Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046070

ARTICLE

Facial Expression Recognition with High Response-Based Local Directional Pattern (HR-LDP) Network

Sherly Alphonse* **and Harshit Verma**

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

*Corresponding Author: Sherly Alphonse. Email: sherly.a@vit.ac.in

Received: 17 September 2023 Accepted: 29 November 2023 Published: 27 February 2024

ABSTRACT

Although lots of research has been done in recognizing facial expressions, there is still a need to increase the accuracy of facial expression recognition, particularly under uncontrolled situations. The use of Local Directional Patterns (LDP), which has good characteristics for emotion detection has yielded encouraging results. An innovative end-to-end learnable High Response-based Local Directional Pattern (HR-LDP) network for facial emotion recognition is implemented by employing fixed convolutional filters in the proposed work. By combining learnable convolutional layers with fixed-parameter HR-LDP layers made up of eight Kirsch filters and derivable simulated gate functions, this network considerably minimizes the number of network parameters. The cost of the parameters in our fully linked layers is up to 64 times lesser than those in currently used deep learning-based detection algorithms. On seven well-known databases, including JAFFE, CK+, MMI, SFEW, OULU-CASIA and MUG, the recognition rates for seven-class facial expression recognition are 99.36%, 99.2%, 97.8%, 60.4%, 91.1% and 90.1%, respectively. The results demonstrate the advantage of the proposed work over cutting-edge techniques.

KEYWORDS

Emotion; classification; CNN; network; HR-LDP

1 Introduction

Human Computer Interaction (HCI) primarily consists of the study of interface design, with its applications concentrating on user-computer interaction. Since computers are used in almost every area of daily life, HCI applications are found in every industry, including social science, psychology, science, industrial engineering for computers, and many more. A crucial field of research in pattern recognition and computer vision is Facial Expression Recognition (FER). FER has emerged as a crucial research area within computer vision and artificial intelligence, offering profound implications for diverse applications, such as human-computer interaction, emotion-aware computing, and affective computing. Automatic emotion recognition from facial expressions is an interesting research topic that has been used in healthcare, social networks, and human-machine interactions, among other domains. To improve computer prediction, researchers in this discipline are working on methods to decode, analyze, and extract these characteristics from facial expressions. The remarkable success of this technology has led to the use of numerous deep-learning architectures to boost

performance [1]. Emotions have a natural influence on human behavior and are important in shaping communication and behavior patterns. Accurately analyzing and interpreting the emotional content of facial expressions is essential for a deeper understanding of human behavior. Computer systems still struggle to accurately identify facial expressions, even though it requires little to no effort for a person to recognize faces and decipher facial emotions. It is believed that analyzing a person's facial features and determining their emotional state are incredibly tough tasks. The main obstacles are the irregularities of the human face and variations in elements such as direction, lighting, shadows, and facial posture. Research has indicated that disparate individuals can identify distinct emotional states within an identical facial expression. FER involves many hurdles, including the need for diverse training data and pictures featuring a range of ethnicities, genders, and nations, among others. Deep learning methods have been researched as a stream of techniques to achieve resilience and provide the required scalability on new forms of data [2]. It is necessary to acquire a proper classification model that is both subtle to minute differences in the appearance of facial emotions and resilient to larger variations to recognize facial expressions under uncontrolled situations. For recognizing facial expressions, a variety of pre-trained deep neural networks can be used. These networks have a great number of parameters that can be learned, yet they were trained and used on quite varied applications. The neural aspects make it challenging to accurately train neural networks for facial emotion recognition. To overcome this, in this study, a big neural network that is trained on extensive facial emotion recognition datasets is chosen which is later used to train a small neural network. The small network has a lesser number of parameters to be learned than the large network. The suggested network is then built using its convolutional layers, and the complete structure is trained with facial expression photos.

The main objective is the construction of neural network models that support the input of the images in the right format and produce an output that can be mapped to a classification of emotion. After the successful building of the model, testing and troubleshooting also have to be done to maximize the accuracy and also to perform analysis via various metrics available to cross-examine the efficiency and the correctness of the model. Another major aim is to try and eliminate problems present in the dataset such as cross-oriented images, wrong facial position, alignment issues, etc. This has to be addressed because the images when they are disoriented, will lead to bad predictions due to unnecessary parallax error and wrong orientation of the images. The next issue is edge detection and the reason for performing edge detection is to enhance the facial features and boost the parts where emotion is displayed, like the position of the mouth, eyebrows, eyes, and even the nose. The alignment problems are rectified using a face detection and alignment method "Chehra" in the proposed work. The proposed High Response-based Local Directional Pattern (HR-LDP) based classification method also uses the Kirch filter which eliminates the noise in images and accurately captures the sharp edges that represent the structure of the face. The major contributions of the proposed work are as follows:

- A novel HR-LDP network-based classification is proposed in this work with a module for eliminating noise using high responses obtained from Kirsch filters that reduce the computation while increasing accuracy.
- The proposed work suggests a novel learnable HR-LDP network that reduces the number of learnable parameters compared to the existing works.
- Compared to existing deep learning-based detection algorithms, the parameters in our fully linked layers can save up to 64 times the cost while outperforming state-of-the-art techniques.

The paper is structured as follows: The state-of-the-art techniques for facial emotion recognition are reviewed in Section 2. Then, in Section 3, the suggested learnable HR-LDP network is presented.

The specifics of the experimental setting are provided and the findings of the detection are then displayed and analyzed in Section 4. Finally, this paper is concluded in Section 5 with the guidelines for future research.

2 Related works

This section presents a detailed survey of the existing works. Table 1 gives a summary of the latest works in literature. In [3], the authors have used the Cohn Kanade (CK+) dataset that is available to the public. They forwarded it through four different Convolutional Neural Networks (CNN) which implement transfer learning. They were VGG-19, ResNet-50, MobileNet and Inception V3. After the image pre-processing and the feature extraction were done, they passed it through the 4 networks and compared the performance of each one with the other. Reference [4] suggested a novel technique called Facial Emotion Recognition using Convolutional neural networks (FERC) and used it for this problem. FERC is a 2-part CNN, one for removing the background of the image and the other for the classification into one of the five emotions set. They tested the algorithm with CK, Caltech, CMU and NIST datasets. In [5], the authors have used deep CNNs with 2 layers that are included with dropouts after each layer. It is passed through an activation function and then to the pooling layer. The same is repeated in the next layer. The final dense layer has 5 units representing each emotion.

