Visual Computing for Industry, Biomedicine, and Art

ORIGINAL ARTICLE Open Access

Robust facial expression recognition system in higher poses

Ebenezer Owusu[1], Justice Kwame Appati[1*] and Percy Okae[2]

Abstract

Facial expression recognition (FER) has numerous applications in computer security, neuroscience, psychology, and engineering. Owing to its non-intrusiveness, it is considered a useful technology for combating crime. However, FER is plagued with several challenges, the most serious of which is its poor prediction accuracy in severe head poses. The aim of this study, therefore, is to improve the recognition accuracy in severe head poses by proposing a robust 3D head-tracking algorithm based on an ellipsoidal model, advanced ensemble of AdaBoost, and saturated vector machine (SVM). The FER features are tracked from one frame to the next using the ellipsoidal tracking model, and the visible expressive facial key points are extracted using Gabor filters. The ensemble algorithm (Ada-AdaSVM) is then used for feature selection and classification. The proposed technique is evaluated using the Bosphorus, BU-3DFE, MMI, CK+, and BP4D-Spontaneous facial expression databases. The overall performance is outstanding.

Keywords: Facial expressions, Three-dimensional head pose, Ellipsoidal model, Gabor filters, Ada-AdaSVM

Introduction

Applications

Facial expression recognition (FER) is the automatic detection of the emotional state of a human face using computer-based technology. The field of study is currently a hotspot of research because it has increasing applications in several domains, such as psychology, sociology, health science, transportation, gaming, communication, security, and business. According to Panksepp [1], facial expressions and emotions guide the lives of people in a variety of ways, and emotions are key aspects that enlighten us in how we should act, from elementary processes to the most intricate acts [2, 3].

The sporadic advancements in the use of facial expressions in neuropsychiatric complications have shown more positive results [4], and current studies are focusing on human behavior and the detection of mental illnesses [5, 6].

FER can also affect data collection in specific research projects. For example, Shergill et al. [7] proposed an intelligent assistant FER framework that could be implemented in e-commerce to determine the product preferences of customers. The system captures the facial data as they browse the e-shop for products to acquire. Based on the facial expression, the systems can automatically suggest more products of possible interest.

Certain physiological features of people have been discovered to be useful as intelligent data in the search for criminals [8, 9]. This theory is based on the tendency for someone with ego to commit a high-profile crime, such as terrorism, exhibits specific emotions such as anger and fear. Consequently, the accurate recognition of these expressions could lead to further security measures in apprehending criminals.

FER can also be valuable during the testing phase of video games. Target groups are frequently invited to play a game for a set amount of time, and their behaviors and emotions are observed as they play. Game developers may acquire more insights and valuable deductions about the emotions recorded during gameplay using FER technology, and incorporate the feedback into production.

*Correspondence: jkappati@ug.edu.gh

[1] Department of Computer Science, University of Ghana, P. O. Box LG 163, Accra, Ghana
Full list of author information is available at the end of the article

Technical issues on the use of two-dimensional facial data

Two-dimensional (2D) FER systems are extremely sensitive to head orientation. Therefore, to achieve good results, the subject must be constantly in a fronto-parallel orientation. The problem resulting from this is that the throughput of most site-access systems is significantly reduced. This implies that subjects are frequently required to perform several verifications to attain an ideal facial orientation. Consequently, surveillance systems operate on luck, hoping the subject faces the camera.

Another problem that arises from the use of 2D technology is the illumination conditions of the surrounding environment. If the subject is in a setting with varying lighting conditions, FER reduces in accuracy because the FER processes are sensitive to the direction of lighting and the ensuing shading pattern. Consequently, cast shadows may obstruct recognition by concealing informative features.

Three-dimensional (3D) FER systems have a higher detection rate than 2D systems because of their higher intensity modality, and they also have more object description geometry information [10, 11]. This demonstrates the importance of pushing FER into higher face orientations to improve its realism and practicality.

Related work

The primary focus of this study is to improve FER accuracy in higher facial orientations.

Yadav and Singha [12] adopted the Viola-Jones descriptor [13] to detect faces and used a combination of local binary patterns (LBP) and the histogram of gradients (HOG) as a feature extraction tool. Subsequently, traditional SVM with the k-means method was employed as a training algorithm. LPB feature extraction techniques, such as Gabor, are orientation-selective, and thus, highly robust in tracking key facial features. However, the Viola-Jones descriptor is computationally demanding and has a low detection accuracy. Furthermore, the conventional SVM described in the study is slow to classify. Consequently, the overall architecture used in the study was computationally expensive. Yao et al. [14] proposed a linear SVM method that used AUs to recognize seven facial expression prototypes in the CK database. The Viola-Jones descriptor was used as the face-detection technique again. Although the goal of the study was to minimize computational complexity and enhance recognition accuracy, the resulting average recognition accuracy of 94.07% for females and 90.77% for males was too low for a viable implementation. Ashir et al. [15] also proposed an SVM-based multiclass classification for detecting seven facial expressions across four prominent databases. The Nyquist–Shannon sampling method [16] was used to compress the extracted facial feature samples. Although the sampling method reduces data loss, it is prone to aliasing issues, particularly when the bandwidth is extremely large. The Nyquist-Shannon sampling technique is difficult to deploy because it assumes the sampled signal is completely band-restricted. In real-world applications, this is a concern because no actual signal is genuinely and completely band-restricted. The compressing sampling [17] paradigm could have been a better option because it is less restrictive. Perez-Gomez et al. [18] recently proposed a 2D–3D FER system that used principal component analysis (PCA) and a genetic algorithm for feature selection, and a k-nearest neighbor (KNN)-multiclass SVM for learning. In this study, the synthetic minority oversampling technique (SMOTE) [19] was used to balance the instances. However, SMOTE creates an equal number of synthetic samples for each minority data sample and relies on the hypothesis performance to update the distribution function. The adaptive synthetic (ADASYN) [20] method tends to generate more synthetic data for minority class samples that are harder to learn than with SMOTE, which is easy to learn. In addition, PCA uses observations from all the extracted features in the projection to the subspace and only considers linear relationships, ignoring the input multivariate structures. Compared to other recent studies, the findings of this study were not positive.

Li et al. [21] proposed a robust 3D local coordinate technique for extracting pose-invariant facial features at key points. The descriptor in this method is a multi-task sparse representation fine-grained matching algorithm. The method was evaluated using the Bosphorus datasets, and an average recognition accuracy of 98.9% was obtained. The success of this study is largely owed to the accurate tracking of 3D key points. This recent study is a primary driving force behind our proposed study.

The following are the significant contributions of this work: (1) A robust head-tracking algorithm that tracks facial features from one frame to the next, accounting for more features in the overall prediction process; (2) A unique ensemble approach that employs AdaBoost for feature selection, and a combination of AdaBoost and SVM for classification. AdaBoost is extremely fast, whereas SVM is extremely accurate. Consequently, the proposed technique becomes extremely fast while also improving the recognition accuracy.

The remainder of this paper is organized as follows. Methods section delves into the proposed strategy. Results and discussion section discusses the findings, debates, and analyses. Finally, Conclusions section concludes the study.

Methods

We robustly tracked the facial features from one frame to the next using 3D facial data. With 3D data, information, such as the size and shape of an object, can be correctly estimated in each frame without prior assumptions.

The first priority is to detect the focal points in each frame. The next step is to search for matching features or objects across all frames. This method addresses the changing behavior of a moving object and the preceding annotations of the scene. In this approach, the location of an object is projected by iteratively updating the object position from previous frames [22, 23].

Architectural framework

Figure 1 presents the framework of this study.

This procedure uploads images and robustly tracks the features across frames using the proposed ellipsoidal model. Subsequently, the Gabor feature-extraction approach was used. Feature points extraction section explains the reason for using Gabor features in this study. Feature selection and classification were executed using the Ada-AdaSVM.

