ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0371

FACIAL EXPRESSION RECOGNITION BASED ON FEATURE ENHANCEMENT AND IMPROVED ALEXNET

Himanshukumar D. Nayak[1] and Ashish K. Sarvaiya[2]

[1]Department of Electronics and Communication Engineering, Gujarat Technological University, India

[2]Department of Electronics and Communication Engineering, Government Engineering College, Bhavnagar, India

Abstract

For interpersonal relations among humans, facial expressions are extremely important. Due to the complications in collecting required features from the frequently changing surroundings, uneven reflection from light sources, and many other aspects, facial expression recognition will encounter significant problems. A novel facial image recognition approach is proposed in this paper. Initially, a face image enhancement framework is created that is capable of enhancing the features of a face in a complicated context for this strategy. The improved Alexnet neural network is then created, which is based on the Alexnet architecture. Multi-scale convolution process is utilised in the improved Alexnet to enhance feature extraction capability. Batch normalisation is used for preventing network overfitting while also improving the model's robustness. The Adabound optimizer and the Relu activation function are used to improve convergence and accuracy. The facial image feature improvement approach is helpful to increasing the capability of the improved Alexnet in trials from many aspects. For face images acquired in the natural surroundings, our technique displays significant stability, serving as a benchmark for the intelligent prediction of other facial images.

Keywords:

Facial Expression Recognition, Deep Learning, Convolutional Neural Network, Improved Alexnet

1. INTRODUCTION

Due to the importance of human-computer interfaces, facial expression detection has captivated the scientific community's interest in recent years. Web services, broadcasting, customer satisfaction profiling, video conferencing, machine vision, robotics, computer games, virtual reality, forensics, and border security systems are some of the applications [1]. For facial expression analysis from image sequences and static images, several approaches have been developed [2]. Human emotion recognition relies heavily on facial expression. The terms emotion recognition and face recognition are interchangeable since facial expressions represent a person's emotions. There have been significant developments in interactive gaming, neuro-marketing, robots, augmented reality, automobile industry, and there is an emerging necessity to improve all features of human-computer interaction, especially in the domain of human emotion identification [3]. The expressions on the face are one of the nonverbal communication tools that can be used to determine a person's mood or mental state.

Emotions can be deduced from facial expressions [4]. Normal, afraid, furious, surprise, sad, and happy are six sorts of human expressions that are widely understood. Emotion recognition can be difficult because it involves behavioural responses, feelings, thoughts, and dissatisfaction or pleasure. The study of computer vision and its applications such as counselling systems, tutoring, deception detection, and assessing a person's mental state has shown that automatic facial expression identification has become increasingly intriguing and hard. Existing facial expression recognition in controlled conditions shows encouraging findings, but performance on real-world datasets is still inadequate [5]. This is due to the wide range of lighting conditions, head position, orientation, expression, attitude, illumination, skin colour, and other factors that affect facial appearance. The facial expression recognition method must have a high level of accuracy in both detecting and regaining the precise images that are similar to the supplied input images from the dataset. Furthermore, facial image retrieval must be quick and reliable.

Facial expression recognition (FER) based on facial features can be classified into two categories: learned feature [6] and handcrafted feature [7]. The creation of a good handcrafted feature relies on domain expertise, which necessitates a significant amount of physical labour. When compared to techniques depending on handcrafted features, deep learning outperforms them in feature learning. As a result, various deep learning techniques depending on Convolutional Neural Network (CNN) exist for FER. A sophisticated CNN, on the other hand, has a significant computational cost due to the enormous number of convolutions, as well as high hardware requirements. Meanwhile, current FER facial features can be classified into two groups based on various domains such as the frequency domain [8] and spatial domain [9]. Face expressions can be estimated in the spatial realm using image gradient and geometry. The high-frequency components in the frequency domain correlate to edges and sounds. Low-frequency components, on the other hand, are a thorough measurement of image intensity. As a result, frequency analysis can be used to calculate image attributes. Frequency domain processing is important in traditional image processing because it allows for the reduction of spatial redundancy and efficient calculation.