Table 1: Summary of literature survey with the algorithms

Work	Algorithms used	Pros
[3]	• CNNs using Transfer-Learning. • VGG-19, ResNet-50, MobileNet, and Inception V3.	• VGG-19–Weights are easily available. • ResNet–Greatly reduced training time. • MobileNet–Guaranteed improvement. • Inception V3–Factorization in to smaller convolutions.
[4]	• 2-level CNN. • Level 1–Background removal. • Level 2–Classification.	As it follows 2 layers, it has improved accuracy over other conventional CNNs.

(Continued)

Table 1 (continued)

Work	Algorithms used	Pros
[5]	Deep CNNs having ReLU as an activation function, max pooling layer, dropout layer, flattening layer, and finally softmax dense layer.	The benefits of ReLU are sparsity and reduced likelihood of vanishing gradients. Also, the sum of Softmax values is always 1 and is better for multiclass.
[6]	VGG-16, ResNet-152, ResNet-18, ResNet-34, ResNet-50, DenseNet-161, VGG-19, and Inception-v3.	8 different methods used to cause the best possible accuracy to be achieved.
[7]	Deep CNN.	Leaky ReLUs mitigate the large weight-handling problems of ReLU.
[8]	Attentional CNNs, max-pooling layers and ReLU, with dropout and fully connected layers.	Better results due to the attentional mechanism and faster computation.
[9]	CNN with VGGNet, SGD optimizer, RLRP learning rate scheduler and cosine annealing schedulers.	VGGNet weights are easily available.
[10]	An enhanced conditional generative adversarial network (im-cGAN).	For GANs to function well, a lot of training data is frequently needed.
[11]	WOA-TLBO and Multi-SVNN.	Optimization can be done effectively.
[12]	CNN with subsampling layers, max pooling, flattening and SoftMax activation.	Reduces reliance on the positioning of features in networks.
[13]	Two distinct CNN models.	Higher accuracy is obtained.
[14]	Monogenic Sobel Directional Pattern (MSDP) and CNN.	Obtains good performance and also reduces noise while extracting edges.

In [6], the authors have developed a FER system, and it has been verified on eight different pre-trained Deep CNN models with the Karolinska Directed Emotional Faces (KDEF) and Japanese Female Facial Expression (JAFFE) facial datasets. On application of a 10-fold cross-validation, the best model uses DenseNet-161. The CNN algorithms [7] are used by several works in literature that have shown superior performance. Among that, the authors in [12] have proposed a CNN-based single classifier that achieved high performance. It also performed the necessary pre-processing. The model

has two Convolution layers, two sub-sampling layers and an output layer. They also used a max-pooling and flattening layer with the final activation function as SoftMax. They got an accuracy of 97.6%. Also, Reference [15] did the necessary pre-processing by taking the mean shape and mapping the dataset with the closeness from the mean shape. Notably, the authors in [16,17] conducted a comprehensive review focused on CNNs for FER. Their study explored various CNN architectures and methodologies, showcasing their effectiveness in capturing spatial hierarchies within facial images. The studies from [18] and [19], have significantly transformed FER. These works highlight the proficiency of CNN in capturing spatial hierarchies and achieving impressive performance, along with the critical contributions of data augmentation and feature extraction in improving FER accuracy and robustness. Despite the remarkable strides made in Facial Emotion Recognition (FER), the field continues to grapple with a series of substantial challenges and limitations that warrant thorough exploration. While FER algorithms [20–24] have shown proficiency in identifying basic emotions, the recognition of nuanced and subtle facial expressions remains an ongoing research frontier. The intricate interplay of various facial muscles and features, especially in complex emotional states, poses a significant challenge for current models. Inside the neural network, the different combinations of layers can accomplish a task with high accuracy. This work proposes a novel HR-LDP network-based classification that helps to attain good accuracy while classifying six datasets and learning a smaller number of parameters. The proposed work is explained in Section 3.

3 The Proposed Work

The architecture of the suggested work is shown in Fig. 1. The three main elements of this network are convolutional layers, a fully connected layer for HR-LDP computation, and another fully connected layer that is proportional to a loss function. This network creates feature maps associated with expression by applying convolutional layers to the input image. The three modules that make up this network are the convolutional layer, HR-LDP layer and loss function layer, as shown in Fig. 1. A classification layer is also used at the end to predict the emotions using a classification algorithm like SVM. The loss function module is used to train the parameters of the network. The convolutional and HR-LDP layers are used to extract simulated HR-LDP features, and classification layers are used to predict the emotion. The main elements of the suggested neural network are thoroughly described in this section.

3.1 Convolution Layer

The faces are detected from the sample images from the dataset using a 'Chehra' [20] face detector. In the proposed work, a face detection and alignment tool 'Chehra' is used to solve the alignment problems. The convolutional feature maps for the original images are created by forward-propagating the unprocessed pixels via the initial module. More precisely, there are three convolutional layers in the initial module: two for convolution, one for pooling, and one for Restricted Linear Units (ReLU). In addition, it reduces the effect of initializing filter parameters. Before the ReLU layer, a batch normalization (BN) layer is used. This is depicted as

$$\hat{I} = {}'\Upsilon.\frac{X-\mu}{\sqrt{\sigma^2+\in}} + \beta \tag{1}$$

where I is the BN layer's input. The mean and variance of I are μ, σ correspondingly. Here, ${}'\Upsilon$ and β are scale and shift factors, respectively, while a constant $\in$ is further added to the variance to account for numerical stability.