Ellipsoidal feature tracking method

Accurate tracking of a human face from the forehead, to the left cheek, to the chin, to the right cheek, and back to the same spot on the forehead where the tracking began unmistakably demonstrates that the human face is best shaped like an ellipse. Thus, considering the 3D facial representation in Fig. 2 with N feature points tracked across frames, we denote:

$$\alpha(t) = \{ f_j(t) \big| 1 \leq j \leq N \} \tag{1}$$

where N represents the most relevant feature points. In this study, we assumed N to be 24. In addition, let $f_j(t) \in \alpha(t)$ denote a facial feature. As the features move from one frame to the next at time $t+1$, the position of feature $f_j(t)$ becomes $f_j(t+1)$. Therefore, $f_j(t+1) \in \alpha(t+1)$. Assuming that Y_j is the position of α_j on the 3D facial model and $\alpha_{j,p}[\varnothing(t+1)]$ represents its back projection on the image plane, the 3D facial orientation at $t+1$ is the vector $\varnothing(t+1)$ that minimizes $\sum_{j=1}^{N} S_j^2$, where:

$$S_j[\phi(t+1)] = ||\alpha_{j,p}[\phi(t+1)] - \alpha_j(t+1)|| \tag{2}$$

This is a multi-view system based on the assumption that cameras are positioned around the subject to capture various rotation movements. Consequently, the facial

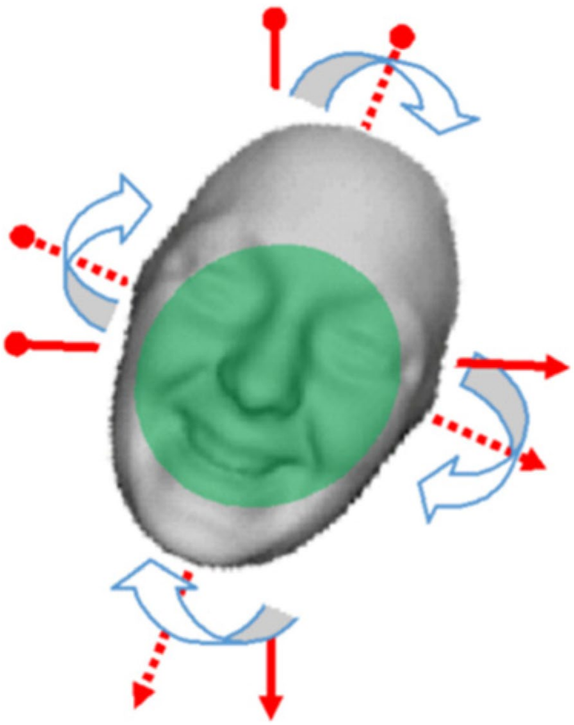


Fig. 2 Tracking of 3D feature points from one frame to another

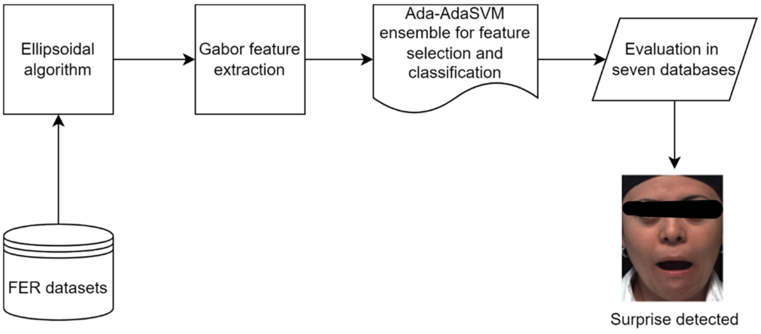

Ellipsoidal algorithm
Gabor feature extraction
Ada-AdaSVM ensemble for feature selection and classification
Evaluation in seven databases
FER datasets
Surprise detected

Fig. 1 Architectural framework of this study

image can be captured with a high degree of precision in any orientation. We extracted the features in the same manner as for 2D images. The right and left eyes, lips, and muscles around the cheeks are important parts of the face to consider. Slight disruptions primarily and severely distort the muscles in these places. The Gabor technique is then used to extract the features of the captured face.

The algorithm models a procedure that chooses a set of features and robustly tracks them from one frame to the next while discarding all other features that are no longer required for tracking. The ellipsoidal 3D face was modelled, as shown in Fig. 3.

Adopting homogeneous coordinates for an ellipsoid of the semi-axis, a, b, and c, states that a point $X_0 = (x_0, y_0, z_0, 1)$ belongs to the surface of the ellipsoid if $X_0^T E_0 X_0 = 0$.

$$E_0 = \begin{bmatrix} b^2c^2 & 0 & 0 & 0 \\ 0 & a^2c^2 & 0 & 0 \\ 0 & 0 & a^2b^2 & 0 \\ 0 & 0 & 0 & -a^2b^2c^2 \end{bmatrix} \tag{3}$$

The algorithm tracks the facial features that are more noticeable by slight deformation from one frame to the next using the brightness change constraint [24]. These muscles are usually near the eyes, mouth, cheeks, and edges, as shown in Fig. 4 and contour τ in Fig. 3.

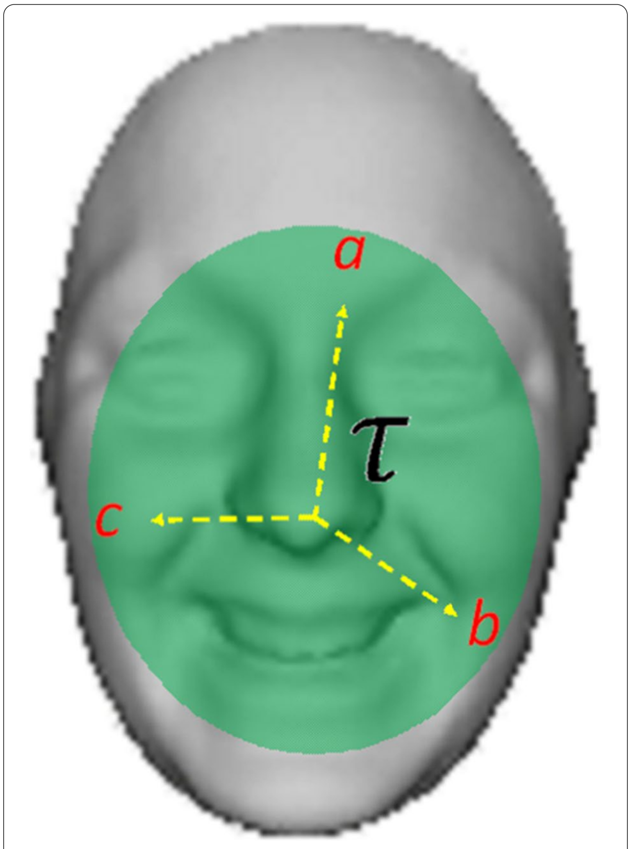


Fig. 3 Ellipsoidal face model

Given that pixel (x, y) with luminance $I(x,y)^T$ moves from position $(x, y)^T$ at frame t to position $(x+u, y+v)^T$ at frame $t+1$ in high frame rates. In this instance, we can deduce that

$$I(x+u, y+v, t+1) = I(x, y, t) \tag{4}$$

By applying Taylor's series, and considering I_x and I_y as gradients and that I_t is a temporal deviation of the image, we can infer that

$$[I_x(x,y,t) I_y(x,y,t)] \begin{pmatrix} u \\ v \end{pmatrix} + I_t(x,y,t) = 0 \tag{5}$$

If a whole window ω_k is considered instead of a single pixel, we deduce that

$$J(u,v) = \left[\sum_{\omega_k} I_x(x,y,t) \sum_{\omega_k} I_y(x,y,t)\right] \begin{pmatrix} u_k \\ v_k \end{pmatrix} + \sum_{\omega_k} I_t(x,y,t) = 0 \tag{6}$$

The solution of Eq. (6) is an optimization problem. By introducing the cost function, it follows that

$$J(u,v) = \left\{ \left[\sum_{\omega_k} I_x(x,y,t) \sum_{\omega_k} I_y(x,y,t)\right] \begin{pmatrix} u_k \\ v_k \end{pmatrix} + \sum_{\omega_k} I_t(x,y,t) \right\}^2 \tag{7}$$

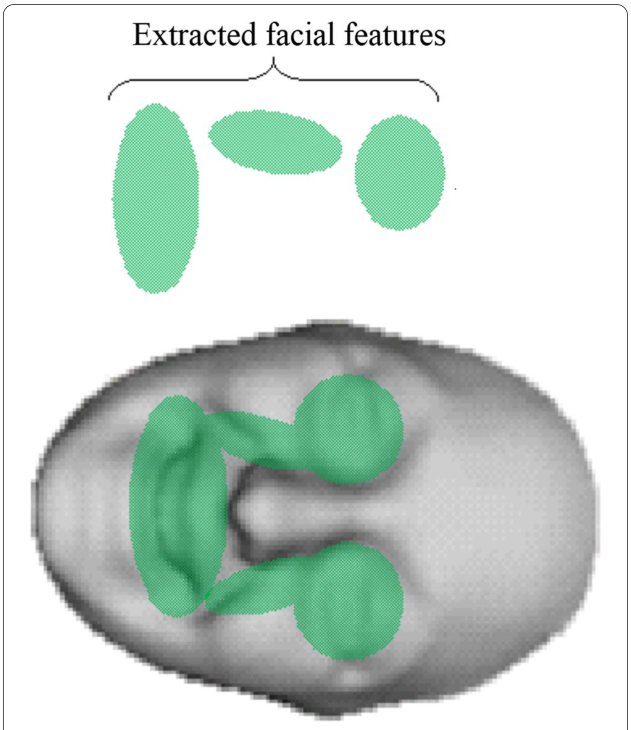


Fig. 4 Model of feature extraction points in 3D

The optimal displacement vector that determines the new position of face ω_k is given by:

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = \arg \underbrace{\min}_{\begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^2} J(u,v) \tag{8}$$

where, (u_k, v_k) represents the image at a new position. By computing the derivative of J with respect to u and v and equating them to zero, we obtain:

$$C_k \begin{pmatrix} u_k \\ v_k \end{pmatrix} + D_k = 0 \tag{9}$$

where $C_k = \begin{pmatrix} \sum_{\omega_k} I_x^2 & \sum_{\omega_k} I_x I_y \\ \sum_{\omega_k} I_x I_y & \sum_{\omega_k} I_y^2 \end{pmatrix}$, and $D_k = \begin{pmatrix} \sum_{\omega_k} I_x I_t \\ \sum_{\omega_k} I_x I_t \end{pmatrix}$. Assuming that $I : [1,m] \times [1,n] \subseteq \mathbb{N}^2 \to [0,1]$ is the matrix of the 3D face, then the j^{th} level of the pyramid description of the face image is expressed by the recursion:

$$I^j(x,y) = \begin{cases} I(x,y)\,, j = 0 \\ \frac{1}{4} I^{j-1}(2x,2y) + \\ \frac{1}{8}[I^{j-1}(2x-1,2y) + I^{j-1}(2x+1,2y) + & ,j \neq 0 \\ I^{j-1}(2x,2y-1) + I^{j-1}(2x,2y+1)] + \\ \frac{1}{16}[I^{j-1}(2x-1,2y-1) + I^{j-1}(2x+1,2y+1) + \\ I^{j-1}(2x+1,2y-1) + I^{j-1}(2x-1,2y+1)] \end{cases} \tag{10}$$

The displacement vector in Eq. (9) can also be rewritten as:

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = -C_k^{-1} D_k \tag{11}$$

The displacement vector in Eq. (10) is computed at the deepest pyramid level j_{max} (in the Newton–Raphson fashion), and the result of the computation is propagated to the upper level $j_{max} - 1$ by the expression:

$$\begin{pmatrix} u_k^{j-1} \\ v_k^{j-1} \end{pmatrix} = 2 \begin{pmatrix} u_k^j \\ v_k^j \end{pmatrix} \tag{12}$$

Equation (12) was used as the initial estimate for the evaluation of the displacement vector of the 3D face. The final displacement vector is given by the expression

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = \sum_{j=0}^{j_{\max}} 2^j \begin{pmatrix} u_k^j \\ v_k^j \end{pmatrix} \tag{13}$$

The visible features of the face can be extracted from any location on the face, similar to any other 2D dimensional face. The extracted features are candidates for predicting the overall expression of the face. The Gabor extraction technique is critical for extracting the maximum amount of information required for the classifier.

Feature points extraction

The 2D Gabor filters are spatial sinusoids localized by the Gaussian window, and because they are orientation-, localization-, and frequency-selective, they are useful in this study. Demonstrate images using Gabor wavelets provides flexibility because the details about their spatial relations are preserved in the process. The general form of the Gabor function is given by:

$$G(x,y,\theta,u,\sigma) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{x^2+y^2}{2\sigma^2} \right\} \exp\left[2\pi i(R_1 + R_2)\right] \tag{14}$$

where $R_1 = uxcos\theta$ and $R_2 = uysin\theta$, u is the spatial frequency of the band pass, θ is the spatial orientation, σ is the standard deviation that the 2D Gaussian envelops, and (x, y) is the position of the light impulse in the visual field. To allow for more robustness in illumination, we set the filter to zero direct current. The Gabor wavelet is then given by:

$$\tilde{G}(x,y,\theta,u,\sigma) \simeq G(i,j,\theta,u,\sigma) = \frac{1}{q}\left[\sum_{i=-n}^{n} \sum_{j=-n}^{n} G(x,y,\theta,u,\sigma) \right] \tag{15}$$

where $\left(x, y, \theta, u, \sigma\right)$ are parameters with (i, j) being the new position of the 2D input point, θ is the scale, u is the orientation of the Gabor kernel, σ is the standard deviation of the Gaussian window in the kernel, n is the maximum size of the face peak, and q is the size of the filter given by $q = (2n+1)^2$. In this study, we used 8 orientations given by $\left\{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}\right\}$ and 5 scales given by $\left\{4, 4\sqrt{2}, 8, 8\sqrt{2}, 16\right\}$. The sample points of the filtered image are coded into two bits (x_1, x_2) such that:

$$G_1 = \begin{cases} x_1 = 1, if \left\{ \Re[\tilde{G}(x,y,\theta,u,\sigma)] * I \right\} \geq 0 \\ x_1 = 0, if \left\{ \Re[\tilde{G}(x,y,\theta,u,\sigma)] * I \right\} < 0 \end{cases} \tag{16}$$

$$G_2 = \begin{cases} x_2 = 1, if \left\{ \Im[\tilde{G}(x,y,\theta,u,\sigma)] * I \right\} \geq 0 \\ x_2 = 0, if \left\{ \Im[\tilde{G}(x,y,\theta,u,\sigma)] * I \right\} < 0 \end{cases} \tag{17}$$

where I is a sub-image of the expressional face; $\Re$ and $\Im$ are the real and imaginary parts of each Gabor kernel, respectively; and the star (*) is the convolution operator. The final magnitude response, representing the feature

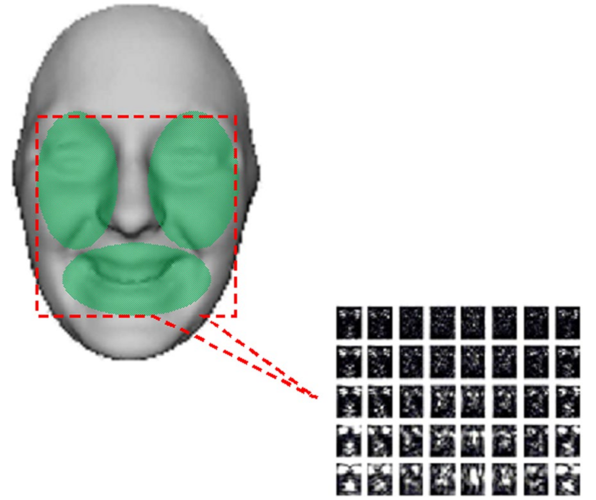


Fig. 5 Gabor magnitude response of the expressive face image: sample image (left), magnitude response image of the whole Gabor filter bank of 40 Gabor filters (right)

vectors, was computed by determining the square root of the sum of the squares of G_1 and G_2. Figure 5 shows the magnitude response of a template image.

Classification using Ada-AdaSVM

For this optimization problem, an SVM with a radial basis function kernel was used as a weak classifier. This weak SVM classifier was trained to produce the optimum Gaussian value for the scale parameter δ and regularization parameter ∂. Typically, the best parameters are $\{'\partial' : 1.0, '\delta' : 0.1\}$. The feature selection hypothesis is then computed from the expression *sgn* $\left[\sum_{t-1}^{T}\omega_t h_t^1\left(\varphi_t^1\right)\right]$, where T is the final iteration, h_t^1 is the hypothesis with the most discriminating information, and ω_t is weights that weigh h_t^1 based on its classification performance. The learning process formulated in our recent study [25] is as follows:

Step 1: Input the training sets, $[(y_1, x_1), (y_2, x_2), \ldots, (y_N, x_N)]$, $N = a + b$; where datasets a and b comprise $y_i = +1$ and $y_i = -1$ datasets, respectively. Initially, $\delta = \delta_{ini}, \delta_{min}, \delta_{step}$. The scale parameter δ, x, and y are the feature vectors selected by the AdaBoost algorithm.

Step 2: Initialize the training set weights, $w_i^{(1)} = 1/2a, \forall(y_i = +1)$ and $w_i^{(1)} = 1/2a, \forall(y_i = -1)$.

Do while $\delta > \delta_{min}$

Step 3: Apply the RBFSVM kernel to train the weighted training datasets by applying the leave-one-subject-out cross validation (LOSOCV) approach and compute the training error for the weak classifier h_t as

$$\xi_t = \sum_{i=1}^{N} w_i^t, y_i \neq h_t(x_i) \tag{18}$$

Step 4: At $\xi_t = 1/2$, reduce δ by a factor of δ_{step} and then jump to Step 1.

