraction

The segmented face from the pre-processing stage is divided into 29 patches with different sizes as shown in Fig. 4. The patches include the whole face, 2×2 and 4×4 equally divided segments from the whole face. In addition, 2×2 patches are also extracted from both the upper and lower part of the face where the upper part includes the facial area from head to eyes and the rest of the area is considered as lower part. For extracting key points from the image, the concept of spatial SIFT is proposed for feature extraction, where SIFT features are extracted independently within each patch. This results in efficient features in all patches that are invariant to different image transformations. The feature vector for each image is obtained by concatenating features extracted locally from the selected patches, thereby, incorporating the spatial information.

For SIFT feature extraction, the scale space of an image is defined by the function $L(p,q,\sigma)$, which is calculated on the basis of convolving input image $I(p,q)$ with a Gaussian function $G(p,q,\sigma)$. The functions of $L(p,q,\sigma)$ and $G(p,q,\sigma)$ are calculated as

$$L(p,q,\sigma) = G(p,q,\sigma) * I(p,q), \tag{5}$$

$$G(p,q,\sigma) = \frac{1}{2\pi\sigma^2} e^{-((x^2+y^2)/2\sigma^2)}, \tag{6}$$

where σ is the variance of the Gaussian distribution. The magnitude of the gradient $m(p,q)$ and orientation $\Theta(p,q)$ are computed using the pixel differences as follows:

$$m(p,q) = \sqrt{((L(p+1,q) - L(p-1,q))^2 + (L(p,q+1) - L(p,q-1))^2}, \tag{7}$$

$$\Theta(p,q) = \tan^{-1}\left(\frac{L(p,q+1) - L(p,q-1)}{L(p+1,q) - L(p-1,q)}\right) \tag{8}$$

A feature descriptor of 1×128 and 1×64 dimension is computed for SIFT and SURF, respectively, with the help of the gradient magnitude and orientation values. These are weighted by a Gaussian on the basis of a subregion of size 4×4 (total of 16 regions) for each detected point.

3.3 Codebook construction

A codebook is used for a compact representation of large feature sets. It groups similar features together into a specified number of clusters. A single mean value of each cluster serves as a representative of the whole cluster resulting in a significant feature reduction. In this work, all SIFT and SURF features extracted from each emotion are quantised for BoF-based codebook generation as shown in Fig. 5. Initially, features extracted from all images belonging to a particular emotion are grouped together, and

© The Institution of Engineering and Technology 2018

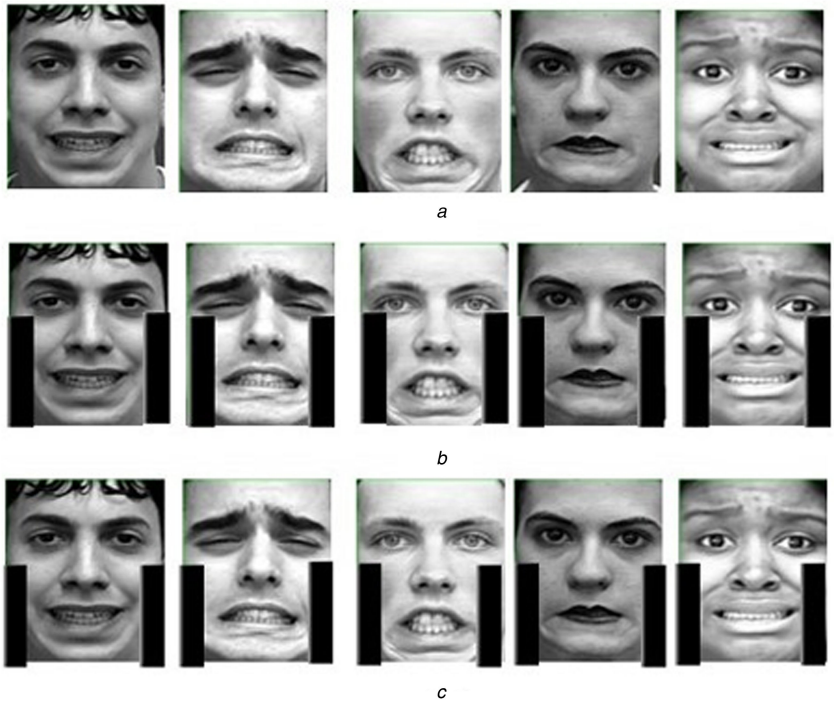

Fig. 3 *Random images for fear expression from CK+ data set*
(a) Images showing variation within a class of facial expressions and irrelevant side area, ***(b)*** With large excluded side area, ***(c)*** With less excluded side area