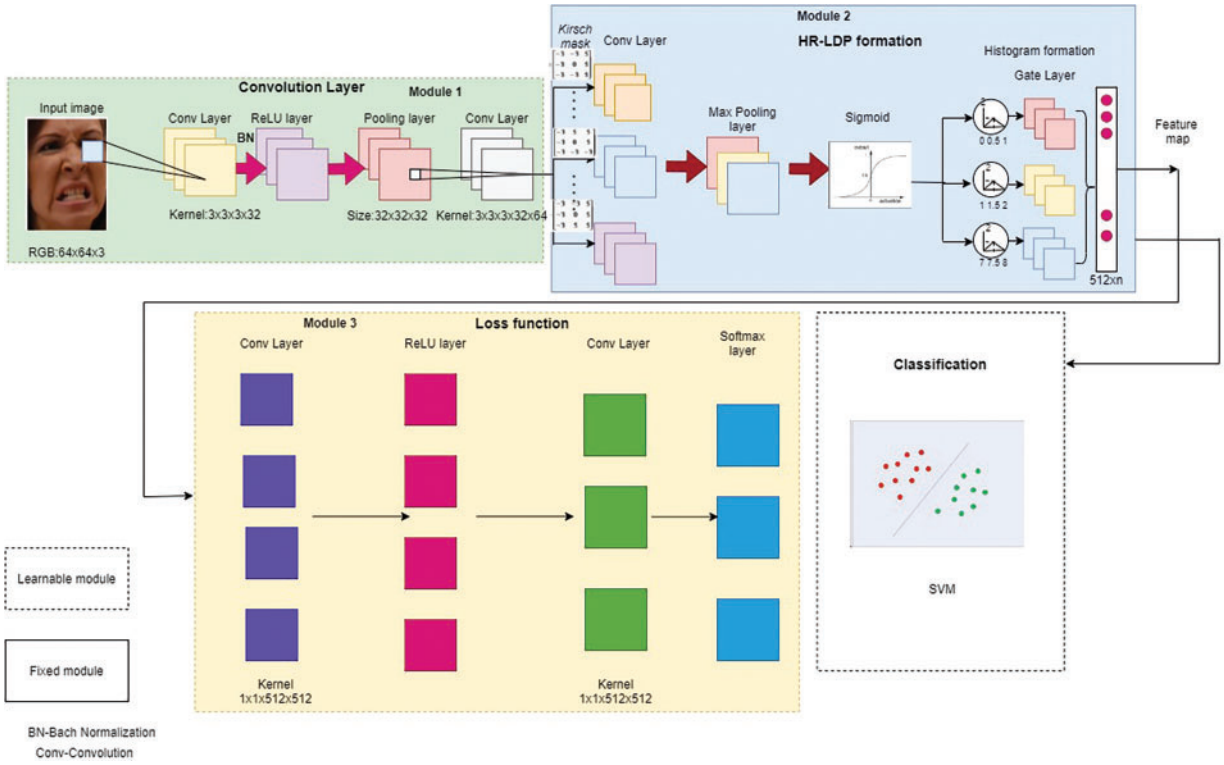

Figure 1: The architecture of the proposed facial expression recognition network

3.2 HR-LDP Layer

The HR-LDP layer performs convolution using Kirsch masks [24] and extracts only the high responses related to shape and texture information which is then normalized using Sigmoid function and then the histograms are extracted using gate functions as in the subsequent sections.

3.2.1 Convolution Using Kirsch Filter Masks

The Kirsch masks in Fig. 2 are applied on the output from the convolution layer and the eight responses are obtained on which max pooling is applied.

$$C(x, y) = \sigma_{argmax}(R_{\theta_i}(x, y) | 0 \leq i \leq 7) \tag{2}$$

$$\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix}$$

$$M_{\theta_0} \quad M_{\theta_1} \quad M_{\theta_2} \quad M_{\theta_3} \quad M_{\theta_4}$$

$$\begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$

$$M_{\theta_5} \quad M_{\theta_6} \quad M_{\theta_7}$$

Figure 2: Kirsch mask

Here σ_{argmax} is obtained by max pooling. The pool_size = 2 and strides = 2 are used when creating a MaxPool2D layer. The MaxPooled output is obtained in tensor form by applying the MaxPool2D layer on the matrix. When it is applied to the matrix, the Max pooling layer will iteratively compute the maximum of each 2 × 2 pool with a 2 jump. The values are then normalized using the sigmoid function and given to the gate functions for histogram formation.

3.2.2 Histogram Calculation

A histogram shows the probability distribution of a quantity in different bins. Different appearance-based feature extraction techniques have been developed, which process the image using either manually applied or learnable filters and a histogram to calculate statistical data. CNN can be thought of as a collection of learnable filters when feature maps are generated at the output of convolutional layers. The feature maps are first flattened, and then they are added to a layer with all connections. A simple method for constructing the histograms of feature maps involves applying specific shifted step activation functions to the obtained feature maps and then aggregating each result as a bin of histograms. However, gradient-based learning is incorrect since the step function's derivative is infinite at its edges and zero everywhere else, and the gate function determines the variable's histogram in the range [0,1].

$$f(x) = \begin{Bmatrix} 2nx & 0 < x < \frac{1}{2n} \\ -2n\left(x - \frac{1}{n}\right) & \frac{1}{2n} \leq x < \frac{1}{n} \\ 0 & ow \end{Bmatrix} \tag{3}$$

where n denotes the histogram's number of bins. The gradient of Eq. (3) in the backpropagation stage is taken to be $2n$ during $0 < x < 1/2n$, and $-2n$ when $1/2n < x < 1/n$, and 0 otherwise. The infrequently occurring discontinuity is comparable to the recognized Rectified Linear Units (ReLU) activation function. The points at $x = 0, 1/n$ and $1/2n$ are disregarded. With this assumption, the ReLU functions adequately in practice. The calculation of histograms in a neural network has been done using Gaussian functions. However, in both backward and forward passes, the computational complexity of the employed gate activation function is significantly lower. This benefit can speed up the neural network's inference and training processes. In contrast to applying gate activation on manually created histograms, it is suggested in this study that gate activation be applied directly to facial expression feature maps to calculate their statistical information. To do this, one can compute an n-bin histogram by shifting each of the n gate activation functions by one. It can also be found by shifting the input signal n times by 1 concerning the previous one. However, the computational complexity of the employed gate activation function in both forward and backward passes is significantly lower. This advantage can hasten the inference and training phases of the neural network. This research proposes to directly use gate activation to face expression feature maps to calculate their statistical information as histograms, in contrast to [21–25], which applied it to a hand-crafted layer to address the issue of spoofing detection. To do this, one can compute an n-bin histogram by shifting each of the n gate activation functions by one. Additionally, as shown in Figs. 3 and 4, it can be discovered by shifting the input signal n times by $1/n$ about the previous one.

$$H_i = \frac{1}{E}\sum f\left(FM - \frac{i}{m}\right), i = 0, \ldots . m - 1 \tag{4}$$

