IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2021, accepted August 13, 2021, date of publication August 26, 2021, date of current version September 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108029

Facial Expression Recognition Using Dynamic Local Ternary Patterns With Kernel Extreme Learning Machine Classifier

SUMEET SAURAV, RAVI SAINI, (Member, IEEE), AND SANJAY SINGH, (Senior Member, IEEE)
Academy of Scientific and Innovative Research, Ghaziabad, Uttar Pradesh 201002, India
CSIR-Central Electronics Engineering Research Institute, Pilani, Rajasthan 333031, India

Corresponding author: Sumeet Saurav (sumeet@ceeri.res.in)

ABSTRACT Rapid growth in advanced human-computer interaction (HCI) based applications has led to the immense popularity of facial expression recognition (FER) research among computer vision and pattern recognition researchers. Lately, a robust texture descriptor named Dynamic Local Ternary Pattern (DLTP) developed for face liveness detection has proved to be very useful in preserving facial texture information. The findings motivated us to investigate DLTP in more detail and examine its usefulness in the FER task. To this end, a FER pipeline is developed, which uses a sequence of steps to detect possible facial expressions in a given input image. Given an input image, the pipeline first locates and registers faces in it. In the next step, using an image enhancement operator, the FER pipeline enhances the facial images. Afterward, from the enhanced images, facial features are extracted using the DLTP descriptor. Subsequently, the pipeline reduces dimensions of the high-dimensional DLTP features via Principal Component Analysis (PCA). Finally, using the multi-class Kernel Extreme Learning Machine (K-ELM) classifier, the proposed FER scheme classifies the features into facial expressions. Extensive experiments performed on four in-the-lab and one in-the-wild FER datasets confirmed the superiority of the method. Besides, the cross-dataset experiments performed on different combinations of the FER datasets revealed its robustness. Comparison results with several state-of-the-art FER methods demonstrate the usefulness of the proposed FER scheme. The pipeline with a recognition accuracy of 99.76%, 99.72%, 93.98%, 96.71%, and 78.75%, respectively, on the CK+, RaF, KDEF, JAFFE, and RAF-DB datasets, outperformed the previous state-of-the-art.

INDEX TERMS Facial expression recognition, dynamic local ternary pattern, principal component analysis, kernel extreme learning machine, cross-dataset, cross-validation.

I. INTRODUCTION

Recently, there has been enormous advancement in assistive technology for industrial, commercial, automobile, and societal applications. Most of these applications require a robust and accurate system for automatic FER to improve the users' adaptability. A system for automatic FER provides crucial clues that reveal a person's actual intention and state of mind. A FER technology embedded robot can execute home services like talking to children and taking care of the elderly [1]. Furthermore, a FER system integrated with the Advanced Driver Assistance System (ADAS) can identify drivers' fatigue levels. These systems can produce a warning alarm when the fatigue level exceeds a pre-defined threshold limit [2]. Production industries can utilize FER technology to determine the worth of consumer products before their actual launch [3]. Besides, such technology can assist recruiters in identifying the hidden emotional state of candidates [4]. In hospitals, a system for automatic FER can assist in remotely monitor the health status of patients [5]. Finally, as demonstrated by Ashwin and Guddeti [6], a FER system integrated with online teaching platforms and classrooms can improve the overall quality of teaching. Therefore, the past decade saw tremendous advancement in FER research to design an efficient and robust system for FER in real-world conditions.

Techniques developed for emotion recognition over the years are categorized based on input modalities as vision-based methods [7], speech-based methods [8], and hybrid methods that use a combination of audio and visual

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Anisetti.

signals [9], [10]. Each of these sensors has its limitations and benefits. However, in general, emotion recognition techniques based on a hybrid of input modalities perform better than methods based on a single input signal. The existing methods for vision-based FER have used RGB image sensor [11] and depth image sensor [12]. The RGB cameras, though very common and cheap, pixel-intensities in the images captured using these sensors rapidly change due to variations in illumination. Meanwhile, the depth sensors capture depth information and are robust against changes in illumination. Also, in contrast to the RGB images, depth images solve privacy issues by hiding persons' identification information.

Moreover, based on the learning scheme, the existing FER techniques are classified as traditional machine learning-based methods and deep learning-based methods. The machine learning-based approach for FER uses a combination of handcrafted feature extractor and the standard machine learning classifiers such as the Support Vector Machine (SVM) [13], Support Vector Neural Network (SVNN) [14], K-Nearest Neighbor (K-NN) [15], Sequential Minimal Optimization (SMO) [16], Classification Trees (CT) [17], Multi-layer Perceptron (MLP) [18], Neural Network (NN) [19], etc. In contrast to the methods based on traditional machine learning, the FER methods based on deep learning techniques are end-to-end trainable. These techniques use convolutional neural networks (CNNs) to extract features automatically and classify facial expressions in static images [20].

Although deep learning-based FER techniques have achieved state-of-the-art results, the traditional machine learning-based approach has also shown substantial performance [21]. It has been due to the absence of large-scale FER datasets. In such situations, the traditional machine learning-based approach sometimes surpasses the deep learning-based techniques. Deep learning techniques are data-driven techniques, and their performance is directly proportional to the amount of data. Nevertheless, one major limitation of the traditional machine learning-based approach for FER is designing an efficient handcrafted feature extractor that requires high skill and expertise. Therefore, current FER research in this domain aims at developing efficient handcrafted feature extractors that can efficiently separate different facial expressions [14].

Designing an efficient FER system for real-world applications is not a trivial task due to several limitations induced due to variations in illumination and facial expressions, partial face occlusion, real-time performance, etc. Therefore, in the last few years, several techniques were developed for real-time and robust FER [22], [23]. But, despite the enormous progress, these systems still have not achieved the desired level of recognition accuracy and computational efficiency. Real-time recognition of facial expressions in complex real-world conditions remains an unsolved problem.

This work examines the effectiveness of Dynamic Local Ternary Pattern (DLTP) in the FER task and introduces a robust and efficient FER pipeline. The proposed pipeline uses different image pre-processing techniques to enhance the facial images before feature extraction using the DLTP descriptor and its uniform variant, named uniform DLTP (uDLTP). In the intermediate step, the pipeline utilizes dimensionality reduction using PCA to reduce the dimensions of the features. Finally, the pipeline classifies the reduced facial features using the K-ELM classifier. Extensive experiments were conducted in single and cross-dataset scenarios on five FER benchmark datasets (CK+, RaF, KDEF, JAFFE, and RAF-DB) using well-known evaluation metrics, namely the recognition accuracy, precision, recall, and F1-score, to validates the performance of the pipeline. In summary, the main contribution of the proposed work are as follows:

- The use of the DLTP descriptor overcomes the manual determination of threshold in the traditional LTP descriptor. The threshold in DLTP is automatically determined using local neighborhood pixel intensities.
- The proposed FER pipeline employs several image pre-processing techniques to enhance the facial images before feature extraction. Such a scheme improves the discriminative power of the descriptor.
- The use of dimensionality reduction via PCA helps to improve the accuracy and computational efficiency of the FER pipeline. PCA reduces the dimensions of the DLTP features without the loss of vital facial information.
- Deployment of the K-ELM classifier improves the recognition accuracy with reduced classification time than the existing classifiers. The K-ELM classifier has not been utilized much in the FER task.
- Extensive experiments are conducted on five benchmark FER datasets using the cross-validation and cross-dataset testing procedures to access the performance of the proposed FER pipeline.

The remaining paper is structured as follows: Section II introduces the existing state-of-the-art techniques for FER. Details of the proposed FER pipeline and its constituent units are provided in Sect. III. Description of the experimental setup and FER datasets makes the content of Sect. IV. Section V provides the experimental analysis results on the FER datasets with related discussions, followed by details on the computation time provided in Sect. VI. Finally, the paper is closed with conclusive remarks given in Sect. VII.

II. RELATED WORK

Based on the feature extraction scheme, the methods for static image-based FER can be classified broadly into appearance-based methods, geometric-based methods, and methods that use hybrid appearance and geometrical features [24]. The former methods can be further sub-classified as the facial texture-based methods, facial shape-based methods, and methods that use the fusion of facial texture and shape features [21]. This section briefly introduces the existing

methods for FER based on traditional machine learning using appearance features.

The FER techniques based on facial texture information employ prominent texture descriptors such as the Local Binary Pattern (LBP) [11], Local Ternary Pattern (LTP) [25], Local Derivative Pattern (LDP) [26], Local Directional Texture Pattern (LDTP) [27], Gradient Local Ternary Pattern (GLTP) [28], Improved Gradient Local Ternary Pattern (IGLTP) [29], Improved Adaptive Local Ternary Pattern (IALTP) [7], and so on. Shan *et al.* [11], in their seminal work, conducted a detailed study to analyze the effectiveness of the LBP and the Boosted-LBP descriptor in the FER task. Although the LBP operator is a powerful and computationally efficient feature descriptor, there is degradation in its performance due to random noise and non-monotonic variation in illumination. Guo *et al.* [30], proposed a robust and efficient facial descriptor named K-ELBP for expression recognition in static facial images. The K-ELBP operator uses an extended variant of the uniform LBP to extract the feature matrix of the facial images. Subsequently, the method use covariance matrix transform in K-L transform (KLT) to reduce the dimensions of the features. The Local Directional Pattern (LDP) developed by Jabid *et al.* [26] has used directional edge response values in contrast to the grey-level intensity values used in LBP. While the LDP descriptor performed better than the LBP; however, like LBP, it also failed to extract vital information from the uniform and near-uniform facial regions.

To mitigate the issues of LBP and related descriptors, Tan and Triggs [25] proposed the Local Ternary Pattern (LTP) descriptor for the texture analysis task. Later on, to utilize the benefits of the Sobel edge detector and the LTP operator, Ahmed and Hossain [28] proposed a new feature descriptor called the Gradient LTP (GLTP) for the FER task. The GLTP operator, instead of directly extracting features from the grayscale facial images, uses Sobel convolved facial images. Holder and Tapamo [29] introduced an improved variant of the GLTP descriptor, named Improved Gradient Local Ternary Pattern (IGLTP). In other work, Saurav *et al.* [7] proposed an improved variant of the Adaptive Local Ternary Pattern (ALTP), named Improved ALTP (IALTP). The ALTP descriptor, originally proposed for face recognition, combines Webers' law and LTP operator to extract enhanced features from the facial images. Inspired by Weber's law, Chen *et al.* [31] proposed the Weber Local Descriptor (WLD) for the FER application. WLD, unlike LBP, is robust against noise and illumination variation. Alhussein [5], proposed a multi-scale variant of WLD called MS-WLD to increase the discriminative power of the descriptor. The MS-WLD, in contrast to the original WLD, extracts finer details from the facial images and thus results in better performance. In other work, Khan *et al.* [32] combined WLD with LBP & Discrete Cosine Transform (DCT) and proposed a novel Weber Local Binary Image Cosine Transform (WLBI-CT) descriptor for the FER task.

The FER technique presented by Mahmood *et al.* [33] has suggested combining the texture and orientation features extracted from the salient facial regions. The dual feature fusion scheme helped the FER pipeline alleviate the adverse effect of noise, illumination, and partial face occlusions. In [34], the authors proposed a novel framework for FER that first selects a few prominent facial patches depending on the position of the facial landmarks. These active patches are further processed to obtain salient patches. Finally, features are extracted from the salient patches using the LBP operator. Eventually, the features are classified using the one-versus-one (OVO) multi-class SVM classifier. In the FER scheme presented in [35], the authors introduced a novel Gradient Local Phase Quantization (GLPQ) descriptor for facial feature extraction. Given an input facial image, the GLPQ descriptor first computes the gradient magnitude image using the Sobel operator. In the second step, from the magnitude facial images divided into multiple regions, local features are extracted using the LPQ descriptor and concatenated to obtain the global facial descriptor. Ryu *et al.* [27], introduced another powerful face descriptor called the Local Directional Ternary Pattern (LDTP). The LDTP descriptor combines LDP and LTP and operates on gradient angle facial images.

In their recent work, Revina and Emmanuel [14] proposed an efficient system for FER that fuses facial features extracted using the Scale-Invariant Feature Transform (SIFT) and a newly proposed Scatter Local Directional Pattern (SLDP) descriptor. The facial features were classified using a new classifier called Whale-Grasshopper Optimization Algorithm based Multi-Support Vector Neural Network (W-GOA-based MultiSVNN). Kar *et al.* [36] proposed an efficient system for automatic FER in static images. Their designed system classifies facial images in three stages. In the first stage, from the input facial images, the framework extract features using the ripplet transform type II (ripplet-II) feature extractor. In the next stage, utilizing the hybrid of PCA and Linear Discriminant Analysis (LDA), the facial feature dimension is reduced to obtain a compact and efficient facial descriptor. In the final stage, the FER pipeline classifies the facial features using the least-square variant of the SVM (LS-SVM) classifier with a radial basis function (RBF) kernel. In [37], the authors proposed a technique for FER that use an improved variant of the Completed Local Ternary Patterns (CLTP) [38] descriptor. Similar to IGLTP [29], the proposed feature extractor has also used the Scharr operator to calculate gradient magnitude image. From the gradient facial images, features are extracted using the CLTP operator and classified using a combination of K-NN and a Sparse Representation Classifier (SRC). The FER system proposed by Revina and Emmanuel [39] has used a novel noise reduction method named the Decision Based Rule-Oriented Median Filter (DBROMF) and a new facial descriptor called Multi-Directional Triangles Pattern (MDTP). Facial features extracted using MDTP are classified

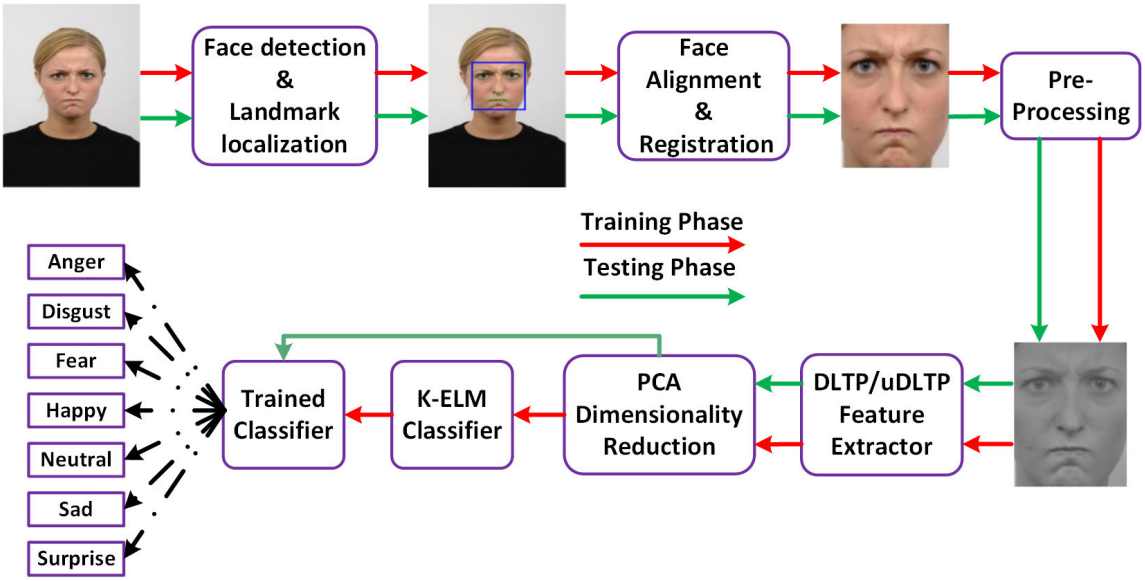

FIGURE 1. Algorithmic pipeline of the proposed facial expression recognition system.

using the SVNN classifier into one of the seven facial expressions.

