A hybrid feature descriptor with Jaya optimised least squares SVM for facial expression recognition

Nikunja Bihari Kar1,3 | Deepak Ranjan Nayak2 | Korra Sathya Babu3 | Yu-Dong Zhang4

1 Department of Computer Science and Engineering, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
2 Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India
3 Department of Computer Science and Engineering, National Institute of Technology, Rourkela, India
4 Department of Informatics, University of Leicester, Leicester, UK

Correspondence
Deepak Ranjan Nayak, Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India.
Email: drnayak@ieee.org

Abstract
Facial expression recognition has been a long-standing problem in the field of computer vision. This paper proposes a new simple scheme for effective recognition of facial expressions based on a hybrid feature descriptor and an improved classifier. Inspired by the success of stationary wavelet transform in many computer vision tasks, stationary wavelet transform is first employed on the pre-processed face image. The pyramid of histograms of orientation gradient features is then computed from the low-frequency stationary wavelet transform coefficients to capture more prominent details from facial images. The key idea of this hybrid feature descriptor is to exploit both spatial and frequency domain features which at the same time are robust against illumination and noise. The relevant features are subsequently determined using linear discriminant analysis. A new least squares support vector machine parameter tuning strategy is proposed using a contemporary optimisation technique called Jaya optimisation for classification of facial expressions. Experimental evaluations are performed on Japanese female facial expression and the Extended Cohn–Kanade (CK+) datasets, and the results based on 5-fold stratified cross-validation test confirm the superiority of the proposed method over state-of-the-art approaches.

1 INTRODUCTION

Human behaviour detection is the backbone of advancement in artificial intelligence, automatic health monitoring system, robotics, surveillance and security [1]. Research on facial expression recognition (FER), a subfield in the vast area of human behaviour detection, drives a significant stride in recent years due to its enormous applications in day to day life. Facial expressions play a significant role in communicating the inner feelings of a human as it does not involve any auditory form.

The development of the automatic FER system by Ekman and Friensen [2] has made a breakthrough in the facial expression analysis field. The basic threefold FER includes the acquisition of the facial image, facial feature representation and expression classification. The acquisition step aims to capture the subject from the target image using various detection techniques [3]. The visual details of each subject are captured using various feature extraction algorithms. The features can be extracted using appearance-based, geometric-based and deep learning techniques. Geometric feature extractors describe different components of the face such as nose, eyebrows and mouth, whereas the appearance-based techniques describe the local appearances and shape of human faces and include histogram of oriented gradients (HOG) features [4–6], local binary pattern (LBP) descriptors, Gabor features [7], and discrete wavelet transform (DWT) [8] etc. The feature extractors based on deep learning algorithms enable to capture both low-level and high-level information from the face images [9, 10] and have gained increasing prevalence in current years. The earlier studies reveal that the increased feature vector length and the disturbances in facial expressions like varying pose, face alignment, variation in the environment, illumination, occlusions [11, 12] strongly affect the robustness and accuracy of the FER methods. To address these challenges, the current research trend is focused on the design of the robust feature descriptor and classifiers.

Shape deformations are vital visual cues used to distinguish between neutral and peak expressions. This information is usually notable in the frequency domain instead of the image domain. Wavelets have incredible spatial-frequency...
localisation characteristics and wavelet features are utilised to recognise multi-scale and multi-directional textural changes. Therefore, many current investigations on FER used wavelet and its variants like discrete wavelet transform (DWT) and stationary wavelet transform (SWT) [13] to derive salient features. Amongst wavelet transforms, SWT enjoys the shift-invariant property and provides a dense approximation to the continuous wavelet transform. A spatial shape descriptor called the pyramid of histograms of orientation gradients (PHOG) has rapidly become a method of choice for image classification tasks since it represents the edges that are distributed spatially and formulates them in a vector form [14]. Also, it generates a smaller size feature vector as compared to other feature descriptors.

The aforementioned reasons have motivated us to design a new and yet effective feature descriptor using both SWT and PHOG. At first, the SWT is applied over the images that yield a set of high-frequency components and a low-frequency component of similar size to the original image. The PHOG features are then computed from the low-frequency component of SWT that allows capturing more facial details.

To perform the classification of different expressions, SVM has been used in most researches although it carries some drawbacks such as high computational overhead and low performance on large scale data. To address these problems, a powerful variant called least squares SVM (LS-SVM) is introduced in [21]. LS-SVM has been shown to be preferred over SVM in many applications [22, 23]. Similar to SVM, LS-SVM with RBF kernel involves the tuning of two key parameters C and γ which greatly regulates its performance and hence a proper setting of these parameters is of great necessity. Most of the previous studies adopt a grid-search technique to find the best values of these parameters; however, it has limitations like high computational burden and getting stuck at local minima. Meta-heuristics algorithms can be a proper choice in place of the grid-search algorithm to mitigate the above issues. Jaya optimisation (JO), a modern meta-heuristic approach has obtained dramatic success in a wide range of applications and has a property of not requiring algorithm-specific parameters [24]. To the best of our knowledge, the capability of the JO algorithm has not been studied for finetuning the parameter of RBF based LS-SVM. Therefore, in this work, we explore the effectiveness of JO for parameter selection and further apply the resultant model “JOLS-SVM” to classify different emotion classes. To verify the efficiency of the proposed JOLS-SVM classifier, it is compared with the competitive methods such as PSO optimised LS-SVM (PSOLS-SVM) and GA optimised LS-SVM (GALS-SVM). The experimental results on two standard datasets confirm the supremacy of the JOLS-SVM method.