Here The H histogram's i[th] bin is designated as H_i. The current feature map is FM. E is the number of feature map (FM) elements, m is the number of histogram bins, and f is the gate activation function mentioned in Eq. (4). The feature map used to calculate the histogram is called FM. In the suggested CNN, $\frac{1}{K}\sum(.)$ executed with average pooling operators. The input variable should fall between 0 and 1 as is expected for histogram calculation with the gate function. However, this presumption might not apply to feature maps. Consequently, the input of the gate activation function needs to be normalized to [0,1] to be used for histogram calculation. The sigmoid function can be utilized for this. Nevertheless, at very large/small values, the sigmoid is saturated. To solve this issue,

$$S(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

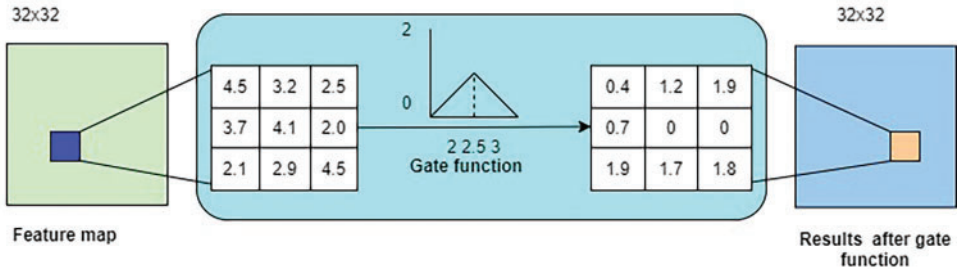

Figure 3: The feature map and gate function

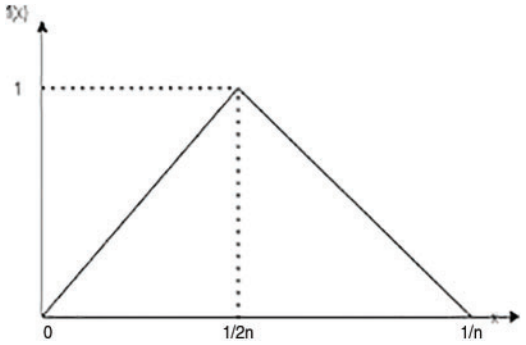

Figure 4: Gate function

As in Fig. 1, the feature values for the histogram computation layer are initially constrained using batch normalization to prevent sigmoid function saturation. The values are then normalized to [0,1] using a sigmoid activation function. The output of the sigmoid function is then shifted n times. Ultimately, n-gate activation functions and the n-bin histograms are calculated via average pooling. The computed histograms show feature-specific statistical data maps for the image input. Convolutional neural networks can employ this feature map histogram computation approach without any issues to the learning process. The generated histograms are then integrated into the completely connected layer of the proposed network which is explained in the following section:

3.2.3 Loss Function

$$F(Y) = \sum_{i=1}^{n} \{log\,(e^{y_{i1}} + e^{y_{i2}} + \ldots + e^{y_{iv}}) + y_{ir}\} \tag{6}$$

The most popular SoftMax loss function is therefore utilized as in Eq. (6) to quantify the classification error following the extraction of HR-LDP features. The SoftMax loss function can optimize the likelihood of the correct class during the training stage and fine-tune the network parameters based on Back Propagation (BP). Here i is the training sample index and n represents the count of training samples. $[Y = Y_1, Y_2, Y_3, \ldots, Y_n]$ is the label set and $[Y_i = y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iv}]$ is the prediction vector of the ith training sample. The predicted value is denoted by y_{iv}, and the number of classes is indicated by v. To combine the data on facial movement during testing, the HRLDP features are taken from a video sequence and the average is calculated and converted into a feature vector. The averaged features are then classified using Support Vector Machine (SVM) classifier. Algorithm 1 describes the basic flow of the classification module.

Algorithm 1: HR-LDP based network

Input:

The training set videos and images are Vtrain and Ltrain, respectively.

The testing face videos are Vtest.

Output:

Seven emotion categories.

Steps:

1. Training of the SVM classifier is done using the following steps:

(i) Each video from Vtrain is divided into frames (Ltrain), signified as $f_i = f_{i1}, f_{i2}, \ldots, f_{in}$, where f_{i1} is the first frame of ith video in Vtrain.

(ii) The f_{i1} is fed into module 1.

(iii) In module 1, the raw face images in the frame would be first aligned and cropped to a size of $64 \times 64 \times 3$ before undergoing the convolution and down-sampling operations.

(iv) Thus, the convolutional feature maps are created from raw images in module 1.

(v) Those convolutional feature maps are then fed into module 2 to extract the texture and edge features as $l_i = l_{i1}, l_{i2}, \ldots, l_{in}$.

(vi) In module 2, the kirsch masks are applied on the convolutional feature maps to create the filtered feature maps.

(vii) After that the max pooling is applied on the filtered feature maps.

(viii) Then the sigmoid function and gate function are applied to obtain the HR-LDP feature vectors.

(ix) The HR-LDP feature vectors thus obtained from Ltrain are used to train the SVM classifier.

(x) The feature vectors can also be classified using SVM.

(xi) The loss function in module 3 helps in fine tuning or training the network parameters and maximizes the probability of right class.

2. Testing of the SVM classifier is done using the following steps:

(i) The HR-LDP features are obtained from the video of the testing set Vtest by following the steps (i) to (viii) used in training.

(ii) The HR-LDP features are then fed into the trained SVM and the classification results are obtained.

The SoftMax loss function, which is based on the BP method, can optimize the likelihood of the correct class during the training stage and fine-tune the network parameters. The given testing sample is classified and the results are given in the next section.

4 Results and Discussion

The suggested approach uses Matlab 2018a for its experiments.

4.1 Datasets

The research makes use of six datasets, including JAFFE [26], Cohn Kanade (CK+) [27], Oulu-CASIA NIR&VIS facial expression database (OULU-CASIA) [28,29], Man Machine Interface (MMI) [30] Multimedia Understanding Group (MUG) [31] and Static Facial Expressions in the Wild (SFEW) [32,33].

4.2 Experimental Analysis

The high computational complexity is a significant limitation for state-of-the-art descriptors like Gabor. The accuracy of every other feature descriptor in literature is far lower, especially under unrestricted circumstances. So, HR-LDP is incorporated into the proposed model which achieves high accuracy under low complexity. SFEW dataset poses significant challenges because it was collected in unrestricted circumstances. Tables 2–7 demonstrate the effectiveness of the suggested strategy by listing both the count of samples that were properly identified and the count of samples that were erroneously classified. The neutral and depressed expressions are confused when predicting other images during the classification of the photos from the JAFFE dataset with the suggested method, as in the confusion matrix given in Table 2. As seen in Table 3, the CK+ dataset's classification accuracy for anger and neutral emotions is significantly lower. Expressions like neutral, happiness, and surprise are mixed up with other emotions in the MUG dataset, as shown in Table 4. The fundamental issue with the SFEW dataset is that the samples of the various classes are out of balance and that the photographs were taken in an unrestricted environment. Therefore, as seen in Table 5, more training data is required to increase accuracy. The suggested method outperforms the other current descriptors in terms of accuracy for SFEW due to its capacity to identify crisp edges and its scale and rotation-invariant characteristics. In comparison to other available datasets, the classification accuracy of the SFEW dataset is lower. When equated to the other descriptors currently used in the literature, however, SFEW obtains a greater accuracy utilizing proposed technique, as shown in Table 5. Most other facial expressions can be mistaken for the disgusted face. As in the confusion matrices provided in Tables 6 and 7, fear and sadness facial emotions cause misunderstanding with the rest of the expressions in the Oulu-CASIA dataset and MMI.