Step 5: Place the weight of the constituent classifier h_t such that

$$h_t : \alpha_t = \ln\left[\frac{1}{\xi_t} - 1\right]^{\frac{1}{2}} \tag{19}$$

Step 6: Update the weights by computing:

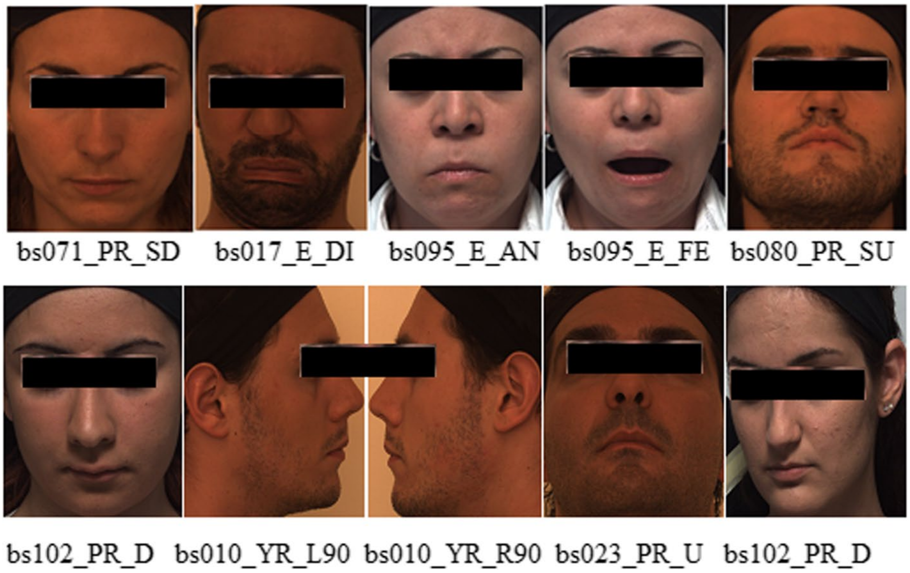


Fig. 6 Sample Bosphorus datasets

Table 1 FER in Bosphorus database

Pose	Expression	Average recognition (%)	Expressions	Average recognition (%)
10^0 Yaw	Neutral	100	Happiness	99.2
20^0 Yaw	Neutral	99.8	Sadness	98.0
30^0 Yaw	Neutral	99.2	Disgust	98.4
L45^0 Yaw	Neutral	97.3	Angry	99.4
R45^0 Yaw	Neutral	97.8	Fear	99.6
L90^0 Yaw	Neutral	63.2	Surprise	99.0
R90^0 Yaw	Neutral	78.2	Overall average	98.9
PR	Neutral	99.7		
CR	Neutral	98.9		

Average recognition accuracy = 92.7%

$$w_i^{t+1} = \frac{w_i^t \exp\left\{-\alpha_t y_i h_t(x_i)\right\}}{N_t} \tag{20}$$

where N_t is a normalization constant and $\sum_{i=1}^{n} w_i^{t+1} = 1$

Step 7: The final classifier is given by

$$H(x) = \operatorname{sgn}\left[\sum\nolimits_{t=1}^{T} \alpha_t h_t(x)\right] \tag{21}$$

The LOSOCV approach is given by the expression: $1/2n = \sum_{t=1}^{n} \left|f_i(x_i) - l_i\right|$, where n represents the total trained data.

Facial expression datasets

The algorithm was trained and tested on five popular datasets: Bosphorus, BU-3DFE, MMI, CK+, and BP4D-Spontaneous, and executed on a (4 CPUs), approximately 2.2 GHz processor with a memory capacity of 8192 MB RAM.

Results and discussion

Experiments on databases

Bosphorus contains 4666 images of 105 subjects [26] comprising 60 men and 5 women, with the majority being Caucasian; 27 of whom were professional actors, in various poses, expressions, and occlusion conditions. In addition to the 6 basic emotional expressions, various systematic head poses (13 yaw and pitch rotations) were present. The texture images have a resolution of 1600 × 1200 pixels, whereas the 3D faces comprise approximately 35,000 vertices [27]. Figure 6 presents sample datasets from Bosphorous. Occlusion images were discarded because they were not the focus of this study. The datasets used comprised 6 poses and 7 expressions. The images were partitioned into training and testing sets using the conventional LOSOCV approach. One specimen from each of the 6 groups of expressions was used as a test dataset during each training run, whereas the rest of the samples were used as a testing set. Table 1 summarizes the FER in Bosphorus.

The BU-3DFE database was created at Binghamton University [28]. There were 100 respondents, ranging in age from 18 to 70 years old. Whites, Blacks, East Asians, Middle East Asians, Indians, and Hispanics are among the ethnic groups. Each participant displayed 7 expressions at 4 intensity levels, including neutral, and 6 archetypal facial expressions. Figure 7 shows sample datasets in the database. The images were separated into training and testing sets using the same LOSOCV method as that used for the Bosphorus datasets, and the average recognition accuracy was 94.56%.

The MMI database comprises over 2900 high-resolution videos submitted by more than 20 students and research staff members, of which 44% are female, ranging in age from 19 to 62 years old. Seventy-five subjects were included in total, and Fig. 8 shows samples. The

Fig. 7 Sample BU3DFE datasets

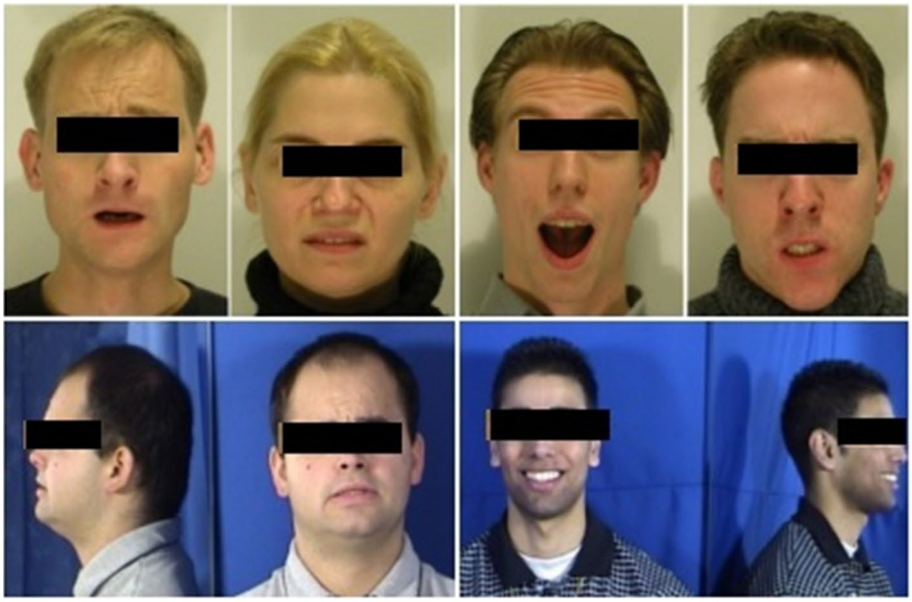


Fig. 8 Sample MMI datasets

datasets are partitioned into training and testing sets using the LOSOCV technique. One sample from each of the 7 types of expressions was used as the test dataset during each training run. The remaining samples were used as training sets. For each training cycle, the samples were repeated with new test samples. The expressions included anger, disgust, fear, happiness, neutral, sadness, and surprise. The average recognition accuracy is 97.2%.

The CK+database is a version of the 210 adult CK database. Participants were 18 to 50 years old, with 69% female, 81% Euro-American, 13% Afro-American, and 6% from other ethnic groups. The expressions included anger, contempt, disgust, fear, happiness, sadness, and surprise. Figure 9 presents sample datasets. A tenfold cross-validation procedure was used to partition the datasets into training and testing sets. The average recognition accuracy is 99.48%.