Several techniques developed for the classification of facial expressions have also employed feature selection techniques. By reducing the dimensions of the features, these techniques achieve the fast classification of expressions besides improving their classification accuracy. Ghosh *et al.* [16] proposed a new feature selection (FS) algorithm based on Late Hill Climbing and Memetic Algorithm (MA) (LHCMA). The LHCMA FS algorithm achieved superior performance than the popular FS algorithms when tested with several facial descriptors, namely the LBP, Histogram of Oriented Gradients (HOG), etc. In other related work, Saha *et al.* [40] introduced the supervised filter harmony search algorithm (SFHSA) for FS in the FER task. The SFHSA algorithm use cosine similarity to remove similar features from feature vectors and minimal-redundancy maximal-relevance (mRMR) to determine the feasibility of the optimal feature subsets using Pearson's correlation coefficient (PCC). Their designed SFHSA algorithm, when tested with five state-of-the-art feature descriptors using the RaF and JAFFE datasets, achieved a notable improvement in the recognition accuracy. The FER technique introduced by Shanthi and Nickolas [41] has fused facial features extracted using the LBP and Local Neighborhood Encoded Pattern (LNEP). The chi-square statistical analysis is used to select the most relevant features from the original high-dimensional feature vectors and is classified using the SVM classifier. Siddiqi *et al.* [42] introduced a system for FER that uses the wavelet transform for feature extraction, a new robust step-wise linear discriminant analysis (SWLDA) feature selection algorithm, and a hidden Markov model (HMM) classifier. Given the facial images, their designed FER system first detects faces using a novel unsupervised technique based on the active contour (AC) model. The FER pipeline proposed by Kumar and Rajagopal [43] has used normalized minimal feature vectors and semi-supervised Twin Support Vector Machine (TWSVM) learning. Li and Wen [44], proposed a sample awareness-based personalized (SAP) FER method that uses the Bayesian learning method to select the optimal classifier from the global perspective and then used the selected classifier to identify the emotional class of each test sample. The authors in [45] proposed a novel sparse modified Marginal Fisher analysis (SMMFA) for the FER task. SMMFA efficiently reduces the dimension of the facial features and thus helps in extracting discriminant features for FER. In another work, Li *et al.* [46] proposed a novel FER scheme that uses a dynamic ensemble pruning method called graph-based dynamic ensemble pruning (GDEP) for the recognition of facial expression in static facial images.

Since their inception, the deep learning techniques have proved their efficacy in solving several computer vision problems like image classification, object detection, speech recognition, etc. Therefore, in the last few years, many works have been proposed for FER in static images using deep learning [20], [47]–[56]. The deep learning algorithms are data-dependent algorithms, whose performance linearly increases with the amount of the dataset. Therefore, on small-scale FER datasets like the CK+, RaF, JAFFE, KDEF, etc., the traditional machine-learning-based FER methods outperform the deep-learning-based FER methods. On the contrary, on large-scale FER datasets, like RAF-DB [57], FER2013 [58], and AffectNet [59], the deep learning CNN models have outperformed the traditional machine-learning-based FER methods [60].

III. PROPOSED METHODOLOGY

Figure 1 shows the block diagram of the proposed FER pipeline. The pipeline consists of six units, namely the face detection & landmark localization, face alignment & registra-

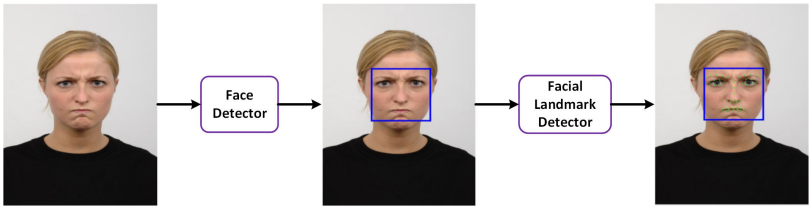

FIGURE 2. Sequence of steps used for face detection & landmark localization.

tion, image enhancement, feature extraction, dimensionality reduction, and classification. Given an input image, the face detection & landmark localization unit detects possible faces and corresponding facial landmarks. In the subsequent step, the face alignment & registration unit register the facial images and crops and scales them to a standard resolution. Afterward, the images are enhanced using an image enhancement operation. Features are subsequently extracted from these enhanced images using the DLTP descriptor. In the next step, the extracted high-dimensional features are passed through the PCA algorithm to reduce their dimension. Finally, the reduced facial features are classified into expression labels using the K-ELM classifier. The following section presents further details of all the constituent units of the pipeline.

A. FACE DETECTION & LANDMARK LOCALIZATION

The face detection & landmark localization unit, as shown in Figure 2 takes an input image and returns the location of all possible faces and their corresponding landmarks. Among the available face detectors, the Viola & Jones frontal face detector [61] has been the popular choice in the FER task. Therefore, the proposed FER pipeline has also used the face detector to detect faces in a static input image. However, instead of using the cascade classifier available in the OpenCV library, the proposed pipeline has employed a more robust and efficient cascade classifier trained on the multi-block LBP (MB-LBP) features [62].

The face coordinates, once available, are passed to the facial landmark detector. The detector, in turn, marks the locations of 68 facial landmarks on the detected facial images. The proposed FER pipeline has used Intraface [63], one of the most widely employed landmark detectors. The Intraface detector uses the Supervised Descent Method (SDM) proposed by Xiong and De la Torre [64] to locate and track the facial landmarks in an input image.

B. FACE ALIGNMENT & REGISTRATION

Once the face and facial landmarks information are available, it is utilized for the face alignment and registration task, as demonstrated in Figure 3. At first, the face alignment & registration unit uses the landmark information of both the left and right eyes to compute their center position and inter-ocular distance (D), and angle between the two eye centers. In the subsequent step, using the angle and the inter-

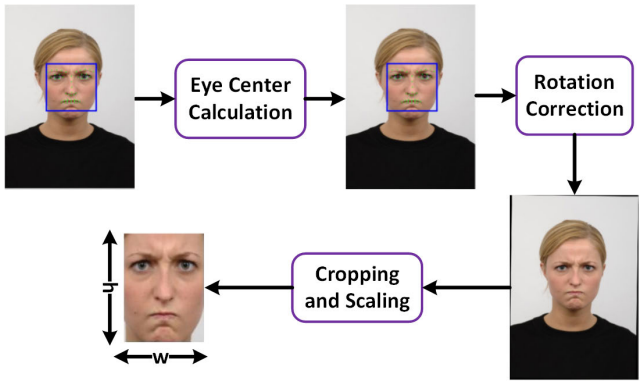

FIGURE 3. Sequence of steps used for face alignment & registration.

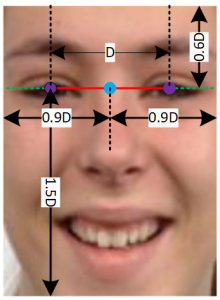

FIGURE 4. Systematic representation of step used for face cropping on a sample facial image from the RaF dataset.

ocular distance, the input facial image is affine transformed such that it gets horizontally aligned. Afterward, from the transformed image, the unit crops the facial area using a pre-defined value obtained as a multiplicative factor of the inter-ocular distance D and calculated from its mid-point, as shown in Figure 4. Finally, the unit crops the face and scales it to a standard size of h × w pixels (empirically determined). The face cropping scheme removes all redundant regions of the face and retains only the relevant area [29]. Additionally, this step ensures spatial consistency of facial parts (nose, eyes, mouth, etc.) and thus delivers enhanced accuracy [65].

C. IMAGE PRE-PROCESSING TECHNIQUES

Atmospheric exposure to digital images makes them ineffective for image processing applications [66], and an intermediate image enhancement step is deemed crucial before further processing. Therefore, the proposed FER scheme

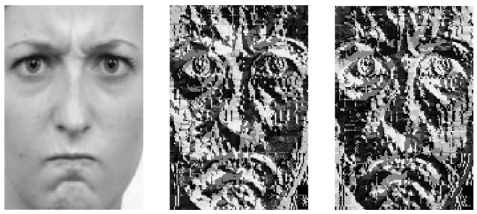

FIGURE 5. DLTP encoded facial images (left to right): Original image, DLTP encoded lower image, and DLTP encoded upper image.

FIGURE 6. Results of Gamma correction image pre-processing operation (left to right): Original image, Gamma corrected image, DLTP encoded lower image, and DLTP encoded upper image.

FIGURE 7. Results of local contrast normalization image pre-processing operation (left to right): Original image, LCN corrected image, DLTP encoded lower image, and DLTP encoded upper image.

has utilized contrast enhancement techniques to enhance the facial images before feature extraction. Depending on their mode of operation, the existing techniques for contrast enhancement are categorized broadly into global, local, or hybrid methods [67]. Global contrast enhancement techniques transform pixel intensities of facial images using a single transformation function. Although these techniques work well in cases where the image is either too dark or too bright, they fail to handle images that require selective enhancement. In such situations, the global techniques may create over or under enhancement problems at some parts of the image [67]. Therefore, to resolve these issues, local contrast enhancement techniques were developed that use the neighboring pixels' information during transformation. Figure 5 shows the DLTP encoded facial images without performing any pre-processing. These images can be used as a reference to visually compare the results obtained after different image enhancement operations.

1) GAMMA CORRECTION (GC)

Gamma correction (GC) is a classic image pre-processing technique employed to enhance the contrast of digital images. By increasing their dynamic range, the technique improves the contrast of images that are either too dark or too bright [68]. Still, GC performs global transformation without considering the local context; it fails in situations where both dark and bright regions are present in the image. Figure 6 shows the DLTP encoded images extracted from the gamma-corrected facial image using (1). In (1), I_{out} and I_{in} are the output and input image intensities, respectively. The variable γ in (1) controls the shape of the transformation function, and its optimal value is determined experimentally.

$$I_{out} = I_{in}^{\gamma} \tag{1}$$

2) LOCAL CONTRAST NORMALIZATION (LCN)

The local contrast normalization (LCN), a local intensity normalization algorithm, is inspired by the computational neuroscience model and has been utilized in several FER works [69]. Mathematically expressed in (2), intensity normalization using LCN requires subtractive and divisive local contrast normalization. Subtractive LCN, as the name implies, subtracts each image pixels from the Gaussian-weighted average (μ) of its neighbors. In contrast, the divisive LCN operation divides image pixels by the standard deviation (σ) of its neighborhood. The neighborhood for both the procedures uses kernels of different sizes. Figure 7 shows DLTP encoded facial images obtained after LCN operation.

$$x_{out} = \frac{x_{in} - \mu}{max(\tau, threshold)} \tag{2}$$

In (2), τ = mean(σ) and threshold = 1e-4. The output pixel value obtained using (2) is min-max normalized with the minimum value α set equal to 0.3 and maximum value β set to 0.7. Finally, the normalized pixel values are multiplied by 255 to get the final enhanced pixel output.

3) GLOBAL CONTRAST NORMALIZATION (GCN)

The global contrast normalization (GCN), as the name indicates, performs contrast normalization by taking complete image pixels' intensity into account [70]. Similar to LCN, GCN computation also proceeds in two steps. The first step of the GCN operation subtracts each pixel from its mean pixel value. In comparison, the second step of the GCN operation divides the mean subtracted pixels by their standard deviation. But, division by standard deviation amplifies sensor noise, and thus to overcome this, LeCun *et al.* [71] introduced a positive regularization parameter λ to add bias to the estimate of standard deviation. Also, to avoid computation errors, a small constant ε is added, as illustrated in (3). Figure 8 shows the GCN pre-processed image and the corresponding DLTP encoded lower and upper facial images.

$$x_{out} = s\frac{x_{in} - x_{mean}}{max(\epsilon, \sqrt{\lambda + (x_{in} - x_{mean})^2}} \tag{3}$$

In (3), the values of the parameters s, λ, and ϵ are set equal to 1, 100, and 1e-4, respectively. Like in LCN, the output pixel value obtained after GCN operation is min-max normalized

FIGURE 8. Results of global contrast normalization image pre-processing operation (left to right): Original image, GCN enhanced image, DLTP encoded lower image, and DLTP encoded upper image.

with the minimum value α set equal to 0.3, and maximum value β set equal to 0.7. Eventually, the normalized pixel values are multiplied by 255 to get the final enhanced pixel output.

D. FEATURE EXTRACTION

For facial feature extraction, the proposed FER pipeline has used the Dynamic Local Ternary Pattern (DLTP) descriptor [72]. In contrast to the popular LTP descriptor, the DLTP descriptor uses an automatic mechanism to determine threshold τ based on Webers' law. Additionally, the descriptor dynamically updates the threshold depending on the pixel intensity values. Webers' law states that the change of a stimulus (e.g., lighting or sound) that will be just noticeable is a constant ratio of the original signal. The form of Webers' law used in DLTP is expressed by (4).

$$\frac{\Delta I}{I} = \tau \tag{4}$$

In (4), ΔI denotes change in intensity I, and τ signifies the proportion that remains constant. In DLTP, ΔI is generalized as $|I_n - I_c|$ when I is considered as I_c and I_n (n=1, 2..., 8) is the neighboring pixel. Thus, the form of Weber's law used for the determination of threshold automatically can be mathematically expressed, as in (5).

$$\frac{|I_n - I_c|}{I_c} = \tau \tag{5}$$

Figure 9 demonstrates pattern encoding scheme using the DLTP descriptor. The threshold τ determined automatically (using (5)) for every neighboring pixel (see Figure 9(b)) is applied around the center pixel value I_c of 3 × 3 neighboring pixels I_n (n=1, 2 . . . , 8). Neighbor pixels that falls in between $I_c + \tau$ and I_c- τ are quantized to 0, while those below I_c - τ to -1 and the remaining above $I_c + \tau$ to 1 using (6). In (6), S_{DLTP} denotes the quantized value of the surrounding neighbors and is shown in Figure 9(c). Similar to LTP, in DLTP also, the generated quantized value is further divided into negative patterns (Figure 9(d)) and positive patterns (Figure 9(e)). The resulting negative and positive binary patterns are then multiplied with fixed weights (see Figure 9(f) and Figure 9(g)) and are summed up to give DLTP encoded lower and upper decimal values, as shown in Figure 9(h) and Figure 9(i), respectively. Mathematical formulation used for the conversion of upper and lower DLTP coded values to

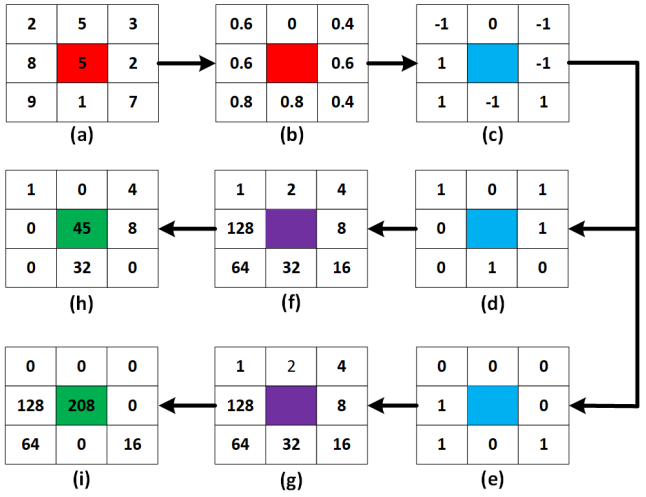

FIGURE 9. Dynamic local ternary pattern calculation (a) a 3 × 3 pixels window (b) automatic dynamic threshold calculation (c) calculation of the sign patterns based on the generated thresholds (d)-(e) lower and upper binary patterns (f)-(g) fixed weight for lower and upper binary pattern multiplication, and (h)-(i) lower and upper encoded decimal value.

positive (upper) P_{DLTP} and negative (lower) N_{DLTP} decimal values are expressed in (7) and (8), respectively.

$$S_{DLTP}(I_c, I_n) = \begin{cases} -1, & \text{if } I_n < I_c - \tau \\ 0, & \text{if } I_c - \tau \le I_n \le I_c + \tau \\ +1, & \text{if } I_n > I_c + \tau \end{cases} \tag{6}$$

$$P_{DLTP} = \sum_{i=0}^{7} S_P(S_{DLTP}(i)) \times 2^i \tag{7}$$

where,

$$S_P(v) = \begin{cases} 1, & \text{if } v > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$N_{DLTP} = \sum_{i=0}^{7} S_N(S_{DLTP}(i)) \times 2^i \tag{8}$$

where,

$$S_N(v) = \begin{cases} 1, & \text{if } v < 0 \\ 0, & \text{otherwise} \end{cases}$$

Figure 10 illustrates the procedure used to capture the textural information from a sample facial image using the DLTP descriptor. Given an input facial image, the procedure extracts the DLTP encoded positive P_{DLTP} and negative N_{DLTP} images by executing the sequence of steps, as demonstrated in Figures 10(a)-(m). Once extracted, the feature extraction scheme divides these images into multiple m × n regions (see Figures 10(n)-(o)). Afterward, the scheme concatenates the local facial features in the form of histograms computed from each of these cells, as shown in Figures 10(p)-(q). Finally, the scheme concatenates the DLTP extracted positive and negative histograms to obtain the final high-dimensional