The contributions of our research can be summarised as follows:

(i) A novel hybrid feature descriptor is proposed based on SWT and PHOG. The SWT is first employed for decomposing the images into several frequency subbands. The PHOG is later applied over the low-frequency SWT component to derive the salient features from the face images. This hybrid feature descriptor exploits both spatial and frequency domain features and also provides better illumination variation and less sensitivity toward the noise.

(ii) A new parameter selection strategy for LS-SVM is proposed using the JO algorithm. The resulted algorithm is called JOLS-SVM. The potential of the JOLS-SVM model is compared against PSOLS-SVM, GALS-SVM, SVM, BPNN, and random forest classifiers.

(iii) The proposed model is evaluated on two benchmark datasets and its performance is compared with a set of existing schemes.

The rest of the paper is composed as follows: A brief overview of the related works is presented in Section 2. Section 3 provides a description of the datasets used. The proposed framework is detailed in Section 4. Section 5 presents the results and discussion and Section 6 concludes the paper.

2 RELATED WORK

FER is one of the most remarkable, natural, and universal way for humans to convey their emotional states and intentions. Mostly FER systems have three necessary steps of operation, namely face detection, facial feature representation, and expression classification. Conventional FER methods can be categorised into two kinds: action unit (AU) based expression recognition and feature-based expression recognition.

AU based recognition method uses different AUs from facial images for emotion detection. The primary motivation behind AU based emotion classification is the facial action coding system (FACS) [2]. The AUs detected from face images are grouped to represent an expression in AU based FER approaches. There exist many algorithms in literature based on the handcrafted features such as Gabor features [25], discrete wavelet transform (DWT) [8], histogram of oriented gradients (HOG) [4], local binary pattern (LBP) [26], local ternary pattern (LTP) [27], curvelet transform (CT) [18], ripplet transform (RT) [28], stationary wavelet transform (SWT) [13] etc. Many of them have been reported as state-of-the-art algorithms. Hence, there is a tradeoff between accuracy and computational efficiency in designing these handcrafted features.

There are few FER systems reported in the literature which extracted features by dividing an input image into patches with various scales [17]. Features derived from image patches represent minute information that formulates a specific expression and thereby, enhancing the recognition accuracy significantly. In a few studies, feature fusion based FER techniques are introduced to improve the comprehensive representation capability. Wang et al. [5] exploited the Weber local descriptor (WLD) in combination with HOG for the extraction of features. They divided the facial image into weighted blocks, and then two features were extracted and fused. The fusion of WLD and HOG was able to represent texture, counter, and shape information efficiently. In [17], autoencoders were used to fuse geometric and LBP features for a comprehensive representation of different facial expressions. Kohonen self-organising map (KSOM) classifier was employed for recognition. Happy and Routray
### TABLE 1 A summary of existing FER systems

| References                  | Feature                                      | Classifier  | Database | Accuracy(%) |
|-----------------------------|----------------------------------------------|-------------|----------|-------------|
| Zhang & Tjondronegoro [7]   | Patch based 3D-Gabor                         | RBF SVM     | CK       | 94.48       |
|                             |                                               |             | JAFFE    | 92.93       |
| Wang et al. [5]             | WLD and HOG                                   | KNN         | CK       | 95.86       |
|                             |                                               |             | JAFFE    | 93.37       |
| Siddiqi et al. [15]        | Stepwise LDA                                  | HCRF        | CK       | 96.83       |
|                             |                                               |             | JAFFE    | 96.33       |
| Happy & Routray [16]       | LBP                                           | SVM         | CK+      | 94.09       |
|                             |                                               |             | JAFFE    | 91.80       |
| Majumder et al. [17]       | Autoencoder based feature fusion (LBP+Geometric) | KSOM        | MMI      | 97.55       |
|                             |                                               |             | CK+      | 98.95       |
| Uccar et al. [18]          | Local Curvelet transform                      | OSELM-SC    | CK+      | 95.1        |
|                             |                                               |             | JAFFE    | 94.65       |
| Zhang et al. [19]          | BWE                                           | Fuzzy SVM   | Own Dataset | 96.77   |
| Wang et al. [20]           | SWE                                           | FNN+Jaya    | Own Dataset | 96.80   |

In many FER systems, multi-resolution techniques have been adopted for feature extraction. A combined feature descriptor using DWT and Fisher linear discriminant analysis (FLD) is suggested in [8]. Zhang and Tjondronegoro [7] considered facial movement and muscle movement features for recognition. Considering salient patches 3D Gabor features were obtained followed by patch matching operation that demonstrated promising outcomes for face registration faults with fast processing time. Kazmi and Jaffar [31] presented a FER approach based on 3 level decomposition of 2D-DWT with PCA and a bank of 7 SVM classifiers, while Siddiqi and Lee [32] reported a novel feature extraction technique based on a symlet wavelet. Later, a novel FER system is proposed by Uccar et al. [18] using curvelet features and online sequential extreme learning machine (OSELM). The image is partitioned into local regions, and the curvelet features are computed from these regions. Zhang et al. [19] formulated a model with the use of fuzzy multiclass SVM and biorthogonal wavelet entropy (BWE) features for emotion recognition. Recently, an effective model is designed in [13] using stationary wavelet entropy (SWE) features and feed-forward neural network. A summary of some notable FER approaches is presented in Table 1.