Finally, the BP4D-Spontaneous dataset is a 3D video collection of spontaneous facial expressions from young individuals. The database comprises 41 subjects

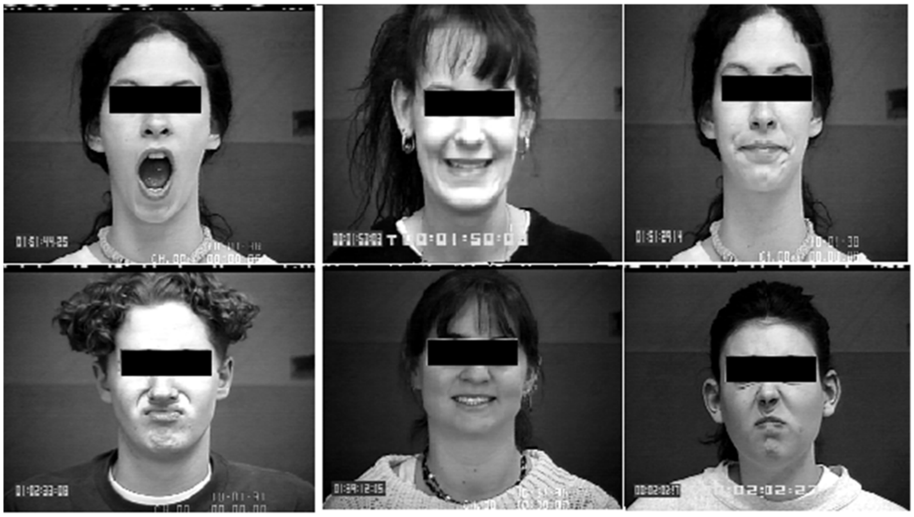


Fig. 9 Sample images in CK+database

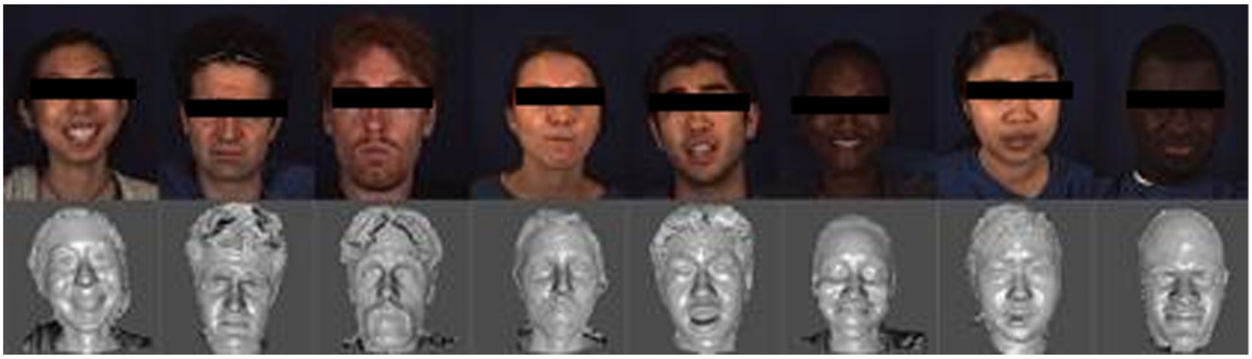


Fig. 10 Sample BP4D-Spontaneous datasets

(23 women and 18 men) ranging in age from 18 to 29 years old, including 11 Asians, 6 African-Americans, 4 Hispanics, and 20 Euro-Americans. Figure 10 shows sample images. We extracted expressions of anger, disgust, fear, pain, happiness, sadness, and surprise. The datasets were partitioned into training and testing sets using tenfold cross-validation. The average recognition accuracy is 97.2%.

Figures 11 and 12 exhibit the respective confusion matrices for facial expressions and pose predictions in the Bosphorus database. Figures 13, 14, 15, and 16 show the rest of the confusion matrices for FERs in BU3DFE, MMI, CK+, and BP4D-Spontaneous, respectively.

Comparison of methods

In Table 2, the proposed method was compared to some recent techniques. These results clearly demonstrated that the proposed method is promising. Figures 17, 18, and 19 show the performance of each of the 7 facial expressions. In the BU3DFE database, many authors failed to report the performance of neutral expressions; thus, the comparison was performed using the other 6. The performance shown in Fig. 17 was encouraging. Figure 18 shows the performance of the CK+database. Although the result, as shown in Fig. 18, depicts fierce rivalry between three current methods [29–31], the overall average recognition shows that the proposed technique is promising. In

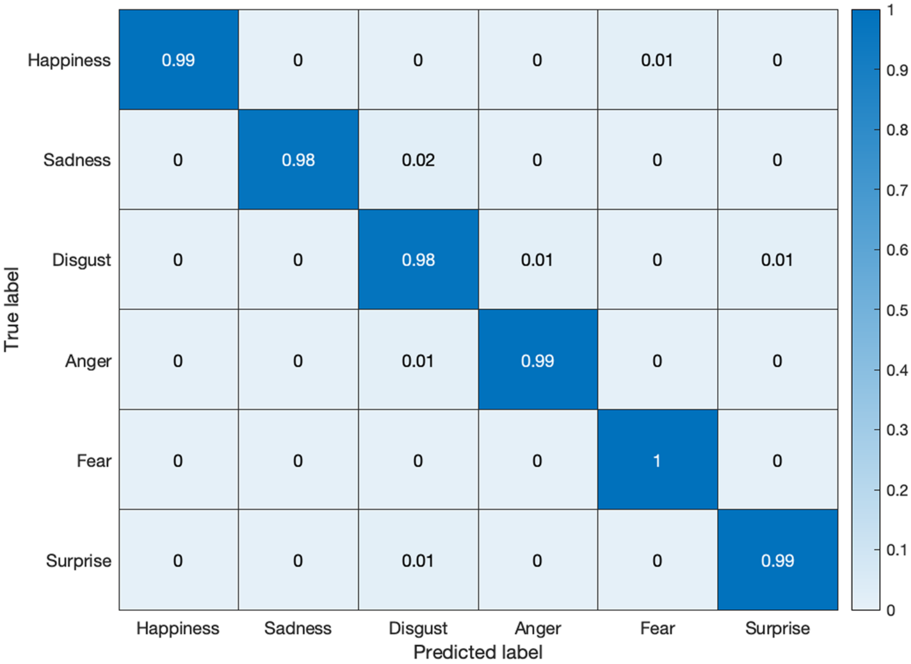
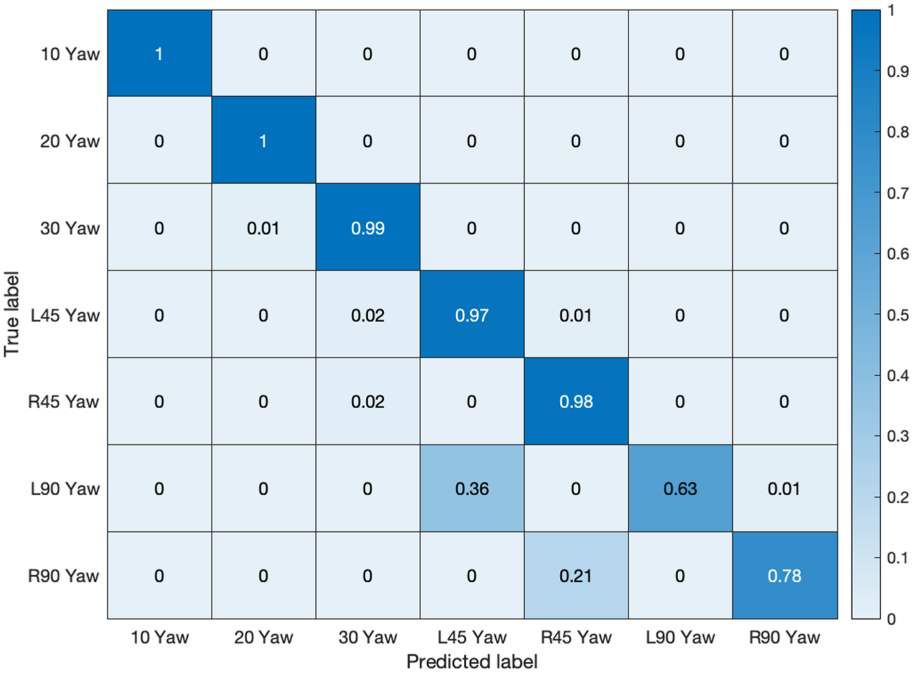
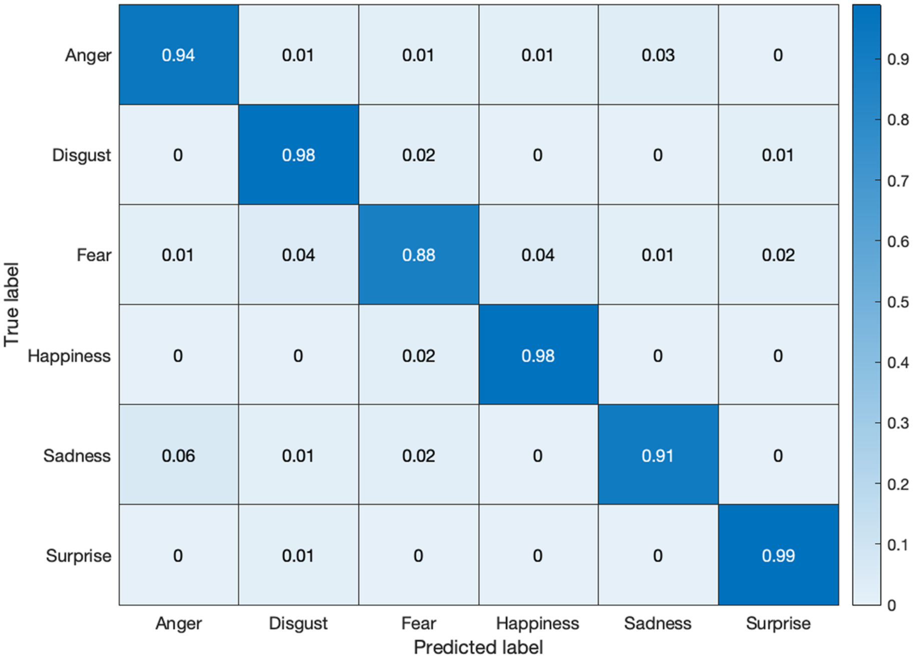