The number of facial expression recognition studies in the spatial domain so far outnumbers those in the frequency domain. The foregoing observations motivate us to take advantage of the benefits of image processing in the frequency domains and also model a FER system using a deep learning approach depending on frequency. Deep learning is currently attracting the interest of a growing number of researchers. It outperforms shallow models in terms of feature extraction and identification. When compared to machine learning-based classifiers, deep learning-based classifiers can considerably reduce feature extraction uncertainty and enhance recognition accuracy. Simultaneously, it has the potential to raise the cost of computing during training and testing. Computer operational capability has dramatically improved in recent years. Deep learning is now being used by a growing number of researchers to create classifiers. Data with Euclidean structures, including images, audio, and video, can only be processed by traditional CNN networks. In real life, however, there are many non-Euclidean structures, such as data networks and social networks.

The structure of the convolutional neural network (CNN) [10] is extensively used in a range of disciplines. The CNN structure can study the features automatically from the training data when the original data on the face image is directly entered into the network. However, its benefits do not guarantee that it can solve problems dependent on facial expression recognition. Despite the absence of manual feature extraction, the indistinct features and presence of noises in the original face images may still influence the accumulation of error, making it difficult to enhance the network recognition accuracy. This research proposes a facial expression recognition approach based on the improved Alexnet method to address the aforementioned issues. To begin, an image enhancement process for face images is proposed, which has the power of enhancing facial features in a complicated context. Following the preceding process, the improved Alexnet neural networks are used for classifying and recognising the facial images. This process is efficient in achieving a greater accuracy of recognition when compared to the classic neural network and feature extraction methods without the features of image enhancement.

2. LITERATURE REVIEW

Many facial expression recognition algorithms have been acquired in recent years, with an improvement in recognition performance. The advent of Deep Learning techniques [11], notably Convolutional Neural Network [12], has accounted for a significant measure of this recent success. Because of the expanding number of data obtainable for training learning techniques and developments in GPU technologies, many approaches have become possible. Many new facial expression recognition algorithms are proposed for on focusing uncontrolled contexts [13]. These studies will emphasise more controlled surroundings and a comparison of diverse and more difficult facial expression recognition scenarios. This section examines recent methods for facial expression recognition that achieve high accuracy utilising a comparable experimental methodology or methods based on deep neural networks.

A new technique termed Boosted Deep Belief Network (BDBN) was proposed by Liu et al. [14]. The BDBN is made up of a group of weak classifiers, as defined by the authors. Each weak classifier is in charge of categorising a single expression. Their method iterates through the three steps of learning (learning the features, selecting the features, and classifier development) in a unique structure. The investigations are carried out with remarkable accuracy utilising two datasets of static images, JAFFE [15] and Cohn-Kanade [16]. They also used a cross-database arrangement to conduct trials under less controlled circumstances, achieving an accuracy of 68%. All images were first pre-processed using the provided eye coordinates, which included cropping and alignment. The training and testing used the classification technique and each expression was classified using a binary classifier. The network needed to be trained for roughly 8 days. The weak classifiers were used to calculate the recognition. Depending on the number of expressions to be detected, they use six or seven classifiers in their method (one for each expression). Each statement was recognised in 30 milliseconds by each classifier, for a total recognition time of 0.21 seconds. The authors used a 6-core 2.4GHz PC to calculate the recognition time.

Song et al. [17] created a face expression detection process that operates on a smartphone and utilises the deep Convolutional Neural Networks. The proposed network has a total of 65,000 neurons and is made up of five layers. When employing a limited number of training datasets and such large networks, overfitting is common. As a result, the authors used data augmentation strategies to enhance the number of the training dataset, as well as the drop-out [18] methodology while training the network. The studies used the CK+ dataset as well as three other datasets produced by the authors. The CK+ dataset images were initially trimmed to focus on regions with facial alterations induced by expressions. The author’s research used a 10-fold cross-validation method, but they never mentioned whether there were images of the same person in more than one-fold. As a result, we believed that participants in the training and test sets were similar. In the CK+ database, 99.2% accuracy was achieved with only five expressions recognised.