The minute variation in the expression plays a vital role in differentiating the emotions. These variations are more significant in the frequency domain than the pixel domain. Therefore, the existing literature covered FER techniques using different frequency-domain feature descriptors. The DWT is widely used as a feature extraction tool in many FER systems despite it is translation-variant. The stationary wavelet transform (SWT) on the other hand holds the translation-invariant property. Due to supplementary benefits, few efforts have been made to propose hybrid feature descriptors, however, the design of a robust and powerful hybrid feature descriptor is still an open research problem. Further, SVM is extensively used as the classifier in most of the studies in spite of possessing shortcomings such as high computational overhead and low performance on large scale data. The LS-SVM overcomes the problems faced by SVM, but requires tuning of two vital parameters $C$ and $\gamma$. Traditionally, these two parameters are set using a grid search technique which suffers from limitations like high computational burden and getting stuck at local minima. Hence, the efficacy of meta-heuristics based optimisation techniques can be investigated to mitigate the above issues. In this study, a novel hybrid feature descriptor is presented using SWT and PHOG which preserves the translation-invariant property and represents the contour and shape information effectively. The JO algorithm is harnessed to fine-tune the parameter of LS-SVM and the
The resultant model JOLS-SVM is applied to predict the emotion class.

3 | DATASET

Two standard datasets such as CK+ [33] and JAFFE [34] are considered to validate the suggested framework. In the CK+ dataset, 123 subjects recorded 593 image sequences, among which 327 image sequences are named with seven crucial emotion groups. While in the JAFFE dataset, 10 subjects exhibit 213 peak expression images. To evaluate our proposed model, we have chosen 213 and 450 peak expressive image samples from JAFFE and CK+ datasets, respectively. The normalised image samples from both the datasets are depicted in Figure 1. Table 2 indicates the selected number of sample images against each emotion class for both the considered datasets.

| Emotion | Database | Anger | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---------|----------|-------|---------|------|-------|---------|-----|----------|
| CK+     |          | 45    | 18      | 59   | 25    | 69      | 123 | 28       | 83      |
| JAFFE   |          | 30    | -       | 29   | 32    | 31      | 30  | 31       | 30      |

4 | PROPOSED METHODOLOGY

The proposed system includes four vital stages: (i) image pre-processing, (ii) feature representation, (iii) facial feature dimension reduction, and (iv) emotion classification. Figure 2 depicts the overall architecture of the proposed framework. The details of each stage are described in the next sections.
4.1 | Preprocessing

At first, all the input images are changed over to grayscale images. At that point, image contrast is increased by saturating 1% of all pixels values from top and bottom. Then, Viola and Jones algorithm [3] has been utilised for identification of the face area. Thereafter, the detected face area is cropped and standardised to obtain 128 × 128 size output images.

4.2 | Feature extraction using SWT and PHOG

The feature extraction is performed using both SWT and PHOG. A brief overview of SWT, PHOG and the proposed hybrid feature descriptor is provided below:

4.2.1 | SWT

Nason and Silverman [35] proposed SWT, which is shift invariant, redundant and provides a dense approximation to the continuous wavelet transform as compared to orthogonal wavelet transform. The DWT lacks translation invariant property that means the DWT of a translated signal is not same as the translated version of DWT of the same signal. Let \( I \) denotes a face image and a translation operator \( \tau \), then

\[
\tau(DWT(I)) \neq DWT(\tau(I)).
\]  

The classical DWT coefficients are obtained by convolving input signal with an appropriate filter by keeping only even indexed elements [36]. The time-variant problem occurs due to the downsampling of odd indexed elements from DWT coefficients. To avoid this problem, we need to keep odd indexed elements by decimating the even indexed elements. At each level decomposition of \( \varepsilon \)-decimated DWT, there is a choice to decimate even or odd indexed elements. If we decompose the input signal up to \( j \)th level and find all possible decompositions, then we have \( 2^j \) distinct decompositions [37]. The SWT of a given signal contains all the \( \varepsilon \)-decimated coefficients. To obtain the approximation and detail level-1 SWT coefficients, the input signal is convolved with a given filter in the absence of downsampling. The size of SWT coefficients at each decomposition level is similar to the input signal. Usually, the approximation and detail coefficients at level \( j \) are obtained by convolving the approximation coefficient from level \( j - 1 \) with a given filter without downsampling the result. The 1D SWT algorithm can be easily extended to 2D SWT.

4.2.2 | PHOG

Pyramid of histogram of gradients (PHOG) [14] is a spatial extension descriptor of the histogram of gradients (HOG) [38]. The HOG has been widely used in the computer-human interface systems to count the number of times the gradient orientation appears in a specific localised part of an image. As an extension to HOG, PHOG represents the edges that are distributed spatially and the vector form is then formulated. The four-fold PHOG extraction process can be described as:

- Extraction of edge contours from the sample image for processing.
- Division of sample image into cells comprising of different levels of a pyramid.
- Each grid of each pyramid at the resolution level is used to compute the HOG.
- The obtained HOG vectors at each resolution level of the pyramid are concatenated for PHOG description. The concatenation of HOG vectors results in the spatial information which is further normalised considering all the pyramid levels.

For the building of the pyramid of the given image at level 1, we need to divide the image into \( 2^c \) cells along with both the directions of 2D axis. Therefore, level 0 is represented by a \( K \)-vector corresponding to the \( K \) bins of the histogram, level 1 is represented by \( 4K \)-vector and so on. Thus, the size of the PHOG descriptor of the entire image \( (F_{\text{vec}}) \) can be formulated as follows:

\[
F_{\text{vec}} = K \times \sum_{j=0}^{L} 4^j.
\]  

In this work, the PHOG descriptor is quantised into eight orientation bins ranging from \([0 - 360^\circ]\) and the values of \( L \) (number of levels) and \( K \) (number of bins) are set to 3 and 8, respectively. Thus, the dimension of the PHOG descriptor is 680.

