Applied Intelligence (2018) 48:2912–2929
https://doi.org/10.1007/s10489-017-1121-y

Accurate and robust facial expression recognition system using real-time YouTube-based datasets

Muhammad Hameed Siddiqi[1]

Published online: 13 January 2018

Abstract

This paper presents an accurate and robust real-time FER system. In this system, an unsupervised technique based on active contour (AC) model is adopted in order to detect and extract the human faces automatically from the facial expression frames. In this model, the combination of two energy functions like Chan-Vese (CV) energy and Bhattacharyya distance functions were exploited that not only minimize the dissimilarities within the object (face) but also maximize the distance between the object (face) and background. Moreover, we extracted the facial features by proposing a new feature extraction method in order to solve the limitations of the previous works of the feature extraction. Similarly, in this system, we also proposed the usage of a robust non-linear feature selection method called stepwise linear discriminant analysis (SWLDA) that focuses on selecting localized features from facial expression images and discriminating their classes based on regression values (i.e., partial *F-test*). Finally, the system has been trained by employing hidden Markov model (HMM) to label the expressions. Unlike most of the previous works that were evaluated using a single dataset in a controlled environment, the performance of the proposed system have been assessed by employing three different spontaneous datasets that have been collected in naturalistic environments. 10–fold cross validation rule has been exploited for the whole experiments. In last, a set of experiments were also performed to assess the effectiveness of each module of the proposed approaches separately. The proposed system achieved weighted average recognition rate (95%) across three different YouTube-based datasets against the existing state-of-the-art methods.

Keywords YouTube · Active contour · Level set · Wavelet transform · SWLDA · HMM · Video surveillance

1 Introduction

Facial expression recognition (FER) has a significant contribution in our daily life communication, which has achieved much attention in many real applications, especially in human computer interaction (HCI) [53, 99], robot control, and driver state surveillance [97, 102], clinical decision-making processes [100], security access control [89], neuroscience, psychology, and cognitive science [12]. These systems are used to improve the communications between the user and computer according to the needs of the user [30, 53]. However, it is still a challenging task for such systems to achieve a robust recognition rate of facial expressions due to the difficulty in precisely extracting the suitable emotional features from the expression images [102]. Most of these features are represented either in static, or dynamic, or point-based geometric, or region-based appearance [102].

The are two types of expression recognition systems: posed-based expression recognition systems [98] and spontaneous expression recognition systems [106]. In the first system, the expressions have been produced artificially, which means that the subjects are forced to performed the expressions [11]. While, in the second system, the expressions were performed spontaneously that are observed on a day-to-day basis, such as during conservations or while watching movies [11]. The focus of this study is posed and spontaneous FER in naturalistic environments which is one of the limitations of the existing works.

There are two types of expression classification. The first one is frame-based classification, in which, only the current frame is used with or without a reference image (neutral face image) to recognize the expressions. On the other hand, in sequence-based classification, the temporal information of

✉ Muhammad Hameed Siddiqi
mhsiddiqi@ju.edu.sa

[1] Department of Computer and Information Sciences, AlJouf University, Sakaka, Kingdom of Saudi Arabia

the sequences are employed to recognize the expressions of one or more frames [90]. In sequence-based methods, the geometrical displacement of facial feature points between the current frame and the initial frame are calculated [76]; while, frame-based methods do not have this property. The temporal information of expression in sequences of frames is important for facial expression analysis [75]; therefore, we employed sequence-based classification that is also one of the limitations of the existing works.

Typical FER system consists of four basic modules: face detection, feature extraction, feature selection, and recognition. Basically, the face contains most of the expression-related information; therefore, in face detection module, first the face in a given image is detected. The feature extraction module deals with getting the distinguishable features from each facial expression shape and quantizing it as a discrete symbol [82, 84]. The feature selection module is used for selecting a subset of relevant features from a large number of features extracted from the input data. In recognition module, a classifier is first trained with training data and then used to generate the labels of the facial expressions contained in the incoming video data [86].

Though, there lots of works have been done to achieve high recognition rates in controlled environments; however, their accuracies consistently degrade in naturalistic environments [63]. Moreover, less amount of works can be found that automatically detect the faces, and extract and selects the features from the facial muscles. So, it is still a challenging task for the existing works to extract the features in a robust way [102] in naturalistic environments. Therefore, the objective of this paper is to propose an unsupervised automatic face detection and extraction model based on active contour (AC) model. The proposed AC model is based on level set that is the combination of two energy functions like Chan-Vese [19] energy and Bhattacharyya distance [45] functions. The proposed AC model is most robust to noise and illuminations, which not only minimize the dissimilarities within the object (face) but also maximize the distance between the two regions such as face and the background.

Once the face has been detected, then for the feature extraction, we proposed a new method based on wavelet transform (especially symlet wavelet). To obtain the feature vectors, symlet wavelet family was tested for which the image was decomposed up to 4 levels. The proposed feature extraction method extracts the most prominent features; however, there might be some redundancy in the features. Therefore, we proposed the usage of a non-linear feature selection method called stepwise linear discriminant analysis (SWLDA) that has been employed to the selected feature space. SWLDA can select the most informative features taking the advantage of the forward selection model and can remove the irrelevant feature by taking the advantage of the backward regression model.

The rest of the paper is organized as follows. Section 2 provides related works and their limitations. Section 3 presents an overview of the proposed approaches. The experimental setup and results with some discussion are presented in Sections 4 and 5 respectively. Finally, the paper has been concluded with some future directions in Section 6.

2 Literature review

2.1 Existing face detection methods

Generally, before recognizing the expressions, the face must be located, detected, and extracted in expression frames. Because it is the face that contains most of the expression-related information. There lots of works have been done for face detection and extraction; however, each of them has its own limitation. Recent works [4, 14, 25, 28] were proposed for the purpose of face detection that utilized artificial neural networks (ANNs). However, the common limitation of ANNs is that the neuron model used in neural networks ignores most of the characteristics of its counterpart [93]. Also computational wise, ANNs are very much expensive and difficult to implement. Similarly, skin-tone based methods [7, 32, 38, 62, 94] were utilized by for face detection and extraction in FER systems. However, skin-tone based methods are very sensitive to illumination like under varying lighting conditions [61], and skin color varies from person to person by employing different cameras at static lighting conditions [71]. Moreover, the presence of the skin color is completely reliant on the brightness and the color temperature of the light source [87], which may cause misclassification.

Some appearance-based methods [24, 34, 40, 65] also proposed for face detection. These methods showed better performance in control environment; however, their performance degrade with wide variation of head pose and illumination in real world environments [23]. Moreover, holistic approaches [77, 78, 105] were employed for face detection which utilized global information rather than local. However, holistic approaches consider the entire face and the detection of facial features are difficult in holistic approaches when there is wide range of rotation, scale, head pose, and illumination variation [13]. The authors of [17, 26, 27, 67] employed invariant feature-based methods for face detection on spontaneous-based datasets and achieved better performance. However, the performance of invariant feature-based methods degrade with the environmental change. Moreover, for these methods, the accurate normalization of the face is required against pose, illumination, scale, and occlusion [33]. Also, computational wise, these methods are much more expensive [22].

2.2 Existing feature extraction methods

Feature extraction is a process that deals with getting the distinguishable features from each facial expression shape and quantizing it as a discrete symbol [81]. A well-defined feature extraction algorithm is that which improves the recognition rate more efficiently and effectively [95]. There are some parts of the face such as eyes, mouth, nose and forehead from where we can extract the most prominent and informative features. According to the face descriptors, there are two types of features, global features and local features. In global features, the features are extracted from the whole face while in local features, the features are extracted from some parts of the face.

The methods that used for global feature extraction named holistic methods include nearest features line-based subspace analysis [64], Eigenfaces and Eigenvector [2, 50] and [46], Fisherfaces [1], global features [60], neural network [1, 74], independent component analysis (ICA) [48, 54], and principal component analysis (PCA) [31, 35, 92]. Moreover, some frequency-based methods and Gabor wavelet were utilized by [57] and [52] respectively. However, all these holistic methods do not know what exact facial features are the most important for FER systems. Moreover, these methods ignore higher order correlation value and might not work if the data sources are dependent [20]. Furthermore, these techniques do not have the capability to handle that data in which the classes are far away from Gaussian and also if there is a small sample size data then these methods have the problem to process it [20]. Most of these methods work well in a control environment for face detection and recognition. However, their performance degrades in FER with the variation of illumination, pose, facial expression, occlusion and aging [18]. Furthermore, complexity-wise, most of these techniques are much more expensive because of considering the whole face and lots of memory are required [21]. These methods preferred only for face recognition because it preserves the interrelations between facial parts information of the face that are very important for FER systems [101].

On the other hand, local feature-based methods have been proposed to compute the local descriptors from some parts of the face and then integrate these information into one descriptor. These methods include local feature analysis (LFA) [42], Gabor features [88], non-negative matrix factorization (NMF) [5, 56], local non-negative matrix factorization (LNMF) [16], and local binary pattern (LBP) [43]. Among these methods, LBP achieved better performance. However, LBP does not provide the directional information of the facial frame [85]. Recent works have been proposed in order to solve the limitations of LBA. These methods include local transitional pattern (LTP) [3], local directional pattern (LDP) [39], and local directional pattern variance (LDPv) [44]. However, most of these methods exploited other information instead of employing intensity to overcome the problems due to noise and illumination change [70]. Moreover, the performance of these methods still degrade in non-monotonic illumination change, noise variation, change in pose, and expression conditions [70]. Similarly, the authors of [69] exploited local fisher discriminant analysis (LFDA) in their systems for FER. However, LFDA might fail to determine the essential assorted structure when the face image space is highly nonlinear [96]. Furthermore, the authors of [72] employed pixel and color segmentation for feature extraction to detect the facial expressions. However, the performance of this approach degrades with the variation of illumination.