Using a neural network method, several algorithms for facial expression classification have been suggested. To recognise a person’s facial emotions, Krithika and Priya [19] created a graph-dependent method for the extraction of features and classifying the images by utilising a hybrid technique. The canthi detection technique was employed by Tai and Chung [20] to accomplish automatic face emotion identification using the Elman neural network. Sreevatsa et al. [21] used a target-oriented technique to detect facial expressions. A neural network with four layers was created, including a single input layer, a double hidden layer, and a single output layer. The data is consumed by a feed-forward neural network that has been trained for classifying emotions such as happiness, sadness, surprise, disgust, fear, and anger. The recognition rate for disgust, fear and sad images was shown to be lower. Moreover, a modern deep learning-based facial emotion identification method using the CNN network is made up of two convolution layers, each followed by a max-pooling layer and four inception layers [22]. Face images are used as input, and they are classified into one of seven expressions using a single component-based architecture.

Deep learning, particularly the traditional neural network, has recently gotten a lot of interest. CNN is a type of deep learning framework that is used to classify digital images. By fine-tuning the parameters of convolutional and pooling layers, CNN can extract image features automatically. Manual feature extraction is a time-consuming process, and researchers frequently need to conduct numerous experiments to determine whether a given feature is adequate for classification. CNN also has the advantage of being able to learn from large datasets. Hundreds of samples are typically used in the training set of traditional machine learning algorithms. When the training sets become larger, these algorithms converge substantially more slowly, if at all. As a result, CNN is currently frequently used in image categorization.

Facial expression detection approaches based on deep learning reduce dependence on pre-processing and face physics processes [23]. CNN convolution layers use a filter collection to convolve the input image. It produces a feature map to recognise facial expressions, which are subsequently merged into fully connected networks. Yolcu et al. [24] suggested an approach based on deep learning for detecting client interest by predicting head pose and analysing facial expressions. Zhao et al. [25] employed the deep regions and multi-label learning technique that utilises the feed-forward network for extracting relevant facial features. It permits the learnt weight to account for the structural face data. A deep learning algorithm extracts the features from a huge amount of controlled face expression datasets. To improve the performance of the algorithm, pre-processing is employed because of the unavailability of any standard method for choosing the design and learning parameters for a neural network.

Yang et al. [26] employed Generative Adversarial Network (GAN) architecture for predicting the face features. In contrast to typical CNN techniques, the ACNN [27] includes an attention mechanism for focusing on the most significant segment of the face. It employed the standard CNN structure, which consists of four convolutional layers for extracting the features. On the significant facial regions, two additional convolutional layers will be applied for implementing the attention processes. A softmax layer and many fully linked layers are utilised for categorization. The Region Attention Network was utilised by Wang et al. [28] for addressing occlusion and pose variation for FER to adaptively extract the relevant region of the face. The methods mentioned above are geared toward extracting spatial features from images. However, a complicated CNN-based technique necessitates a substantial amount of processing due to the enormous number of convolutional kernels. Unlike earlier techniques, we concentrate on researching frequency-domain feature learning and increasing FER performance.

In the process of recognition, deep learning has achieved significant progress [29]. The relevant feature representations will be extracted from a huge number of input images. CNN is widely considered to be the most effective algorithm for extracting high-level semantic characteristics. The CNN can produce higher image classification performance by extracting more abstract information layer by layer from images. As a result, it outperforms the competition in image classification. VGG [30], ResNet [31], Google Net [32], and other CNNs are examples. The network structure of these models, on the other hand, is quite complex, necessitating a vast number of images. Obtaining a large number of images is also difficult because the networks based on CNN are very deep to train with only a minimal amount of data. To overcome this problem, a new better Alexnet architecture is proposed. The proposed method extends the network structure based on LeNet, learns rich and high dimensional features in images, with 3 fully connected layers, 5 convolutional layers, Dropout for suppressing over-fitting, and Relu as the activation function.