4.2.3 | Proposed hybrid feature descriptor

The potential of both SWT and PHOG has motivated us to design a hybrid feature extractor and the steps for feature extraction are discussed in the following. Firstly, the level-1 SWT is exploited over input images to obtain the decomposed frequency components. Secondly, the PHOG features are extracted from the approximation coefficients \( (A_1) \) of 1-level SWT decomposition. The steps can be mathematically represented as

\[
[A_j, D_j] = SWT(I),
\]

\[
F_{\text{vec}} = PHOG(A_j),
\]

where, \( A_j \) and \( D_j \) denote the approximation and detail SWT coefficients at level \( j \) and \( F_{\text{vec}} \) represents the final feature vector. The vital information of the image that is essential for representing different emotions lies in the approximation coefficient of the SWT decomposition. The PHOG captures the shape.
information and it is robust against illumination and orientation. The SWT features ensure the translation-invariant property and PHOG retrieves the shape information of facial images. The information contained in the local block is crucial for emotion detection. Therefore, in this work, we hybridise the effectiveness of both SWT and PHOG features that facilitates to provide more robust and relevant features.

4.3 Dimensionality reduction using LDA

The derived hybrid features are further analysed using LDA to get more discriminant features to improve recognition performance. That is, a set of hybrid features of \( N \) samples \( \{p_1, p_2, \ldots, p_N\} \) in \( d \)-dimensional space \( p_i \in \mathbb{R}^d \) belong to \( C \) classes of \( \{q_i\}_{i=1,2,\ldots,C} \) are subjected to LDA. The LDA technique [39] has been used to increase the emotion class separability.

The majority of discriminant feature vectors map from high dimension to lower dimension feature space. The projection vectors from LDA assist all project samples in forming maximum between-class separability, and minimum inter-class scatter. Note that there should be at most \( C - 1 \) non-zero generalised eigenvalues for the \( d \)-dimensional feature matrix, where \( C \) represents the number of classes in the database. The database employed for this work has 8 and 7 emotion categories for CK+ and JAFFE dataset, respectively. Therefore, the resultant feature vector contained 7 and 6 numbers of features for CK+ and JAFFE dataset, respectively.

4.4 Proposed Jaya optimised LS-SVM

In this section, we discuss the Jaya optimisation, LS-SVM and the proposed JOLS-SVM method.

4.4.1 Jaya optimisation

Jaya is a simple and robust optimisation technique for solving different kinds of optimisation problems efficiently [40]. Due to its simplicity, it has gained a lot of attention among researchers. This algorithm doesn’t need any algorithm-specific parameters (\( AxP \)); however, it requires the common control parameters (CCP) like population size, generation number, and termination criterion to find the solution. Jaya algorithm is shown to give preferable outcomes over other optimisation algorithms [24, 41, 42]. The conceptual thought behind this algorithm is that it generally pushes the obtained solution toward the best solution by avoiding the worst solution.

The overall steps involved in Jaya algorithm are illustrated in Figure 3. Let \( f(p) \) be the objective function to be minimised, and \( x_i, i \epsilon \) be the index of variable, iteration, and candidate solution, respectively. At any iteration \( i \), suppose there are “\( d \)” number of variables (i.e. dimension \( x = 1, 2, 3, \ldots, d \)), with “\( n \)” number of candidate solutions (i.e. population size \( c = 1, 2, 3, \ldots, n \)). Assume \( P_{x,i,c} \) indicates the value of \( x^c \)th variable for \( i \)th iteration. The modified solution \( P'_{x,i,c} \) can be defined as follows:

\[
P'_{x,i,c} = P_{x,i,c} + r_{x,i,1} \left( P_{x,i,Best} - P_{x,i,c} \right) - r_{x,i,2} \left( P_{x,i,Worst} - P_{x,i,c} \right),
\]

where, \( P_{x,i,Best} \) and \( P_{x,i,Worst} \) are the values of \( x^c \)th variable in \( i \)th iteration for the best and worst candidate, respectively, \( r_{x,i,1} \) and \( r_{x,i,2} \) are two random variables in the range \([0, 1]\) for \( x^c \)th variable in \( i \)th iteration. The term \( r_{x,i,1} \left( P_{x,i,Best} - P_{x,i,c} \right) \) helps the candidate to approach toward the best solution, while the term \( r_{x,i,2} \left( P_{x,i,Worst} - P_{x,i,c} \right) \) helps the candidate to move away from worst solution. The value of \( P'_{x,i,c} \) is only accepted if it produces better function value. All the accepted values at the end of each iteration are stored and are fed as the input to the next iteration.