2.3 Existing feature selection and dimension reduction methods

Dimension reduction or feature selection is the essential part before the classification. The dimensions of a feature space can be reduced by extracting discriminating features that are reliant on exploiting the total distribution of the data, while lessening the differences within the classes. that the feature values for the six classes are highly merged, which can result in a high misclassification rate. The use of inappropriate coefficients results in high within-class differences and low between-class differences. Therefore, a method is required to address the aforementioned problem and to reduce the dimension space and to increase the class separability. This idea is employed by various feature selection methods of machine learning, mainly Principal Component Analysis (PCA) [15], Linear Discriminant Analysis (LDA) [58], kernel discriminant analysis (KDA) [59], and Generalized Discriminant Analysis (GDA) [8].

2.3.1 Principal component analysis (PCA)

PCA is one of the most well-known methods for feature selection and dimension reduction. PCA is a second-order approach that offers an easy way of reducing a complex set of data by assigning it onto a space with a small dimension, while protecting as much of the unpredictability as possible. PCA fabricates the best linear least-squares decomposition of a training set. This method has the benefit of being linear and makes no hypothesis concerning the data distribution. The role of PCA is to estimate the original data with lower dimensional features, which represents the data economically. It also focuses on the global features of the gray-scale faces. In this case, there is a strong correlation among observed variables. So, for this work, the main purpose for using PCA was to express the large 1D vector of pixels constructed from the 2D image into the

compact principal components of the feature space. This is called eigenspace projection. The primary job of PCA is to compute the eigenvectors of the covariance data matrix M and then, by the combination of a few top eigenvectors, the approximation is done. It is the most common feature extraction technique that has been widely employed in FER systems. However, PCA has poor discriminating power within the class and computational wise it is much expensive method. Therefore, Linear Discriminant Analysis (LDA) has been exploited to resolve the shortcomings of PCA.

2.3.2 Linear discriminant analysis (LDA)

LDA maximizes the ratio of between-class variance to within-class variance in any particular data set, thereby guaranteeing maximal separability. The use of LDA for data classification is applied to classification problems in speech recognition [29]. LDA produces an optimal linear discriminant function that maps the input into the classification space on which the class identification of the samples is decided. LDA easily handles the case in which the within-class frequencies are unequal and their performances have been examined on randomly generated test data. Thus, LDA maximizes the total scatter of the data while minimizing the within scatter of the classes. The use of LDA, however, failed in resolving the overlap or low between-class variance among the facial expressions. LDA is a linear technique, which limits its flexibility when applied to complex datasets. Moreover, the assumption made by LDA that all classes share the same within-class covariance matrix is not a valid one. In addition, large amounts of data are necessary to generate robust transforms for LDA, and there may be insufficient data to robustly estimate transforms to separate the classes. Moreover, linear discriminant analysis (LDA)-based methods suffer from the limitations that their optimality criteria are not directly associated to the classification capability of the achieved feature representation [55]. Moreover, LDA is much sensitive than PCA on partial occlusion. For more details on LDA, please refer to [10]. Thus, we believe that the use of LDA will not essentially yield an improvement in the performance of the FER systems. Moreover, LDA cannot provide better classification rate due to the aforementioned limitations. Therefore, Kernel Discriminant Analysis (KDA) has proposed to solve the limitations of LDA.

2.3.3 Kernel discriminant analysis (KDA)

KDA is a non-linear discriminating approach, which seeks non-linear discriminating features using kernel techniques. However, KDA does not have the capability to provide better performance in the case if the face images belong to the same subjects are scattered rather than dispersed as clusters [66]. Moreover, during the model evaluation, KDA needs the entire data set for training that is inappropriate for FER systems. A recent non-linear feature selection method such as Generalized Discriminant Analysis (GDA) has been proposed in order to solve the shortcomings of KDA.

2.3.4 Generalized discriminant analysis (GDA)

GDA plots the input data (training data) into a classification (high dimensional feature) space by generating an optimum discriminant function. GDA builds a nonlinear feature space for discrimination that is not separable by linear methods. The discriminant eigenvector has been calculated by GDA in the feature space in an efficient way. However, some eigenvectors may be degenerated by GDA in small sample size data case [104]. Furthermore, if there are slight changes in the training data then the solution of GDA might not be stable and perhaps are not optimal in terms of the discriminant ability [104]. Moreover, GDA takes lots of time during the data training and testing [6].

Therefore, to come up with the limitations of PCA, LDA, KDA, and GDA, we proposed the use of a robust technique called Stepwise Linear Discriminant Analysis (SWLDA) for the FER systems with the aim of extracting localized features from the some parts of the face that previous feature selection and dimension reduction techniques were limited in analyzing. The proposed technique is based on two processes: forward and backward regression processes. In the forward process, the most correlated features from the response are selected from the regression model, and the selected features are based on partial *F-test* values from the feature space. In the backward process, the least significant values are removed from the regression model (i.e., the lowest *F-test* values). In both processes, *F-test* values are calculated on the basis of defined class labels.

3 Proposed facial expression recognition (FER) system

3.1 Proposed face detection and extraction algorithm

Mostly, the performance of the FER systems reliant on accurate face detection. In the field of image segmentation, since it was first introduced by [49], active contour (AC) model has attracted much attention.

An active contour model is a deformable spline influenced by constraint and image forces that pull it toward object contours. It tries to move into a position where its energy is minimized. Active contour tries to improve

by imposing desirable properties such as continuity and smoothness to the contour of the object, which means that the active contour approach adds a certain degree of prior knowledge for dealing with problem of finding the object contour.

Recently, Chan-Vese (CV) proposed in [19] a novel form of active contour for object segmentation based on level set framework. Its energy function is defined by.

$$\mathrm{F(C)} = \int\limits_{\mathrm{inside(C)}} |\mathrm{I(x)} - \mathrm{c_{in}}|^2 \mathrm{dx} + \int\limits_{\mathrm{outside(C)}} |\mathrm{I(x)} - \mathrm{c_{out}}|^2 \mathrm{dx} \quad (1)$$

where c_{in} and c_{out} are respectively the average intensities inside the variable curve C. Compared to the other AC contour models, the CV active model can detect the faces more exactly sice it does not need to smooth the initial facial image (via the edge function $g|\nabla I_\sigma|^2$, even if it is very noisy. In other words, this model is more robust to noise. However, the convergence of CV active contour model depends on the homogeneity of the segmented faces. When the inhomogeneity becomes large, the CV active model provides unsatisfactory results. Moreover, unlike other active contour models which rely much on the gradient of the image as the stopping term and thus have unsatisfactory performance in noisy images. The reason CV active contour model fails is that, a segment is represented by only its mean value, which is not sufficient for a highly inhomogeneous object. Moreover, the CV active contour model does not use the edge information but utilizes the difference between the regions inside and outside of the curve, making itself one of the most robust and thus widely used techniques for image segmentation, especially, in the area of face detection. Moreover, the global minimum of the above energy functional does not always guarantee the desirable results. The unsatisfactory result of the CV AC in this case is due to the fact that it is trying to minimize the dissimilarity within each segment but does not take into account the distance between different segments.

The proposed methodology is to incorporate an evolving term based on the Bhattacharyya distance to the CV energy functional that minimizes the dissimilarities within the object and maximizes the distance between the two regions. The proposed energy function is:

$$E_0(C) = \beta F(C) + (1-\beta)B(C) \quad (2)$$

where $\beta \in [0, 1]$. Note that to be comparable to the $F(C)$ term, in practice, $B(C)$ is multiplied by the area of the facial image (frame) because its value is always within the interval $[0, 1]$ whereas $F(C)$ is calculated based on the integral over the facial image plane. As usual [19], one regularizes the solution by constraining the length of the curve and the area of the region inside ir, yielding the total energy functional as

$$\begin{aligned} E(C) &= \gamma Length(C) + \eta Area(inside(C)) \\ &\quad + \beta F(C) + (1-\beta)B(C) \end{aligned} \quad (3)$$

where γ and η are non-negative constants.

The intuition behind the proposed model, in $E(C)$, is that we seek for a curve which is regular (the first two terms) and partitions the facial image into regions such that the differences within each region are minimized (the $F(C)$ term) like reducing environmental effects and the distance between the two regions (like human face and background) is maximized (the term $B(C)$ term).

For the level-set formulation, let us define ϕ as the level-set function, $I : \Omega \to \mathrm{Z} \subset R^n$ as a certain image feature such as intensity, color, texture, or a combination thereof, and $H(\bullet)$ and $\delta_0(\bullet)$ as the Heaviside and the Dirac function respectively.

$$H(u) = \begin{cases} 1, \ if \ u \geq 0 \\ 0, \ if \ u < 0 \end{cases} \qquad \delta_0(u) = \frac{d}{du} H(u) \quad (4)$$

The energy function can then be rewritten as

$$\begin{aligned} E(\phi) &= \gamma \int_\Omega |\nabla H(\phi(x))| dx + \eta \int_\Omega H(-\phi(x)) \\ &\quad + \beta \begin{bmatrix} \int_\Omega |I(x) - c_{in}|^2 H(-\phi(x)) \\ + \int_\Omega |I(x) - c_{out}|^2 H(\phi(x)) \end{bmatrix} \\ &\quad + (1-\beta) \int_Z \sqrt{p_{in}(z) p_{out}(z)} dz \end{aligned} \quad (5)$$

where

$$\begin{aligned} p_{in}(z) &= \frac{\int_\Omega \delta_0(z - I(x)) H(-\phi(x)) dx}{\int_\Omega H(-\phi(x)) dx} \\ p_{out}(z) &= \frac{\int_\Omega \delta_0(z - I(x)) H(\phi(x)) dx}{\int_\Omega H(\phi(x)) dx} \end{aligned} \quad (6)$$