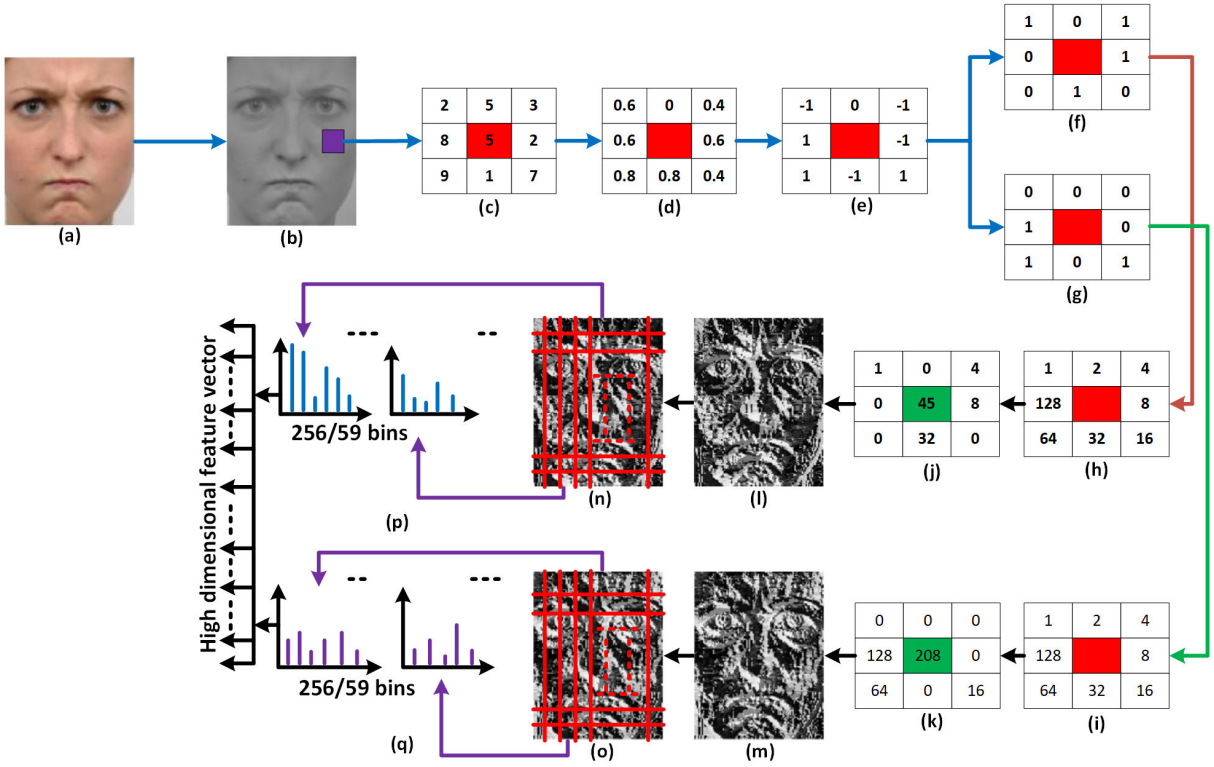

FIGURE 10. Systematic representation of feature extraction scheme using DLTP descriptor.

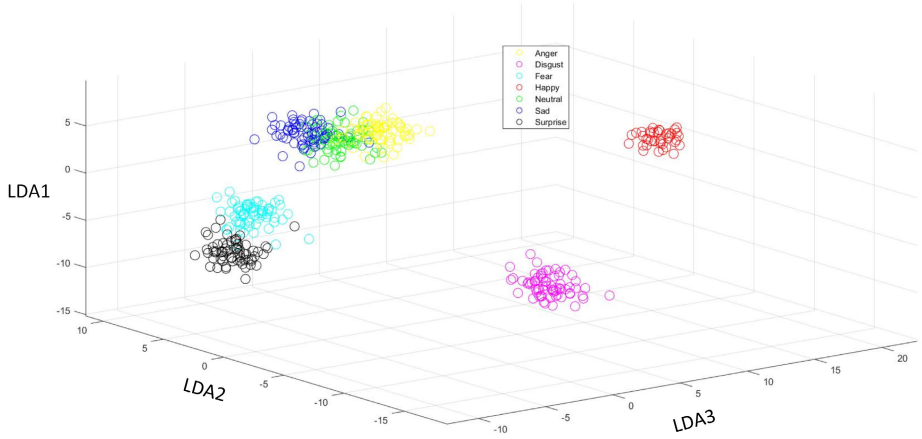

FIGURE 11. Scatter plot showing DLTP features of different facial expressions in lower dimensional space.

facial feature. Similar to uniform LBP, histograms corresponding to the uniform variant of the DLTP descriptor has 59-bins. Positive $H_{P_{DLTP}}$ and negative $H_{N_{DLTP}}$ histograms for both DLTP and uDLTP are computed for each facial image region using (9) and (10), respectively.

$$H_{P_{DLTP}}(\tau) = \sum_{r=1}^{m}\sum_{c=1}^{n} f\left(P_{DLTP}(r,c),\tau\right) \tag{9}$$

$$H_{N_{DLTP}}(\tau) = \sum_{r=1}^{m}\sum_{c=1}^{n} f\left(N_{DLTP}(r,c),\tau\right) \tag{10}$$

where,

$$f(a,\tau) = \begin{cases} 1, & \text{if } a = \tau \\ 0, & \text{otherwise} \end{cases}$$

In (9)-(10), m and n denotes the width and height of the DLTP and uDLTP encoded facial image region, respectively. The value of τ ranges from 0-58 and 0-255 in the case of uDLTP and DLTP, respectively.

The DLTP extracted facial features are high-dimensional, and a major fraction of these features are redundant. These high-dimensional features hamper the performance of the

classifier and increase its computational cost. Therefore, this work has utilized dimensionality reduction using PCA to reduce the dimensions of the features. Figure 11 shows the scatter plot of the high-dimensional DLTP features obtained from the RaF category-2 7-expression dataset. The PCA reduced DLTP features are projected to three-dimensional feature space using LDA for visualization. A closer look at the scatter plot reveals that the DLTP features corresponding to fear, disgust, happiness, and sadness expressions are separable. Still, for the rest of the three classes, namely anger, sadness, and neutral, some overlaps exist in the reduced feature space. It may be due to some extent of similarities among the facial expression images belonging to anger, sadness, and neutral.

E. PRINCIPAL COMPONENT ANALYSIS (PCA)

In machine learning applications, it is often desirable to reduce the number of input features. High-dimensional features dramatically impact the performance of machine learning classifiers. Technically, in the machine learning community, the problem is referred to as the curse of dimensionality. It states that having a large feature vector may not always be useful [73]. Therefore, over the years, several dimensionality reduction techniques were developed [74]. These techniques aim to reduce the feature space of the high-dimensional feature vectors without any adverse impact on the classifiers' performance.

Principal component analysis (PCA) is one of the most widely used dimensionality reduction techniques [75]. It comes under the category of unsupervised machine learning techniques and strives to find the PCA space. For input high-dimensional data, the PCA space consists of orthonormal and uncorrelated principal components (PCs). The PCs indicate the direction of maximum variance, and their optimal number is a hyperparameter that is determined experimentally.

The high-dimensional features obtained from the DLTP and uDLTP descriptors contain a lot of redundant information. Feature vectors containing too many features slow down the classification process. It also leads to degradation in the classification accuracy of the classifier. Therefore, dimensionality reduction via PCA not only reduces the computational and memory budget of the FER system but enhances its recognition accuracy as well. In literature, there are two methods used to compute the principal components (PCs) of data [76]. The first method uses the covariance matrix, while the second uses SVD (singular value decomposition). The covariance matrix-based method computes PCs in two steps, in which the first step calculates the covariance matrix of the feature matrix. The second step calculates the eigenvalues and eigenvectors of the covariance matrix, and thus, the PCs. The SVD-based method calculates the PCs of the PCA space using the SVD method [76].

F. KERNEL-EXTREME LEARNING MACHINE (K-ELM) CLASSIFIER

This work has used the Kernel Extreme Learning Machine (K-ELM) classifier for the multi-class classification of facial expressions. The K-ELM classifier is the kernelized variant of the extreme learning machine (ELM) classifier [77]. The naive ELM classifier is a single-layer feed-forward neural network (SLFN). Because of its fast training compared to the traditional back-propagation-based neural networks, the ELM classifier has been used in several existing works related to pattern classification [78]. The ELM and K-ELM classifiers are not iterative and use a simple matrix inversion operation during training to compute the output weights, as discussed by Huang *et al.* [79]. Additionally, in the training phase, the input weights (value of connections between the input layer and the hidden layer) of the ELM classifier are randomly generated and kept fixed.

To understand the working of the K-ELM classifier, one can refer to the internal details of the ELM classifier shown in Figure 12. The first layer of the ELM termed the input layer is connected to the n-dimensional DLTP feature vector $\mathbf{x} \in R^d$. The second layer, named the hidden layer, transforms input features from the original feature space to a higher dimensional feature space. With L hidden neurons, the hidden layer transforms n-dimensional feature vectors into an L-dimensional transformed feature vector. Each hidden neuron receives feature vectors as input, and the necessary computation that takes place inside a neuron indexed by i is given by (11).

$$g(\mathbf{x}; \mathbf{w}_i, b_i) = g(\mathbf{x}.\mathbf{w}_i + b_i) \tag{11}$$

In (11), the function g is called the activation function, $\mathbf{w}_i$ is the input weight vector that reflects the strength of connection between all the input neurons and the i^{th} hidden neuron, the bias of the i^{th} node is denoted by b_i. The value of i ranges from 1 to L (number of hidden neurons). Although there are numerous activation functions, this work has used the sigmoid activation function, and thus, using the sigmoid activation function, operation in (11) is re-written, as in (12).

$$g(\mathbf{x}; \mathbf{w}_i, b_i) = \frac{1}{1 + exp[-(\mathbf{x}.\mathbf{w}_i + b_i)]} \tag{12}$$

The transformed feature vector obtained from all the hidden neurons for a single n-dimensional facial feature vector $\mathbf{x}$ is mathematically expressed, as in (13).

$$h(\mathbf{x}) = [g(\mathbf{x}; \mathbf{w}_1, b_1), \ldots, g(\mathbf{x}; \mathbf{w}_L, b_L)] \tag{13}$$

The third layer of the ELM, named the output layer, contains C neurons equal to the number of facial expressions. Let $\beta_{i,j}$ denote the output weight between the i^{th} hidden node and the j^{th} output node. The mathematical expression involved in the computation of the value of an output node j is given by (14).

$$f_j(\mathbf{x}) = \sum_{i=1}^{L} \beta_{i,j} \times g(\mathbf{x}; \mathbf{w}_i, b_i) \tag{14}$$

Also, in the vectorized form, for a facial image with feature vector $\mathbf{x}$, the output vector obtained from the output node is

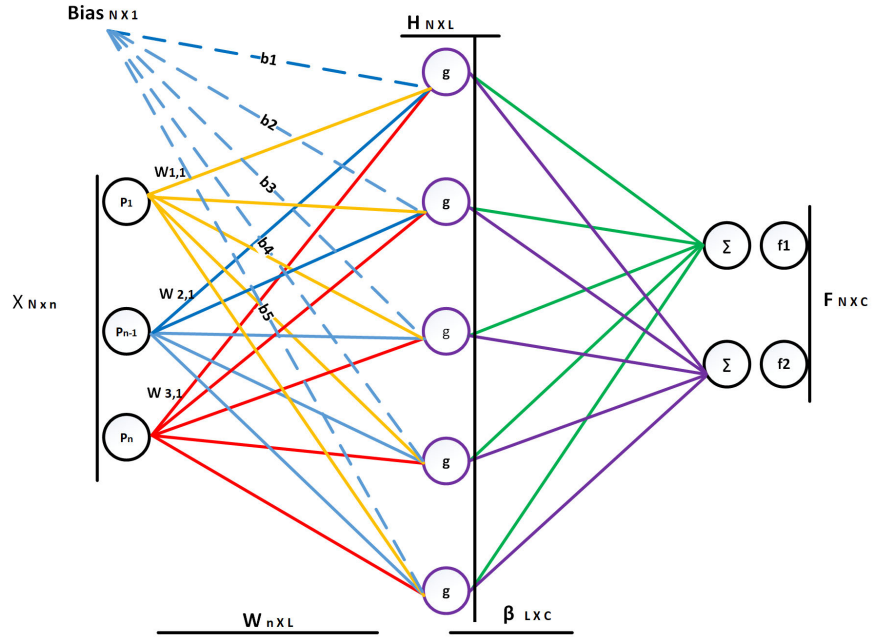

FIGURE 12. Internal details of the extreme learning machine (ELM) classifier.

written as

$$\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_C(\mathbf{x})] = \mathbf{h}(\mathbf{x})\beta \tag{15}$$

where,

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_L \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,C} \\ \beta_{2,1} & \cdots & \beta_{2,C} \\ \vdots & \ddots & \vdots \\ \beta_{L,1} & \cdots & \beta_{L,C} \end{bmatrix}$$

Once the classifier is trained, during the recognition process, for a sample test facial image with feature vector $\mathbf{x}$, its corresponding category is determined as expressed in (16), i.e., the classifier selects the output node that has the maximum magnitude as the output class.

$$label(\mathbf{x}) = arg_{j=1,\ldots,C} max f_j(\mathbf{x}) \tag{16}$$

Training an ELM classifier requires a labeled FER dataset. Let's assume there are N training sample pairs (images with their corresponding labels) in the dataset. At first, for each facial image, feature vector $\mathbf{x}$ is obtained using the DLTP descriptor. Also, labels of all the facial images can form a matrix denoted as $\mathbf{T} = [\mathbf{l}_1, \ldots, \mathbf{l}_N]^T$. Here $\mathbf{l}_1$ represents the one-hot encoded binary label of an input facial image. During training, the ELM classifier only determines the optimal values of the output weight matrix $\boldsymbol{\beta}_{i,j}$ where j=1,...,C, as the input weights and biases $\{\mathbf{w}_i, b_i\}_{i=1,\ldots,L}$ are randomly generated and remains fixed.

Let us assume that there are N training samples in the FER dataset. Also, for each image, $\mathbf{x}_k$ (k=1,2,...,N) denotes the DLTP extracted feature vector. Also, let $\mathbf{y}_k$ denotes the predicted output label and $\mathbf{T} = [\mathbf{l}_1, \ldots, \mathbf{l}_N]^T$ be the one-hot encoded actual facial labels, then in the matrix form, (14) can be written as

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{Y} \tag{17}$$

where,

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x_1}) \\ \vdots \\ \mathbf{h}(\mathbf{x_N}) \end{bmatrix} = \begin{bmatrix} g(\mathbf{x}_1; \mathbf{w}_1, b_1) & \cdots & g(\mathbf{x}_1; \mathbf{w}_L, b_L) \\ \vdots & \ddots & \vdots \\ g(\mathbf{x}_N; \mathbf{w}_1, b_1) & \cdots & g(\mathbf{x}_N; \mathbf{w}_L, b_L) \end{bmatrix}$$

and

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,C} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,C} \end{bmatrix}$$

During training phase, the ELM classifier tries to minimize the training error $\|\mathbf{T} - \mathbf{H}\boldsymbol{\beta}\|^2$ and the norm of output weight $\|\boldsymbol{\beta}\|$. It can be formulated as a constrained-optimization problem [79] and mathematically expressed as

$$\begin{aligned} &\text{minimize: } \psi(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\|\boldsymbol{\xi}\|^2 \\ &\text{subject to: } \mathbf{H}\boldsymbol{\beta} = \mathbf{T} - \boldsymbol{\xi} \end{aligned} \tag{18}$$

The constant C, in (18) is a regularization parameter. It aids in improving the generalization performance of the classifier, and its optimal value is determined experimentally.

To solve the constrained optimization problem expressed in (18), the Lagrange multiplier technique is used [80]. The technique determines the value of output weight $\boldsymbol{\beta}$ based on the nature of the matrix $(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H})$. In case, the matrix

$(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H})$ is not singular, the value of $\boldsymbol{\beta}$ is obtained as

$$\boldsymbol{\beta} = \left(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H}\right)^{-1} \mathbf{H}^T\mathbf{T} \tag{19}$$

In case, the matrix $(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H})$ is singular, the value of $\boldsymbol{\beta}$ is determined as

$$\boldsymbol{\beta} = \mathbf{H}^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T\right)^{-1} \mathbf{T} \tag{20}$$

Upon closely examining (19) - (20), one can find that the matrix dimensions of $(\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H})$ is $L \times L$ while that of $(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)$ is $N \times N$. Therefore, depending on the sample size of the dataset, the solutions in (19)-(20) is used to determine the values of the output weight $\boldsymbol{\beta}$ [79].