Fig. 11 Confusion matrix of facial expressions in Bosphorus



True label \ Predicted label	10 Yaw	20 Yaw	30 Yaw	L45 Yaw	R45 Yaw	L90 Yaw	R90 Yaw
10 Yaw	1	0	0	0	0	0	0
20 Yaw	0	1	0	0	0	0	0
30 Yaw	0	0.01	0.99	0	0	0	0
L45 Yaw	0	0	0.02	0.97	0.01	0	0
R45 Yaw	0	0	0.02	0	0.98	0	0
L90 Yaw	0	0	0	0.36	0	0.63	0.01
R90 Yaw	0	0	0	0	0.21	0	0.78

Fig. 12 Confusion matrix of pose prediction in Bosphorus



True label \ Predicted label	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0.94	0.01	0.01	0.01	0.03	0
Disgust	0	0.98	0.02	0	0	0.01
Fear	0.01	0.04	0.88	0.04	0.01	0.02
Happiness	0	0	0.02	0.98	0	0
Sadness	0.06	0.01	0.02	0	0.91	0
Surprise	0	0.01	0	0	0	0.99

Fig. 13 Confusion matrix of facial expressions in BU3DFE database

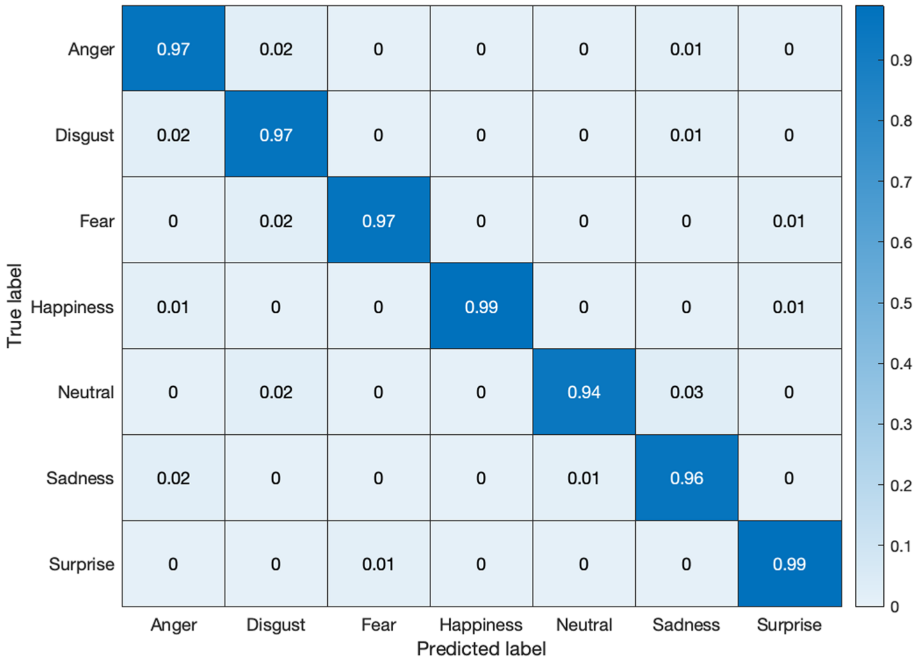


Fig. 14 Confusion matrix of facial expressions in MMI database

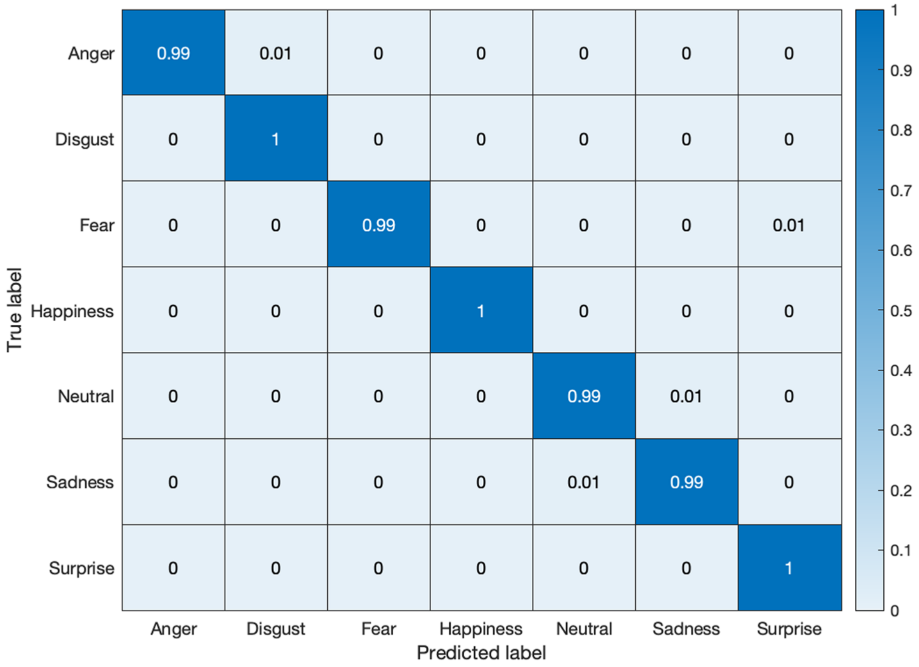


Fig. 15 Confusion matrix of facial expressions in CK + database

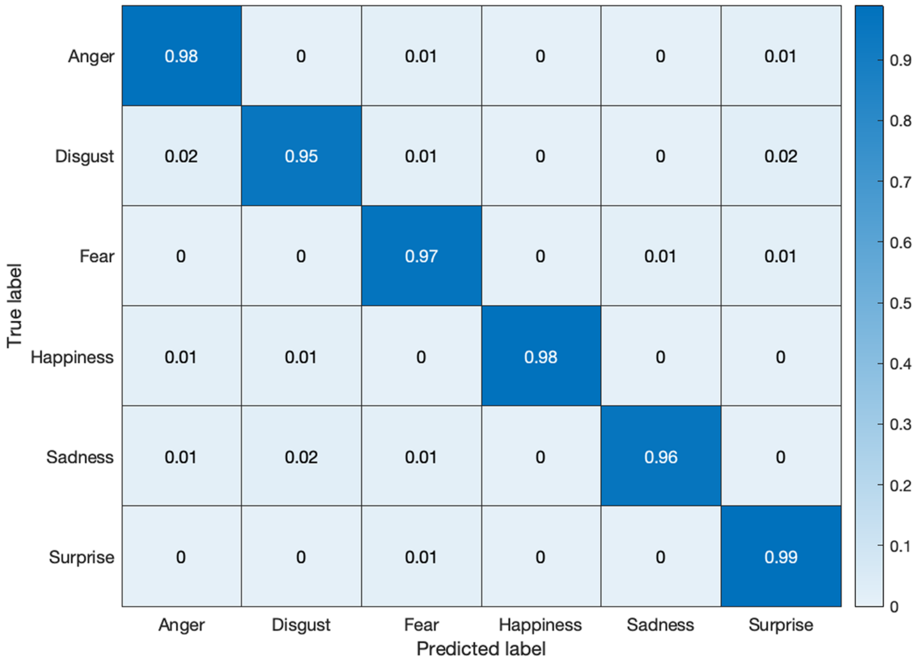


Fig. 16 Confusion matrix of facial expressions in BP4D-Spontaneous datasets

Table 2 Comparison of results on different methods

Method	Database	Recognition (%)	Ref
Twin support vector machines classifier	MMI	92.56 ± 3.02	[32]
DBM-DACNN with entropy loss	MMI	79.25	[33]
Deep learning neural network-regression	CK+	97.27	[30]
Deep learning + random forest	CK+	99.00	[31]
Twin support vector machines classifier	CK+	93.42 ± 3.25	[32]
DBM-DACNN with entropy loss	CK+	96.46	[33]
Geotopo+	BP4D-Spontaneous	88.56	[34]
Two-phase weighted collaborative representation classification	BP4D-Spontaneous	100	[35]
Fine-grained matching of 3D keypoint descriptors	Bosphorus	98.90	[21]
Kernel methods on Riemannian manifold	Bosphorus	86.70	[36]
SVM with EPE	Bosphorus	84.00	[37]
Two-phase weighted collaborative representation classification	Bosphorus	98.90	[35]
Kernel methods on Riemannian manifold	BU-3DFE	92.62	[36]
SVM with EPE	BU-3DFE	85.81	[37]
Manifold CNN	BU-3DFE	86.67	[38]
CNN model	BU-3DFE	92.57	[39]
Proposed method	MMI	97.20	This study
Proposed method	CK+	98.20	This study
Proposed method	BP4D-Spontaneous	97.20	This study
Proposed method	Bosphorus	98.90	This study
Proposed method	BU-3DFE	93.50	This study