In general form, it reads

$$E(\phi) = \int_\Omega \underbrace{f\left(\phi, \phi_{x_1}, \phi_{x_2}, \ldots, \phi_{x_n}\right) dx}_{\overline{F}(\phi)} + (1-\beta)B(\phi) \quad (7)$$

where $X = [x_1, x_2, \ldots, x_n] \in R^n$, $\phi_{x_i} = \frac{\partial \phi}{\partial_{x_i}}$, $i = \overline{1..n}$, $B(\phi) = \int_z \sqrt{p_{in}(z) p_{out}(z)} dz$. The first variation (w.r.t $\phi(x)$) is given by

$$\frac{\delta E}{\delta \phi} = \frac{\delta \bar{F}}{\delta \phi} + (1-\beta)\frac{\delta B}{\delta \phi} \quad (8)$$

Using Euler-Lagrange equation, one has

$$\frac{\delta \bar{F}}{\delta \phi} = \frac{\partial f}{\partial \phi} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \frac{\partial f}{\partial \phi x_i} = \delta_0(\phi)[-\eta - \beta(I - c_{in})^2 + \beta(I - c_{out})^2 - \gamma k] \quad (9)$$

On the other hand,

$$\frac{\delta B}{\delta \phi} = \frac{1}{2} \int_z \left(\begin{array}{l} \frac{\partial p_{in}(z)}{\partial \phi} \sqrt{\frac{p_{out}(z)}{p_{in}(z)}} \\ + \frac{\partial p_{out}(z)}{\partial \phi} \sqrt{\frac{p_{in}(z)}{p_{out}(z)}} \end{array} \right) dz \quad (10)$$

where $p_in(z)$ and $p_out(z)$ are given in (6). Differentiating them w.r.t $\phi(x)$, one obtains

$$\begin{aligned} \frac{\partial p_{in}(z)}{\partial \phi} &= \frac{\partial_0(\phi)}{A_{in}} [p_{in}(z) - \delta_0(z - I)] \\ \frac{\partial p_{out}(z)}{\partial \phi} &= \frac{\partial_0(\phi)}{A_{out}} [\delta_0(z - I) - p_{out}(z)] \end{aligned} \quad (11)$$

where A_in and A_out are respectively the areas inside and outside the contour and are given by

$$A_{in} = \int_\Omega H(-\phi(x))dx \qquad A_{out} = \int_\Omega H(\phi(x))dx \quad (12)$$

Substituting (11) into (10) and taking some simple modification, one obtains

$$\frac{\delta B}{\delta \phi} = \delta_0(\phi) V(x) \quad (13)$$

where

$$V(x) = \frac{B}{2} \left(\frac{1}{A_{in}} - \frac{1}{A_{out}} \right) + \frac{1}{2} \int_z \frac{\delta_0(z - I(x))}{\left(\frac{1}{A_{out}} \sqrt{\frac{p_{in}(z)}{p_{out}(z)}} - \frac{1}{A_{in}} \sqrt{\frac{p_{out}(z)}{p_{in}(z)}} \right)} dz \quad (14)$$

Combining (8), (9), and (13), one can derive the first variation of $E(\phi)$ as

$$\frac{\partial E}{\partial \phi} = \delta_0(\phi) \left[\begin{array}{l} -\gamma k - \eta - \beta (I - c_{in})^2 \\ + \beta (I - c_{out})^2 + (1 - \beta) V \end{array} \right] \quad (15)$$

Hence, the evaluation flow associated with minimizing the energy functional in (5) is given as

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E}{\partial \phi} = \delta_0(\phi) \left\{ \begin{array}{l} \gamma k + \eta + \beta[(I - c_{in})^2 + (I - c_{out})^2] \\ -(1 - \beta) \left[\begin{array}{c} \frac{B}{2} \left(\frac{1}{A_{in}} - \frac{1}{A_{out}} \right) \\ \delta_0(z - 1) \\ + \frac{1}{2} \int_z \left(\begin{array}{l} \frac{1}{A_{out}} \sqrt{\frac{p_{in}}{p_{out}}} \\ - \frac{1}{A_{in}} \sqrt{\frac{p_{out}}{p_{in}}} \end{array} \right) dz \end{array} \right] \end{array} \right\} \quad (16)$$

where A_in and A_out are respectively the areas inside and outside the curve C. Thus, the proposed AC model overcame the limitation of conventional CV AC model in the area of face detection.

3.2 Proposed feature extraction technique

Once the faces have been detected and extracted then the wavelet transform is applied for feature extraction. In wavelet transform, we used the decomposition process for which the video frames were in gray scale. The reason for converting from RGB to gray scale was to improve the efficiency of the proposed algorithm. The wavelet decomposition could be interpreted as signal decomposition in a set of independent feature vectors. Each vector consists of sub-vectors like

$$V_0^{2D} = V_0^{2D-1}, V_0^{2D-2}, ..., V_0^{2D-n} \quad (17)$$

where V represents the 2D feature vector. If we have 2D expression frame X, and it is decomposed into orthogonal sub images corresponding to different visualization. The following equation shows one level of decomposition.

$$X = A_1 + D_1 \quad (18)$$

where X indicates the decomposed image and A_1 and D_1 show approximation and detailed coefficient vectors respectively. If the expression frame is decomposed up to multilevel, then, the (18) can then be written as

$$X = A_j + [D_j + D_{j-1} + D_{j-2} + ... + D_2 + D_1] \quad (19)$$

where j represents the level of decomposition. Mostly, the detail coefficients consist of noise; therefore, only the approximation were utilized for feature extraction. During the decomposition process, each frame is decomposed up to four levels of decomposition, i.e., $j = 4$, because by exceeding the value of $j = 4$ the image loses lots of information due to which the informative coefficients cannot be detected properly and might cause misclassification. The detail coefficients further consist of three sub-coefficients. So the (19) can be written as

$$\begin{aligned} X &= A_4 + [D_4 + D_3 + D_2 + D_1] \\ &= A_4 + [(D_h)_4 + (D_v)_4 + (D_d)_4] \\ &\quad + [(D_h)_3 + (D_v)_3 + (D_d)_3] \\ &\quad + [(D_h)_2 + (D_v)_2 + (D_d)_2] \\ &\quad + [(D_h)_1 + (D_v)_1 + (D_d)_1] \end{aligned} \quad (20)$$

Or simply, the (20) can be written as

$$X = A_4 + \sum_{j=4}^{1} [(D_h)_j + (D_v)_j + (D_d)_j] \quad (21)$$

where D_h, D_v, and D_d indicate horizontal, vertical and diagonal coefficients respectively. We can observe from

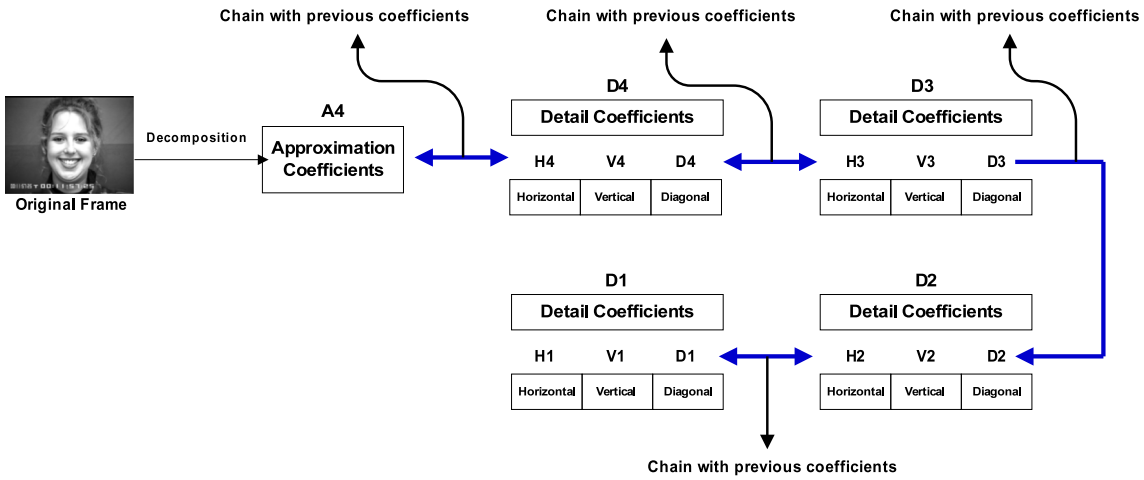

Fig. 1 All the coefficients are connected with one after another like performing head to tail rule in vector addition that produces one dimensional matrix, due to which the coefficients are extracted easily

(20) or (21), that all the coefficients are connected with each other like a chain, through which we can easily extract the prominent features. These coefficients graphically is represented by Fig. 1. In each decomposition step, the approximation and detail coefficient vectors are obtained by passing the signal through the low-pass and high-pass filters.

After the decomposition process, the feature vector is created by taking the average of all the frequencies of the expression frames. In a specified time window the frequency of each expression frame has been estimated by analyzing the corresponding frame by utilizing the wavelet transform [91].

$$C(a_i, b_j) = \frac{1}{\sqrt{a_i}} \int_{-\infty}^{\infty} y(t) \psi^*_{f.e} \left(\frac{t - b_j}{a_i} \right) dt \tag{22}$$

where a_i is the scale of the wavelet between the lower and upper frequency bounds to get higher decision for the frequency estimation, b_j is the position of the wavelet from the start to end of the time window with the spacing of signal sampling period, t is the time, $\psi_{f.e}$ is the wavelet function used for frequency estimation, and $C(a_i, b_j)$ are the wavelet coefficients with the specified scale and position parameters, which is converted to mode frequency as given below.

$$f_1 = \frac{f_a\left(\psi_{f.e}\right)}{a_m\left(\psi_{f.e}\right).\Delta} \tag{23}$$

where $f_a\left(\psi_{f.e}\right)$ is the average frequency of the wavelet function, and Δ is the signal sampling period. So the feature vector is obtained by taking the average of the whole frame frequencies for each expression that is given as

$$f_{avg} = \frac{(f_1 + f_2 + f_3 + ... + f_K)}{N} \tag{24}$$

where f_{avg} indicates the average frequency of each expression, which is a feature vector for that expression, K is the last frame of the current expression, and N represents the whole number of the frames in each expression.