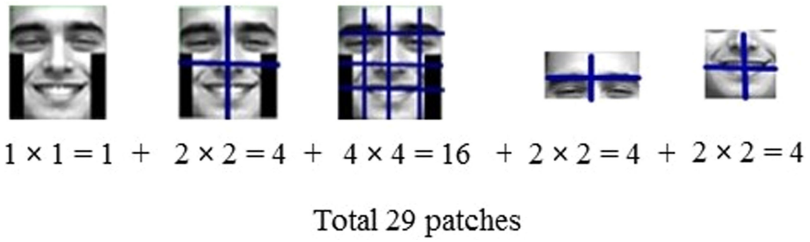

Fig. 4 *Face segmentation for spatial SIFT feature extraction to incorporate spatial information*

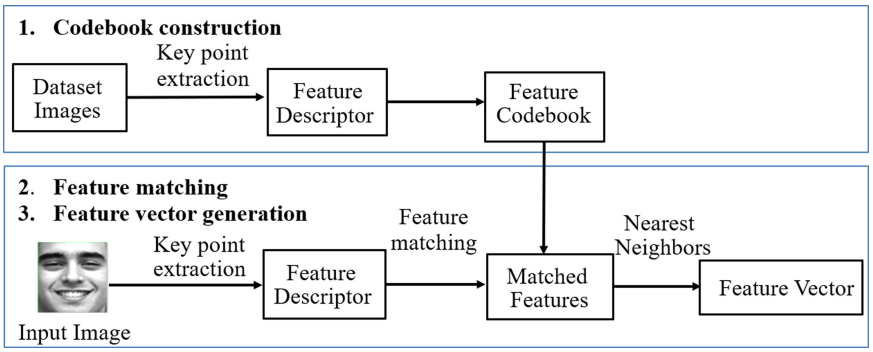

Fig. 5 *Image representation with SBoFs*

distributed among c specified clusters using K-means clustering. The individual codebooks for all emotions are then concatenated together for final codebook generation with C clusters. The value of c is set to 150 and 70 for CK+ and JAFFE data set, respectively.

The next two steps as shown in Fig. 5 deal with the generation of image feature vector for the purpose of classifier training and testing. For feature matching, nearest neighbours in codebook for all extracted features are determined on the basis of fast library for approximate nearest neighbours that quickly computes the nearest neighbours for large data sets. The computation time is reduced with codebook, due to a reduction in the number of feature comparisons. Each 1×128 feature vector is now compared with C cluster centres instead of comparing with each single feature of each image in the training data set. Finally, an M-bin histogram-based feature vector is generated for each image.

3.4 Classification

Fig. 6 shows the training phase of classifier. Image feature vector is assigned a corresponding emotion label for the generation of training data set. An $n \times C$ labelled feature vector is used for supervised training, where n represents the total number of training images in data set and C is the corresponding label. In testing phase, an emotion label is assigned to the test image for the histogram bin where the maximum features of the test image are mapped. For classifying basic emotions, two different classifiers, including support vector machine (SVM) and K-nearest neighbour, are trained and tested on the extracted features.

3.4.1 Support vector machine: It is an algorithm that finds a hyperplane that fits the training data and separates it into classes. SVMs are supervised learning models and are best suited for