The updated candidate in the \((i + 1)\)th iteration can be termed as follows:

\[
P_{x,i+1,c} = \begin{cases} 
P'_{x,i,c} & \text{if } f(P'_{x,i,c}) < f(P_{x,i,c}) \\ P_{x,i,c} & \text{if } f(P'_{x,i,c}) \geq f(P_{x,i,c}) \end{cases},
\]

where \( f(p) \) represents the fitness function. The above equation indicates that \( P_{x,i+1,c} \) is assigned to \( P'_{x,i,c} \) (the modified candidate) if its fitness value is better than \( P_{x,i,c} \) (the current candidate), otherwise, it is assigned to \( P_{x,i,c} \).
4.4.2 | LS-SVM

The performance of the conventional SVM is low and also the computational overhead is high when it is exposed to massive datasets. To reduce computational overhead and improve performance, LS-SVM [21] is generally utilised which is a more capable variant of SVM. The LS-SVM classifier utilises linear or non-linear hyperplanes to classify samples belong to two or more hyperplane. In general, extensive experiments are carried out utilising RBF, linear, and polynomial kernels with LS-SVM to find the best among them. The RBF kernel is mostly used along with LS-SVM as compared to polynomial and linear kernel. Let \( N \) be the number of samples \( \{p_i, q_i\}_{i=1}^{N} \), where \( p_i \in \mathbb{R}^n \) denoted the \( i \)th input data and \( q_i \in \mathbb{R} \) denotes the \( i \)th output label. The decision function of the LS-SVM classifier can be defined as,

\[
f(p) = \text{sign} \left[ \sum_{i=1}^{N} \alpha_i q_i \kappa(p, p_i) + b \right],
\]

where, \( \alpha_i \) is the Lagrange multiplier, and \( \kappa(\cdot, \cdot) \) represents the kernel function. The RBF kernel requires tuning of two crucial parameters: penalty parameter \( C \) and bandwidth \( \gamma \). In RBF kernel, \( \gamma \) is responsible for the non-linear mapping of low-dimensional feature space into high-dimensional output space, while \( C \) controls the tradeoff between the model complexity and minimisation of the fitting error.

4.4.3 | Jaya-based LS-SVM

Several studies illustrate that the performance of RBF kernel based LS-SVM is solely dependent on parameters \( C \) and \( \gamma \). Therefore, these parameters need to be initialised precisely before applying to real-life problems. The grid search technique has been served as the conventional procedure for obtaining the best results despite it possessing limitations like high computational burden and getting stuck at local minima. Recently, it has been observed that the optimisation methods such as moth-flame optimisation (MFO) algorithm [43] and genetic algorithm (GA) [6] are widely used to obtain the global best solution as compared to the grid-search method. Therefore, in this study, the simplest and yet effective optimisation technique, Jaya, is introduced to identify the parameters of RBF based LS-SVM, and the resultant model is referred to as JOLS-SVM that is used to predict the emotion class.

The general framework of the proposed JOLS-SVM is illustrated in Figure 4. The procedure of JOLS-SVM is carried out in two steps: (1) parameter optimisation and (2) classification.
evaluation. In the parameter optimisation phase, a 5-fold stratified cross-validation (SCV) strategy is applied to 4-folds of training data. Once the parameter optimisation is completed, the optimal parameters are supplied to the LS-SVM classifier for the classification of emotions on the original test data. It is worth noting that the classification task is carried out using the remaining 1-fold data. To design the fitness function, the classification accuracy has been taken into consideration and the accuracy is evaluated using the 5-fold SCV adopted for parameter optimisation.

5 | RESULTS AND DISCUSSION

The proposed scheme was evaluated on two standard datasets such as CK+ and JAFFE. The experiments were carried out on an Intel Core i7 processor with 3.4 GHz clock speed with primary memory of 8 GB machine. The MATLAB tool running on Windows OS was used to perform all the experiments.

5.1 Feature extraction and reduction

The SWT with a level-1 Haar wavelet was used to decompose the facial images of size $128 \times 128$ and obtain four frequency components: approximation, diagonal, horizontal, and vertical. These four components are named as $A_1$, $D_1$, $H_1$ and $V_1$. The coefficients of the $A_1$ component was chosen for the extraction of PHOG features. The number of $A_1$ coefficients for one image is found to be $128 \times 128 = 16,384$. After applying the PHOG descriptor over the $A_1$ coefficients, an aggregate feature vector length of 680 is obtained. It is worth noting that the parameters $L$ and $K$ in PHOG was set to 3 and 8, respectively. The hybrid feature vector length is still large, so there is a need for feature dimension reduction. The dimension of the features was reduced using LDA and the number of features after the reduction was only 7 and 6 for CK+ and JAFFE dataset, respectively. Note that there must be at most $C-1$ non-zero generalised eigenvalues for a $d$-dimensional feature matrix, where $C$ represents the number of classes in the dataset. The datasets considered in this work has 8 and 7 emotion categories for CK+ and JAFFE datasets, respectively and hence, the resultant feature vector contains 7 and 6 number of features for CK+ and JAFFE dataset.

5.2 Classification

In our experiments, a 5-fold SCV technique is used to make the classifier stable and more generalised for independent datasets. The two datasets, CK+ and JAFFE contain a different number of samples for each emotion category. The stratified technique splits each fold in such a way that each fold contains an equal proportion of images from each emotion category.

The classification results of the proposed method is evaluated using eight expressions of CK+ dataset and seven expressions of JAFFE dataset. The sensitivity ($S_{en}$), specificity ($S_{pc}$), and accuracy ($A_{cc}$) are the performance metrics that are employed to evaluate the efficiency of the proposed framework. All the parameter settings were kept constant throughout the experiment.

The effectiveness of Jaya optimisation with LS-SVM is rigorously compared with two widely used optimisation techniques such as PSO and GA, resulting in algorithms named as JOLS-SVM, PSOLS-SVM, and GALS-SVM. These resulted algorithms were implemented from scratch on the MATLAB environment. Figures 5a and 5b show the average evolution results of the best validation fitness obtained by JOLS-SVM, PSOLS-SVM, and GALS-SVM across 5 folds on CK+ and JAFFE datasets, respectively. It can be seen from the figure that the Jaya optimisation has exhibited best convergence curve compared to PSO and GA.