The original ELM classifier described above works very well in situations where the type of activation function to be used is known. However, when the feature vectors are not linearly separable, and none of the activation functions work, there comes the role of the K-ELM classifier. Also, to achieve satisfactory results, the classical ELM classifier requires a large number of hidden nodes, which results in higher computational complexity and a longer training time [81]. The K-ELM classifier uses kernels that maps the features into higher dimensional space. Also, the RBF kernels required is much less in K-ELM than the hidden nodes in a conventional ELM classifier. These properties of the K-ELM classifier enhance the recognition accuracy and computational efficiency of the FER system. The performance of the K-ELM classifier is insensitive against the randomness of parameters than the counterpart ELM classifier.

A K-ELM classifier makes use of the kernel technique which states that given two input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the dot product of their mapped features represented by $\mathbf{h}(\mathbf{x_i}) \cdot \mathbf{h}(\mathbf{x_j})$ can be replaced by a kernel function $\phi(\mathbf{x}_i, \mathbf{x}_j)$. It is based on the Mercers' condition and the output vector $\mathbf{f}(\mathbf{x})$ of a K-ELM can be represented as

$$\begin{aligned}\mathbf{f}(\mathbf{x}) &= \mathbf{h}(\mathbf{x})\boldsymbol{\beta} = \mathbf{h}(\mathbf{x})\mathbf{H}^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T\right)^{-1} \mathbf{T} \\ &= \begin{bmatrix} \phi(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ \phi(\mathbf{x}, \mathbf{x}_{N_k}) \end{bmatrix} \left(\frac{\mathbf{I}}{C} + \boldsymbol{\Phi}\right)^{-1} \mathbf{T} \end{aligned} \tag{21}$$

where,

$$\boldsymbol{\Phi} = \mathbf{H}\mathbf{H}^T = \begin{bmatrix} \phi(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \phi(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \phi(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

where, N denotes the number of training samples selected randomly from the training set. This work has used the Gaussian function as the kernel ϕ, which is expressed as

$$\phi(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right) \tag{22}$$

In (22), the parameter σ denotes the spread (i.e., standard deviation) of the Gaussian function. The K-ELM classifier has two hyperparameters, namely the Gaussian kernel spread (kernel parameter) σ and the regularization factor C. In this study, the optimal values of these hyperparameters are determined manually using the grid-search procedure.

FIGURE 13. Sample prototypical facial image from the CK+ 7-expression dataset (left to right): Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

IV. EXPERIMENTAL SETUP

This section discusses the evaluation results of the experiments performed on various FER datasets to determine the optimal values of the hyperparameters in the proposed FER pipeline. These include determining the optimal facial image and the cell size, the right number of principal components (PCs), and the values of the kernel and regularization parameter of the K-ELM classifier. It also provide details of different FER benchmark datasets used in these experiments. These experiments were carried out in the MATLAB 2015a environment running on a laptop with 16GB RAM and Windows 10 operating system.

A. DATASET DETAILS

This section provides details of the five FER datasets, namely the CK+, JAFFE, RaF, KDEF, and RAF-DB, used in the experiments.

1) CK+ DATASET

The first dataset used in experiments is the extended Cohn-Kanade (CK+) dataset [82]. The dataset contains emotion sequences of people from different age range (18 to 30 years), origin (African-American, Asian or Latino), and sex (male and female) displaying eight facial expressions, namely anger, contempt, disgust, fear, happiness, neutral, sad, and surprise. However, the proposed experimental setup considers classifying only seven prototypical facial image samples belonging to anger, disgust, fear, happiness, neutral, sad, and surprise. These images are taken from 309 labeled video sequences in the dataset and belong to 106 subjects. Following the standard procedure in static image FER, from the video sequences of expressions, the custom CK+ dataset uses only the last frame for neutral class, and the three peak frames in the case of the other six facial expressions [29], [56]. Thus, the final CK+ dataset has anger (135), disgust (177), fear (75), happiness (207), neutral (309), sad (84), and surprise (249), resulting in 1236 images. Figure 13 shows sample facial images of seven prototypical expressions from the dataset.

2) JAFFE DATASET

The next dataset used in the experiments is the Japanese Female Facial Expression (JAFFE) dataset [83]. This dataset

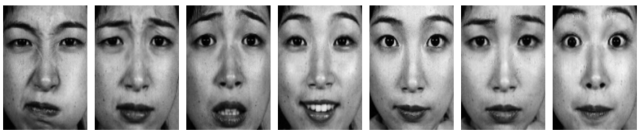

FIGURE 14. Sample prototypical facial images from the JAFFE 7-expression dataset (left to right): Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

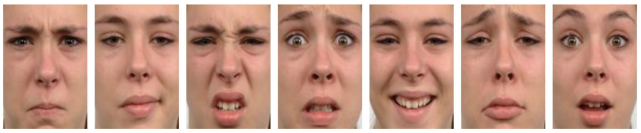

FIGURE 15. Sample prototypical facial images from the RaF category-1 7-expression dataset (left to right): Anger, Contempt, Disgust, Fear, Happy, Sad, and Surprise.

FIGURE 16. Sample prototypical facial images from the RaF category-2 7-expression dataset (left to right): Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

FIGURE 17. Sample prototypical facial images from the RaF category-3 8-expression dataset (left to right): Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

contains facial images of seven facial expressions, namely anger, disgust, fear, happiness, neutral, sad, and surprise. There are 213 images in the dataset created with the participation of ten Japanese female actresses. Figure 14 shows sample facial images from the JAFFE dataset.

3) RaF DATASET

The performance of the proposed FER pipeline is validated on yet another FER dataset. The dataset, named the Radboud Faces (RaF) dataset has facial images of eight facial expressions belonging to 67 subjects [84]. Same as the CK+ dataset, during the RaF dataset preparation, too, the subjects displayed anger, disgust, fear, happiness, contemptuous, sadness, surprise, and neutral facial expressions, looking straight, slightly left, and right. There are 201 images per facial expression for each of the three gaze directions in the dataset. Experiments were performed on the facial images belonging frontal and combined three gaze direction. From the original RaF dataset, four categories of the dataset are prepared and used in the experiments.

The first category, named the RaF category-1 dataset, contains images belonging to anger, contempt, disgust, fear, happiness, sadness, and surprise. There are 469 ($=67 \times 7$) images in the dataset with 67 images each from the seven facial expressions. Sample images from the RaF category-1 dataset is shown in Figure 15.

The second category, named the RaF category-2, also consists of facial images of seven expressions (anger, disgust, fear, happiness, neutral, sad, and surprise) and has a distribution similar to the RaF category-1 dataset. However, the contempt class present in category-1 is replaced by the neutral in this category. Figure 16 displays sample images from this category of the RaF dataset.

The third category of the RaF dataset, named the RaF category-3, contains facial images belonging to all the eight prototypical expressions (anger, contempt, disgust, fear, happy, neutral, sad, and surprise) and has 536 ($=67 \times 8$) images. This category is more challenging as there is more resemblance between the contempt and disgust expressions, which might confuse the classifier and degrades its performance. Figure 17 shows registered facial expression images from this category of the RaF dataset.

The proposed work has utilized another variant of the RaF dataset named the RaF category-4 dataset for fair comparison of the performance of the proposed FER techniques with the deep learning-based FER scheme introduced by Sun *et al.* [48]. The RaF category-4 dataset is an enhanced variant of the RaF category-2 dataset. However, in contrast to category-2, it contains frontal facial images with three gaze directions and has 2-times more images than the RaF category-2 dataset.

4) KDEF DATASET

The fourth in-the-lab dataset used in the experiments is the Karolinska Directed Emotional Face (KDEF) dataset [85]. A total of 70 actors (35 females and 35 males) in the age group of 20 and 30 years, wearing a particular type of gray T-shirt, participated in the dataset creation. Each actor displayed seven emotional expressions while being photographed by cameras placed at five different angular locations. Moreover, during the photo sessions, the female actors did not wear earrings, eyeglasses, and make-up. The male actors, on the other hand, did not have beards and mustaches. The dataset was created in two sessions and contained images from both sessions. This study evaluated the performance of the proposed FER pipeline on the 980 frontal expression images with distribution: anger (140), disgust (140), neutral (140), fear (140), happy (140), sad (140), and surprise (140). Figure 18 shows sample prototypical facial images from the KDEF 7-expression FER dataset.

5) RAF-DB DATASET

The RAF-DB dataset is a real-world FER dataset gathered using image URLs obtained from Flicker [57]. The dataset is very challenging as it has facial images captured under different illumination conditions and has images with partial occlusion. The study used 12,271 images as the training set

FIGURE 18. Sample prototypical facial images from the KDEF 7-expression dataset (left to right): Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

FIGURE 19. Sample prototypical facial images from the RAF-DB 7-expression dataset (left to right): Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

and 3,068 images as test data in the experiments. Facial expression images from the dataset displayed in Figure 19 clearly show that the images in the dataset are from different ethnicities and poses and are thus challenging compared to in-the-lab FER datasets.

B. PARAMETER SELECTION

Designing an efficient FER pipeline involves tuning several hyperparameters. It is essential to find the optimal values of these hyperparameters to achieve better performance. Therefore, the initial experiments were performed to determine the optimal values of different hyperparameters, namely the facial image and cell size, values of the regularization parameter (C) and kernel parameter (γ) of the K-ELM classifier, and the number of principal components (PCs).

1) DETERMINATION OF OPTIMAL FACIAL IMAGE AND CELL SIZE

A series of experiments using eight different combinations of facial image and cell sizes were performed on the RaF category-3 8-expression dataset using both the DLTP and uniform DLTP (uDLTP) descriptor. Intuitively, there can be several combinations of face and cell sizes; however, out of them, these eight combinations maintain a balance between the computational cost and the recognition accuracy. Essentially, the face and cell size determine the dimensions and effectiveness of the feature vectors. From the facial image cells, the designed FER pipeline first extracts the DLTP and uDLTP histogram features. Subsequently, these features are L2-normalized and concatenated to represent the complete facial information. The experiments have used ten rounds of 10-fold cross-validation (CV) and measure the performance in terms of accuracy, precision, and F1-score. The regularization parameter C and kernel parameter (γ) of the K-ELM classifier were fixed to a constant value of 100 and 200, respectively.

Tables 1 and 2 summarizes the analysis result of the experiments conducted using the DLTP and uDLTP features, respectively. Analyzing the results of Table 1, it becomes clear that a high-dimensional feature vector may not always enhance the recognition accuracy of a FER system. On the contrary, an optimal combination of image and cell size often boosts recognition accuracy. Out of eight, the FER pipeline achieved the best performance using a facial image and cell sizes of 156 $\times$ 106 and 14 $\times$ 13, respectively. Using the 45056-dimensional DLTP feature vector on the RaF category-3 dataset, the ten runs of 10-fold CV achieved a mean recognition accuracy of 95.52$\pm$0.38%. Out of the ten, the best 10-fold CV achieved recognition accuracy, precision, recall, and F1-score of 96.10%, 96.14%, 96.08%, and 96.07%, respectively.

Similar to DLTP, the uDLTP variant also achieved optimal performance using an image size of 156 $\times$ 106 and a cell size of 14 $\times$ 13 (see Table 2). On the RaF category-3 dataset using 10384-dimensional uDLTP feature vector, the ten runs of 10-fold CV using K-ELM classifier achieved mean recognition accuracy of 95.07$\pm$0.46% with the best 10-fold CV achieving recognition accuracy, precision, recall, and F1-score of 95.72%, 95.78%, 95.71%, and 95.70%, respectively.

2) DETERMINATION OF K-ELM CLASSIFIERS' PARAMETERS & OPTIMAL IMAGE ENHANCEMENT TECHNIQUE

The next set of experiments aimed at determining the optimal values of the K-ELM parameters, namely the regularization coefficient C and the kernel parameter (γ). A total of eight grid-search experiments were performed corresponding to each of the FER datasets using the different combinations of the two feature extractors (DLTP and uDLTP) and three image enhancement operators, namely the Gamma correction (GC), Local Contrast Normalization (LCN), and Global Contrast Normalization (GCN). These grid-search experiments use the optimal facial image and cell size determined in the previous experiments. The values of C and γ in the grid-search experiments were taken in the range of 1 to 10 in the logarithmic scale of base 2.

Figure 20 shows results of the experiments conducted on the RaF category-3 8-expression dataset. For Gamma correction (GC), the value of Γ used is 3.5 (empirically determined), and the LCN operation has used a Gaussian filter of size 21 $\times$ 21 and 25 $\times$ 25 in its first and the second steps of computation, respectively. Analyzing the results of Figure 20, one can find that on the RaF category-3 8-expression dataset, features extracted using the uDLTP operator from the GC enhanced facial images achieved the optimal performance. And it corresponds to the value of the K-ELM classifier C and γ equal to 1024 and 512, respectively. On the RaF category-3 dataset, the ten runs of the 10-fold CV using the GC-uDLTP feature achieved mean recognition accuracy of 96.55$\pm$0.37%, whereas the best 10-fold CV attained a recognition accuracy of 96.99%.

Figures 21 and 22 show the analysis results of experiments on RaF category-1 and category-2 datasets, respectively. On the RaF category-1 dataset, DLTP features extracted from GCN enhanced facial images achieved optimal performance

TABLE 1. Performance evaluation of DLTP features using different combinations of image and cell sizes on the RaF category-3 8-expression dataset. (Bold: best result.)

Performance Metrics	[128,92] [7,6]	[138,100] [8,7]	[146,106] [9,8]	[152,110] [10,9]	[156,112] [11,10]	[158,112] [12,11]	[158,110] [13,12]	[156,106] [14,13]
Avg. accuracy of 10 runs	86.26	89.79	91.06	92.49	93.23	93.55	93.54	**95.52**
of 10-fold CV (Std-Deviation)	(0.66)	(0.50)	(0.82)	(0.56)	(0.53)	(0.46)	(0.55)	(0.38)
Feature Dimension	138240	121856	106496	92160	78848	66560	55296	45056
Avg. accuracy of best 10-fold	87.31	90.71	92.77	93.70	93.87	94.21	94.26	96.10
Avg. precision of best 10-fold	87.53	90.73	92.88	93.81	93.93	94.36	94.32	96.14
Avg. recall of best 10-fold	87.31	90.67	92.72	93.66	93.84	94.22	94.22	96.08
Avg. F1-Score of best 10-fold	87.09	90.57	92.73	93.63	93.83	94.24	94.22	96.07

TABLE 2. Performance evaluation of uDLTP features using different combinations of image and cell sizes on the RaF category-3 8-expression dataset. (Bold: best result.)

Performance Metrics	[128,92] [7,6]	[138,100] [8,7]	[146,106] [9,8]	[152,110] [10,9]	[156,112] [11,10]	[158,112] [12,11]	[158,110] [13,12]	[156,106] [14,13]
Avg. accuracy of 10 runs	78.65	82.48	85.79	89.39	90.99	92.00	92.86	**95.07**
of 10-fold CV (Std-Deviation)	(0.99)	(0.56)	(0.80)	(0.67)	(0.40)	(0.55)	(0.74)	(0.46)
Feature Dimension	31860	28084	24544	21240	18172	15340	12744	10384
Avg. accuracy of best 10-fold	79.87	83.38	87.15	90.32	91.57	92.77	93.71	95.72
Avg. precision of best 10-fold	80.41	83.42	87.44	90.42	91.68	92.84	93.74	95.78
Avg. recall of best 10-fold	79.85	83.40	87.13	90.30	91.60	92.72	93.66	95.71
Avg. F1-Score of best 10-fold	79.34	83.01	87.28	90.22	91.53	92.71	93.66	95.70

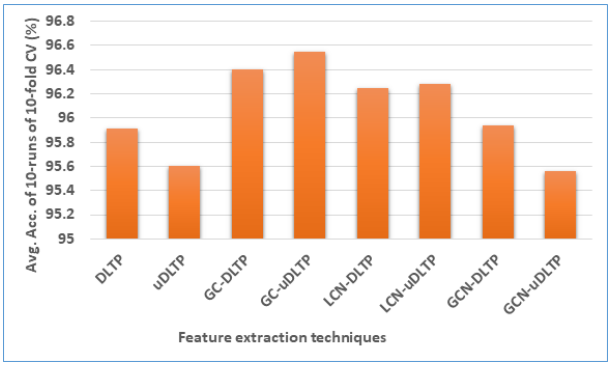

FIGURE 20. Performance of pre-processing operations and feature extractors on the RaF category-3 8-expression dataset.