© The Institution of Engineering and Technology 2018

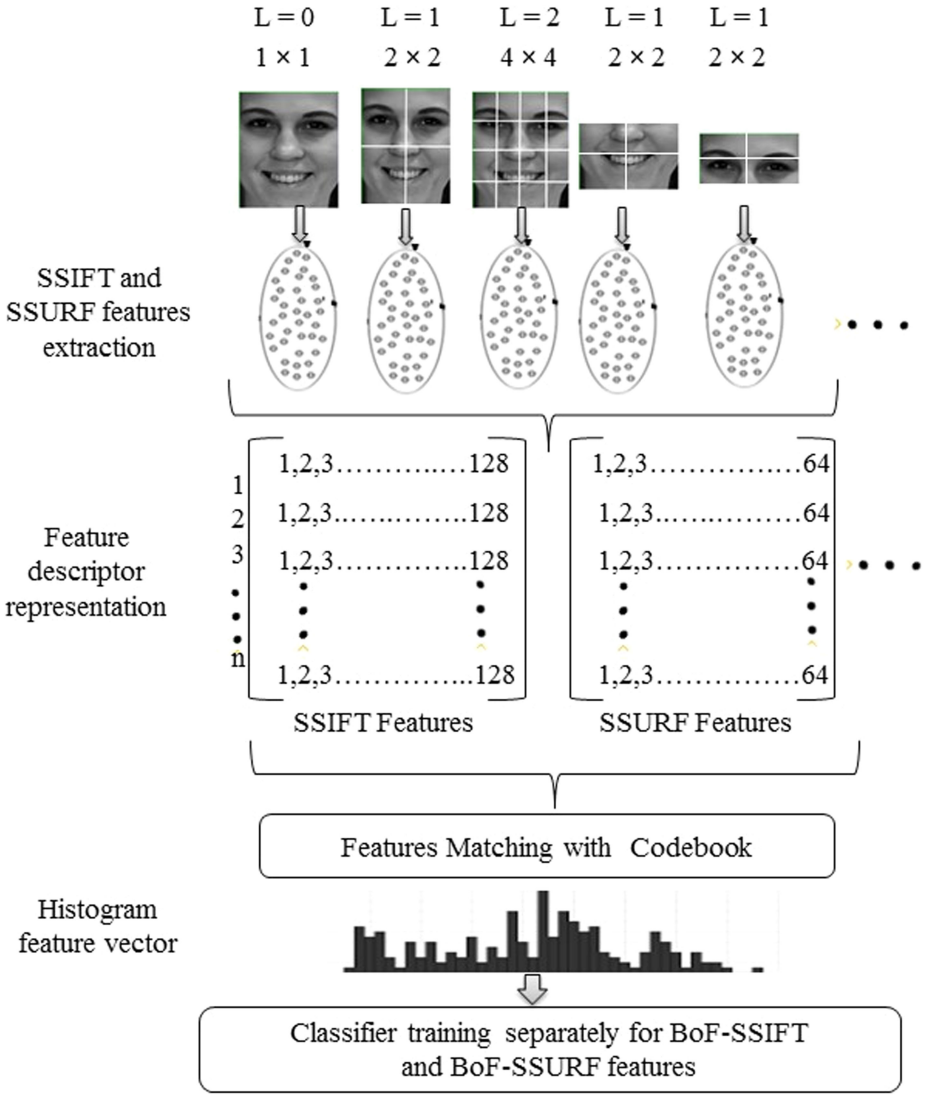

Fig. 6 *Steps involved in SBOF-SSIFT and SBOF-SSIFT-based classifier training*

binary classification problems. SVM can also be extended to incorporate multi-class classification using multi-class SVM. The performance of SVM classifiers depends on the type of kernel e.g. linear, polynomial, and so on. In this study, multi-class SVM classifiers are trained on the basis of spatial bag of features with spatial scale-invariant feature transform (SBoF-SSIFT) and spatial bag of features with spatial speeded up robust transform (SBoF-SSURF) features using linear, radial basis function (RBF), and polynomial kernels. Each single feature vector is matched with all C codebook entries and the histogram bin corresponding to the highest matching point is incremented. The histogram with frequency values against each cluster centre is then used as a feature vector of the image for classifier training and testing.

The detailed process for classifiers training on the basis of extracted features is presented in Fig. 6. A total of N classifiers are trained on the training data set for SBoF-SSIFT and SBoF-SSURF features, where N represents total number of classes. The discriminating points detected from segmented patches are described by SIFT and SURF descriptor of 1×128 and 1×64 dimension, respectively. These descriptors are then mapped to a codebook for histogram-based feature vector generation. At the end, SVM and KNN classifiers are trained separately for both SSIFT and SSURF feature vectors. In order to predict an emotion, the trained SVM classifiers are loaded in a sequence and the test image is predicted to represent the class where a match is found. The process continues until the image is classified with a certain class label or labelled as 'not predicted' when all the classifiers fail to predict the emotion.