The effectiveness of the proposed JOLS-SVM classifier is tested against competent methods over the same datasets. The
TABLE 3  Performance evaluation of JO, PSO, and GA with LS-SVM

| Method       | CK+       |            |          | JAFFE     |            |          |
|--------------|-----------|------------|----------|-----------|------------|----------|
|              | Sen (%)   | Spe (%)    | Acc (%)  | Sen (%)   | Spe (%)    | Acc (%)  |
| JOLS-SVM     | 94.08     | 99.51      | 98.38    | 95.23     | 99.61      | 98.99    |
| PSOLS-SVM    | 97.99     | 91.57      | 91.96    | 93.65     | 99.36      | 98.52    |
| GALS-SVM     | 93.41     | 98.27      | 97.49    | 89.49     | 98.66      | 97.31    |

TABLE 4  Detailed classification results of RBF JOLS-SVM over CK+ dataset

| Emotion   | TP  | TN  | FP  | FN  | Sen (%) | Spe (%) | Acc (%) |
|-----------|-----|-----|-----|-----|---------|---------|---------|
| Anger     | 45  | 405 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Contempt  | 18  | 432 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Disgust   | 57  | 390 | 1   | 2   | 96.61   | 99.74   | 99.33   |
| Fear      | 25  | 425 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Happy     | 69  | 380 | 1   | 0   | 99.73   | 99.77   | 99.77   |
| Sad       | 28  | 421 | 1   | 0   | 100.00  | 99.76   | 99.77   |
| Surprise  | 58  | 362 | 5   | 25  | 69.87   | 98.63   | 93.33   |
| Neutral   | 106 | 321 | 11  | 16  | 86.17   | 98.16   | 94.88   |
| Average   |     |     |     |     | 94.08   | 99.51   | 98.38   |

TABLE 5  Detailed classification results of linear LS-SVM over CK+ dataset

| Emotion   | TP  | TN  | FP  | FN  | Sen (%) | Spe (%) | Acc (%) |
|-----------|-----|-----|-----|-----|---------|---------|---------|
| Anger     | 45  | 405 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Contempt  | 18  | 432 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Disgust   | 55  | 386 | 5   | 4   | 93.22   | 98.72   | 98.00   |
| Fear      | 25  | 425 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Happy     | 69  | 380 | 3   | 0   | 99.21   | 99.33   | 99.33   |
| Sad       | 28  | 421 | 1   | 0   | 100.00  | 99.76   | 99.77   |
| Surprise  | 53  | 347 | 20  | 30  | 63.85   | 94.55   | 88.88   |
| Neutral   | 107 | 316 | 11  | 16  | 86.99   | 96.63   | 94.00   |
| Average   |     |     |     |     | 93.01   | 98.61   | 97.50   |

TABLE 6  Detailed classification results of polynomial LS-SVM over CK+ dataset

| Emotion   | TP  | TN  | FP  | FN  | Sen (%) | Spe (%) | Acc (%) |
|-----------|-----|-----|-----|-----|---------|---------|---------|
| Anger     | 45  | 405 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Contempt  | 15  | 431 | 1   | 3   | 83.33   | 99.76   | 99.11   |
| Disgust   | 57  | 390 | 1   | 2   | 96.61   | 99.74   | 99.33   |
| Fear      | 25  | 425 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Happy     | 69  | 380 | 1   | 0   | 99.73   | 99.77   | 99.77   |
| Sad       | 27  | 421 | 1   | 1   | 96.42   | 99.76   | 99.77   |
| Surprise  | 58  | 360 | 7   | 25  | 69.87   | 98.09   | 92.88   |
| Neutral   | 108 | 321 | 6   | 15  | 87.80   | 98.16   | 95.33   |
| Average   |     |     |     |     | 91.76   | 99.41   | 98.25   |

TABLE 7  Detailed classification results of RBF JOLS-SVM over JAFFE dataset

| Emotion   | TP  | TN  | FP  | FN  | Sen (%) | Spe (%) | Acc (%) |
|-----------|-----|-----|-----|-----|---------|---------|---------|
| Anger     | 30  | 183 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Disgust   | 25  | 181 | 1   | 3   | 83.33   | 99.76   | 99.11   |
| Fear      | 32  | 181 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Happy     | 30  | 181 | 1   | 1   | 96.77   | 99.45   | 99.06   |
| Neutral   | 29  | 181 | 2   | 1   | 96.66   | 98.90   | 98.59   |
| Sad       | 28  | 181 | 1   | 1   | 96.32   | 99.45   | 98.12   |
| Surprise  | 29  | 183 | 0   | 0   | 100.00  | 100.00  | 100.00  |
| Average   |     |     |     |     | 95.23   | 99.61   | 98.99   |