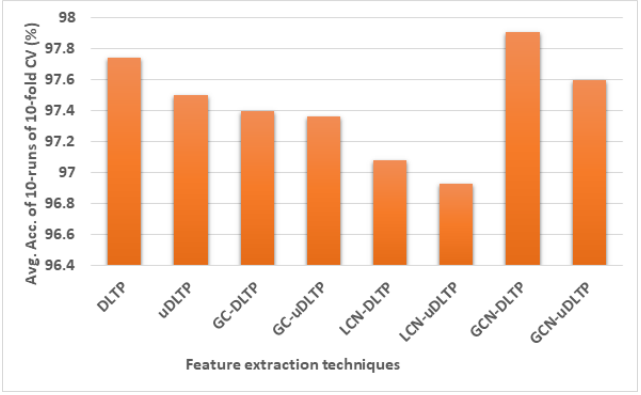

FIGURE 21. Performance of image pre-processing operators and feature extractors on the RaF category-1 7-expression dataset.

using the K-ELM classifier having the value of C and γ equal to 64 and 1024, respectively. Ten runs of the 10-fold CV achieved a mean recognition accuracy of 97.91±0.29%,

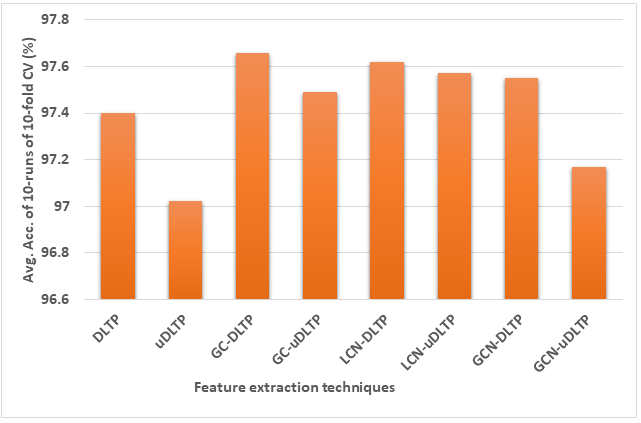

FIGURE 22. Performance of image pre-processing operators and feature extractors on the RaF category-2 7-expression dataset.

while the best 10-fold CV achieved recognition accuracy of 98.50%. Moreover, on the RaF category-2 dataset, as shown in Figure 22, features extracted by the DLTP descriptor using GC enhanced facial images displayed the best performance. It corresponds to the values of C and γ of the K-ELM classifier equal to 1024 and 512, respectively. The mean accuracy of the 10-runs of 10-fold CV recorded on this category of the RaF dataset is 97.66±0.28%, with the best 10-fold CV producing a recognition accuracy of 98.08%.

Grid-search experiments using 10-runs of the 10-fold CV were also conducted on the CK+ dataset to determine the optimal values of parameters of the K-ELM classifier and the suitable image enhancement technique. On the CK+ dataset, the DLTP features obtained from the GC (having Γ value of 3) enhanced facial images achieved the best performance (see Figure 23). The ten runs of 10-fold CV using K-ELM

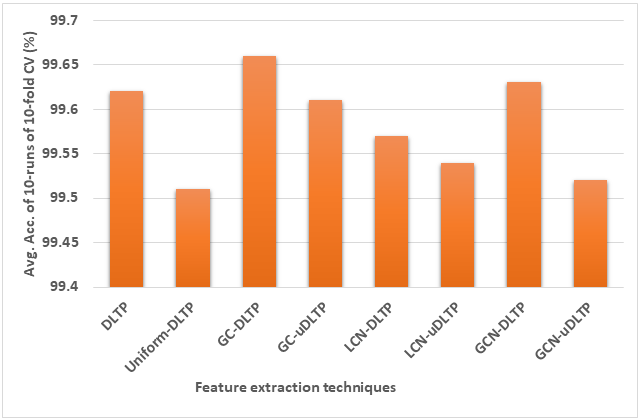

FIGURE 23. Performance of image pre-processing operators and feature extractors on the CK+ 7-expression dataset.

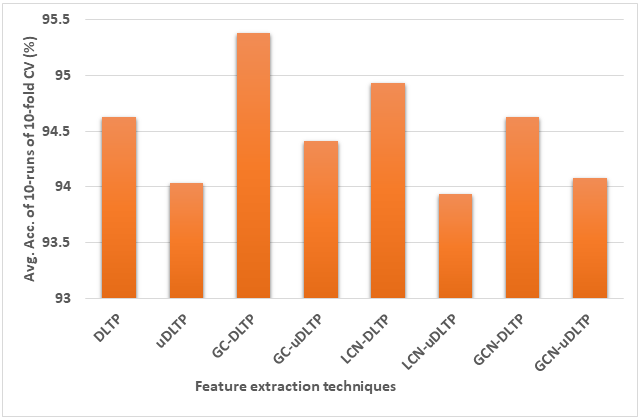

FIGURE 25. Performance of image pre-processing operators and feature extractors on the JAFFE 7-expression dataset.

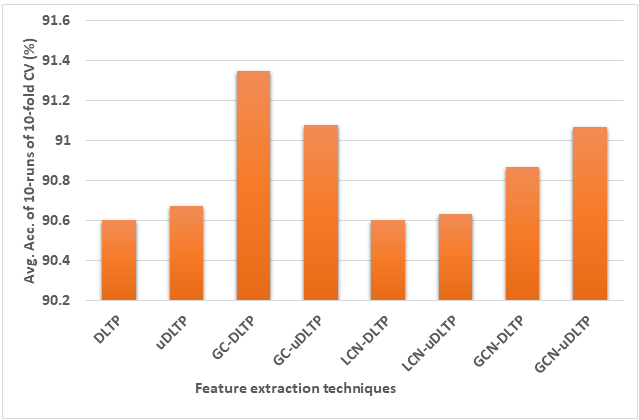

FIGURE 24. Performance of image pre-processing operators and feature extractors on the KDEF 7-expression dataset.

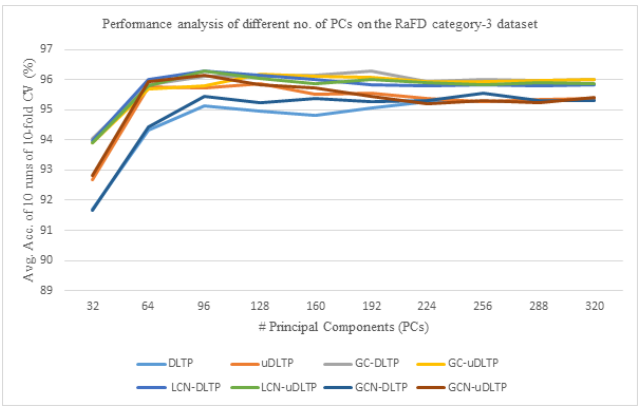

FIGURE 26. Curves showing variations in accuracy with number of PCs on the RaF category-3 8-expression dataset. (Best viewed in color.)

classifier with C and γ values of 512 and 1024, respectively, achieved mean recognition accuracy of 99.66±0.06%. At the same time, on the CK+ dataset, the best 10-fold CV attained recognition accuracy of 99.76%.

Figure 24 displays the analysis results of the experiments conducted on the KDEF 7-expression dataset. Referring to the results of Figure 24, one can find that on the KDEF dataset, too, the DLTP features extracted from the GC enhanced facial images demonstrate the best performance. The optimal value of K-ELM classifier parameters obtained from the grid-search analysis is 1024 and 128 for C and γ, respectively. On the KDEF dataset, the 10-runs of 10-fold CV achieved a mean recognition accuracy of 91.35±0.37%. At the same time, the best 10-fold CV attained recognition accuracy of 91.84%.

The final set of experiments are performed on the JAFFE dataset to find the optimal values of the hyperparameters of the K-ELM classifier. Figure 25 presents the analysis results of different feature extraction schemes plotted against the mean recognition accuracy of the 10-runs of the 10-fold CV. A closer look at Figure 25 shows that the combination of GC image pre-processing operation and the DLTP feature extraction scheme achieved the best performance on the JAFFE dataset. The optimal values of the K-ELM parameters that resulted in the best performance are equal to 128 for C and 256 for γ. Also, on this dataset, the 10-runs of the 10-fold CV achieved mean recognition accuracy of 95.38± 1.00%. Further, the best 10-fold CV achieved recognition accuracy of 96.62% on the JAFFE dataset.

3) DETERMINATION OF PRINCIPAL COMPONENTS (PCs)

Dimensionality reduction experiments using PCA were also performed on the FER datasets to determine the optimal number of principal components (PCs). These experiments use features extracted using eight different combinations of the feature extractor and image enhancement techniques, namely the DLTP, uDLTP, GC-DLTP, GC-uDLTP, GCN-DLTP, GCN-uDLTP, LCN-DLTP, and LCN-uDLTP. Further, these experiments utilize the optimal values of the K-ELM classifier parameters determined in the earlier experiments. In all these experiments, the values of PCs were varied from 32 to 320 at an equal interval of 32.

On the RaF category-3 8-expression dataset, as shown in Figure 26, the PCA-reduced GC-DLTP features with

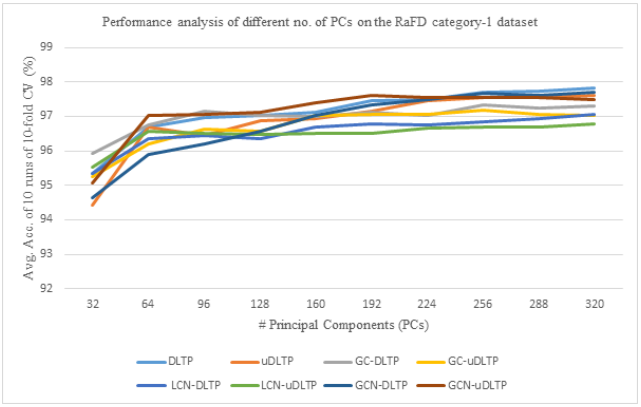

FIGURE 27. Curves showing variations in accuracy with number of PCs on the RaF category-1 7-expression dataset. (Best viewed in color.)

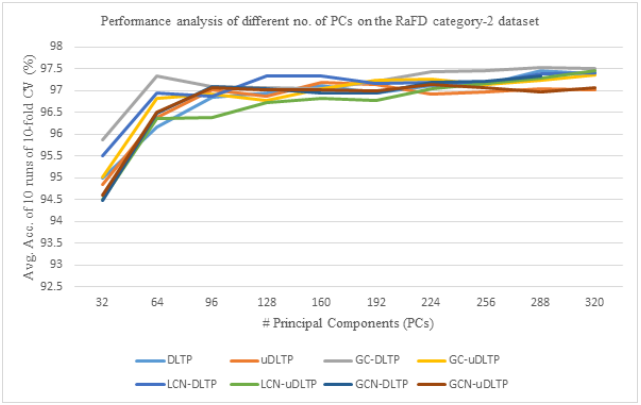

FIGURE 28. Curves showing variations in accuracy with number of PCs on the RaF category-2 7-expression dataset. (Best viewed in color.)

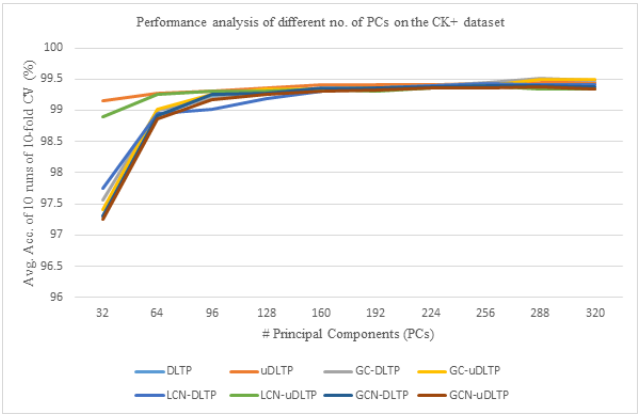

FIGURE 29. Curves showing variations in accuracy with number of PCs on the CK+ 7-expression dataset. (Best viewed in color.)

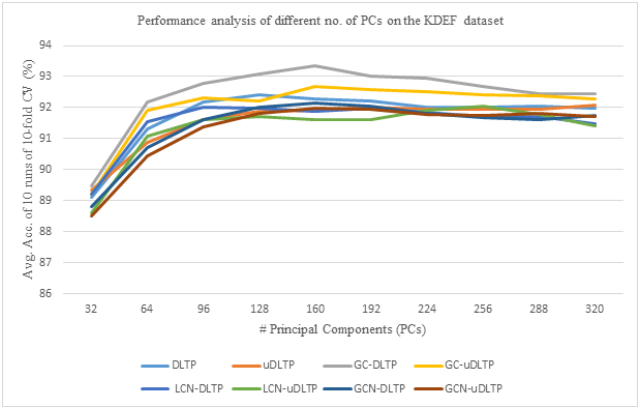

FIGURE 30. Curves showing variations in accuracy with number of PCs on the KDEF 7-expression dataset. (Best viewed in color.)

192 PCs achieved the best mean recognition accuracy of 96.29±0.49%. Among the ten runs, the best 10-fold CV attained recognition accuracy of 97.01%. Figure 27 shows the analysis results of experiments on the RaF category-1 dataset. On the RaF category-1 dataset, with 320 PCs, the DLTP feature extraction scheme without any image pre-processing technique achieved maximum performance. Also, the ten runs of 10-fold CV achieved mean recognition accuracy of 97.81±0.19%, and the corresponding best 10-fold CV attained recognition accuracy of 98.08%. Finally, on the RaF category-2 7-expression dataset, as shown in Figure 28, the GC-DLTP features achieved superior performance with just 288 PCs. The 10-runs of 10-fold CV, on this variant of the RaF dataset, attained a mean recognition accuracy of 97.53±0.27%, and the best 10-fold CV achieved recognition accuracy of 97.89%.

Upon closely examining the analysis results of Figure 29, one can find that on the CK+ dataset, with just 288 PCs, the PCA-reduced GC-DLTP features using 10-runs of the 10-fold CV achieved maximum mean recognition accuracy of 99.52±0.16%. At the same time, the best 10-fold CV on the CK+ dataset achieved average recognition accuracy of 99.76%. Figure 30 shows the analysis results of dimensionality reduction experiments performed on the KDEF dataset. These experiments have utilized the features obtained using different combinations of image enhancement and feature extraction techniques and the values of the K-ELM classifier parameters determined previously. Like the CK+ dataset, on the KDEF dataset, too, the PCA-reduced GC-DLTP features with 160 PCs achieved the best performance. The mean recognition accuracy of 10-runs of 10-fold CV on the KDEF dataset is 93.34±0.53%. The best 10-fold CV, on the other hand, achieved recognition accuracy of 93.98%.

The final set of dimensionality reduction experiments is conducted on the JAFFE dataset to determine the optimal number of PCs. A closer look at the performance evaluation results shown in Figure 31, one can find that on the JAFFE dataset, with just 128 PCs, the GC-DLTP features achieved maximum performance. On this dataset, the ten runs of 10-fold CV achieved mean recognition accuracy of 94.78±1.5%, and the corresponding best 10-fold achieved recognition accuracy of 96.71%.

V. RESULTS AND DISCUSSION

This section provides detailed evaluation results of the proposed FER pipeline on the CK+, JAFFE, KDEF, RaF, and RAF-DB datasets. It also outlines the details of the evaluation

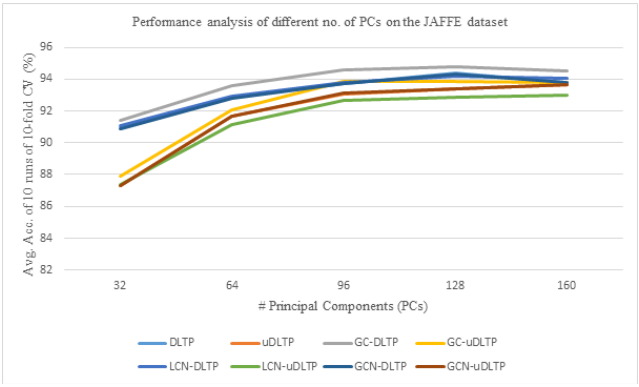

FIGURE 31. Curves showing variations in accuracy with number of PCs on the JAFFE 7-expression dataset. (Best viewed in color.)

procedures adopted to evaluate the performance of the proposed pipeline. Finally, it reports comparative analysis results with the existing works on the FER datasets.