Table 2: Matrix showing the confusion in the JAFFE dataset

Emotion (Em)	Angry (An)	Disgust (Dis)	Fear (Fea)	Happy (Happ)	Neutral (Neut)	Sadness (Sa)	Surprise (Sur)
An	98.67					1.33	
Dis		99.8			0.2		
Fea			92.44		2.0	3.50	2.06
Happ		1.5		96.8	0.2	1.5	
Neut					98.2	1.8	
Sa					1.57	98.43	
Sur			0.8				99.2

Table 3: Matrix showing the confusion in the CK+ dataset

Em	An	Dis	Fea	Happ	Neut	Sa	Sur
An	98		1.5		0.5		
Dis		97.3	2.20		0.50		
Fea		0.4	99.1	0.5			
Happ				99.75	0.25		
Neut			0.4		99	0.6	
Sa				1.0		99	
Sur					0.37		99.63

Table 4: Matrix showing the confusion in the MUG dataset

Em	An	Dis	Fea	Happ	Neut	Sa	Sur
An	93.17	1.83		1.3	0.44	0.26	
Dis	0.03	97.97		1.00	1.00		
Fea		0.6	98.44	0	0.51	0.35	0.10
Happ		0.17	0.52	99.07			0.24
Neut				2.23	97.77		
Sa	0.62		0.38		2.18	96.81	
Sur			1.76	0.24			98.00

Table 5: Matrix showing the confusion in the SFEW dataset

Em	An	Dis	Fea	Happ	Neut	Sa	Sur
An	61.7	2.2	3.4	6.7	13	8	5
Dis	7	63.3	1.7	5	6.9	8.1	8
Fea	13		83	2.5	3	0.5	0
Happ	2	19.0	1.8	54.9	1	18	4
Neut	1	12.4	18		44	14.6	10
Sa	15.4		4.3	16.4	15.2	39.7	9
Sur	4.1		3.7	14	1.9	0	76.3

Table 6: Matrix showing the confusion in the Oulu-CASIA dataset

Em	An	Dis	Fea	Happ	Neut	Sa	Sur
An	95.67		4.33				
Dis		91.67		4	2.33		

(Continued)

Table 6 (continued)

Em	An	Dis	Fea	Happ	Neut	Sa	Sur
Fea	13	0	75.94		8	11.03	3.03
Happ		0		100			0
Neut		0.2		9.8	88.1	0.1	1.8
Sa		0.6	10	0.4	4.67	84.0	0.33
Sur		0.60		2.39			97.1

Table 7: Matrix showing the confusion in the MMI dataset

Em	An	Dis	Fea	Happ	Neut	Sa	Sur
An	94.77	0	0.3	1.53		0.4	3
Dis		90.87		5	4.13		
Fea	13		82.84		0.7	0.03	3.03
Happ		0		99.2			0.8
Neut		1.1	15	7.8	75.1	0.1	0.9
Sa		0.7	7	0.3	5.57	85.2	1.33
Sur		0.62		2.38		2.7	94.3

Figs. 5–10 compare the recognition outcomes. In comparison to more current methods like inter-category distinction feature fusion network [34–38] and ROI-guided deep architecture [39–42], the suggested study attains greater accuracy. Because there is less likelihood of overfitting [43–45], less data noise, improved discriminating, and improved data visualization, the proposed approach performs better. The recommended feature extraction technique automatically chooses only the relevant data needed for this activity. This work suggested using a new HR-LDP network to tackle the detection of emotions. The suggested network mixes deep learning and manually created features, and it can minimize the network parameters by producing statistical histograms. Numerous tests using the databases produced intriguing findings. Furthermore, unlike the majority of modern techniques, this suggested approach produces reliable performance. The VGG-face network [46] is chosen as the reference network for comparing the efficiency of our network in terms of time and memory intake. VGG-face is fine-tuned for face expression detection because it is utilized for face recognition. To be fair, identical training data is employed, as training parameters, and loss function in both the VGG-face network and our suggested LBP network. On an HP workstation set up as follows, the comparison experiments are conducted. Matlab 2018a, 64 G of RAM, two Intel E-52620 v3 CPUs, and one NVIDIA GeForce GTX 1080 Ti GPU are all included with the Windows 10 Enterprise Edition operating system. The results of the comparison of time and memory are shown in Table 8. The table shows that, when training, our suggested network requires just 132 MB of memory, which is up to 25 times less memory than the VGG-face network. Furthermore, in training rounds, the proposed network outperforms the VGGface network. Depending on the input's size the suggested network should be significantly faster than the VGG-face network due to the size of the proposed network.