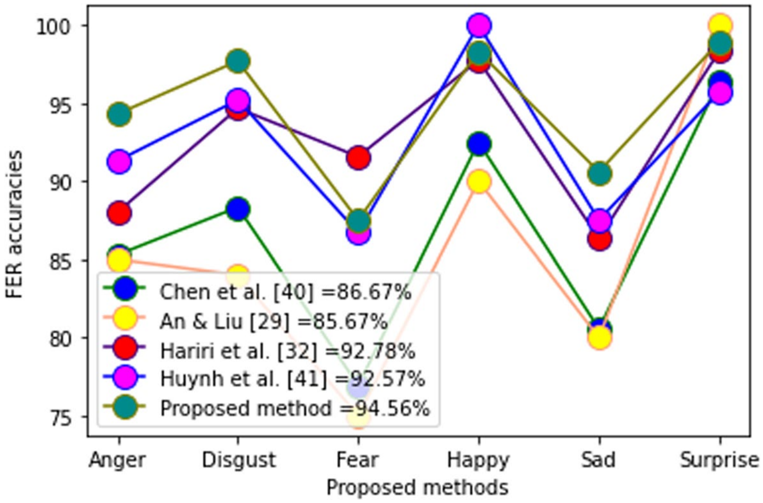


Fig. 17 Performance of 6 FER prototypes in BU3DFE database

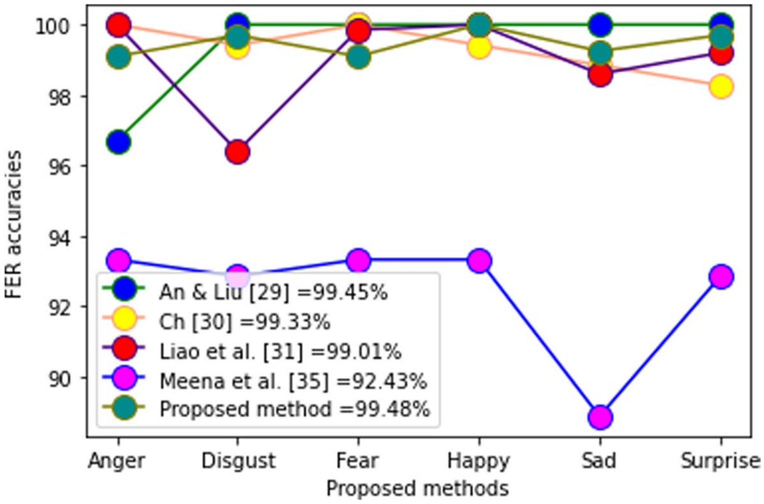


Fig. 18 Performance of 6 FER prototype in CK + database

the Bosphorus database, the proposed method outperformed the most recent methods (Fig. 19). A comparison of the performances of the individual FER prototypes in the MMI and BP4D-Spontaneous databases could not be executed because there were no reported data for comparison at the time of compilation. Statistical analysis using ANOVA shows the following performance results:

In the Bosphorus database, an analysis of variances demonstrated statistically significant differences between the proposed technique and the following: Hariri et al. [36] ($p = 0.001$), Azazi et al. [37] ($p = 0.000$), and Moeini A and Moeini H [40] ($p = 0.013$). In addition, the outcome is the same as in the BU3DFE: the variance analysis shows that a statistically significant difference ($p < 0.05$) exists between the proposed method and all other methods. However, in the CK + FER database, the statistical analysis shows that, except ref. [41], where a statistically significant difference ($p < 0.05$) exists, the remaining datasets show no statistically significant differences ($p > 0.05$). The proposed method compared to yields from An and Liu [29] ($p = 0.847$), Ch [30] ($p = 0.909$), and Liao et al. [31] ($p = 0.991$). Although the analysis appears to reveal a balanced performance between the proposed methodology and the last three techniques, the average

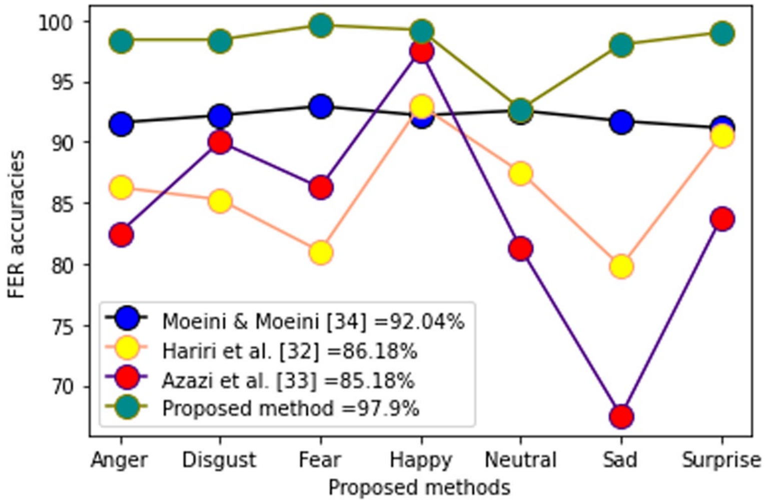


Fig. 19 Performance of 7 FER prototypes in Bosphorus database

recognition accuracy of the proposed method against any of them, as shown in Fig. 18, indicates that the proposed method is superior.

Conclusions

This study improves the FER performance in higher poses. 2D pose conversion schemes have been established to handle pose-invariant FER problems successfully, within a small-scale pose variation. However, they often flop for large-scale, in-depth face variations because of the disjointedness of the image. Human face geometry is ellipsoidal; therefore, the feature points are robustly tracked from one frame to next using an ellipsoidal model. We use the Gabor feature extraction technique for the salient visible features, mostly around the cheeks, eyes, mouth, and nose ridges. The Gabor feature extraction algorithm is useful for this study because it is selective toward orientation, localization, and frequency. We then used an ensemble classification technique, which combines SVM and AdaBoost, for feature selection and classification. The proposed technique outperforms the most recent and popular methods. In the future, we intend to investigate this problem using other feature extraction methods such as LBP and LBP + HOG.

Abbreviations

FER: Facial expression recognition; SVM: Saturated vector machine; LBP: Local binary patterns; HOG: Histogram of gradients; PCA: Principal component analysis; KNN: K-nearest neighbor; SMOTE: Synthetic minority oversampling technique; 2D: Two-dimensional; 3D: Three-dimensional; LOSOCV: Leave-one-subject-out cross validation.

Acknowledgements

Not applicable.

Authors' contributions

All authors drafted this manuscript. Ideation was proposed by EO. EO and JKA developed the proposed solution. PO performed the experimentation. All authors finally discussed and analyzed the results from the experimentation. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All data used for this data are publicly available on request from the original authors.

Declarations

Competing interests

All authors declare that there is no known competing interest.

Author details

[1]Department of Computer Science, University of Ghana, P. O. Box LG 163, Accra, Ghana. [2]Department of Computer Engineering, University of Ghana, P. O. Box LG 77, Accra, Ghana.