3. PROPOSED WORK

3.1 IMAGE PRE-PROCESSING

The first stage in the suggested method is image pre-processing. The facial images are first pre-processed for removing any noise, redundant or data distortion. Preprocessing the main purpose is to improve an image with undesirable distortions or characteristics and prepare it for processing further. Many phases of pre-processing will be done for improvement. In computer vision, extracting facial features is crucial, and locating the fiducial points in the face is the initial step [33]. Face landmark point recognition is the initial stage in detecting critical spots in the human faces and mapping the input images to certain facial structures. Many facial landmark points are automatically recognised from an input face image, followed by face alignment [34]. Image Resize aims to normalise all of the faces in a dataset to address the diversity of faces caused by different genders, races, and data collecting environment factors. This technique ensures that the magnitudes of geometric feature elements are distributed in a stable manner in feature space.

Face alignment is the process of rotating the face in the image plane in an upright position based on the eye location. The uttermost landmark point in the right and left ends are used to calculate the face centre point. Similarly, the centre points for the mouth, eyes, and nose are determined in those areas by averaging landmark points. The size of the input image is reduced to a single dimension with similar width and height for all of the face image datasets before being supplied as input to the feature extraction procedure. As a result, the facial landmark detection stage is critical for face identification, expression analysis, age prediction, and gender categorization since it aids in the alignment of facial images to a meaningful form. For removing the noises from the images, a nonlinear technique known as the median filter is applied in this process. The edge pixels are preserved when the images are smoothed. This filter minimises the intensity variations by switching the pixels in the image with the average value [35]. After the input image has been pre-processed, it is sent to the feature extraction step as an input.

3.2 FEATURE EXTRACTION USING ALEXNET

By raising both network parameters and network depth, deep learning usually enhances recognition accuracy. However, just increasing the network size can result in overfitting and increased computing complexity. To overcome this issue, the Multi-scale convolution model is designed. The model will become inefficient only if the convolution cores of varied sizes are utilised, due to the large number of new parameters introduced. A conventional neural network consists of a series of convolution layers, each with a single convolution core size. In general, a single feature map can capture features of various scales using several convolution kernels of varying sizes before combining them. Deep CNN architecture learning capacity was limited by hardware limitations, which limited their size. Alexnet was trained in parallel to overcome hardware limitations to make use of deep CNN representational capacity.

AlexNet is regarded as the first deep CNN architecture to achieve breakthrough results in image classification and recognition. Krizhevsky et al. [36] proposed AlexNet, which improved the CNN learning capacity by making it deeper and utilising multiple parameter optimization procedures. CNN depth was increased from 5 to 8 layers in AlexNet, allowing it to be used for a wide range of image types. Even though increasing depth increases generalisation for various image resolutions, the fundamental disadvantage of increasing depth is overfitting. To deal with this problem, the method skips some transformational units at random during training to drive the model to learn more robust features. Furthermore, ReLU was used as a non-saturating activation function to boost the convergence rate by somewhat relieving the vanishing gradient problem. In comparison to prior proposed networks, the employment of large size filters at the initial layers was also changed.

AlexNet is significant in the new generation of CNNs due to its efficient learning approach, and it has ushered in a new era of study in CNN architecture developments. To reduce the complexity of the network and to increase the learning process, multi-scale convolution neural networks depending on Alexnet are presented. The Adabound optimizer and the Relu activation function are used to create the Alexnet. The Alexnet can help to increase the network model robustness and identification accuracy and also reduce over-fitting. The Alexnet network structure contains three fully-connected layers, a Multi-scale convolution module, and five convolution layers. The initial convolution layer employs void convolution to enable a huge range of feature extraction. Multi-scale convolution uses a multi-scale convolution kernel for obtaining characteristics at several scales while lowering network computation costs. The classification function is provided by the three fully linked layers, while the final convolution layer incorporates the previously extracted properties. This model contains three fully connected layers, one multi-scale convolution module, and five convolution layers. The BN (Batch normalisation) layers follow each convolution layer, and they can progress the generalisation ability of the model, prevent over-fitting, and raise the robustness of the network. For obtaining a broader receptive field, the initial convolution layer employs dilated convolution. The multi-scale convolution model is positioned before the final convolution layer depending on the collective experiences.