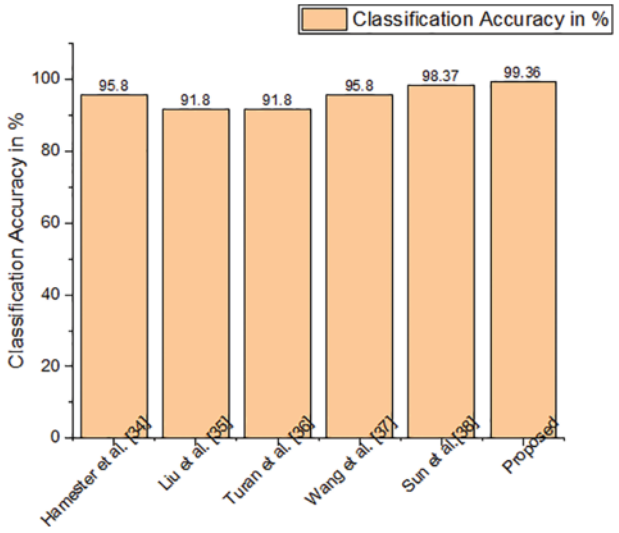

Figure 5: Classification accuracy of JAFFE dataset

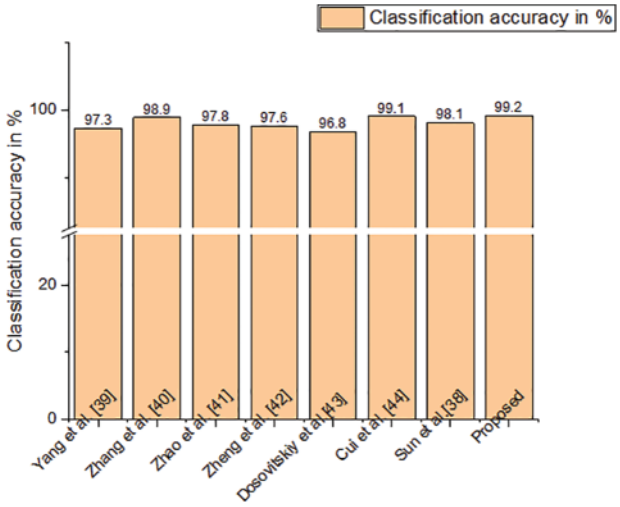

Figure 6: Classification accuracy of CK+ dataset

Table 9 represents the different parameters used. The Stochastic Gradient Descent (SGD) approach is used for optimization during the training phase, with learning rate = 0.01 and momentum = 0.9. There are 100 training epochs, and from the thirty-first to the last epoch, the learning rate drops by 0.99 in each epoch. This setting of 0.5 for the dropout prevents over-fitting. The margin hyper-parameter is set to 0.2. The dimension of the feature vector obtained at the output of the histogram computation layer is $512 \sim 10 = 5120$ since the count of histogram bins in HR-LDP is initialized to 10. Ten percent of the training in each trial is selected at random and utilized for validation. Table 10 represents the results obtained using different classifiers in the proposed work. The proposed work achieves higher accuracy when using SVM, deep learning techniques and CNN in the final classification layer of the proposed architecture. However, the SVM in the final layer has a lesser number of parameters, saving the computational cost and achieving higher accuracy.

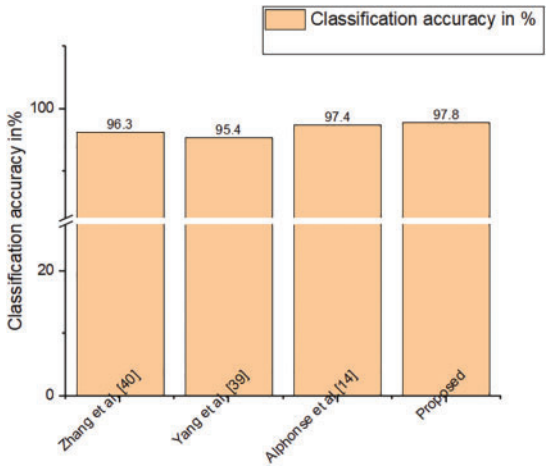

Figure 7: Classification accuracy of MUG dataset

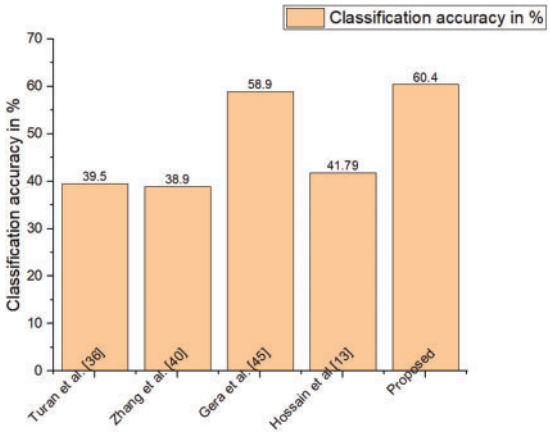

Figure 8: Classification accuracy of SFEW dataset

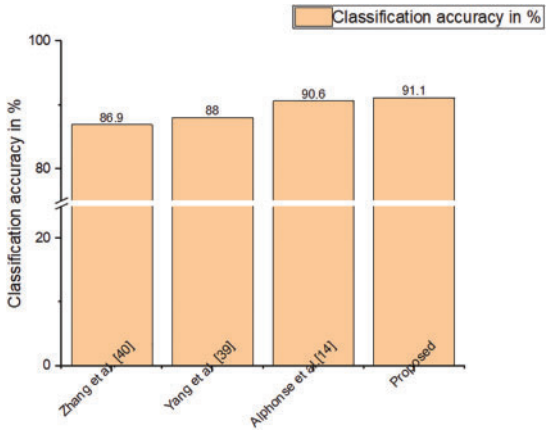

Figure 9: Classification accuracy of OULU-CASIA dataset

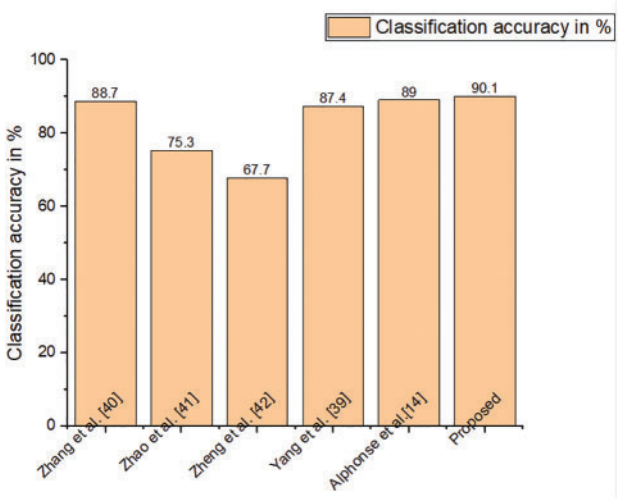

Figure 10: Classification accuracy of MMI dataset

Table 8: Comparison of mean time and cost of memory

	Training		Testing	
20 pt network	Time in s	Memory in MB	Time in s	Memory in MB
VGG-Face	1203.4	3176	0.022	1232
Proposed work	768	132	0.021	90

Table 9: Parameters used in the proposed model

Layers-Li	L1	L2	L3	L4	L5	L6	L7	L8
Modules-Mod I	Mod 1	Mod 1	Mod 1	Mod 1	Mod 1	Mod 2	Mod 2	Mod 2
Types	Convolution	Batch normalization	ReLU	Max pool	Convolution	Convolution	Max pool	Sigmoid
Support	3	1	1	2	3	3	2	1
Filter dimension	3	–	–	–	3	–	–	–
Number of filters	32	–	–	–	64	–	–	–
Number of strides	1	1	1	2	1	1	2	1
Padding	1	0	0	0	1	1	0	1
Layers-li	L9	L10	L11	L12	L13	L14	L15	L16
Modules-modi	Mod 2	Mod 2	Mod 2	Mod 3	Mod 3	Mod 3	Mod 3	Mod 3
Types	Sum	Gate layer	Average pooling	Fully connected	Batch normalization	ReLU	Fully connected	Softmax
Support	1	2	32	1	1	1	1	1
Filter dimension	–	–	–	1	–	–	1	–
Number of filters	–	–	–	512	–	–	512	–
Number of strides	1	2	0	1	1	1	1	1
Padding	0	0	0	0	0	0	0	0