Received: 18 November 2021 Accepted: 19 April 2022
Published online: 16 May 2022

References

1. Panksepp J (2005) Affective consciousness: Core emotional feelings in animals and humans. Conscious Cogn 14(1):30-80. https://doi.org/10.1016/j.concog.2004.10.004
2. Plutchik R (2001) The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Amer Scient 89(4):344-350. https://doi.org/10.1511/2001.4.344
3. Zautra AJ (2003) Emotions, stress, and health. Oxford University Press, Oxford.
4. Kohler CG, Martin EA, Stolar N, Barrett FS, Verma R, Brensinger C et al (2008) Static posed and evoked facial expressions of emotions in

schizophrenia. Schizophr Res 105(1-3):49-60. https://doi.org/10.1016/j.schres.2008.05.010

5. Ambron E, Foroni F (2015) The attraction of emotions: irrelevant emotional information modulates motor actions. Psychon Bull Rev 22(4):1117-1123. https://doi.org/10.3758/s13423-014-0779-y
6. Kumari J, Rajesh R, Kumar A (2016) Fusion of features for the effective facial expression recognition. Paper presented at the international conference on communication and signal processing, IEEE, Melmaruvathur, 6–8 June 2016. https://doi.org/10.1109/ICCSP.2016.7754178
7. Shergill GS, Sarrafzadeh A, Diegel O, Shekar A (2008) Computerized sales assistants: the application of computer technology to measure consumer interest-a conceptual framework. J Electron Commer Res 9(2):176-191.
8. Tierney M (2017) Using behavioral analysis to prevent violent extremism: Assessing the cases of Michael Zehaf-Bibeau and Aaron Driver. J Threat Assessm Manag 4(2):98-110. https://doi.org/10.1037/tam0000082
9. Nonis F, Dagnes N, Marcolin F, Vezzetti E (2019) 3D approaches and challenges in facial expression recognition algorithms - A literature review. Appl Sci 9(18):3904. https://doi.org/10.3390/app9183904
10. Sandbach G, Zafeiriou S, Pantic M, Rueckert D (2011) A dynamic approach to the recognition of 3D facial expressions and their temporal models. Paper presented at the ninth IEEE international conference on automatic face and gesture recognition, IEEE, Santa Barbara, 21–25 March 2011. https://doi.org/10.1109/FG.2011.5771434
11. Vieriu RL, Tulyakov S, Semeniuta S, Sangineto E, Sebe N (2015) Facial expression recognition under a wide range of head poses. Paper presented at the 11th IEEE international conference and workshops on automatic face and gesture recognition, IEEE, Ljubljana, May 4–8, 2015. https://doi.org/10.1109/FG.2015.7163098
12. Yadav KS, Singha J (2020) Facial expression recognition using modified Viola-John's algorithm and KNN classifier. Multimed Tools Appl 79(19):13089-13107. https://doi.org/10.1007/s11042-019-08443-x
13. Jones M, Viola P (2003) Fast multi-view face detection. Mitsubishi Electric Research Laboratories, Cambridge.
14. Yao L, Wan Y, Ni HJ, Xu BG (2021) Action unit classification for facial expression recognition using active learning and SVM. Multimed Tools Appl 80(16):24287-24301. https://doi.org/10.1007/s11042-021-10836-w
15. Ashir AM, Eleyan A, Akdemir B (2020) Facial expression recognition with dynamic cascaded classifier. Neural Comput Appl 32(10):6295-6309. https://doi.org/10.1007/s00521-019-04138-4
16. Farrow CL, Shaw M, Kim H, Juhás P, Billinge SJL (2011) Nyquist-Shannon sampling theorem applied to refinements of the atomic pair distribution function. Phys Rev B 84(13):134105. https://doi.org/10.1103/PhysRevB.84.134105
17. Li F, Cornwell TJ, de Hoog F (2011) The application of compressive sampling to radio astronomy. I. Deconvolution. Astron Astrophys 528:A31. https://doi.org/10.1051/0004-6361/201015045
18. Perez-Gomez V, Rios-Figueroa HV, Rechy-Ramirez EJ, Mezura-Montes E, Marin-Hernandez A (2020) Feature selection on 2D and 3D geometric features to improve facial expression recognition. Sensors 20(17):4847. https://doi.org/10.3390/s20174847
19. Duan J (2019) Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. J Franklin Inst 356(8):4716-4731. https://doi.org/10.1016/j.jfranklin.2019.01.046
20. Kurniawati YE, Permanasari AE, Fauziati S (2018) Adaptive synthetic-nominal (ADASYN-N) and adaptive synthetic-KNN (ADASYN-KNN) for multiclass imbalance learning on laboratory test data. Paper presented at the 4th international conference on science and technology, IEEE, Yogyakarta, 7–8 August 2018. https://doi.org/10.1109/ICSTC.2018.8528679
21. Li HB, Huang D, Morvan JM, Wang YH, Chen LM (2015) Towards 3D face recognition in the real: a registration-free approach using fine-grained matching of 3D keypoint descriptors. Int J Comput Vis 113(2):128-142. https://doi.org/10.1007/s11263-014-0785-6
22. Comaniciu D, Ramesh V, Meer P (2003) Kernel-based object tracking. IEEE Trans Pattern Anal Mach Intell 25(5):564-577. https://doi.org/10.1109/TPAMI.2003.1195991
23. Hao GT, Du XP, Chen H, Song JJ, Gao TF (2015) Scale-unambiguous relative pose estimation of space uncooperative targets based on the fusion of three-dimensional time-of-flight camera and monocular camera. Opt Eng 54(5):053112. https://doi.org/10.1117/1.OE.54.5.053112
24. Dibeklioglu H, Salah AA, Akarun L (2008) 3D facial landmarking under expression, pose, and occlusion variations. Paper presented at the IEEE second international conference on biometrics: theory, applications and systems, IEEE, Washington, 29 September-1 October 2008. https://doi.org/10.1109/BTAS.2008.4699324
25. Owusu E, Wiafe I (2021) An advance ensemble classification for object recognition. Neural Comput Appl 33(18):11661-11672. https://doi.org/10.1007/s00521-021-05881-3
26. Dharavath K, Laskar RH, Talukdar FA (2013) Qualitative study on 3D face databases: A review. Paper presented at the annual IEEE India conference, IEEE, Mumbai, 13–15 December 2013. https://doi.org/10.1109/INDCON.2013.6726093
27. Sandbach G, Zafeiriou S, Pantic M, Yin LJ (2012) Static and dynamic 3D facial expression recognition: A comprehensive survey. Image Vision Comput 30(10):683-697. https://doi.org/10.1016/j.imavis.2012.06.005
28. Quan W, Matuszewski BJ, Shark LK, Ait-Boudaoud D (2009) Facial expression biometrics using statistical shape models. EURASIP J Adv Signal Process 2009:261542. https://doi.org/10.1155/2009/261542
29. An FP, Liu ZW (2020) Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM. Vis Comput 36:483-498. https://doi.org/10.1007/s00371-019-01635-4
30. Ch S (2021) An efficient facial emotion recognition system using novel deep learning neural network-regression activation classifier. Multimed Tools Appl 80(12):17543-17568. https://doi.org/10.1007/s11042-021-10547-2
31. Liao HB, Wang DH, Fan P, Ding L (2021) Deep learning enhanced attributes conditional random forest for robust facial expression recognition. Multimed Tools Appl 80(19):28627-28645. https://doi.org/10.1007/s11042-021-10951-8
32. Kumar MP, Rajagopal MK (2019) Detecting facial emotions using normalized minimal feature vectors and semi-supervised twin support vector machines classifier. Appl Intell 49(12):4150-4174. https://doi.org/10.1007/s10489-019-01500-w
33. Li S, Deng WH (2019) Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. Int J Comput Vis 127(6):884-906. https://doi.org/10.1007/s11263-018-1131-1
34. Danelakis A, Theoharis T, Pratikakis I, Perakis P (2016) An effective methodology for dynamic 3D facial expression retrieval. Pattern Recogn 52:174-185. https://doi.org/10.1016/j.patcog.2015.10.012
35. Lei YJ, Guo YL, Hayat M, Bennamoun M, Zhou XZ (2016) A two-phase weighted collaborative representation for 3D partial face recognition with single sample. Pattern Recogn 52:218-237. https://doi.org/10.1016/j.patcog.2015.09.035
36. Hariri W, Tabia H, Farah N, Benouareth A, Declercq D (2017) 3D facial expression recognition using kernel methods on Riemannian manifold. Eng Appl Artif Intell 64:25-32. https://doi.org/10.1016/j.engappai.2017.05.009
37. Azazi A, Lutfi SL, Venkat I, Fernández-Martínez F (2015) Towards a robust affect recognition: Automatic facial expression recognition in 3D faces. Expert Syst Appl 42(6):3056-3066. https://doi.org/10.1016/j.eswa.2014.10.042
38. Chen ZX, Huang D, Wang YH, Chen LM (2018) Fast and light manifold CNN based 3D facial expression recognition across pose variations. Paper presented at the 26th ACM international conference on multimedia, ACM, Seoul, 22–26 October 2018. https://doi.org/10.1145/3240508.3240568
39. Huynh XP, Tran TD, Kim YG (2016) Convolutional neural network models for facial expression recognition using BU-3DFE database. In: Kim K, Joukov N (eds) Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering, vol 376. Springer, Singapore, pp 441–450. https://doi.org/10.1007/978-981-10-0557-2_44
40. Moeini A, Moeini H (2015) Real-world and rapid face recognition toward pose and expression variations via feature library matrix. IEEE Trans Inform Forensics secur 10(5):969-984. https://doi.org/10.1109/TIFS.2015.2393553
41. Meena HK, Sharma KK, Joshi SD (2020) Effective curvelet-based facial expression recognition using graph signal processing. Signal Image Video Process 14(2):241-247. https://doi.org/10.1007/s11760-019-01547-9

Publisher's Note