The results are presented in terms of several metrics like FP, TP, FN, and TN along with accuracy, specificity, and sensitivity. It can be observed that the JOLS-SVM with RBF kernel achieved the highest average classification accuracy compared to LS-SVM with polynomial and linear kernel. Similarly, the classification performance on JAFFE dataset are given in Tables 8 and 9 for RBF, linear and polynomial kernels. A higher accuracy is obtained with RBF kernel based JOLS-SVM classification method in comparison to linear and polynomial kernel based LS-SVM classifiers.
**TABLE 9** Details classification results of polynomial LS-SVM over JAFFE dataset

| Emotion | TP | TN | FP | FN | $S_{en}$ (%) | $S_{pe}$ (%) | $A_{cc}$ (%) |
|---------|----|----|----|----|--------------|--------------|-------------|
| Anger   | 30 | 183| 0  | 0  | 100.00       | 100.00       | 100.00      |
| Disgust | 22 | 177| 7  | 7  | 75.86        | 96.19        | 93.42       |
| Fear    | 30 | 179| 2  | 2  | 93.75        | 98.89        | 98.12       |
| Happy   | 28 | 178| 4  | 3  | 90.32        | 97.80        | 96.71       |
| Neutral | 29 | 182| 1  | 1  | 96.66        | 99.45        | 99.06       |
| Sad     | 29 | 178| 4  | 2  | 93.54        | 97.80        | 97.18       |
| Surprise| 28 | 178| 5  | 2  | 93.33        | 97.26        | 96.71       |
| Average |    |    |    |    | 91.93        | 98.20        | 97.31       |

Table 10 shows the training and testing time comparison (in seconds) among three classifiers such as JOLS-SVM, PSOLS-SVM, and GALS-SVM for both CK+ and JAFFE datasets. It can be noticed that the JOLS-SVM classifier obtained better accuracy with comparatively less training and testing time. It is worth noting here that the execution time exhibited has been computed for the entire set of training and testing samples.

### 5.3 Comparison with other classification methods

The performance of the proposed JOLS-SVM classifier is compared against its competent methods such as Lin LS-SVM and Poly LS-SVM and some standard classifiers such BPNN and RF and the results on CK+ and JAFFE datasets are listed in Table 11. It can be seen that the classification results obtained by JOLS-SVM+RBF classifier is superior to other classification methods. The accuracies achieved by Lin LS-SVM, Poly LS-SVM, BPNN and RF classifiers are 97.50%, 98.25%, 98.23% and 96.90%, respectively, for CK+ dataset, whereas the accuracies are 98.79%, 97.31%, 97.88% and 97.88% for JAFFE dataset which is less than the accuracy obtained by JOLS-SVM+RBF classifier. It is worth addressing here that the FER model with good specificity and sensitivity leads to a better performance. It is also observed that the JOLS-SVM yielded higher specificity (i.e. 99.51% for CK+ and 99.61% JAFFE dataset) and a comparable sensitivity (i.e. 94.08% for CK+ and 95.23% for JAFFE dataset) than other techniques.

The parameters involved with different classifiers and their settings are listed in Table 12. It can be noted that the CCP of optimisation techniques were kept similar to derive a fair comparison among them.

### 5.4 Comparison with state-of-the-art methods

Table 13 presents the performance comparison results of the proposed method with other state-of-the-art approaches over CK+ and JAFFE datasets. It is evident from the table that the proposed system leads to achieving the highest classification
TABLE 13  Comparison of proposed scheme with state-of-the-art methods

| Reference | Facial feature       | Classifier   | CK (%) | JAFFE (%) |
|-----------|----------------------|--------------|--------|-----------|
| Zhang et al. [7] | Patch based Gabor     | SVM          | 94.48  | 99.23     |
| Wang et al. [5]  | HOG + WLD            | KNN          | 93.97  | 95.86     |
| Mlakar and Potocinik [6] | HOG difference vector | SVM          | 95.64  | 87.82     |
| Happy and Routray [16] | Salient facial patches | RBF SVM     | 94.09  | 91.79     |
| Siddiqui et al. [15] | SWLDA                | HCRF         | 96.83  | 96.33     |
| Uccar et al. [18] | Local curvlet transform | OSELM-SC | 95.17  | 94.65     |
| Guo et al. [44]  | ELBP + KLT           | SVM          | 92.30  | 93.30     |
| Hu et al. [30]   | Feature Fusion of CS-LSMP | SVM      | 82.86  | 98.15     |
| Proposed method  | SWT + PHOG + LDA     | JOLS-SVM     | 98.38  | 98.99     |

TABLE 14  Classification performance evaluation (in %) of the proposed method in presence of random Gaussian noise

| Method                        | CK (%) | JAFFE (%) |
|-------------------------------|--------|-----------|
| Proposed method without noise | 94.08  | 99.51     |
| Proposed method with Noise    | 95.95  | 96.86     |

5.5  Performance evaluation in presence of the noise

The effectiveness of the proposed hybrid feature descriptor method is verified in presence of the noise. Table 14 exhibits the results obtained by the proposed method with noise in terms of accuracy, sensitivity and specificity for CK+ and JAFFE datasets. It is worth mentioning that the random Gaussian noise with a variance between 0.01 and 0.1 has been taken into consideration for this purpose. It is observed that the presence of noise led to a small decrease in classification performance that demonstrates the robustness of the suggested method.

6  CONCLUSION

In this work, we formulated an effective and accurate methodology for facial expression recognition based on a hybrid feature descriptor and a new classifier. The hybrid feature descriptor was designed using SWT and PHOG. Initially, the images were decomposed into several frequency components using SWT. The PHOG features were then computed from the approxima-

tion SWT coefficients to preserve the edge and shape features. The LDA method was subsequently employed to derive a reduced and discriminant feature set. Finally, the classification of different emotions was performed using a Jaya optimised LS-SVM (JOLS-SVM) classifier. The Jaya optimisation algorithm was served as a parameter tuning method and facilitated to maximise the generalisation capability of the standard LS-SVM classifier. The experimental results on the benchmark datasets showed that the results obtained by the proposed system are promising compared to the state-of-the-art methods. The proposed model achieved an accuracy of 98.38% and 98.99% over CK+ and JAFFE datasets. Further, the proposed JOLS-SVM method was found superior than other classifiers such as PSOLS-SVM, GALS-SVM, BPNN and RF.