3.3 Proposed feature selection model

In the this step, the most informative features are selected by using SWLDA, which maximizes the ratio of between-class variance to within-class variance in any particular data set, thereby guaranteeing maximal separability. Its forward and backward selection techniques enable SWLDA to effectively reduce the dimensions of the feature space.

In the forward step, the most correlated features are selected based on partial F-test values from the feature space. On the other hand, in the backward step, the least significant values are removed from the regression model i.e. lower F-test values. In both processes, the F-test values are calculated on the basis of defined class labels. The advantage of this method is that it is very efficient in seeking the localized features.

Procedure In the beginning, there is no predictor in the model. Based on the significance test, i.e., partial *F-test* (the *t-test*), predictor is either entered or removed from the model in each iteration. Two predictors Alpha-to enter and Alpha-to remove are defined for significance level test. Alpha-to enter $a_e = 0.10$ and Alpha-to-remove $a_\gamma = 0.20$ are set as threshold parameters. These values are chosen based on various experiments. These threshold parameters show the significance level of the predictors which are entered or removed from the model, respectively. The algorithm stops when there are no more predictors to enter or remove from the stepwise model.

We present an example in which we have three independent predictors: x_1, x_2, and x_3, and an output (response) y. Each predictor fits into the model using a regression; that is, we regress y on x_1, x_2, ..., and x_{p-1}, where p is the total number of predictors ($p = 3$ in this case). The first predictor to enter into the stepwise model is the predictor that has the smallest *t-test* p-value (i.e., below a_e). This will continue until the stopping criterion is met (i.e., if there is no predictor with a p-value less than a_e). Now suppose x_1 is the best predictor. Then fit each of the two predictor models that includes x_1 in the model, i.e., the model regresses y on (x_1, x_2), regress y on (x_1, x_3)...y on (x_1, x_{p-1}). The second predictor to enter into the stepwise model is the predictor that has the smallest p-value. If again there is no p-value less than a_e, the iteration stops.

Suppose this time x_2 is the best second predictor in the model. The analysis procedure then steps back and checks the p-value for $\beta_1 = 0$ (i.e., criterion for the removal of the predictor from the model). In this case, if the p-value is above a_γ for $\beta_1 = 0$, then the predictor is not significant compared to the new entry, and x_1 is removed from the stepwise model.

In contrast, suppose both x_1 and x_2 have made it into the two-predictor stepwise model. The analysis procedure then fits each of the three-predictor models with x_1 and x_2 in the model, i.e., it regresses y on (x_1, x_2, x_3), regresses y on $(x_1, x_2, x_4), \ldots,$ and regresses y on (x_1, x_2, x_{p-1}). The third predictor that enters the stepwise model is the predictor that has the smallest p-value less than a_e. The stopping criterion is met when there is no p-value less than a_e. In this case, the analysis checks the p-values $\beta_1 = 0$. If either p-value has not become significant (i.e., above a_γ), the predictor is removed from the stepwise model. This procedure will stop when adding an additional predictor does not yield a p-value below a_e. For more details on SWLDA, please refer to a previous study [51].

3.4 Hidden Markove model (HMM) based classification

Commonly, the function of HMMs is to provide a statistical model λ for a set of observation sequences. Sometimes, the observations are called "frames" in facial expression recognition applications. Suppose there are sequences of observations of length T that are denoted by $O_1, O_2, \ldots, O_T$. An HMM also consists of particular sequences of states, S, whose lengths range from 1 to N ($S = S_1, S_2, \ldots, S_N$), where N is the number of states in the model, and the time t for each state is denoted $Q = q_1, q_2, \ldots, q_N$. The states are connected by arcs, and each time that a state j is entered, an observation is generated according to the multivariate Gaussian distribution $b_j(O_t)$ with the mean value μ_j and covariance matrix V_j correlated with that state. The arcs also have transition probabilities correlated with them such that probability a_{ij} is the resultant transition probability from state i to state j. The initial model probability for the state j is Π_j. An HMM can be defined by this set of parameters, such as $\lambda = A, B, \Pi$, where A indicates the probability of the state transition such that $A = a_{ij}$, $a_{ij} = Prob(q_{t+1} = S_j | q_t = S_i)$, $1 \leq i, j \leq N$, where B represents the probability of observations such that $B = b_j(O_t)$, $b_j = Prob(O_t | q_t = S_j) 1 \leq j \leq N$, and the initial state probability is indicated by Π such that $\Pi = \Pi_j$, $\Pi_j = Prob(q_1 = S_1)$. All the equations are based on the work by [68] and make use of the initial state probability distribution.

In the training step, for a given model λ, the multiplication of each transition probability by each output probability at each step t provides the joint likelihood of a state sequence Q and the corresponding observation O that is calculated as:

$$P(O, Q|\lambda) = \pi_{q1} b_{q1}(O_1) \left[\prod_{t=2}^{T} a_{qt-1,q_t} b_{qt}(O_t) \right] \tag{25}$$

Basically, the above equation cannot be evaluated, because in practice, the state sequence is hidden. Therefore, the likelihood $P(O|\lambda)$ can be evaluated by summing over all possible state sequences:

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) \tag{26}$$

A simple procedure for finding the parameters λ that maximize the above equation in HMMs, introduced in [9], depends on the forward and backward algorithms $\alpha_t(j) = P(O_1 \ldots O_t, q_t = j|\lambda)$ and $\beta_t(j) = P(O_(t+1) \ldots O_T | q_t = j, \lambda)$, respectively, such that these variables can be initiated inductively by the following three processes:

$$\alpha_1(j) = \pi_j b_j(O_1), 1 \leq j \leq N \tag{27}$$

$$\beta_T(j) = 1, 1 \leq j \leq N \tag{28}$$

The first process defined in (27) and (28) is known as initialization, and the second is known as the induction process and can be written as:

$$\alpha_{1+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(j) \alpha_{ij} \right] b_j(O_{t+1}), 1 \leq t \leq T-1 \ \ and \ \ 1 \leq j \leq N \tag{29}$$

$$\beta_t(i) = \sum_{j=1}^{N} \alpha_{ij} b_j(O_{t+1}) \beta_{t+1}(j), t = T-1, T-2, \ldots 1 \ \ and \ \ 1 \leq i \leq N \tag{30}$$

The last process, known as the termination process, can be written as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \beta_1(i) \tag{31}$$

Due to the multiplication of the forward and backward probabilities a new set of HMM parameters $\gamma_t(j)$ for each state j can be found by computing weighted averages, and it can be simply calculated as

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(j)\beta_t(j)} \tag{32}$$

The model illustrated by the new set of parameters is described as $\overline{\lambda} = \overline{\pi}.\overline{A}.\overline{B}$. A related quantity $\xi_t(i, j)$ is used to estimate the transition parameters and is identified as the probability of state i in time t and in state j in $t + 1$. The observation sequence and the model are given as

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j|O, \lambda) \tag{33}$$

According to the forward and backward probabilities, (33) can be written as

$$\xi_t(i, j) = \frac{\alpha_t(i)\alpha_{ij}b_j(O_{t+1})\beta_{t+j}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\alpha_{ij}b_j(O_{t+1})\beta_{t+j}(j)} \tag{34}$$

By using the above concept, the new parameter $\overline{\lambda}$ can be re-estimated as

$$\overline{\pi} = \gamma_1(i) \tag{35}$$

$$\overline{\alpha}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{36}$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T-1} \gamma_t(i).O_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{37}$$

$$\overline{V}_i = \frac{\sum_{t-1}^{T=1} \gamma_t(i).(O_t - \overline{\mu_i})(O_t - \overline{\mu_i})'}{\sum_{t=1}^{T} \gamma_t(i)} \tag{38}$$

where prime denotes vector transpose, $\overline{\alpha_{ij}}$ is the estimated transition probability from the state i to state j, and $\overline{\mu_i}$ and $\overline{V_i}$ are the estimates of the mean and the covariance matrix of the Gaussian output probability function for state i. The $\overline{\lambda}$ is calculated iteratively in place of λ based on the above estimations, and this process is replicated until the parameter approximately meets at a critical point that is a local maximum of $P(O|\lambda)$.

During testing, the appropriate HMMs can then be determined by mean of likelihood estimation for the sequence observations O calculated based on the trained λ as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{39}$$

The maximum likelihood for the observations provided by the trained HMMs indicates the recognized label. For more details on HMM, please refer to [73].

4 Experimental environment

We assessed the performance of the proposed FER system in real world environment by utilizing three real time YouTube-based datasets [79, 80] such as emulated, semi-naturalistic, and naturalistic datasets. Each dataset is described as below.

4.1 Emulated dataset

In this dataset, expressions from different subjects belonging to different colors, age, and ethnicity are collected. The dataset includes six basic expressions such as happy, sad, angry, normal, disgust, and fear. The subjects age ranges from childish like 4 years to eldest subjects such as 60 years. In some of the cases, the images in some expressions are rotated using the camera for better accuracy of the system. The subjects include both male and females. Each expression has at least 165 images. The images used in the dataset are of size 240×320 and 320×240 pixels with facial frame.

4.2 Semi-naturalistic dataset

In this dataset, the expressions have been collected from the actors and actresses of the Hollywood, Bollywood, and Lollywood while performing the in their respective movies. From all the subjects, six expressions including angry, normal, disgust, happy, fear, and sad are collected based on basic two variant captured and analyzed. Different views from different angles with glasses, hair open, wearing hat, and other obvious actions are collected in this dataset. In the whole dataset, each expression consists of at least 165 images. The dataset has the images of size 240×320 and 320×240 pixels with facial frame.