3.4.2 K-nearest neighbour: It is used in pattern recognition problems both for classification and regression. The algorithm depends on the value of K, and for a chosen value of K, operates by voting to decide which class a test case belongs. The algorithm is a simplistic classification algorithm but is found to give competitive performance when compared with complex algorithms in most classification problems. The algorithm is dependent on the underlying data structure and could fail for more complex data distributions. As the predicted label value vary with changing number of specified neighbours (K), experiments are performed with varying K values, i.e. $K = 1, 3, 5$ for both JAFFE and CK+ data sets.

4 Experimental results and discussion

In this section, a detailed description of the data sets used, experimental results, and their discussion is presented.

4.1 Data sets

The proposed system is evaluated by performing experiments on CK+ [25] and JAFFE [26] data sets that consists of a sequence of images for each subject showing different expressions. The detail of these data sets is as follows.

4.1.1 Extended Cohn–Kanade data set (CK+): This data set comprises of 593 image sequences collected for 123 subjects. The age range of subjects is 18–50 years, where 31% are men and 69% are women. The image sequences represent seven different

© The Institution of Engineering and Technology 2018

expressions (sad, surprise, happy, fear, anger, disgust, and neutral) and cover the basic human emotions.

4.1.2 JAFFE data set: This data set consists of a total of 213 images for 10 females' facial expressions including six basic (sad, surprise, happy, fear, anger, disgust) and one neutral expression. The images have been rated by 60 Japanese subjects to belong to a particular expression. The top voted expression is assigned to the image as its class label.

In both data sets, each image sequence for a particular subject starts from neutral expression frame and ends with the peak expression frame. In this work, six classifiers are trained on the basis of corresponding expression images, i.e. sadness, happiness, surprise, disgust, fear, and anger for classification. The images contained in the data sets are randomly divided such that 70 and 30% of these are used for training and testing purpose, respectively. Five-fold cross-validation is also used during the training phase to avoid classifier over-fitting. The emotion recognition accuracy of the proposed system is analysed by varying following parameters,

i. The ROI selected from face for feature selection.
ii. The segmentation levels that are used for acquiring spatial information related to facial expressions.
iii. The number of clusters used for feature codebook generation.
iv. The kernel function that is used for SVM and the value of K for KNN algorithm.

These parameters are carefully chosen after experimentation on the basis of scaled and wavelet transformed images. The average accuracy of emotion recognition by using the wavelet transformed images is comparable to the scaled images with substantial reduction in computation time. An extensive set of experiments is performed for the preprocessed face images with complete information and excluding smaller and larger side information. The results obtained for the whole image have an accuracy of 83.5% which is improved up to 85.7% when side areas are excluded. The features in these side areas are not discriminative and hence cause misclassification. The amount of side and top image that is cropped for selecting the significant facial is controlled by α and β. The optimal values of these parameters for a random set of images are shown in Table 1.

Further experiments are performed by varying facial image segmentation levels for incorporation of spatial information and are presented in Table 2. The recognition accuracy is increased for image segmentation up to four levels (i.e. $L = 3$) and after that it started decreasing. When an image of pixel dimension 48×64 is divided into a 16×16 patch, the resulting patches are not big enough for good feature point detection. The optimal recognition accuracy is achieved by segmenting face into upper eyes part and lower part of face comprising nose and mouth. The feature vector for every single image is obtained by concatenating features extracted from patches that comprise the whole face and segmented up to a level of 4×4. A total of 29 patches of comparatively large size are considered for emotion recognition as discussed in Section 3.2.

After selecting the patches from segmented face, experiments are performed with varying number of clusters for the purpose of creating codebook vocabulary. Table 3 demonstrates that for CK+ data set, the recognition accuracy increases with an increase in the number of clusters until K equals 150 for each emotion. After that it remains constant until K equals 200 and starts decreasing afterwards. This decrease in accuracy is due to over-fitting of the classifier on the training set having a very large feature vector. For JAFFE data set, an optimal recognition accuracy of 98.3% is achieved when K equals 70 for each emotion. A smaller value of K suffices for JAFFE data set since it contained lesser number of images. The final experimental results presented are based upon SBoF-SSIFT feature vector length of 900 for CK+ and 420 for JAFFE data set.