A. EVALUATION PROCEDURES

For performance validation, the proposed FER pipeline has used two evaluation procedures viz the cross-validation [29], and cross-dataset [11]. Moreover, for a fair comparison with the existing works, the performance has been evaluated using four metrics (recognition accuracy, precision, recall, and F1-score). The following section provides details of different evaluation procedures and metrics.

1) CROSS-VALIDATION

The pattern recognition tasks usually use K-fold cross-validation (CV) to measure the performance of a classifier in two scenarios: (a) available data is not sufficient, and (b) distribution of the dataset into training and test set is not known. In these scenarios, K-fold CV is performed by randomly dividing the data roughly into K equal parts. For each fold, a classifier is trained on the (K-1) data parts and tested on the remaining. Afterward, the test accuracy obtained on each fold of the 10-fold CV is summed up and divided by K to get the average accuracy. Since the dataset is divided randomly in the 10-fold CV, its multiple runs each time give different average accuracy. Therefore, as suggested by Holder and Tapamo [29], the proposed FER testing protocol has utilized ten runs of 10-fold CV and uses their mean accuracy as the final measure of the performance.

2) CROSS-DATASET EVALUATION

Another evaluation procedure though not widely used but often accompany the K-fold CV is the cross-dataset evaluation procedure. As the name indicates, this evaluation procedure uses one of the FER datasets as a training set and the other one as the testing set. The cross-dataset evaluation procedure, by default, the best performance evaluation procedure, is utilized to access the generalization capability and robustness of the FER systems.

TABLE 3. Confusion matrix on the RaF category-3 8-expression dataset using PCA reduced GC-DLTP features.

	An	Co	Di	Fe	Ha	Ne	Sa	Su	Recall
An	67	0	0	0	0	0	0	0	100.00
Co	0	65	0	0	0	2	0	0	97.01
Di	0	0	67	0	0	0	0	0	100.00
Fe	0	0	0	60	0	1	1	5	89.55
Ha	0	1	0	0	66	0	0	0	98.51
Ne	0	2	0	0	0	64	1	0	95.52
Sa	1	0	0	1	0	0	65	0	97.01
Su	0	0	0	0	0	1	0	66	98.51
Precision	98.53	95.59	100.00	98.36	100.00	94.12	97.01	92.96	
F1-Score	99.26	96.30	100.00	93.75	99.25	94.81	97.01	95.65	

Avg. performances: accuracy=97.01, recall=97.01, precision=97.07, F1-score=97.00 (*An* Anger, *Co* Contempt,*Di* Disgust, *Fe* Fear, *Ha* Happy, *Ne* Neutral, *Sa* Sad, *Su* Surprise).

3) DETAILS OF THE PERFORMANCE METRICS

In the classification task, the recognition accuracy alone is not enough to measure the robustness of a classifier. Therefore, over the years, researchers developed several metrics to evaluate the robustness of the classifier. As suggested by Carcagni *et al.* [65], this study has also adopted various performance metrics viz precision, recall, and F1-score, other than recognition accuracy, to evaluate the robustness of the proposed FER pipeline. These metrics are briefly discussed below for the sake of completeness.

Recognition accuracy is usually employed as the initial measure of the performance and is calculated by dividing the number of correct by the total number of predictions. The mathematical formula used for the calculation of the recognition accuracy, is expressed as

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{23}$$

where, FN, TP, TN, and FP denotes the number of false negatives, true positives, true negatives, and false positives in the prediction results, respectively.

Precision is another metric used to evaluate the robustness of the classifier and indicates its exactness. Its low value indicates a large number of false positives (FP) in the prediction. The precision, as expressed in (24), is defined as the ratio of correct predictions (TP) to all predictions (sum of TP & FP).

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

The recall is another widely used evaluation metric and obtained by dividing the number of correctly classified samples by the total number of data samples. The mathematical formulation used to compute recall is expressed, as in (25).

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

In contrast to precision, recall is the indicator of classifiers' completeness, and its low value indicates many false negatives in the prediction. It led to the development of yet another metric called the F1-score. The F1-score measures the balance between precision and recall and is obtained by (26).

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \tag{26}$$

TABLE 4. Confusion matrix on the RaF category-1 7-expression dataset using GCN-DLTP features.

	An	Co	Di	Fe	Ha	Sa	Su	Recall
An	67	0	0	0	0	0	0	100.00
Co	0	67	0	0	0	0	0	100.00
Di	0	0	67	0	0	0	0	100.00
Fe	0	0	0	64	0	1	2	95.52
Ha	0	1	0	0	66	0	0	98.51
Sa	1	0	0	1	0	65	0	97.01
Su	0	0	0	1	0	0	66	98.51
Precision	98.53	98.53	100.00	96.97	100.00	98.48	97.06	
F1-Score	99.26	99.26	100.00	96.24	99.25	97.74	97.78	

Avg. performances: accuracy=98.51, recall=98.51, precision=98.51 F1-score=98.50 (*An* Anger, *Co* Contempt, *Di* Disgust, *Fe* Fear, *Ha* Happy, *Sa* Sad, *Su* Surprise).

TABLE 5. Confusion matrix on the RaF category-2 7-expression dataset using GC-DLTP features.

	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	67	0	0	0	0	0	0	100.00
Di	0	67	0	0	0	0	0	100.00
Fe	0	0	62	0	1	2	2	92.54
Ha	0	0	0	67	0	0	0	100.00
Ne	0	0	0	0	67	0	0	100.00
Sa	2	0	0	0	0	65	0	97.01
Su	0	0	1	0	1	0	65	97.01
Precision	97.10	100.00	98.41	100.00	97.10	97.01	97.01	
F1-Score	98.53	100.00	95.38	100.00	98.53	97.01	97.01	

Avg. performances: accuracy=98.08, recall=98.08, precision=98.09, F1-score=98.07 (*An* Anger,*Di* Disgust, *Fe* Fear, *Ha* Happy, *Ne* Neutral, *Sa* Sad, *Su* Surprise).

TABLE 6. Confusion matrix on the RaF category-4 7-expression dataset using PCA-reduced DLTP features.

	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	201	0	0	0	0	0	0	100.00
Di	0	201	0	0	0	0	0	100.00
Fe	0	0	199	0	0	0	2	99.00
Ha	0	0	0	201	0	0	0	100.00
Ne	0	0	0	0	201	0	0	100.00
Sa	1	0	1	0	0	199	0	99.00
Su	0	0	0	0	0	0	201	100.00
Precision	99.50	100.00	99.50	100.00	100.00	100.00	99.01	
F1-Score	99.75	100.00	99.25	100.00	100.00	99.50	99.50	

Avg. performances: accuracy=99.72, recall=99.72, precision=99.72, F1-score=99.72 (*An* Anger,*Di* Disgust,*Fe* Fear,*Ha* Happy,*Ne* Neutral, *Sa* Sad,*Su* Surprise).

B. RESULTS ON THE RaF DATASET

On the RaF category-3 8-expression dataset, utilizing the 10-fold CV testing procedure, the proposed FER pipeline using GC-uDLTP features achieved recognition accuracy, recall, precision, and F1-score equal to 97.01%, 97.01%, 97.06%, and 97.01%, respectively (see Figure 20). Also, on this dataset, as shown in Figure 26, the GC-DLTP feature with PCA reduced dimensions achieved 10-fold CV recognition accuracy, recall, precision, and F1-score of 97.01%, 97.01%, 97.07%, and 97.00%, respectively. Therefore, compared to the original GC-uDLTP features, the recognition accuracy of PCA-reduced GC-DLTP features are comparatively similar. However, PCA-reduced DLTP features with 192 PCs are significantly smaller in size than the 10384-dimensional uDLTP features. Upon closely examining the confusion matrix results of Table 3, one can observe that the K-ELM classifier trained on PCA-reduced GC-DLTP features is very efficient in recognizing facial images belonging to anger, disgust, happiness, and surprise. However, it failed to classify a few facial images belonging to contempt, fear, neutral, and sad. The failed cases might be due to some extent of resemblance between the facial expression pairs, which may have deceived the classifier.

Figure 21 shows the performance analysis results of the proposed FER pipeline on the RaF category-1 7-expression dataset. On this dataset, out of different combinations of image enhancement operations and feature extractors, the DLTP features extracted from the GCN pre-processed facial images performed best. The combination achieved 10-fold CV recognition accuracy, recall, precision, and F1-score of 98.51%, 98.51%, 98.51%, and 98.50%, respectively. The results of dimensionality reduction using PCA on the RaF category-1 7-expression dataset shown in Figure 27 illustrate that the PCA-reduced DLTP features performed well compared to other variants. Using just 320 PCs, the 10-fold CV achieved recognition accuracy of 98.08%, recall of 98.08%, precision of 98.10%, and F1-score of 98.07%. Looking at the confusion matrix results of Table 4, one can find that the classifier trained on GCN-DLTP features correctly classified all the facial images belonging to anger, contempt, and disgust classes. However, the pipeline misclassified a few sample facial images from the fear, happiness, sadness, and surprise classes.

As shown in Figure 22, on the RaF category-2 7-expression dataset, the GC-DLTP features achieved the best performance. On this dataset, out of the ten, the best 10-fold CV run using GC-DLTP feature and K-ELM classifier achieved recognition accuracy, recall, precision, and F1-score of 98.08%, 98.08%, 98.09%, and 98.07%, respectively. The proposed FER scheme using GC-DLTP + K-ELM correctly classified all the facial images belonging to the anger, disgust, happiness, and neutral facial expressions (see Table 5). However, out of the 67 samples from the fear, sadness, and surprise classes, the classifier misclassified 5, 2, and 2 facial samples, respectively. The analysis report of the experiments conducted on PCA reduced features (see Figure 28) shows that on the RaF category-2 dataset, with just 288 PCs, the ten runs of 10-fold CV achieved mean recognition accuracy of 97.53±0.27%. There is a marginal gap of 0.21% in the accuracy of the classifier trained on the original 45056-dimensional GC-DLTP feature vector and the PCA-reduced GC-DLTP features having 288 PCs.

On the RaF category-4 dataset, the 10-runs of 10-fold CV using DLTP features attained a mean recognition accuracy of 99.43±0.09% using the pre-determined values of the K-ELM regularization parameter (C) and kernel parameter γ values equal to 256 and 1024, respectively. The uDLTP variant, on the other hand, attained a mean recognition accuracy of 99.45±0.08% using the K-ELM regularization parameter (C) and kernel parameter γ value equal to

TABLE 7. Comparison results with state-of-the-art FER methods on the RaF dataset. (Bold: best result.)

Reference	Technique	Feature dimensions	Accuracy	RaF Database
2015 [65]	HOG + SVM	–	94.90	RaF category-1
2015 [65]	HOG + SVM	–	92.90	RaF category-3
2019 [48]	Deep learning	–	99.17	RaF category-4
Proposed	GCN-DLTP + K-ELM	45056	98.51	RaF category-1
Proposed	DLTP + PCA + K-ELM	320	98.08	RaF category-1
Proposed	GC-DLTP + K-ELM	45056	98.08	RaF category-2
Proposed	GC-DLTP + PCA + K-ELM	288	97.87	RaF category-2
Proposed	GC-uDLTP + K-ELM	45056	97.01	RaF category-3
Proposed	GC-DLTP + PCA + K-ELM	192	97.01	RaF category-3
Proposed	uDLTP + K-ELM	10384	99.64	RaF category-4
Proposed	DLTP + PCA + K-ELM	288	**99.72**	RaF category-4

TABLE 8. Confusion matrix on the CK+ 7 expression dataset using PCA reduced GC-DLTP features.

	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	135	0	0	0	0	0	0	100.00
Di	0	177	0	0	0	0	0	100.00
Fe	0	0	75	0	0	0	0	100.00
Ha	0	0	0	207	0	0	0	100.00
Ne	0	0	0	1	308	0	0	99.68
Sa	0	0	0	0	0	84	0	100.00
Su	0	0	0	0	2	0	247	99.20
Precision	100.00	100.00	100.00	99.52	99.35	100.00	100.00	
F1-Score	100.00	100.00	100.00	99.76	99.52	100.00	99.60	

Avg. performances: accuracy=99.76, recall=99.84, precision=99.84, F1-score=99.84 (*An* Anger, *Di* Disgust, *Fe* Fear, *Ha* Happy, *Ne* Neutral, *Sa* Sad, *Su* Surprise).

32 and 256, respectively. The best 10-fold CV using uDLTP achieved recognition accuracy, recall, precision, and F1-score of 99.64%, 99.57%, 99.58%, and 99.57%, respectively. Furthermore, on the RaF category-4 dataset, the 10-runs of 10-fold CV using PCA-reduced DLTP and uDLTP features with 288 and 160 PCs achieved the mean recognition accuracy of 99.62±0.09% and 99.55±0.01%, respectively. Therefore, on the RaF category-4 dataset, the PCA-reduced DLTP features performed well compared to PCA-reduced uDLTP features. The best 10-fold CV results of PCA-reduced DLTP features achieved recognition accuracy, recall, precision, and F1-score of 99.72% (see Table 6) and the trained classifier correctly classified most of the facial images belong to all the seven expressions.

Table 7 summarizes the comparison results of the proposed FER pipeline on the RaF dataset. As per the standard, the table compares the 10-fold CV recognition accuracy with related state-of-the-art FER methods [48], [65]. The FER technique proposed by Carcagni *et al.* [65] using HOG + SVM has achieved 10-fold CV accuracy of 94.90% and 92.90% on the RaF category-1 and category-3 FER datasets, respectively. Using a similar testing protocol, on the RaF category-1 dataset, the proposed GCN-DLTP + K-ELM achieved recognition accuracy of 98.51%. Also, on the RaF category-3 8-expression dataset, the proposed FER pipeline using PCA-reduced GC-DLTP features achieved recognition accuracy of 97.01% and thus surpassed the recognition accuracy (92.90%) achieved by the HOG features by a substantial margin (4.11%). Besides, on the RaF category-4 7-expression, the proposed FER pipeline has surpassed the recognition accuracy of 99.17% reported by Sun *et al.* [48]. On this dataset, the proposed DLTP + PCA + K-ELM obtained recognition accuracy of 99.72%. In summary, on the RaF dataset, the proposed FER method achieved superior performance than the state-of-the-art machine learning and deep learning-based FER methods.

C. RESULTS ON THE CK+ DATASET

The CK+ FER dataset is the most popular dataset utilized to evaluate the performance of the static image-based FER methods. Based on the analysis results displayed in Figure 23, one can find that on the CK+ dataset, the GC-DLTP features performed well compared to other feature extraction techniques. On this dataset, ten runs of the 10-fold CV achieved mean recognition accuracy of 99.66±0.06%. The best 10-fold CV, on the other hand, achieved recognition accuracy, precision, recall, and F1-score of 99.76%, 99.86%, 99.80%, and 99.83%, respectively. Figure 29 shows the performance analysis results obtained by varying the number of principal components (PCs) on the CK+ dataset. With an increase in the number of PCs, the pipeline registers an increment in the mean recognition accuracy. Among the several combinations of the pre-processing and feature extraction scheme, the PCA-reduced GC enhanced DLTP descriptor achieved the best mean recognition accuracy (99.52±0.16%) using 288 PCs. Table 8 shows the confusion matrix corresponding to the best performing 10-fold CV, along with the values of different performance metrics, namely precision, recall, F1-score. On the CK+ dataset, the proposed FER scheme achieved recognition accuracy, precision, recall, and F1-score of 99.76%, 99.84%, 99.84%, and 99.84%, respectively. With just 288 PCs, the proposed FER pipeline successfully classified all the facial images from the anger, disgust, fear, happy, and sad classes. However, the K-ELM classifier trained on the PCA-reduced GC-DLTP descriptor failed to correctly classify few sample images belonging to the surprise (2 out of 249) and neutral (1 out of 309) expression classes.

Table 9 summarizes the comparison results of the proposed FER pipeline on the CK+ dataset. The proposed FER scheme has achieved competitive performance similar to several state-of-the-art traditional machine learning-based FER

TABLE 9. Comparison results with state-of-the-art FER methods on the CK+ 7-expression dataset. (**Bold**: best result.)