4.3 Naturalistic dataset

In this dataset of facial expression, variety of subjects from various parts of the world, races, and ethnicities have been selected. In this dataset of facial expressions, six basic universal expressions including normal, happy, sad, angry, fear, and disgust have been captured from real world scenarios that include mainly from real world talk-shows, interviews, and YouTube natural videos such as news and real world incidents. The total of 165 images have been considered for each expression. The age range of the subjects are from 18 to 50 years. Images used in the dataset are of size 240×320 and 320×240 pixels with facial frame.

4.4 Setup

For a thorough validation, the following series of experiments were performed in Matlab using Intel® Pentium® Core^{TM} i7-6700 (3.4 GHz) with a RAM capacity of 16 GB.

- In the first experiment, the performance of the proposed face detection and extraction method was analyzed on each dataset.
- In the second experiment, the proposed FER system was tested and validated on all the datasets using 10−fold cross-validation scheme. Which means that out of 10 subjects data from a single subject was used as the testing data, whereas data for the remaining 9 subjects were used as the training data. This process was repeated 10 times with data from each subject used exactly once as the testing data.
- In the third experiment, the performance of the proposed FER system was validated under the absence of the proposed methods. For each modules (for feature extraction and selection), we used existing well-known statistical methods such as ICA and LDA for feature extraction and selection instead of using the proposed methods such as wavelet transform and SWLDA.
- In the fourth experiment, the performance was analyzed across the datasets in order to show the robustness of the proposed system. In other words, from the three datasets, two datasets were used as testing datasets, whereas one dataset was used as the training data. This process was repeated three times, with data from each dataset used exactly once as the training data.
- Finally, in the fifth experiment, the performance of proposed FER system was compared against the previous state-of-the-art works.

5 Experimental results and discussion

5.1 First experiment

The proposed face detection and extraction method is validated using all the the three datasets. In certain frames, the proposed AC model is performed independently; which means that the face detection and extraction is performed frame-by-frame. In this model, first, an ellipse with x-axis of length 15 and y-axis of length 15 is selected as the initial contour. In this experiment, the initial shape was the same for all frames, and only the center location varied. We manually segmented the first frame by placing the initial contour which must be closer to the face. Then from the second frame, the position of the initial contour's center in the current frame is the mean value of the points along the final contour in the previous frame. By this way, the only information utilized from the previous frame is the final contour obtained in the previous frame. This information is used to determine the initial position of the active contour in the current frame.

Table 1 Classification results of the proposed FER system on emulated dataset of facial expressions (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	97	0	1	0	1	1
Sad	0	98	2	0	0	0
Anger	0	0	100	0	0	0
Disgust	1	1	0	95	2	1
Fear	0	2	0	0	96	2
Normal	0	2	0	1	1	96
Average			97.00			

5.2 Second experiment

This experiments show the performance of the proposed FER system in naturalistic environments. Therefore, the system was tested and validated on the three datasets such as emulated, semi-naturalistic, and naturalistic datasets separately. The overall results on the three datasets are shown in Tables 1, 2, and 3, and in Figs. 2, 3, and 4 respectively. It is clear from Tables 1, 2, and 3 that the proposed system constantly better performance and achieved high recognition rates when applied on all the datasets separately. This means that, unlike existing methods, the proposed FER system is more robust, i.e., it provided high recognition rates not only on one dataset but using all the three datasets. The reason is that the proposed feature extraction and feature selection methods are more robust to the real life scenarios.

5.3 Third experiment

In this experiment, a set of sub-experiments were performed in order to show the importance of sub-components in the proposed FER system, i.e., wavelet transform with optical

Table 2 Classification results of the proposed FER system on semi-naturalistic dataset of facial expressions (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	95	2	1	1	1	0
Sad	1	93	2	1	1	2
Anger	1	1	96	0	0	2
Disgust	1	2	0	94	2	1
Fear	1	1	2	1	95	0
Normal	2	0	0	0	0	98
Average			95.17			

Table 3 Classification results of the proposed FER system on naturalistic dataset of facial expressions (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	94	0	2	1	3	0
Sad	1	92	2	2	1	2
Anger	0	0	97	2	1	0
Disgust	1	3	3	90	2	1
Fear	1	2	2	1	93	1
Normal	0	1	2	0	0	97
Average			93.83			

flow, SWLDA, and HCRF. For this purpose, nine sub-experiments were performed on the spontaneous dataset using the 10-fold validation rule.

5.3.1 Results while removing the proposed feature extraction technique

In the first three sub-experiments, ICA (a well-known local feature extraction technique) was utilized with SWLDA and HMM instead of the proposed feature extraction method (i.e., wavelet transform). The overall results for the these sub-experiments on emulated, semi-naturalistic, and naturalistic datasets are shown in Tables 4, 5, and 6, respectively.

For the feature extraction, we utilized symlet wavelet transform. So, it can be seen from Tables 4, 5, and 6 that without the proposed feature extraction method, the system is unable to show better performance. It is because symlet wavelet can extract the most prominent information in the form of frequency from expression frames, and also it is a compactly supported wavelet on frames with the least asymmetry and highest number of vanishing moments for a given support width. The symlet wavelet has the capability to support the characteristics of orthogonal, biorthogonal, and reverse biorthogonal of gray scale images. That's why it provides better classification results. The frequency-based assumption is supported in our experiments and we measure the statistic dependency of wavelet coefficients for all expression frames. Joint probability of a frame is computed by collecting geometrically aligned frames of the expression for each wavelet coefficient. Mutual information for the wavelet coefficients computed using these distributions is used to estimate the strength of statistical dependency between the two frames. Moreover, symlet wavelet transform is capable to extract prominent features from expression frames with the aid of locality in frequency, orientation and in space as well. Since wavelet is a multi-resolution that helps us to efficiently find the images in coarse-to-find way.

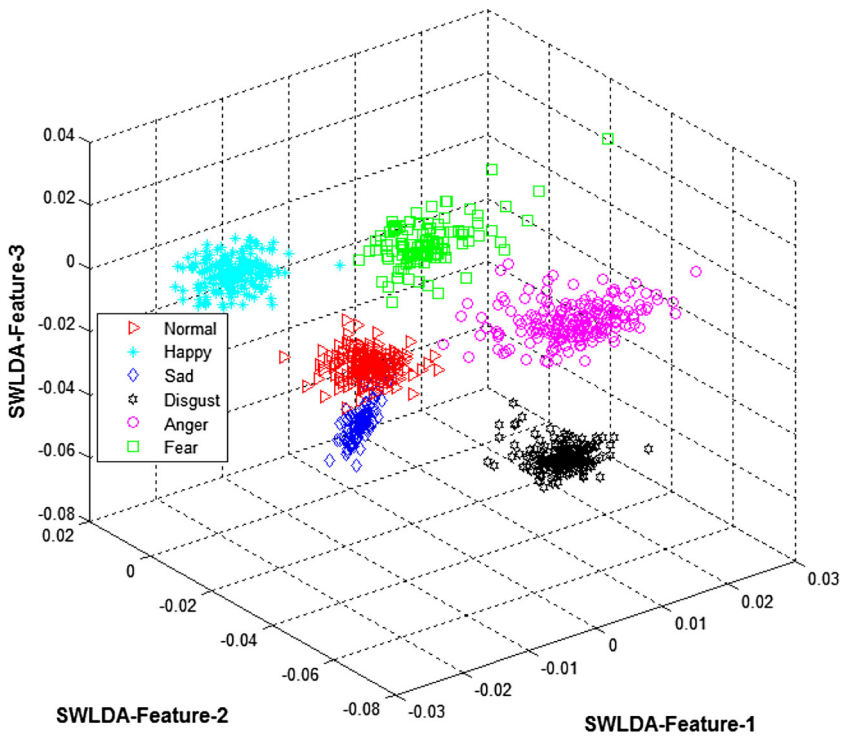

Fig. 2 3D feature plots for the six expressions after applying the proposed FER system on emulated dataset

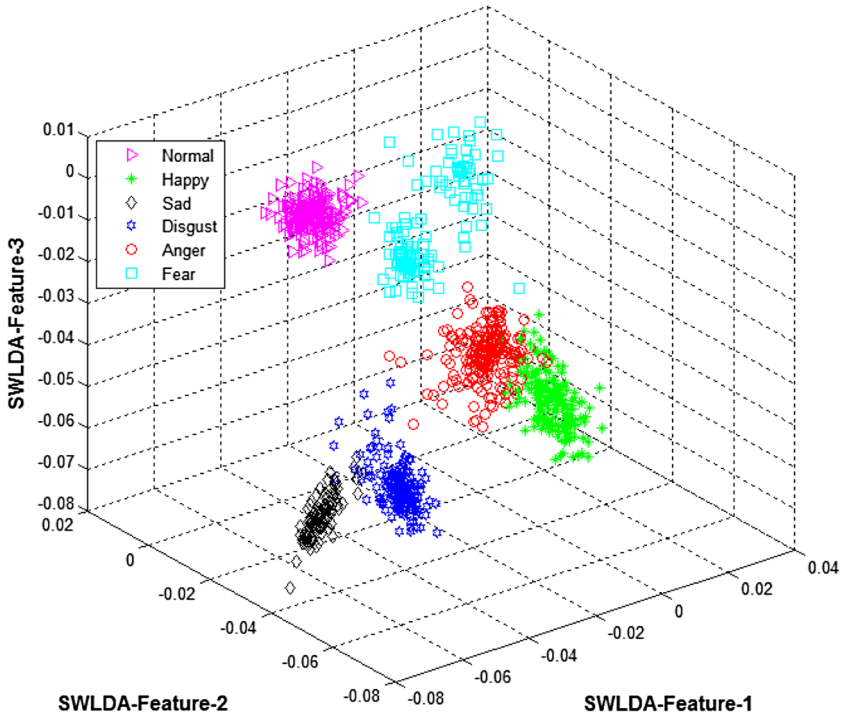

Fig. 3 3D feature plots for the six expressions after applying the proposed FER system on semi-naturalistic dataset

5.3.2 Results while removing the proposed feature selection technique

In the next three sub-experiments, wavelet transform was coupled with LDA (a well-known discriminant analysis approach) before feeding the features to the proposed HMM. The results for the these sub-experiments on emulated, seminaturalistic, and naturalistic datasets are presented in Tables 7, 8, and 9, respectively.