4.2 Performance analysis for CK+ data set

The accuracy is evaluated by using SVM classifiers with different kernels. The performance of linear, RBF, and polynomial kernels are evaluated. The resulting SBoF-SSIFT feature performance is not good when used with RBF and polynomial kernel. As, in SSIFT each feature point is described by 128 dimensions resulting in high discrimination power. These kernels are used to map data from lower to higher dimensional space and resulted in an accuracy of 69.5 and 73.7% for RBF and polynomial kernel, respectively. With a feature vector length of 900 for SBoF-SSIFT, linear kernel successfully discriminates the input feature space. The confusion matrix for CK+ data set using SBoF-SSIFT with a linear kernel SVM and a feature vector length of 900 is shown in Table 4.

4.3 Performance analysis for JAFFE data set

A maximum recognition accuracy of 98.33% is achieved for a feature vector length of 420 on the basis of SBoF-SSIFT with SVM linear kernel. The confusion matrix for JAFFE data set using these parameter settings is presented in Table 5.

The results on the basis of varying feature descriptors and classifiers are presented in Table 6. SVM classifier with RBF and polynomial kernel does not perform well on the basis of SBoF-SSIFT and SBoF-SSURF features because of their high dimensionality, which further increases when mapped by these kernels to higher dimensional space resulting in less discrimination power. However, good recognition accuracy of SBoF-SSIFT for

Table 1 Accuracy achieved with varying facial ROI from random images

	Ha	Sa	Su	Di	Fe	An	Average
scaled image	94	79	85.1	82	75	82.5	83.9
wavelet transformed image (complete face)	93.9	78.7	84.9	81.5	75.9	82.2	83.5
wavelet transformed image ($\alpha = 0.15, \beta = 0.15$)	95.1	78.7	87.7	90.8	60.4	86.7	84.1
wavelet transformed image ($\alpha = 0.10, \beta = 0.12$)	97.6	80.3	95.9	86.2	67.2	77.8	**85.7**
wavelet transformed image ($\alpha = 0.7, \beta = 0.10$)	96.1	77.1	91.3	82.5	60.5	75.7	80.5

Ha, happy; Sa, sad; Su, surprise; Di, disgust; Fi, fear; An, anger.

The bold values represent the parameter values that give the best results and are selected to achieve the final results.

Table 2 Accuracy with varying number of patches selected with different segmentation levels

Segmentation levels	Ha	Sa	Su	Di	Fe	An	Average
1×1 (orderless)	88.8	70	90	87.5	60	0	64.9
1×1, 2×2	100	70	100	87.5	50	0	66.7
1×1, 2×2, 4×4	88.9	90	100	87.5	70	0	71.9
1×1, 2×2, 4×4, 8×8	100	80	100	87.5	70	40	80.7
1×1, 2×2, 4×4, 8×8, 16×16	90	70	100	67.5	60	40	73.7
whole face patches ($L = 1, 2$), upper and lower face patch ($L = 1$)	100	90	100	100	90	100	96.5

© The Institution of Engineering and Technology 2018

Table 3 Accuracy with varying number of clusters for CK+ and JAFFE data set

	Clusters per emotion	Ha	Sa	Su	Di	Fe	An	Average
CK+ data set	60	92.6	89.3	91.4	87.6	74.1	85.3	86.8
	70	95.9	89.3	98.3	88.5	76	89	89.6
	80	99.1	96.4	97.4	93.8	83.9	87.1	93.1
	90	97.5	94.6	97.4	99.1	92.8	94.5	96.1
	150	98.4	98.2	100	100	97.3	97.2	**98.5**
	200	99.1	99.1	98.2	100	99.1	95.4	98.5
JAFFE data set	60	100	80	100	87.5	80	100	91.2
	70	100	80	100	87.5	80	100	**98.3**
	80	100	60	100	100	70	100	87.7

The bold values represent the parameter values that give the best results and are selected to achieve the final results.