Reference	Technique	Feature dimensions	Accuracy	Testing Protocol
2015 [65]	HOG + SVM	–	98.50	10-fold CV
2016 [5]	WLD + SVM	–	98.82	7-fold CV
2017 [86]	Multi-gradient EQP + SVM	–	99.36	10-fold CV
2017 [21]	LBP + HOG + SVM	11,636	98.30	10-fold CV
2017 [29]	GLTP+SVM	–	96.90	10-fold CV
2017 [29]	IGLTP+SVM	256	97.60	10-fold CV
2017 [87]	LTP + HOG + SVM	–	96.06	10-fold CV
2018 [47]	WMDNN	–	97.02	10-fold CV
2018 [50]	CNN Ensemble	–	95.36	10-fold CV
2019 [35]	Gradient LPQ+SVM	–	97.05	10-fold CV
2019 [37]	ICLTP+K-NN+SRC	–	97.80	10-fold CV
2019 [48]	Deep learning technique	–	98.38	10-fold CV
2019 [88]	DAM-CNN	–	95.88	10-fold CV
2020 [89]	Deep Learning technique	–	97.38	10-fold CV
Proposed	GC-DLTP + K-ELM	45056	99.76	10-fold CV
Proposed	GC-DLTP + PCA + K-ELM	288	**99.76**	10-fold CV

TABLE 10. Confusion matrix on the KDEF 7-expression dataset using PCA reduced GC-DLTP features.

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**	Recall
An	133	4	2	0	0	1	0	95.00
Di	3	134	0	0	0	3	0	95.71
Fe	5	1	114	1	3	5	11	81.43
Ha	0	0	1	137	2	0	0	97.86
Ne	0	0	0	0	140	0	0	100.00
Sa	0	2	6	0	2	130	0	92.86
Su	0	0	5	0	2	0	133	95.00
Precision	94.33	95.04	89.06	99.28	93.96	93.53	92.36	
F1-Score	94.66	95.37	85.07	98.56	96.89	93.19	93.66	

Avg. performances: accuracy=93.98, recall=93.98, precision=93.94, F1-score=93.92 (*An* Anger, *Di* Disgust, *Fe* Fear, *Ha* Happy, *Ne* Neutral, *Sa* Sad, *Su* Surprise).

methods [5], [21], [29], [35], [37], [65], [86], [87] and deep-learning-based FER methods [47], [48], [50], [56], [88], [89]. On the CK+ dataset, the proposed FER pipeline employing the person-independent (PI) 10-fold CV setting achieved recognition accuracy of 99.76% using both GC-DLTP and PCA-reduced GC-DLTP features. The previous best recognition accuracy of 99.68% on this dataset has been reported by the LBF-NN method [56]. Also, the CNN introduced by Li *et al.* [89] has achieved a 10-fold CV accuracy of 97.38% on the dataset. In summary, on the CK+ dataset, the proposed FER pipeline using the DLTP descriptor performed well than several state-of-the-art machine-learning and deep-learning methods for FER in static images.

D. RESULTS ON THE KDEF DATASET

A closer look at the performance analysis results of Figure 24, one can find that on the KDEF 7-expression dataset, the DLTP features extracted from GC enhanced facial images attained optimal performance. On the dataset, the 10-runs of 10-fold CV produced a mean recognition of 91.35±0.37%, while the best 10-fold CV obtained recognition accuracy, recall, precision, and F1-score of 91.84%, 91.84%, 91.78%, and 91.69%, respectively.

As shown in Figure 30, on the KDEF 7-expression dataset, the PCA-reduced GC-DLTP features with 160 PCs achieved the best performance. The 10-runs of 10-fold CV achieved mean recognition accuracy of 93.34±0.53%, and the corresponding best 10-fold CV accomplished recognition accuracy, recall, precision, and F1-score of 93.98%, 93.98%, 93.94%, and 93.93%, respectively. Upon closely examining the confusion matrix results of Table 10, one can find that the PCA-reduced GC-DLTP features performed satisfactorily in classifying sample facial images belonging to fear and sadness. Out of 26 misclassified samples from the fear class, 11 got classified to surprise, 5 to sad, 5 to anger, 1 to happy, 3 to neutral, and the final one to disgust. The trained classifier successfully classified more than 95% of samples belonging to anger, disgust, happiness, neutral, and surprise. Overall, in contrast to the CK+ and RaF datasets, the pipeline performed poorly on the KDEF dataset. One possible reason can be the under-exposed facial images in the dataset. Such lightening conditions might have adversely affected the effectiveness of the facial features.

Table 11 compares the performance of the proposed FER pipeline to other state-of-the-art techniques on the KDEF dataset. On this dataset, the state-of-the-art FER method using LTP + HOG + SVM has achieved recognition accuracy of 93.34% [87]. Though the technique achieved good recognition accuracy, the high dimensionality of the fused features might have increased the overall computational cost of the system. Meanwhile, the proposed FER pipeline using GC-DLTP features achieved recognition accuracy of 91.84%, while the PCA-reduced GC-DLTP descriptor with only 128 PCs achieved recognition accuracy 93.98%. Therefore, on the KDEF dataset, the proposed FER pipeline registers a 0.64% improvement in the recognition accuracy with a multi-fold improvement in the execution speed. The PCA-reduced GC-DLTP features with much smaller feature dimensions achieved better performance than the original GC-DLTP features. It indicates the effectiveness of dimensionality reduction via PCA in the proposed FER scheme.

E. RESULTS ON THE JAFFE DATASET

On the JAFFE 7-expression dataset, as shown in Figure 25, the DLTP features extracted from the GC enhanced facial

TABLE 11. Comparison results with state-of-the-art FER methods on the KDEF 7-expression dataset. (Bold: best result.)

Reference	Technique	Feature dimensions	Accuracy	Testing Protocol
2017 [87]	LTP + HOG + SVM	NA	93.34	10-fold CV
Proposed	GC-DLTP + K-ELM	45056	91.84	10-fold CV
Proposed	GC-DLTP + PCA + K-ELM	160	**93.98**	10-fold CV

TABLE 12. Confusion matrix on the JAFFE 7-expression dataset using PCA-reduced GC-DLTP features.

	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	30	0	0	0	0	0	0	100.00
Di	0	28	0	0	0	1	0	96.55
Fe	0	1	30	0	0	0	1	93.75
Ha	0	0	0	31	0	0	0	100.00
Ne	0	0	0	0	30	0	0	100.00
Sa	0	0	0	1	0	30	0	96.77
Su	0	0	2	1	0	0	27	90.00
Precision	100.00	96.55	93.75	93.94	100.00	96.77	96.43	
F1-Score	100.00	96.55	93.75	96.88	100.00	96.77	93.10	

Avg. performances: accuracy=96.71, recall=96.73, precision=96.78, F1-score=96.72 (*An* Anger, *Di* Disgust, *Fe* Fear, *Ha* Happy, *Ne* Neutral, *Sa* Sad, *Su* Surprise).

images achieved the best performance. Ten runs of 10-fold CV achieved a mean recognition accuracy of 95.38±1.00% with the best 10-fold CV registering recognition accuracy of 96.71% along with the precision of 96.74%, recall of 96.74%, and F1-score of 96.73%. Also, analyzing the performance analysis results obtained by varying the number of PCs on the JAFFE 7-expression dataset (see Figure 31), one can find that on this dataset, the PCA-reduced GC-DLTP features also achieved optimal performance. With just 128 PCs, the FER pipeline using PCA-reduced GC-DLTP features achieved mean recognition accuracy of 94.78±1.5% using 10-runs of the 10-fold CV. Table 12 shows the confusion matrix results corresponding to the best-performing 10-fold CV using PCA-reduced GC-DLTP features. The trained K-ELM classifier achieved average recognition accuracy, precision, recall, and F1-score of 96.71%, 96.78%, 96.73%, and 96.72%, respectively. Looking at the table results, it becomes apparent that with only 128 PCs, the pipeline correctly classified all sample facial images from anger, happiness, and neutral classes. However, the classifier got entangled and wrongly classified a few sample images from disgust, fear, sadness, and surprise classes.

Table 13 presents comparative analysis results of the proposed FER pipeline with the related state-of-the-art techniques on the JAFFE 7-expression dataset. Out of the existing methods, the FER scheme using the IALTP descriptor [7] has achieved the highest 10-fold CV accuracy of 97.60%. Nevertheless, the IALTP descriptor requires manual determination of the threshold and thus may not be realistic. The proposed FER scheme using the DLTP descriptor, on the other hand, does not have such constraints and is thus fit for real-world applications. The FER pipeline introduced by Alhussein [5] using MS-WLD + SVM has achieved 7-fold CV accuracy of 97.00%. However, the pipeline is not computationally efficient than the proposed FER pipeline using PCA-reduced GC-DLTP features.

One possible reason for the low accuracy of the DLTP descriptor on the JAFFE dataset can be the non-optimal values of the hyperparameters (optimal size of the cropped face and the facial regions). These hyperparameters were determined on the RaF dataset and kept fixed for the rest. Nevertheless, the proposed FER pipeline using the DLTP descriptor achieved a significant boost in the recognition accuracy compared to the related IGLTP descriptor [29] and other competitive descriptors [30], [32], [35], [37], [47]. Finally, the pipeline also demonstrated competitive performance similar to the state-of-the-art deep learning-based FER techniques [50], [89]. In summary, on the JAFFE 7-expression dataset, the proposed FER pipeline achieved competitive performance compared to other related techniques based on hand-crafted and deep-learned features.

F. RESULTS ON THE RAF-DB DATASET

Further, to test the robustness of the proposed FER pipeline in complex real-world conditions, the pipeline is trained and tested on the RAF-DB dataset. Researchers widely use the RAF-DB dataset to test the robustness of the FER system in real-world conditions of partial face occlusion, illumination variation, etc.

The RAF-DB dataset has been utilized extensively in the FER works based on deep learning techniques or the convolutional neural network (CNN). Only a few works are available in the literature that has utilized the dataset to evaluate the performance of FER methods based on traditional machine learning. On the RAF-DB dataset, the baseline results reported by Li and Deng [57] has achieved recognition accuracy of 72.71%, 74.35%, and 77.28% using LBP + SVM, HOG + SVM, and Gabor + SVM, respectively. However, as reported in Table 14, the proposed FER scheme has performed well and achieved an accuracy of 78.75% and 78.46%, using the DLTP + K-ELM and uDLTP + K-ELM, respectively. Therefore, compared to the traditional texture and shape descriptor, the proposed DLTP descriptor is robust against illumination and partial face occlusion. Hence, it is suitable for FER in complex real-world conditions.

G. CROSS-DATASET PERFORMANCE EVALUATION

A FER pipeline trained on one FER dataset often performs poorly on another dataset. Degradation in the performance might be due to the variation in the feature distribution of emotions across datasets. Therefore, besides cross-validation, the study has used the cross-dataset testing procedure to evaluate the generalization performance of the proposed FER

TABLE 13. Comparison results with state-of-the-art FER methods on the JAFFE 7-expression dataset. (Bold: best result.)

Reference	Technique	Feature dimension	Accuracy (%)	Testing protocol
2016 [5]	WLD + SVM	–	97.00	7-fold CV
2017 [30]	K-ELBP + SVM	–	93.30	Train-test Split
2017 [29]	GLTP + SVM	256	74.40	10-fold CV
2017 [29]	IGLTP + SVM	256	81.70	10-fold CV
2018 [47]	WMDNN	–	92.21	10-fold CV
2018 [32]	LBI-CT	192	94.50	10-fold CV
2018 [50]	CNN Ensemble	–	96.57	10-fold CV
2019 [35]	Gradient LPQ + SVM	–	92.19	10-fold CV
2019 [37]	ICLTP + K-NN + SRC	–	92.10	10-fold CV
2020 [89]	Deep Learning	–	97.18	10-fold CV
2020 [7]	IALTP + K-ELM	–	**97.60**	10-fold CV
Proposed (Ours)	GC-DLTP + K-ELM	45056	96.71	10-fold CV
Proposed (Ours)	GC-DLTP + PCA + K-ELM	128	96.71	10-fold CV

TABLE 14. Comparison results with state-of-the-art FER methods on the RAF-DB 7-expression dataset. (Bold: best result.)

Reference	Technique	Feature dimension	Accuracy (%)
2018 [57]	LBP + SVM	–	72.71
2018 [57]	Gabor + SVM	–	74.35
2018 [57]	HOG + SVM	–	77.28
Proposed (Our)	DLTP + K-ELM	45056	**78.75**
Proposed (Our)	uDLTP + K-ELM	10384	78.46

pipeline. In contrast to cross-validation, the cross-dataset evaluation directly measures the discriminative power of the descriptor [48], [87].

As the name indicates, in the cross-dataset testing procedure, one of the FER datasets is used as training and the other as the test dataset. Table 15 reports the cross-dataset evaluation results conducted on the possible train and test combinations of the four FER datasets. The proposed FER pipeline has achieved competitive test accuracy than the state-of-the-art FER methods. The pipeline trained on the KDEF dataset achieved a classification accuracy of 83.33% on the CK+ test dataset, which is much better than the previously reported test accuracy of 78.85% [87].

Moreover, the proposed FER pipeline employing the DLTP descriptor attained better test accuracy of 86.17% in contrast to the previously reported 75.13% [48], using the RaF category-4 dataset as the training set and CK+ dataset as the test set. Furthermore, the FER technique introduced by Shan *et al.* [11] has achieved test recognition accuracy of 41.30%, using the CK+ dataset as the training and the JAFFE dataset testing set. On a similar dataset configuration, the proposed FER pipeline using uDLTP descriptor has also achieved competitive test recognition accuracy of 42.25%. It demonstrates the usefulness of the DLTP/uDLTP feature extractor and the K-ELM classifier. Upon closely examining the results of Table 15, one can find that the DLTP descriptor, in contrast to the uDLTP descriptor, has performed well on most of the train-test combinations of the FER datasets. It indicates that dimensionality reduction, though it works well on standalone FER datasets, it results in loss of useful information that, in most cases, results in low cross-dataset test accuracy. Nevertheless, trade-offs between accuracy and speed exist between the 45056-dimensional DLTP and 10384-dimensional uDLTP descriptors. Depending on the application, the FER pipeline using uDLTP might be suitable than the pipeline using the DLTP descriptor.

Low cross-dataset test accuracy in the FER task has remained a challenging problem due to the obvious biases caused by diverse subjects and diverse data collection conditions, i.e., the training and testing dataset are not independent and identically distributed. One can adopt the domain adaptation technique that might enhance the cross-dataset test accuracy by learning invariant representations across domains (datasets). Further, training instances from different datasets can be combined for training and evaluating the models during cross-dataset testing.

VI. COMPUTATION TIME

Table 16 shows the execution time (in milliseconds) taken by the feature extractor and the K-ELM classifier (with and without the dimensionality reduction) on the CK+ dataset. The feature extraction time per image and the classification time with and without dimensionality reduction are calculated in a MATLAB 2015a environment running on a laptop with 16GB RAM and Intel i9-8950HK processor running at 2.90 GHz. The analysis results of Table 16 show that the uniform variant of the DLTP (uDLTP) descriptor, as expected, consumes less time in feature extraction than the original DLTP. Furthermore, the classifier attains 25 and 13 times boost in the execution speed using PCA-reduced DLTP and uDLTP features, respectively. Different components of the proposed pipeline have also attained multi-fold improvement in the execution speed than those reported in the literature [41]. The results demonstrate the usefulness of dimensionality reduction via PCA in improving the execution speed of the proposed FER pipeline without much degradation in its recognition accuracy. In summary, the proposed FER pipeline is computationally efficient and suitable for applications that demand real-time classification of facial expressions.

VII. CONCLUSION

This research conducted an extensive study to analyze the role of dynamic local ternary patterns (DLTP) for FER in

TABLE 15. Results of cross-dataset evaluation using different combinations of the FER datasets. (**Bold**: best result.)