Similarly, it is also apparent from Tables 7, 8, and 9 that without the proposed feature selection method (SWLDA), the system was also unable to achieve high classification

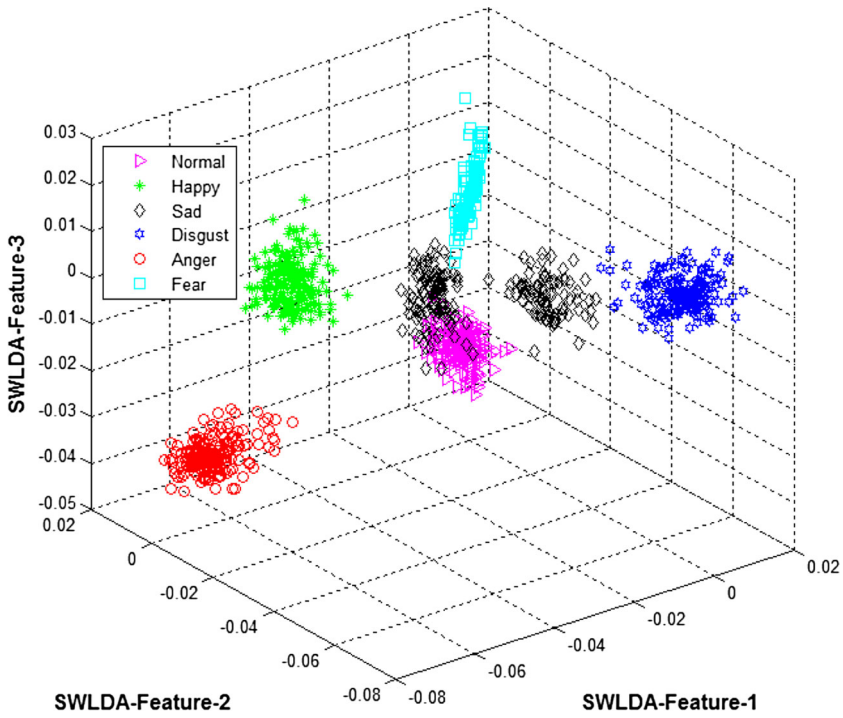

Fig. 4 3D feature plots for the six expressions after applying the proposed FER system on naturalistic dataset

Table 4 Classification rates of ICA+SWLDA with HMM on emulated dataset of facial expressions, while removing the proposed feature extraction (wavelet transform) method (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	87	1	3	4	2	3
Sad	3	85	4	2	2	4
Anger	2	1	89	3	2	3
Disgust	1	2	2	90	3	2
Fear	2	3	2	0	91	2
Normal	4	2	3	3	4	84
Average			87.66			

Table 5 Classification rates of ICA+SWLDA with HMM on semi-naturalistic dataset of facial expressions, while removing the proposed feature extraction (wavelet transform) method (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	88	2	3	2	2	3
Sad	3	84	3	2	4	4
Anger	4	3	82	4	4	3
Disgust	2	3	4	86	2	3
Fear	3	3	4	2	85	3
Normal	1	1	3	1	4	90
Average			85.83			

Table 6 Classification rates of ICA+SWLDA with HMM on naturalistic dataset of facial expressions, while removing the proposed feature extraction (wavelet transform) method (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	85	4	3	2	2	4
Sad	5	80	5	2	5	3
Anger	4	3	79	5	5	4
Disgust	1	4	2	88	3	2
Fear	3	2	1	4	86	4
Normal	4	3	4	2	4	83
Average			83.50			

Table 7 Classification rates of the wavelet transform+LDA with HMM using emulated dataset of facial expressions, while removing the proposed feature selection (SWLDA) method (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	88	3	1	4	2	2
Sad	4	84	3	2	3	4
Anger	2	3	89	3	1	2
Disgust	4	2	3	90	1	0
Fear	4	1	3	3	87	2
Normal	4	3	2	4	5	82
Average			86.66			

Table 8 Classification rates of the wavelet transform+LDA with HMM using semi-naturalistic dataset of facial expressions, while removing the proposed feature selection (SWLDA) method (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	89	2	2	3	1	3
Sad	4	80	3	4	5	4
Anger	3	2	85	3	4	3
Disgust	3	2	2	90	1	2
Fear	3	5	4	4	79	5
Normal	4	4	3	2	3	84
Average			84.50			

Table 9 Classification rates of the wavelet transform+LDA with HMM using naturalistic dataset of facial expressions, while removing the proposed feature selection (SWLDA) method (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	79	4	5	3	5	4
Sad	4	83	2	4	4	3
Anger	3	3	82	4	5	3
Disgust	3	5	4	79	3	6
Fear	3	3	4	3	85	2
Normal	1	2	4	2	3	88
Average			82.66			

Table 10 Classification rates of the proposed FER system training on emulated dataset and testing on semi-naturalistic and naturalistic datasets (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	80	3	4	5	4	4
Sad	3	85	2	3	3	4
Anger	2	3	86	4	1	4
Disgust	3	3	3	82	5	4
Fear	3	2	3	7	81	4
Normal	1	2	3	2	2	90
Average			84.00			

Table 11 Classification rates of the proposed FER system training on semi-naturalistic and testing on emulated, naturalistic datasets (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	81	4	3	4	5	3
Sad	1	89	3	3	0	4
Anger	4	5	80	4	3	4
Disgust	3	3	4	84	2	4
Fear	3	3	4	5	79	6
Normal	4	4	6	5	4	77
Average			81.67			

Table 12 Classification rates of the proposed FER system training on naturalistic dataset and testing on emulated and semi-naturalistic datasets (Unit: %)

Expressions	Happy	Sad	Anger	Disgust	Fear	Normal
Happy	79	4	3	6	4	4
Sad	3	83	5	4	1	4
Anger	0	6	81	3	4	6
Disgust	3	5	6	76	6	4
Fear	3	4	4	6	78	5
Normal	5	3	5	3	4	80
Average			79.50			

rate. This is because SWLDA not only provides dimension reduction, it also increases the low between-class variance to increase the class separation before the features are fed to the classifier. The low within class variance and high between class variance are achieved because of the forward and backward regression models in the SWLDA.

5.4 Fourth experiment

For this experiment, $n-$fold cross-validation rule based on dataset was performed (in our case $n = 3$). The overall results for this experiment are presented in Tables 10, 11, and 12, respectively.

It is clear from Table 10 that the proposed FER system achieved a high recognition rate when it was trained using the emulated dataset and tested on semi-naturalistic and naturalistic datasets. Similarly, it is also apparent from Table 11 that the system achieved slightly better performance when it was trained using the semi-naturalistic dataset and trained on emulated and naturalistic datasets (as shown in Table 11). However, the system achieved low accuracy when it was trained on the naturalistic dataset and tested on emulated and semi-naturalistic datasets (shown in Table 12). This might be because the datasets have different facial features and different environment. For instance, the subjects of emulated dataset performed the expressions in a posed manner, that is, each subject tried to copy or mimic the instructor, so there were little variations from subject-to-subject and in timings. However, the variation in capturing of expression from various angles (placing camera at variant angles) gave us the ability to test the proposed algorithm on the maximum possible alterations/variations in the images. Moreover, emulated dataset images are mostly front-faced, right-sided, and left-sided with up and down orientations. Likewise, the expressions in semi-naturalistic dataset were collected from the movie/drama scenes of professional actors and actresses, where we had no control on expression timings, camera, lighting and background settings. Hence, these expressions are semi-naturalistic expressions collected under dynamic settings. The performance of the system degrades when trained on naturalistic dataset. This is because the expressions in naturalistic dataset were recorded from real world talk shows, news, and interviews. Hence, these expressions are spontaneous expressions collected in natural and dynamic settings. The dataset includes both indoor and outdoor subjects with varying and dynamic backgrounds. In this dataset, different views from different angles with glasses, hair open, wearing hat, and other complex scenarios with obvious actions and things were included. Moreover, the images in this dataset were collected in real life setting such as a variety of backgrounds, unintentional expressions of the subjects, some variant/orientation angles of the face of the subjects, and lighting variations. These were some factors which may cause misclassification. Nevertheless, the results are very encouraging and this suggests that the proposed FER system is robust, i.e., the system not only achieved a high recognition rate on one dataset, but also provided good recognition rates when used across multiple datasets.

5.5 Fifth experiment

In this experiment, the recognition rates for the proposed spontaneous FER system was compared against some of the existing FER systems. The overall results of these systems along with the proposed spontaneous FER system are summarized in Table 13.

It can be seen from Table 13 that the proposed spontaneous FER system outperformed the existing methods. Thus, the proposed system shows significant potential in its ability to accurately and robustly recognize human facial expressions in naturalistic scenarios.

Furthermore, the proposed FER system has been compared with one of the recent FER systems like [37]. In the proposed system, we utilized the same dataset of [37] for fair comparison. The proposed FER system achieved 99% while, the accuracy of [37] is 96%. As can be seen that the proposed FER system showed significant performance than of the existing state-of-the-art works.