Table 10: Classification outcomes using different classifiers in the classification layer of the proposed work

The technique used for feature extraction	Classifier	JAFFE	CK+	MUG	SFEW	OULU-CASIA	MMI
HR-LDP	K-nearest neighbor	78.0	88.2	78.5	40.2	80.4	76.2
	Bayes classifier	76.4	75.6	74.3	44.3	82.3	80.2
	Support vector machine	95.6	96.5	88.6	52.3	88.4	81.2
	Extreme learning machine	96.5	98.7	90.2	55.6	89.2	82.3
	Multi-layer perceptron	90.2	92.3	84.5	56.2	87.2	80.1
	Stacked restricted Boltzmann machine	97.3	98.4	96.8	59.3	89.7	88.5
	CNN	99.2	99.1	97.6	60.4	91.0	90.2
	SVM	99.3	99.2	97.8	60.4	91.1	90.1

4.3 Ablation Study: Analysis of Several Proposed Model Components

(i) *By eliminating the histogram formation layer in the suggested work:* At the output of module 2 in Fig. 1 of the experiment, a max pooling layer is utilized to construct a 5120-dimensional feature vector. Next, using a loss function, the network is trained for seven classes of face emotion identification.

(iii) *Changing from SoftMax loss function to chi-squared distance-based loss function:* The loss function is defined in Eq. (6) as a SoftMax function. This is changed as an improved chi-squared distance-based loss function [47] as in Eq. (7).

$$l = \sum_{i=1}^{N} \left[\chi_L^2 \left(X_i^a - X_i^p \right) - \chi_L^2 \left(X_i^a - X_i^n \right) + \alpha \right]_+ \tag{7}$$

where N denotes the count of the triplets present in the set of training samples, X_i^a represents the anchor sample, X_i^p represents the set of positive samples, X_i^n is the set of negative samples and χ_L^2 represents the distance between two samples, α denotes the margin hyper parameter.

(iii) *Using the whole proposed work for emotion classification:* In this experiment, face expression recognition is accomplished by using the whole HR-LDP and SVM. The results from three different cases are given in Fig. 11.

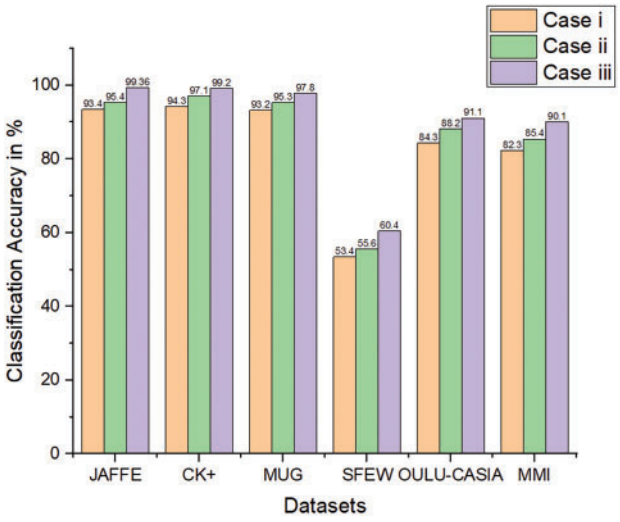

Figure 11: Ablation study using three different cases

5 Conclusion

This novel HR-LDP network is suggested to tackle facial expression recognition. The suggested network mixes deep learning and manually created features, and it can minimize the network parameters by producing statistical histograms. Numerous tests using the seven databases produced intriguing findings. Furthermore, unlike the majority of modern techniques, this suggested approach produces reliable performance. Concerning SFEW photos with significant blur and occlusions, the suggested technique obtains greater classification accuracy compared to other methodologies in the literature, it achieves good accuracy. The results show that the suggested strategy improves classification accuracy across six datasets. Future research will concentrate on micro-expressions and the analysis of dynamic emotions in videos.

Acknowledgement: We thank Vellore Institute of Technology, Chennai for supporting us with the APC.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design, draft manuscript preparation: Sherly Alphonse. analysis and interpretation of results: Harshit Verma. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Both CK and JAFFE are openly accessible datasets. On request, more datasets from specific authors are available. Access the MUG dataset at https://mug.ee.auth.gr/fed/. Access the Oulu-CASIA dataset at https://paperswithcode.com/dataset/oulu-casia. Access the MMI dataset at https://mmifacedb.eu/. Visit https://paperswithcode.com/dataset/sfew to get the SFEW dataset.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. R. Adyapady and B. Annappa, "A comprehensive review of facial expression recognition techniques," *Multimed. Syst.*, vol. 29, no. 1, pp. 73–103, 2023. doi: 10.1007/s00530-022-00984-w.

[2] M. Sajja *et al.,* "A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines," *Alex. Eng. J.*, vol. 68, pp. 817–840, 2023. doi: 10.1016/j.aej.2023.01.017.

[3] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human-computer interaction applications," *Neural. Comput. Appl.*, pp. 1–18, 2021. doi: 10.1007/s00521-021-06012-8.

[4] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Appl. Sci.*, vol. 2, no. 3, pp. 446, 2020. doi: 10.1007/s42452-020-2234-1.

[5] E. Pranav, S. Kamal, C. S. Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *2020 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, IEEE, 2020, pp. 317–320.

[6] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, pp. 1036, 2021.

[7] S. Modi and M. H. Bohara, "Facial emotion recognition using convolution neural network," in *2021 5th Int. Conf. Intell. Syst. Comput.*, IEEE, 2021, pp. 1339–1344.

[8] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, pp. 3046, 2021. doi: 10.3390/s21093046.

[9] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," arXiv preprint arXiv:2105.03588, 2021.

[10] Z. Sun, H. Zhang, J. Bai, M. Liu, and Z. Hu, "A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition," *Pattern Recognit.*, vol. 135, pp. 109157, 2023. doi: 10.3390/s21093046.