Train dataset	Test dataset	Method	Precision (%)	Recall (%)	F1-score (%)	Test Accuracy (%)	C	γ
CK+	RaF	DLTP (Proposed)	91.18	89.98	89.60	**89.98**	512	1024
CK+	RaF	uDLTP (Proposed)	90.91	89.55	89.25	89.55	1024	128
CK+	RaF	Deep learning [48]	–	–	–	86.80	–	–
CK+	KDEF	DLTP (Proposed)	69.14	69.90	67.49	69.90	512	1024
CK+	KDEF	uDLTP (Proposed)	69.81	70.00	67.91	**70.00**	1024	128
CK+	JAFFE	DLTP (Proposed)	49.23	40.84	36.65	40.85	512	1024
CK+	JAFFE	uDLTP (Proposed)	57.93	42.39	40.60	**42.25**	1024	128
CK+	JAFFE	LBP-SVM [11]	–	–	–	41.30	–	–
RaF	CK+	DLTP (Proposed)	85.43	83.41	84.14	**86.17**	256	1024
RaF	CK+	uDLTP (Proposed)	83.65	82.20	82.68	85.36	32	256
RaF	CK+	Deep learning [48]	–	–	–	75.13	–	–
RaF	KDEF	DLTP (Proposed)	73.45	73.57	73.02	73.57	256	1024
RaF	KDEF	uDLTP (Proposed)	73.50	73.88	73.23	**73.88**	32	256
RaF	JAFFE	DLTP (Proposed)	53.55	40.20	37.40	40.38	256	1024
RaF	JAFFE	uDLTP (Proposed)	55.47	41.12	37.90	**41.31**	32	256
KDEF	CK+	DLTP (Proposed)	83.96	76.51	78.11	**83.33**	1024	128
KDEF	CK+	uDLTP (Proposed)	82.85	76.16	77.65	82.12	1024	128
KDEF	CK+	LTP + HOG [87]	–	–	–	78.85	–	–
KDEF	RaF	DLTP (Proposed)	87.24	86.14	85.06	**86.14**	1024	128
KDEF	RaF	uDLTP (Proposed)	86.73	85.71	84.62	85.71	1024	128

TABLE 16. The computation time in milliseconds (ms) for feature extraction and classification on a single facial image (DR: Dimensionality reduction and FS: Feature selection). (**Bold**: best result.)

Descriptor	Feature extraction (per image)	Classification without DR/FS	Classification with DR/FS
LBP [41]	28.9	76.7	34.5
LDP [41]	106.9	81.4	37.2
LTP [41]	51.4	146.5	36.5
LDN [41]	132.9	73.2	33.8
LNEP [41]	22.9	76.7	34.6
DLTP (Ours)	19.36	0.756	0.029
uDLTP (Ours)	**12.29**	**0.193**	**0.014**

static images. An efficient FER pipeline was proposed, which uses a sequence of steps to classify facial expressions. Firstly, using an image preprocessing operation, the FER pipeline enhances the facial images. In the subsequent step, features are extracted from the enhanced facial image using the DLTP descriptor. Afterward, the pipeline reduces the features' dimensions via the principal component analysis (PCA). Finally, the proposed pipeline classified the reduced features using the K-ELM classifier. The study utilized both the cross-validation and cross-database testing procedures to evaluate the performance of the proposed pipeline. The proposed FER pipeline using 10-fold CV achieved a mean recognition accuracy of 99.76%, 99.72%, 93.98%, and 96.71%, on the CK+, RaF, KDEF, and JAFFE datasets, respectively. Also, tested on the validation set of the RAF-DB dataset, the pipeline attained a classification accuracy of 78.75%. The cross-dataset evaluation using different combinations of the FER datasets also reflected the discriminative power of the DLTP descriptor. Comparative analysis with state-of-the-art FER methods showed the usefulness of the proposed FER pipeline using the DLTP descriptor. The designed pipeline is robust and computationally efficient and thus suitable for real-world applications. Future work will develop a more robust gradient DLTP descriptor and design the FER method using the fusion of facial texture and shape features extracted using DLTP and HOG, respectively.

REFERENCES

[1] T.-H.-S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "CNN and LSTM based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019.

[2] M. R. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, no. 12, p. 4270, Dec. 2018.

[3] M. S. Bouzakraoui, A. Sadiq, and N. Enneya, "A customer emotion recognition through facial expression using POEM descriptor and SVM classifier," in *Proc. 2nd Int. Conf. Big Data, Cloud Appl.*, Mar. 2017, pp. 1–6.

[4] Y. S. Su, H. Y. Suen, and K. E. Hung, "Predicting behavioral competencies automatically from facial expressions in real-time video-recorded interviews," *J. Real-Time Image Process.*, vol. 19, pp. 1011–1021, Jan. 2021.

[5] M. Alhussein, "Automatic facial emotion recognition using weber local descriptor for e-Healthcare system," *Cluster Comput.*, vol. 19, no. 1, pp. 99–108, Mar. 2016.

[6] T. Ashwin and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Educ. Inf. Technol.*, vol. 25, pp. 1–29, Mar. 2019.

[7] S. Saurav, S. Singh, R. Saini, and M. Yadav, "Facial expression recognition using improved adaptive local ternary pattern," in *Proc. 3rd Int. Conf. Comput. Vis. Image Process.* Singapore: Springer, 2020, pp. 39–52.

[8] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.

[9] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 975–985, 2019.

[10] S. Oh, J.-Y. Lee, and D. K. Kim, "The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals," *Sensors*, vol. 20, no. 3, p. 866, Feb. 2020.

[11] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[12] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[13] J. Chen, R. Xu, and L. Liu, "Deep peak-neutral difference feature for facial expression recognition," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29871–29887, Nov. 2018.

[14] I. M. Revina and W. R. S. Emmanuel, "Face expression recognition with the optimization based multi-SVNN classifier and the modified LDP features," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 43–55, Jul. 2019.
[15] M. Nazir, Z. Jan, and M. Sajjad, "Facial expression recognition using histogram of oriented gradients based transformed features," *Cluster Comput.*, vol. 21, no. 1, pp. 539–548, Mar. 2018.
[16] M. Ghosh, T. Kundu, D. Ghosh, and R. Sarkar, "Feature selection for facial emotion recognition using late hill-climbing based memetic algorithm," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25753–25779, Sep. 2019.
[17] S. Zhou, G. Feng, and J. Xie, "Facial expression recognition based on classification tree," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2014, pp. 128–135.
[18] H. Boughrara, M. Chtourou, C. B. Amar, and L. Chen, "Facial expression recognition based on a mlp neural network using constructive training algorithm," *Multimedia Tools Appl.*, vol. 75, no. 2, pp. 709–731, 2016.
[19] T. Li, C. Du, T. Naren, Z. Chen, S. Liu, J. Zhou, and X. Xu, "Using feature points and angles between them to recognise facial expression by a neural network approach," *IET Image Process.*, vol. 12, no. 11, pp. 1951–1955, Nov. 2018.
[20] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.
[21] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors*, vol. 17, no. 4, p. 712, Mar. 2017.
[22] K. Bahreini, W. van der Vegt, and W. Westera, "A fuzzy logic approach to reliable real-time recognition of facial emotions," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 18943–18966, Jul. 2019.
[23] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2018.
[24] A. R. Rivera, R. Castillo, and O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2012.
[25] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.
[26] T. Jabid, M. H. Kabir, and O. Chae, "Facial expression recognition using local directional pattern (LDP)," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1605–1608.
[27] B. Ryu, A. R. Rivera, J. Kim, and O. Chae, "Local directional ternary pattern for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 6006–6018, Dec. 2017.
[28] F. Ahmed and E. Hossain, "Automated facial expression recognition using gradient-based ternary texture patterns," *Chin. J. Eng.*, vol. 2013, pp. 1–8, Dec. 2013.
[29] R. P. Holder and J. R. Tapamo, "Improved gradient local ternary patterns for facial expression recognition," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 42, Dec. 2017.
[30] M. Guo, X. Hou, Y. Ma, and X. Wu, "Facial expression recognition using ELBP based on covariance matrix transform in KLT," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2995–3010, 2017.
[31] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, Sep. 2010.
[32] S. A. Khan, A. Hussain, and M. Usman, "Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1133–1165, Jan. 2018.
[33] A. Mahmood, S. Hussain, K. Iqbal, and W. S. Elkilani, "Recognition of facial expressions under varying conditions using dual-feature fusion," *Math. Problems Eng.*, vol. 2019, pp. 1–12, Aug. 2019.
[34] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2014.
[35] S. Kherchaoui and A. Houacine, "Facial expression identification using gradient local phase," *Multimedia Tools Appl.*, vol. 78, no. 12, pp. 16843–16859, Jun. 2019.
[36] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, "Face expression recognition system based on ripplet transform type II and least square SVM," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 4789–4812, 2019.
[37] Y. Luo, X.-Y. Liu, Y. Zhang, X.-F. Chen, and Z. Chen, "Facial expression recognition based on improved completed local ternary patterns," *Optoelectronics Lett.*, vol. 15, no. 3, pp. 224–230, May 2019.
[38] T. H. Rassem and B. E. Khoo, "Completed local ternary pattern for rotation invariant texture classification," *Sci. World J.*, vol. 2014, pp. 1–10, Apr. 2014.
[39] I. M. Revina and W. R. S. Emmanuel, "MDTP: A novel multi-directional triangles pattern for face expression recognition," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 26223–26238, Sep. 2019.
[40] S. Saha, M. Ghosh, S. Ghosh, S. Sen, P. K. Singh, Z. W. Geem, and R. Sarkar, "Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm," *Appl. Sci.*, vol. 10, no. 8, p. 2816, 2020.
[41] P. Shanthi and S. Nickolas, "An efficient automatic facial expression recognition using local neighborhood feature fusion," *Multimedia Tools Appl.*, vol. 80, no. 7, pp. 1–26, 2020.
[42] M. H. Siddiqi, R. Ali, M. Idris, A. M. Khan, E. S. Kim, M. C. Whang, and S. Lee, "Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection," *Multimedia Tools Appl.*, vol. 75, no. 2, pp. 935–959, 2016.
[43] M. P. Kumar and M. K. Rajagopal, "Detecting facial emotions using normalized minimal feature vectors and semi-supervised twin support vector machines classifier," *Int. J. Speech Technol.*, vol. 49, no. 12, pp. 4150–4174, Dec. 2019.
[44] H. Li and G. Wen, "Sample awareness-based personalized facial expression recognition," *Int. J. Speech Technol.*, vol. 49, no. 8, pp. 2956–2969, Aug. 2019.
[45] Z. Wang, L. Zhang, and B. Wang, "Sparse modified marginal Fisher analysis for facial expression recognition," *Int. J. Speech Technol.*, vol. 49, no. 7, pp. 2659–2671, Jul. 2019.
[46] D. Li, G. Wen, X. Li, and X. Cai, "Graph-based dynamic ensemble pruning for facial expression recognition," *Int. J. Speech Technol.*, vol. 49, no. 9, pp. 3188–3206, Sep. 2019.
[47] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.
[48] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognit. Lett.*, vol. 119, pp. 49–61, Mar. 2019.
[49] A. Ullah, J. Wang, M. S. Anwar, U. Ahmad, U. Saeed, and Z. Fei, "Facial expression recognition of nonlinear facial variations using deep locality de-expression residue learning in the wild," *Electronics*, vol. 8, no. 12, p. 1487, Dec. 2019.
[50] A. Sun, Y. Li, Y.-M. Huang, Q. Li, and G. Lu, "Facial expression recognition using optimized active regions," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 33, Dec. 2018.
[51] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocompting*, vol. 355, pp. 82–92, Aug. 2019.
[52] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 1–14, 2019.
[53] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2020.
[54] D. Li, X. Zhao, G. Yuan, Y. Liu, and G. Liu, "Robustness comparison between the capsule network and the convolutional network for facial expression recognition," *Appl. Intell.*, vol. 51, no. 4, pp. 1–10, 2020.
[55] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. neural Inf. Process. Syst.*, 2017, pp. 3856–3866.
[56] I. Gogić, M. Manhart, I. S. Pandžić, and J. Ahlberg, "Fast facial expression recognition using local binary features and shallow neural networks," *Vis. Comput.*, vol. 36, no. 1, pp. 1–16, 2018.
[57] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2018.
[58] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Springer, 2013, pp. 117–124.
[59] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2017.
[60] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020.
[61] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[62] K. Martin, "Efficient metric learning for real-world face recognition," Graz Univ. Technol., Graz, Austria, Tech. Rep., 2013. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.494.9447&rep=rep1&type=pdf
[63] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.
[64] X. Xiong and F. D. la Torre, "Supervised descent method for solving nonlinear least squares problems in computer vision," 2014, *arXiv:1405.0601*. [Online]. Available: http://arxiv.org/abs/1405.0601
[65] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: A comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 645, Dec. 2015.
[66] S.-C. Huang, B.-H. Chen, and W.-J. Wang, "Visibility restoration of single hazy images captured in real-world weather conditions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1814–1824, Oct. 2014.
[67] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, "An adaptive gamma correction for image enhancement," *EURASIP J. Image Video Process.*, vol. 2016, no. 1, pp. 1–13, Dec. 2016.
[68] J. G. G. Salas and J. L. Lisani, "Local color correction," *Image Process. Line*, vol. 1, pp. 260–280, Sep. 2011.
[69] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.
[70] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with preprocessing stage in automatic emotion recognition," *Procedia Comput. Sci.*, vol. 116, pp. 523–529, Jan. 2017.
[71] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[72] S. Parveen, S. Ahmad, N. Abbas, W. Adnan, M. Hanafi, and N. Naeem, "Face liveness detection using dynamic local ternary pattern (DLTP)," *Computers*, vol. 5, no. 2, p. 10, May 2016.
[73] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 13th Annu. ACM Symp. Theory Comput. (STOC)*, 1998, pp. 604–613.
[74] L. V. D. Maaten, E. O. Postma, and H. J. V. D. Herik, "MATLAB toolbox for dimensionality reduction," MICC, Maastricht Univ., Maastricht, The Netherlands, Tech. Rep., 2007.
[75] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 1997, pp. 583–588.
[76] J. Shlens, "A tutorial on principal component analysis," 2014, *arXiv:1404.1100*. [Online]. Available: http://arxiv.org/abs/1404.1100
[77] A. Iosifidis, A. Tefas, and I. Pitas, "On the kernel extreme learning machine classifier," *Pattern Recognit. Lett.*, vol. 54, pp. 11–17, Mar. 2015.
[78] Deepika, S. Vashisth, and S. Saurav, "Histogram of oriented gradients based reduced feature for traffic sign recognition," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 2206–2212.
[79] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2011.
[80] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 920–933, Apr. 2016.
[81] Y. Zeng, X. Xu, D. Shen, Y. Fang, and Z. Xiao, "Traffic sign recognition using kernel extreme learning machines with deep perceptual features," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1647–1653, Jun. 2017.
[82] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.
[83] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese female facial expression (JAFFE) database," in *Proc. 3rd Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 14–16.
[84] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognit. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.
[85] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces (KDEF)," CD ROM from Dept. Clin. Neurosci., Psychol. Sect., Karolinska Institutet, Solna, Sweden, 1998, p. 2, vol. 91, no. 630.
[86] S. A. M. Al-Sumaidaee, M. A. M. Abdullah, R. R. O. Al-Nima, S. S. Dlay, and J. A. Chambers, "Multi-gradient features and elongated quinary pattern encoding for image-based facial expression recognition," *Pattern Recognit.*, vol. 71, pp. 249–263, Nov. 2017.
[87] K. Lekdioui, R. Messoussi, Y. Ruichek, Y. Chaabi, and R. Touahni, "Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier," *Signal Process., Image Commun.*, vol. 58, pp. 300–312, Oct. 2017.
[88] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognit.*, vol. 92, pp. 177–191, Aug. 2019.
[89] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 391–404, Feb. 2020.

SUMEET SAURAV received the master's degree from the Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India, in 2014. He is currently pursuing the Ph.D. degree. He is also working as a Scientist with the CSIR-Central Electronics Engineering Research Institute, Pilani, Rajasthan, India. He is involved in various projects sponsored by the Government of India on Artificial Intelligence. His research interests include computer vision, machine learning, deep learning architectures for vision-based applications, and FPGA-based real-time implementation of computer vision algorithms.

RAVI SAINI (Member, IEEE) received the master's degree in electronics from DAVV, Indore, India, in 2000, the M.Tech. degree from Panjab University, India, in 2002, and the Ph.D. degree in electronics from Kurukshetra University, Kurukshetra, India. He is currently working as a Senior Scientist with the CSIR-Central Electronics Engineering Research Institute, Pilani, Rajasthan, India. His research interests include VLSI Architectures, ASIC and ASIP Design, HDLs, and FPGA prototyping.

SANJAY SINGH (Senior Member, IEEE) received the M.Sc. and M.Tech. degrees and the Ph.D. degree in electronics from Kurukshetra University, Kurukshetra, India, in 2005, 2007, and 2015, respectively. He is currently working as a Senior Scientist with the CSIR-Central Electronics Engineering Research Institute (CEERI), Pilani, Rajasthan, India, where he is also the Head of the Cognitive Computing Group. He is actively handling several industrial and the Government of India sponsored projects related to computer vision and machine learning.