Table 13 The weighted average classification results of the proposed FER system with some existing state-of-the-art systems (Unit: %)

Existing Works	[47]	[83]	[39]	[81]	[41]	[103]	[36]	[70]	Proposed System
Recognition Rates	85	81	76	84	85	65	78	86	95

6 Conclusion and future direction

In naturalistic environments, facial expression recognition (FER) has received lots of attention. So, for this purpose, several FER systems have been proposed; however, recognizing the expressions accurately in naturalistic environment is still a major concern for most of these systems. Therefore, in this study, we proposed an accurate and robust facial expression recognition system that is capable of exhibiting high recognition rate in naturalistic scenarios. In this system, an unsupervised face detection and extraction model is proposed. In this model, two energy functions such as Chan-Vese (CV) energy and Bhattacharyya distance functions were exploited that not only minimize the dissimilarities within the object (face) but also maximize the distance between the object (face) and background. Furthermore, in this system, we also proposed a new feature extraction method based on symlet wavelet transform, which has the capability to extract the most prominent information in the form of frequency from the expression frames. Though, the proposed feature extraction technique extracts the most informative features; however, there might be some redundancy in the features. Therefore, in this work, we also proposed the usage of a robust non-linear feature selection called stepwise linear discriminant analysis (SWLDA). This method selects the most informative features taking the advantage of the forward selection model and can remove the irrelevant feature by taking advantage of the backward regression model. The proposed system has been tested and validated using three different YouTube-based datasets. These datasets have been collected from YouTube, real world talk shows and interviews, and daily conversations. Each dataset is consisted of six facial expressions like happy, sad, anger, disgust, fear, and normal. For the proposed system, we utilized 10–fold cross validation scheme for each datasets. The system achieved weighted average recognition rate (95%) against the existing FER systems. That is a significant contribution in accuracy in naturalistic environment.

All these experiments were performed in laboratory. In near future, we will employ the proposed FER in smartphone, due to which normal users can check their mental states during their daily routine.

Acknowledgements The author would like to thank Faculty of Computer and Information Sciences, AlJouf University, Sakaka, Kingdom of Saudi Arabia.

References

1. Abidin Z, Harjoko A (2012) A neural network based facial expression recognition using fisherface. Int J Comput Appl 59(3):30–34
2. Aguilar-Torres G, Toscano-Medina K, Sanchez-Perez G, Nakano-Miyatake M, Perez-Meana H (2009) Eigenface-gabor algorithm for feature extraction in face recognition. Intern J Comput 3(1):20–30
3. Ahsan T, Jabid T, Chong UP et al (2013) Facial expression recognition using local transitional pattern on gabor filtered facial images. IETE Tech Rev 30(1):47–52
4. Al-Allaf ONA (2014) Review of face detection systems based artificial neural networks algorithms. arXiv:1404.1292
5. Ali HB, Powers DMW, Jia X, Zhang Y (2015) Extended non-negative matrix factorization for face and facial expression recognition. Intern J Mach Learn Comput 5(2):142
6. Azeem A, Sharif M, Raza M, Murtaza M (2013) A survey: face recognition techniques under partial occlusion. Intern Arab J Inform Technol 11(1):1–10
7. Lal Banchhor B, Sharma T (2014) Hybrid approach for face detection using skin color based segmentation and edge detection. Skin 3(6):7252–7257
8. Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Comput 12(10):2385–2404
9. Baum LE (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. Inequalities 3:1–8
10. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
11. Bettadapura V (2012) Face expression recognition and analysis: the state of the art. arXiv:1203.6722
12. Bettadapura VK (2009) Face expression recognition and analysis: the state of the art. Emotion:1–27
13. Bhati D, Gupta V (2015) Survey–a comparative analysis of face recognition technique. Neural Comput 3(2):597–609
14. Bhutani Tarika, Arora Rashmi, Juneja Ronak (2014) Face recognition using artificial neural network. Intern J Res 1(8):1374–1377
15. Boutsidis C, Mahoney MW, Drineas P (2008) Unsupervised feature selection for principal components analysis. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 61–69
16. Buciu I, Pitas I (2004) Application of non-negative and local non negative matrix factorization to facial expression recognition. In: Proceedings of the 17th international conference on pattern recognition, ICPR 2004, vol 1. IEEE, pp 288–291
17. Cai Y Invariant local features for face detection. Department of Computer Science, University of British Columbia
18. Rojas Castillo JA, Ramirez Rivera A, Chae O (2012) Robust facial recognition based on local gaussian structural pattern. Int J Innov Comput Appl, Inf Control 8(12):8399–8413
19. Chan TF, Vese LA (2001) Active contours without edges. IEEE Trans Image Process 10(2):266–277
20. Chitra S, Balakrishnan G (2012) A survey of face recognition on feature extraction process of dimensionality reduction techniques. J Theor Appl Inform Technol 36(1):92–100
21. Cynthia Christabel S, Annalakshmi M, Prince Winston D et al (2013) Facial feature extraction based on local color and texture for face recognition using neural network. Intern J Sci Eng Appl 2(4):78–82
22. Dai B, Zhang D, Liu H, Sun S, Li K (2009) Evaluation of face recognition techniques. In: International conference on photonics and image in agriculture engineering (PIAGENG 2009). International Society for Optics and Photonics, pp 74890M–74890M
23. Datta AK, Datta M, Banerjee PK (2015) Face detection and recognition: theory and practice. CRC Press, Boca Raton

24. Delac K, Grgic M, Liatsis P (2005) Appearance-based statistical methods for face recognition. In: 47th international symposium ELMAR-2005, pp 151–158
25. Deotale N, Vaikole SL, Sawarkar SD (2010) Face recognition using artificial neural networks. In: 2010 the 2nd international conference on computer automation engineering (ICCAE), vol 2. IEEE, pp 446–450
26. Ding C, Tao D (2015) A comprehensive survey on pose-invariant face recognition. arXiv:1502.04383
27. Gooya DPES, Gripon V (2015) Automatic face recognition using sift and networks of tagged neural cliques. In: The seventh international conference on advanced cognitive technologies and applications, pp 57–61
28. Endeshaw S, Raimond K (2015) Face recognition using artificial neural network. Zede J 25:43–52
29. Erdogan H (2004) Subspace kernel discriminant analysis for speech recognition. In: COST278 and ISCA tutorial and research workshop (ITRW) on robustness issues in conversational interaction
30. Fu Z, Huang F, Sun X, Vasilakos A, Yang C-N (2016) Enabling semantic search based on conceptual graphs over encrypted outsourced data. IEEE Trans Serv Comput
31. Garg A, Choudhary V (2012) Facial expression recognition using principal component analysis. International Journal of Scientific Research Engineering & Technology (IJSRET) 1(3):39–42
32. Ghimire D, Lee J (2013) A robust face detection method based on skin color and edges. J Inf Process Syst 9(1):141–156
33. Gou G, Huang D, Wang Y (2012) A hybrid local feature for face recognition. In: PRICAI 2012: trends in artificial intelligence. Springer, pp 64–75
34. Hajiarbabi M, Askari J, Sadri S, Saraee M (2008) A new linear appearance-based method in face recognition. In: Advances in communication systems and electrical engineering. Springer, pp 579–587
35. Halder A, Jati A, Singh G, Konar A, Chakraborty A, Janarthanan R (2012) Facial action point based emotion recognition by principal component analysis. In: Proceedings of the international conference on soft computing for problem solving (SocProS 2011) December 20–22, 2011. Springer, pp 721–733
36. Happy SL, Routray A (2015) Robust facial expression classification using shape and appearance features. In: 2015 eighth international conference on advances in pattern recognition (ICAPR). IEEE, pp 1–5
37. Hernandez-Matamoros A, Bonarini A, Escamilla-Hernandez E, Nakano-Miyatake M, Perez-Meana H (2016) Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach. Knowl-Based Syst 110:1–14
38. Huang Z-L, Lu Z-M, Yu F-X, Zhang Y (2014) A fast face detection scheme utilizing hsv-based skin color model with kl transform. Inf Technol J 13(1):183
39. Jabid T, Kabir H, Chae O (2010) Robust facial expression recognition based on local directional pattern. ETRI J 32(5):784–794
40. Jang U, Lee EC (2015) Comparative analysis on performance of pixel-based face recognition methods by considering various facial poses. In: Advanced science and technology letters, vol 116. Springer, pp 278–281
41. Jia Q, Gao X, Guo H, Luo Z, Wang Y (2015) Multi-layer sparse representation for weighted lbp-patches based facial expression recognition. Sensors 15(3):6719–6739
42. Jia Q, Liu Y, Guo H, Luo Z, Wang Y (2011) A sparse representation approach for local feature based expression recognition. In: 2011 international conference on multimedia technology (ICMT). IEEE, pp 4788–4792
43. Jonnalagedda MV, Doye DD (2015) Radially defined local binary patterns for facial expression recognition. Int J Comput Appl 119(21)
44. Kabir H, Jabid T, Chae O (2010) A local directional pattern variance (ldpv) based face descriptor for human facial expression recognition. In: 2010 seventh IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 526–532
45. Kailath T (1967) The divergence and bhattacharyya distance measures in signal selection. IEEE Trans Commun Technol 15(1):52–60
46. Kalita J, Das K (2013) Recognition of facial expression using eigenvector based distributed features and euclidean distance based decision making technique. Int J Adv Comput Sci Appl (IJACSA) 4(2):196–202
47. Kapoor R, Gupta R (2014) Morphological mapping for non-linear dimensionality reduction. IET Comput Vis 9(2):226–233
48. Karande KJ, Talbar SN, Inamdar SS (2012) Face recognition using oriented laplacian of gaussian (olog) and independent component analysis (ica). In: 2012 second international conference on digital information and communication technology and it's applications (DICTAP). IEEE, pp 99–103
49. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Int J Comput Vis 1(4):321–331
50. Kittusamy SRV, Chakrapani V (2012) Facial expressions recognition using eigenspaces. J Comput Sci 8(10):1674–1679
51. Krusienski DJ, Sellers EW, McFarland DJ, Vaughan TM, Wolpaw JR (2008) Toward enhanced p300 speller performance. J Neurosci Methods 167(1):15–21
52. Kumbhar M, Jadhav A, Patil M (2012) Facial expression recognition based on image feature. Intern J Comput Commun Eng 1(2):117–119
53. Lajevardi SM, Wu HR (2012) Facial expression recognition in perceptual color space. IEEE Trans Image Process 21(8):3721–3733
54. Long F, Wu T, Movellan JR, Bartlett MS, Littlewort G (2012) Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. Neurocomputing 2(1):126–132
55. Lu J, Plataniotis KN, Venetsanopoulos AN (2003) Face recognition using lda-based algorithms. IEEE Trans Neural Netw 14(1):195–200
56. Lu X, Kong L, Liu M, Zhang X (2015) Facial expression recognition based on gabor feature and src. In: Biometric recognition. Springer, pp 416–422
57. Ma L (2008) Facial expression recognition using 2-d dct of binarized edge images and constructive feedforward neural networks. In: IEEE international joint conference on neural networks, 2008. IJCNN 2008. (IEEE world congress on computational intelligence). IEEE, pp 4083–4088
58. Mika S (2002) Kernel fisher discriminants. PhD thesis, Universitätsbibliothek
59. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR (1999) Fisher Discriminant analysis with kernels. In: Proceedings of the IEEE signal processing society workshop neural networks for signal processing IX, 1999. IEEE, pp 41–48
60. Mistry VJ, Goyani MM (2013) A literature survey on facial expression recognition using global features. Intern J Eng Adv Technol 2(4):653–657
61. Nam MY, Rhee PK (2005) Human face detection using skin color context awareness and context-based bayesian classifiers. In: Knowledge-based intelligent information and engineering systems. Springer, pp 298–307
62. Nanni L, Lumini A, Dominio F, Zanuttigh P (2014) Effective and precise face detection based on color and depth data. Appl Comput Inform 10(1):1–13