Table 4 Confusion matrix for SBoF-SSIFT for CK+ data set using linear kernel SVM for a feature vector of length 900

	Ha	Sa	Su	Di	Fe	An	Not predicted
Ha	98.4	0	0	0	0	0	0.7
Sa	0	98.2	0	0	0.9	0	1.6
Su	0	0	100	0	0	0	0
Di	0	0.2	0	100	0	0	0
Fe	0.9	0	0	0	97.3	0	1.8
An	0	0	0	0	0	97.2	2.8

Table 5 Confusion matrix for SBoF-SSIFT for JAFFE data set using linear kernel SVM for a feature vector of length 420

	Ha	Sa	Su	Di	Fe	An	Not predicted
Ha	100	0	0	0	0	0	0
Sa	0	95	0	0	0	0	5
Su	0	0	100	0	0	0	0
Di	0	0	0	100	0	0	0
Fe	0	0	0	0	95	0	5
An	0	0	0	0	0	100	0

Table 6a Recognition accuracy of the proposed feature descriptors for SVM and KNN classifiers

		JAFFE data set	
		SBoF-SSIFT	SBoF-SSURF
		Feature vector length = 420	
SVM classifier	linear kernel	**98.3**	**95.4**
	RBF	71.5	78.2
	polynomial	73.7	80.7
KNN classifier	$K = 1$	**94.7**	**93.6**
	$K = 3$	87.7	85.1
	$K = 5$	77.2	75.8

The bold values represent the parameter values that give the best results and are selected to achieve the final results.

Table 6b

		CK+ data set	
		SBoF-SSIFT	SBoF-SURF
		Feature vector length = 900	
SVM classifier	linear kernel	**98.5**	**97.4**
	RBF	69.5	57.3
	polynomial	73.0	62.0
KNN classifier	$K = 1$	95.7	94.9
	$K = 3$	**96.1**	**95.0**
	$K = 5$	95.8	89.6

The bold values represent the parameter values that give the best results and are selected to achieve the final results.

CK+ is achieved because of relatively large size of CK+ data set when compared with the feature dimension. SBoF-SSURF also performs well with a linear SVM kernel for both CK+ and JAFFE data set. For KNN classifier, best performance is achieved for a *K*-value of 1 and 3 for JAFFE and CK+ data set, respectively.

4.4 Computational performance

The computational efficiency analysis of an algorithm can be performed either theoretically or empirically. In empirical analysis, computational parameters from experimentation on an available system are used, but these can vary with a change in system

© The Institution of Engineering and Technology 2018

Table 7 Comparison of the proposed solution with state-of-the-art methods for six basic emotions

Approaches	Data set	Average accuracy
DSIFT, SVM, bag of words [28]	CK+	95.8
SWT, ANN [29]	CK+, JAFFE, self-recorded	96.6, 98.8, 94.1
HoG, SVM [30]	CK+	95.8
AAM-SIFT, fuzzy *C*-means with SVM [31]	BU-3DFE	81.4
proposed	CK+, JAFFE	98.5, 98.3

computational power. Whereas, in theoretical analysis, the size of input determines the computational power required [27]. In the proposed system, theoretical analysis is performed for the computational time analysis. The original image size after segmenting for the core central facial area is reduced by wavelet transform to 48×64 pixels. The total number of input pixels that are processed after preprocessing is significantly reduced. A reduction in the input size translates into reduced computational burden. This computational analysis is independent of the system used to run the system as well the optimality of the program code.

4.5 Performance comparison

The proposed algorithm has been compared with state-of-art algorithms that have been reported in the literature and are based on CK+ and JAFFE data sets with six basic emotions. The comparative analysis is summarised in Table 7 with a clear demonstration that the proposed solution resulted in improved emotion recognition accuracy. In [28], DSIFT is used for facial emotion recognition that extracts SIFT descriptor for every second pixel. The algorithm divides a 96×96 face image up to five levels, i.e. $L = 0, 1, 2, 3, 4$ each with 2^L patches resulting in 341 segments of a single image. After concatenation, all segment features are classified by using SVM and resulted in a recognition accuracy of 95.8% on CK+ data set. The use of DSIFT translates into larger feature vector that has been addressed in our proposed method by using spatial SIFT. In [31], SIFT features are extracted for the facial feature points that are detected using AAM technique. Those points are used that are located near the mouth boundary and are effected by the deformation in the mouth shape during a facial expression. Different facial regions are then weighted by using the fuzzy *C*-means clustering algorithm for the computation of final feature vector, which is used as an input to the SVM classifier. Experiments are performed on BU-3DFE database [32] resulting in an average recognition accuracy of 81.4%. An emotion recognition system is developed [30] using HoG features to detect happiness, anger, surprise, and disgust emotions. The number of histogram bins per cell and cell dimension are tuned for optimum recognition accuracy. An average accuracy of 95.8% is achieved by using a cell size of seven pixels and seven orientation bins. In [29], frequency domain features are used on three different data sets including one locally generated using Microsoft kinect sensor. The algorithm can be computationally expensive and does not account for the image transformations that could occur in most real-world applications.