Despite yielding higher performance, the proposed model needs to be tested over datasets containing large and diverse samples. In the future, the potency of the non-handcrafted features could be studied along with the proposed features to recognise emotions.

ORCID

Deepak Ranjan Nayak  https://orcid.org/0000-0002-8929-5778
Yu-Dong Zhang  https://orcid.org/0000-0002-4870-1493

REFERENCES

1. Bousmalis, K., Mehu, M., Pantic, M.: Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. Image and Vision Computing 31(2), 203–221. Elsevier, New York (2013)https://doi.org/10.1016/j.imavis.2012.07.003
2. Friesen, E., Ekman, P: Facial action coding system: a technique for the measurement of facial movement, vol. 3. Consulting Psychologists Press, Palo Alto, CA (1978)
3. Viola, P., et al.: Rapid object detection using a boosted cascade of simple features. Conference on Computer Vision and Pattern Recognition (CVPR) 1, 511–518 (2001). https://doi.org/10.1109/CVPR.2001.990517
4. Kar, N.B., Babu, K.S., Jena, S.K.: Face expression recognition using histograms of oriented gradients with reduced features. In: International Conference on Computer Vision and Image Processing, pp. 209–219. Springer, Berlin (2017)https://doi.org/10.1007/978-3-319-1017-7_19
5. Wang, X., et al.: Feature fusion of HOG and WLD for facial expression recognition. In: International Symposium on System Integration, pp. 227–232. IEEE, Piscataway, NJ (2013) https://doi.org/10.1109/SI.2013.6766664

6. Milakar, U., Potočnik, B: Automated facial expression recognition based on histograms of oriented gradient feature vectors. Signal, Image Video Process. 9(1), 245–253 (2015) https://doi.org/10.1007/s11760-015-0810-4

7. Zhang, L., Tjondronegoro, D: Facial expression recognition using facial movement features. IEEE Trans. Affective Comput. 2(4), 219–229 (2011) https://doi.org/10.1109/TAFFC.2011.13

8. Dongchong, S., Jieqing, J.: The method of facial expression recognition based on DWT-PCA/LDA. In: 3rd International Congress on Image and Signal Processing, vol. 4, pp. 1970–1974. IEEE, Piscataway, NJ (2010) https://doi.org/10.1109/CISP.2010.5648166

9. Wang, X., Guo, R., Kambhamettu, C.: Deeply-learned feature for age estimation. In: Winter Conference on Applications of Computer Vision, pp. 534–541. IEEE, Piscataway, NJ (2015) https://doi.org/10.1109/WACV.2015.77

10. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 34–42. IEEE, Piscataway, NJ (2015) https://doi.org/10.1109/IIWAT.2018.8369721

11. Wang, C., et al.: Rapid face detection algorithm of color images under complex background. In: International Symposium on Neural Networks. pp. 356–363. Springer, (2011) https://doi.org/10.1007/978-3-642-21090-7_42

12. Ramanan, D., Zhu, X.: Face detection, pose estimation, and landmark localization in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE, Piscataway, NJ (2012) https://doi.org/10.1109/CVPR.2012.6480104

13. Wang, S.H., et al.: Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm. Neurocomputing 272, 668–676 (2018) https://doi.org/10.1016/j.neucom.2017.08.015

14. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: 6th International Conference on Image and Video Retrieval, pp. 401–408. ACM, New York (2007) https://doi.org/10.1145/1282280.1282340

15. Siddiqi, M.H., et al.: Human facial expression recognition using wavelet transform and hidden Markov model. In: International Workshop on Ambient Assisted Living, pp. 112–119. Springer, New York (2013) https://doi.org/10.1007/978-3-642-21090-7_42

16. Lucey, P., et al.: The Extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion specified expression. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101. IEEE, Piscataway, NJ (2010) https://doi.org/10.1109/CVPRW.2010.5543262

17. Lyons, M., et al.: Coding facial expressions with Gabor wavelets. In: 3rd International Conference on Automatic Face and Gesture Recognition (FG), pp. 200–205. IEEE, Piscataway, NJ (1998) https://doi.org/10.1109/AFGR.1998.670940

18. Nason, G.P., Silverman, B.W.: The stationary wavelet transform and some statistical applications. In: Wavelets and Statistics, pp. 281–299. Springer, Berlin (1995) https://doi.org/10.1007/978-1-4612-2544-7_17

19. Zhang, Y., et al.: Feature extraction of brain MRI by stationary wavelet transform. In: International Conference on Biomedical Engineering and Computer Science, pp. 1–4. IEEE, Piscataway, NJ (2010) https://doi.org/10.1109/S2013839010003652

20. Misiti, M., et al.: Wavelet Toolbox User’s Guide. The MathWorks, Inc. (1996)
43. Zhang, L., et al.: Intelligent facial emotion recognition using moth-firefly optimization. Knowledge-Based Syst. 111, 248–267 (2016) https://doi.org/10.1016/j.knosys.2016.08.018

44. Guo, M., et al.: Facial expression recognition using ELBP based on covariance matrix transform in KLT. Multimed. Tools Appl. 76(2), 2995–3010 (2017) https://doi.org/10.1007/s11042-016-3282-9

How to cite this article: Kar NB, Nayak DR, Babu KS, Zhang Y-D. A hybrid feature descriptor with Jaya optimised least squares SVM for facial expression recognition. IET Image Process. 2021;15:1471–1483. https://doi.org/10.1049/ipr.2.12118