63. Pan Z, Lei J, Zhang Y, Sun X, Kwong S (2016) Fast motion estimation based on content property for low-complexity h. 265/hevc encoder. IEEE Trans Broadcast 62(3):675–684
64. Pang Yanwei, Yuan Yuan, Li X (2009) Iterative subspace analysis based on feature line distance. IEEE Trans Image Process 18(4):903–907
65. Pardeshi SA, Talbar SN (2009) Face description with local invariant features: application to face recognition. In: 2009 2nd international conference on emerging trends in engineering and technology (ICETET). IEEE, pp 248–251
66. Park SW, Savvides M (2010) A multifactor extension of linear discriminant analysis for face recognition under varying pose and illumination. EURASIP J Adv Signal Process 2010:6
67. Purandare V, Talele KT (2014) Efficient heterogeneous face recognition using scale invariant feature transform. In: International conference on circuits, systems, communication and information technology applications (CSCITA), 2014. IEEE, pp 305–310
68. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
69. Rahulamathavan Y, Phan R, Chambers J, Parish D (2013) Facial expression recognition in the encrypted domain based on local fisher discriminant analysis. IEEE Trans Affective Comput 4(1):83–92
70. Ramirez Rivera A, Rojas Castillo J, Chae O (2013) Local directional number pattern for face analysis: face and expression recognition. IEEE Trans Image Process 22(5):1740–1752
71. Roy S, Bandyopadhyay SK (2013) Face detection using a hybrid approach that combines hsv and rgb. Intern J Comput Sci Mobile Comput 2(3):127–136
72. Patil CS, Patil AJ (2013) A review paper on facial detection technique using pixel and color segmentation. Int J Comput Appl 62(1):21–24
73. Samaria FS (1994) Face recognition using hidden Markov models. PhD thesis, University of Cambridge
74. Saudagare PV, Chaudhari DS (2012) Facial expression recognition using neural network–an overview. International Journal of Soft Computing and Engineering (IJSCE) 2(1):224–227
75. Shan C, Braspenning R (2010) Recognizing facial expressions automatically from video. In: Handbook of ambient intelligence and smart environments. Springer, pp 479–509
76. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816
77. Sharif A, Sharif MI, Riaz S, Shaheen A, Badini MK (2014) Face recognition: holistic approaches an analytical survey. Sci Intern 26(2):639–644
78. Sharma R, Patterh MS (2015) Face recognition using face alignment and pca techniques: a literature survey. Neural Comput 17(4):17–30
79. Siddiqi MH, Ali M, Eldib MEA, Khan A, Banos O, Khan AM, Lee S, Choo H (2017) Evaluating real-life performance of the state-of-the-art in facial expression recognition using a novel youtube-based datasets. Multimed Tools Appl 1–21
80. Muhammad HS, Ali M, Idris M, Banos O, Lee S, Choo H (2016) A novel dataset for real-life evaluation of facial expression recognition methodologies. In: Canadian conference on artificial intelligence. Springer, pp 89–95
81. Siddiqi MH, Ali R, Idris M, Khan AM, Kim ES, Whang MC, Lee S (2016) Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection. Multimed Tools Appl 75(2):935–959
82. Siddiqi MH, Ali R, Khan AM, Kim ES, Kim GJ, Lee S (2015) Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. Multimed Syst 21(6):541–555
83. Siddiqi MH, Ali R, Khan AM, Park Y-T, Lee S (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans Image Process 24(4):1386–1398
84. Siddiqi MH, Ali R, Sattar A, Khan AM, Lee S (2014) Depth camera-based facial expression recognition system using multilayer scheme. IETE Tech Rev 31(4):277–286
85. Siddiqi MH, Farooq F, Lee S (2012) A robust feature extraction method for human facial expressions recognition systems. In: Proceedings of the 27th conference on image and vision computing New Zealand. ACM, pp 464–468
86. Siddiqi MH, Lee S, Lee Y-K, Khan AM, Truc PTH (2013) Hierarchical recognition scheme for human facial expression recognition systems. Sensors 13(12):16682–16713
87. Stoerring M, Andersen HJ, Granum E (1999) Skin colour detection under changing lighting conditions. In: 7th symposium on intelligent robotics systems. Citeseer
88. Suresh AJ (2015) Face expression recognition using gabor features and probabilistic neural network. Int J Adv Res Comput Sci 6(5):86–90
89. Tian Q, Chen S (2017) Cross-heterogeneous-database age estimation through correlation representation learning. Neurocomputing 238:286–295
90. Tian Y-L, Kanade T, Cohn JF (2005) Facial expression analysis. In: Handbook of face recognition. Springer, pp 247–275
91. Turunen J et al (2011) A wavelet-based method for estimating damping in power systems
92. Md ZU, Kim T-S, Song BC (2013) An optical flow feature-based robust facial expression recognition with hmm from video. Int J Innov Comput Inf Control 9(4):1409–1421
93. Uwechue OA, Pandya AS (2012) Human face recognition using third-order synthetic neural networks, vol 410. Springer Science & Business Media
94. Verma A, Achyut Raj S, Midya A, Chakraborty J (2014) Face detection using skin color modeling and geometric feature. In: 2014 international conference on informatics, electronics & vision (ICIEV). IEEE, pp 1–6
95. Wang X, Paliwal KK (2003) Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recogn 36(10):2429–2439
96. Wang Z, Sun X (2012) Manifold adaptive kernel local fisher discriminant analysis for face recognition. J Multimed 7(6):387–393
97. Wei W, Song H, Li W, Shen P, Vasilakos A (2017) Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. Inf Sci 408:100–114
98. Wu X, Zhao J (2010) Curvelet feature extraction for face recognition and facial expression recognition. In: 2010 sixth international conference on natural computation (ICNC), vol 3. IEEE, pp 1212–1216
99. Xia Z, Wang X, Sun X, Liu Q, Xiong N (2016) Steganalysis of lsb matching using differences between nonadjacent pixels. Multimed Tools Appl 75(4):1947–1962
100. Xia Z, Wang X, Zhang L, Qin Z, Sun X, Ren K (2016) A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing. IEEE Trans Inf Forensics Secur 11(11):2594–2608
101. Zhang L, Tjondronegoro D (2010) Feature extraction and representation for face recognition. In: Oraves M (ed) Face recognition, pp 1–20
102. Zhang L, Tjondronegoro D (2011) Facial expression recognition using facial movement features. IEEE Trans Affect Comput 2(4):219–229

103. Zhao X, Shi X, Zhang S (2015) Facial expression recognition via deep learning. IETE Technical Review, (ahead-of-print):1–9
104. Zheng W, Zhao L, Zou C (2004) A modified algorithm for generalized discriminant analysis. Neural Comput 16(6):1283–1297
105. Zhou Z, Yang C-N, Chen B, Sun X, Liu Q, Jonathan QM (2016) Effective and efficient image copy detection with resistance to arbitrary rotation. IEICE Trans Inf Syst 99(6):1531–1540
106. Zhu Z, Ji Q (2006) Robust real-time face pose and facial expression recovery. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 1. IEEE, pp 681–688

Muhammad Hameed Siddiqi is currently working as an Assistant Professor in Faculty of Computer Science and Information, AlJouf University, Sakaka, Kingdom of Saudi Arabia. He was a Postdoctoral Research Scientist at the Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, South Korea from March 2016 to August 2016. He has completed his Bachelor of Computer Science (Hons) from Islamia College university of Peshawar, KPK, Pakistan in 2007, and Master and PhD from Ubiquitous Computing (UC) Lab, Department of Computer Engineering, Kyung Hee University, Suwon, South Korea by 2012 and 2016, respectively. He was a Graduate Assistant at Universiti Teknologi PETRONAS, Malaysia from 2008 to 2009. He published more than 50 articles in high reputable international journals and conferences. His research interest is Image Processing, Pattern Recognition, Machine Intelligence, Activity Recognition, and Facial Expression Recognition.