In comparison, our proposed method relies on image spatial information and is most suited for data sets with varying images as is evident from the performance on CK+ data set. The results are compared with both model and featured-based approaches. On CK + data set, the proposed scheme improves the classification accuracy and the best results are achieved for spatial SIFT when used with spatial BoF and SVM with linear kernel. Spatial SURF and KNN algorithms also achieve better or comparative performance when compared with these methods.

5 Conclusion

In this paper, an automatic facial emotion recognition system is proposed on the basis of hybrid SBoF-SSIFT and SBoF-SSURF feature descriptors. A detailed analysis is performed and it is concluded that SBoF-SSIFT hybrid descriptor suites better for emotion recognition using facial images. An accuracy of 98.33 and 98.5% is achieved on JAFFE and CK+ data sets, respectively. These high accuracies are achieved after careful selection of relevant facial regions, and hyper-parameters for hybrid feature descriptors. Extensive experiments are performed using SBoF-SSIFT with scaled, wavelet transformed, with side irrelevant area and without irrelevant area. It is concluded that the accuracy increases with the elimination of the face area that remains unchanged during facial expression. If these areas are not removed, feature overlapping occurs and results in performance degradation. It has been shown that the recognition performance depends on the number of clusters for codebook generation, number of detected features, levels for image segmentation, and size of training data set. It has also been shown that proposed methods based on SBoF-SSIFT and SBoF-SSURF are highly robust and powerful for emotion recognition. The intensity value-related constraints are not hard coded in the proposed system; hence, it can be tuned according to different environments with the help of training data set. In future, the system can be extended to develop a recommender system that will be able to recommend multimedia content to user on the basis of recognised emotions.

6 References

[1] Chakraborty, A., Konar, A., Chakraborty, U.K., *et al.*: 'Emotion recognition from facial expressions and its control using fuzzy logic', *IEEE Trans. Syst. Man Cybern. A Syst. Humans*, 2009, **39**, (4), pp. 726–743

[2] Pantic, M., Rothkrantz, L.J.M.: 'Automatic analysis of facial expressions: the state of the art', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (12), pp. 1424–1445

[3] Cornelius, R.R.: 'Theoretical approaches to emotion'. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000

[4] Busso, C., Deng, Z., Yildirim, S., *et al.*: 'ACM'. Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information, 2004, pp. 205–211

[5] Vijayakumari, V.: 'Face recognition techniques: a survey', *World J. Comput. Appl. Technol.*, 2013, **1**, (2), pp. 41–50

[6] Mishra, B., Fernandes, S.L., Abhishek, K., *et al.*: 'Facial expression recognition using feature based techniques and model based techniques: a survey'. 2015 2nd Int. Conf. Electronics and Communication Systems (ICECS), 2015, pp. 589–594

[7] Lartillot, O., Toiviainen, P., Eerola, T.: 'A matlab toolbox for music information retrieval'. Data Analysis, Machine Learning and Applications, 2008, pp. 261–268

[8] Bhatti, A.M., Majid, M., Anwar, S.M., *et al.*: 'Human emotion recognition and analysis in response to audio music using brain signals', *Comput. Hum. Behav.*, 2016, **65**, pp. 267–275

[9] Lee, Y.H., Han, W., Kim, Y., *et al.*: 'Robust emotion recognition algorithm for ambiguous facial expression using optimized AAM and k-NN', *Int. J. Secur. Appl.*, 2014, **8**, (5), pp. 203–212

[10] Shbib, R., Zhou, S.: 'Facial expression analysis using active shape model', *Int. J. Signal Process. Image Process. Pattern Recogn.*, 2015, **8**, (1), pp. 9–22