[11] A. V. Lakshmi and P. Mohanaiah, "WOA-TLBO: Whale optimization algorithm with teaching-learning-based optimization for global optimization and facial emotion recognition," *Appl. Soft. Comput.*, vol. 110, pp. 107623, 2021. doi: 10.1016/j.asoc.2021.107623.

[12] 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, pp. 391–404, 2020. doi: 10.1007/s00371-019-01627-4.

[13] S. Hossain, S. S. Umer, R. K. Rout, and M. Tanveer, "Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling," *Appl. Soft. Comput.*, vol. 134, pp. 109997, 2023. doi: 10.1016/j.asoc.2023.109997.

[14] A. S. Alphonse, S. Abinaya, and K. S. Arikumar, "A novel monogenic Sobel directional pattern (MSDP) and enhanced bat algorithm-based optimization (BAO) with Pearson mutation (PM) for facial emotion recognition," *Electronics*, vol. 12, no. 4, pp. 836, 2023. doi: 10.3390/electronics12040836.

[15] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 101–106, 2018. doi: 10.1016/j.patrec.2018.04.010.

[16] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3238–3247. doi: 10.48550/arXiv.1708.03474.

[17] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Patt. Recog.*, Honolulu, HI, USA, 2017, pp. 2272–2281.

[18] A. Uçar, "Deep convolutional neural networks for facial expression recognition," in *2017 IEEE Int. Conf. on Innov. in Intell. Sysy. App. (INISTA)*, Orlando, FL, USA, IEEE, 2017, pp. 371–375.

[19] H. Sikkandar and R. Thiyagarajan, "Deep learning based facial expression recognition using improved cat swarm optimization," *J. Ambient. Intell. Humaniz. Comput.*, vol. 12, pp. 3037–3053, 2021. doi: 10.1007/s12652-020-02463-4.

[20] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit.*, Columbus, OH, USA, 2014, pp. 1859–1866.
[21] H. Sadeghi and A. A. Raie, "HistNet: Histogram-based convolutional neural network with Chi-squared deep metric learning for facial expression recognition," *Inf. Sci.*, vol. 608, pp. 472–488, 2022. doi: 10.1016/j.ins.2022.06.092.
[22] F. J. Xu, V. Naresh Boddeti, and M. Savvides, "Local binary convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit.*, Honolulu, HI, USA, 2017, pp. 19–28.
[23] L. Li, X. Feng, Z. Xia, X. Jiang, and A. Hadid, "Face spoofing detection with local binary pattern network," *J. Vis. Commun. Image Represent.*, vol. 54, pp. 182–192, 2018. doi: 10.1016/j.jvcir.2018.05.009.
[24] T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (LDP) for face recognition," in *2010 Dig. Tech. Pap. Int. Conf. Consum. Elec. (ICCE)*, Las Vegas, NV, USA, IEEE, 2010, pp. 329–330.
[25] A. R. Rivera, J. R. Castillo, and O. Chae, "Local directional texture pattern image descriptor," *Pattern Recognit. Lett.*, vol. 51, pp. 94–100, 2015. doi: 10.1016/j.patrec.2014.08.012.
[26] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. Third IEEE Int. Conf. on Auto. Face and Gesture Recognit.*, Nara, Japan, IEEE, 1998, pp. 200–205.
[27] 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 *2010 IEEE Comp. Soc. Conf. Comp. Vis. Pattern Recognit.-Work.*, San Francisco, CA, USA, IEEE, 2010, pp. 94–101.
[28] G. Zhao, X. Huang, S. Z. Li M.Taini, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011. doi: 10.1016/j.imavis.2011.07.002.
[29] M. Pantic, M. Valstar, R. Rademaker and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE Int. Conf. Multimed. Expo*, Amsterdam, Netherlands, IEEE, 2005, pp. 5. doi: 10.1016/j.imavis.2011.07.002.
[30] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the mmi facial expression database," in *Proc. 3rd Int. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, vol. 10, 2010, pp. 65–70.
[31] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG facial expression database," in *11th Int. Workshop on Image Analysis for Multimedia Interactive Services WIAMIS*, IEEE, vol. 10, 2010, pp. 1–4.
[32] A. Dhall, R. Goecke, J. Joshi, K. Sikka and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proc. the 16th Int. Conf. Multimodal Interaction*, 2014, pp. 461–466. doi: 10.1145/2663204.2666275.
[33] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial expression databases from movies," *IEEE Multimed.*, vol. 19, pp. 34–41, 2012. doi: 10.1145/2663204.2666275.
[34] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *2015 Int. Jt. Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, IEEE, 2015, pp. 1–8.
[35] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. the IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1805–1812.
[36] C. Turan, K. M. Lam, and X. He, "Soft locality preserving map (SLPM) for facial expression recognition," arXiv preprint arXiv:1801.03754, 2018.
[37] W. Wang *et al.,* "A fine-grained facial expression database for end-to-end multi-pose facial expression recognition," arXiv preprint arXiv:1907.10838, 2019.
[38] X. Sun, P. Xia, L. Zhang, and L. Shao, "A ROI-guided deep architecture for robust facial expressions recognition," *Inf. Sci.*, vol. 522, pp. 35–48, 2020.
[39] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. the IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 2168–2177.
[40] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, pp. 550–569, 2018. doi: 10.48550/arXiv.1609.0642.
[41] R. Zhao, T. Liu, J. Xiao, D. P. Lun, and K. M. Lam, "Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing," in *2020 25th Int. Conf. on Pattern Recognit. (ICPR)*, Milan, Italy, IEEE, 2021, pp. 4412–4419.

[42] H. Zheng *et al.,* "Discriminative deep multi-task learning for facial expression recognition," *Inf. Sci.*, vol. 533, pp. 60–71, 2020. doi: 10.1016/j.ins.2020.04.04.
[43] A. Dosovitski *et al.,* "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[44] Y. Cui, Y. Ma, W. Li, N. Bian, G. Li, and D. Cao, "Multi-EmoNet: A novel multi-task neural network for driver emotion recognition," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 650–655, 2020. doi: 10.1016/j.ifacol.2021.04.155.
[45] D. Gera and S. Balasubramanian, "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition," *Pattern Recognit. Lett.*, vol. 145, pp. 58–66, 2021. doi: 10.1016/j.patrec.2021.01.029.
[46] Z. Qawaqneh, A. A. Mallouh and B. D. Barkana, "Deep convolutional neural network for age estimation based on VGG-face model," arXiv preprint arXiv:1709.01664, 2017.
[47] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. the IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 815–823.