[11] Huang, X., Zhao, G., Pietikäinen, M., *et al.*: 'Dynamic facial expression recognition using boosted component-based spatiotemporal features and multi-classifier fusion'. Int. Conf. Advanced Concepts for Intelligent Vision Systems, 2010, pp. 312–322

[12] Whitehill, J., Omlin, C.W.: 'Local versus global segmentation for facial expression recognition'. FGR 2006. 7th Int. Conf. Automatic Face and Gesture Recognition, 2006, 2006, pp. 357–362

[13] Cao, Y., Wang, C., Li, Z., *et al.*: 'Spatial-bag-of-features'. 2010 IEEE Conf. IEEE Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3352–3359

[14] Mele, K., Suc, D., Maver, J.: 'Local probabilistic descriptors for image categorisation', *IET Comput. Vis.*, 2009, **3**, (1), pp. 8–23

[15] Mikolajczyk, K., Schmid, C.: 'A performance evaluation of local descriptors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1615–1630

[16] Lowe, D.G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110

[17] Sivic, J., Zisserman, A.: 'Video Google: a text retrieval approach to object matching in videos', 2003, pp. 1470–1477

[18] Koller, D., Sahami, M.: '*Hierarchically classifying documents using very few words*' (Stanford InfoLab, California, 1997)

[19] Zhao, X., Zhang, S.: 'Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding', *EURASIP J. Adv. Signal Process.*, 2012, **2012**, (1), p. 20

© The Institution of Engineering and Technology 2018

[20] Tie, Y., Guan, L.: 'Automatic landmark point detection and tracking for human facial expressions', *EURASIP J. Image Video Process.*, 2013, **2013**, (1), p. 8
[21] Kamarol, S.K.A., Jaward, M.H., Parkkinen, J.*, et al.*: 'Spatiotemporal feature extraction for facial expression recognition', *IET Image Process.*, 2016, **10**, (7), pp. 534–541
[22] Donia, M.M., Youssif, A.A., Hashad, A.: 'Spontaneous facial expression recognition based on histogram of oriented gradients descriptor', *Comput. Inf. Sci.*, 2014, **7**, (3), p. 31
[23] Carcagnì, P., Del Coco, M., Leo, M.*, et al.*: 'Facial expression recognition and histograms of oriented gradients: a comprehensive study', *SpringerPlus*, 2015, **4**, (1), p. 645
[24] Viola, P., Jones, M.: 'Rapid object detection using a boosted cascade of simple features'. Proc. 2001 IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 2001. CVPR 2001, 2001, vol. 1, pp. I–I
[25] 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 (CVPRW), 2010, pp. 94–101
[26] Lyons, M.J., Akamatsu, S., Kamachi, M.*, et al.*: 'The Japanese female facial expression (JAFFE) database'. Proc. Third International Conf. Automatic Face and Gesture Recognition, 1998, pp. 14–16
[27] Kao, M.Y.: '*Encyclopedia of algorithms*' (Springer Science & Business Media, Berlin, 2008)
[28] Sikka, K., Wu, T., Susskind, J.*, et al.*: 'Exploring bag of words architectures in the facial expression domain'. Computer Vision–ECCV 2012. Workshops and Demonstrations, 2012, pp. 250–259
[29] Qayyum, H., Majid, M., Anwar, S.M.*, et al.*: 'Facial expression recognition using stationary wavelet transform features', *Math. Probl. Eng.*, 2017, **2017**, 9854050
[30] Del Coco, M., Carcagnì, P., Palestra, G.*, et al.*: 'Analysis of hog suitability for facial traits description in FER problems'. Int. Conf. Image Analysis and Processing, 2015, pp. 460–471
[31] Ren, F., Huang, Z.: 'Facial expression recognition based on AAM–sift and adaptive regional weighting', *IEEJ Trans. Electr. Electron. Eng.*, 2015, **10**, (6), pp. 713–722
[32] Savran, A., Alyüz, N., Dibeklioğlu, H.*, et al.*: 'Bosphorus database for 3d face analysis', Proc. European Workshop on Biometrics and Identity Management, Roskilde, Denmark, May 2008, pp. 47–56

© The Institution of Engineering and Technology 2018