3.2.1 Batch Normalization:

Batch normalisation is a common technique for training deep neural networks (DNNs) more quickly and consistently. Deep Neural Network training is complicated since the distribution of each input layer alters during training as there will be changes in the parameters of the previous layers. Because this needs careful parameter setup and lower learning rates, training models with saturating nonlinearities is notoriously difficult. Internal covariate shift is the term we use to describe this phenomenon, and we solve the problem by normalising layer inputs. The effectiveness of our strategy comes from including normalisation into the model architecture and performing it for each training mini-batch. Batch Normalization enables us to use much greater learning rates and be less cautious with initialization, and it can even eliminate the necessity for Dropout in some circumstances.

Gradient flow across the network is further improved by batch normalisation since gradients are less dependent on the scale of their initial values or the parameters. This eliminates the possibility of divergence by allowing us to employ considerably larger learning rates. Furthermore, batch normalisation makes the model more consistent and eliminates the necessity for Dropout. Finally, Batch Normalization allows saturating nonlinearities to be used since it prevents the network from becoming stuck in saturated modes. Face image characteristics are complicated and changeable. Neural networks have a slow or non-existent learning rate. Meanwhile, the hidden layer data distribution has seen fluctuations and significant changes as the neural network structure deepens, which harms the network stability. In this work, the BN approach is used to normalise each layer of data to a standard deviation of 1 and a mean of 0. This is done to ensure data stability, make deep network model training easier and more stable, and improve network generalisation capabilities. The BN method estimates the variance and means of each batch sample as shown in Eq.(1).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \tag{2}$$

where σ and μ represent the standard deviation and mean of each batch sample X. It will be followed by batch normalization.

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} \tag{3}$$

where ε is defined as the constant for preventing the fractal from failing if the standard deviation falls to zero.

3.3 ACTIVATION FUNCTION

In the facial image, the range of intensity of pixel signal is particularly large due to interference induced by dust, fog and sunlight. The sigmoid function is frequently utilised in traditional networks, can delay the gradient changes in the saturated zone and causes it to progressively approach zero, resulting in gradient disappearance. In neural networks, the activation function was employed to give nonlinearity. As a result, typical activation function options include logistic, tanh, and arctan functions, among others.

However, in deep models, these functions are prone to the gradient vanishing problem, as the gradient is only a large value when the input is within a narrow range. A novel activation function, the rectified linear unit, was employed to solve this problem. The Relu function converges quicker than the Sigmoid and Tanh functions. The output is the fundamental quality in the ReLU activation function of the alexnet method for facial images. It is repeated several times till the images have been spatially merged to a small size. ReLU is the Half-wave rectifier method that can significantly speed up the training stages while also avoiding over-fitting of the output images update. The Eq.(4) explains how this activation function works.

$$A_{RELU}\left(m\right) = \begin{cases} m & if\ m > 0 \\ 0 & else \end{cases} \tag{4}$$

If the input is not less than 0, the ReLU gradient is always 1. Deep networks utilising ReLU as an activation function have been shown to converge quicker than tanh units. The training was substantially aided by this acceleration. In each hidden layer of the neural network, this function serves as an activation function, as well as a classification capability in the network’s final layer. Each ResBlock module simply contains a ReLU because the network will learn changes to bilinear up the sampled Image. The network will learn changes to bilinear up the sampled Image later with the main convolution, which could be destructive. The layers of DCNN will be improved as well as the exceptional restoration qualities of medical images, by batch normalising and optimising its coordinates in the Alexnet approach.

Then, to avoid over-fitting, dropout was used. It is most common in fully connected layers. Only a portion of the neurons was taught in each repetition of dropout. When the ratio is set to 50%, for example, just half of the parameters are taught in each iteration. Dropout forces a neuron to work along with others, which lowers joint adaptation and enhances generalisation. The network can be divided into multiple sub-networks, each with its dropout. Though each sub-network may be over-fitted to some extent, they all have the same loss function. The output of the total network was calculated by averaging the output of the sub-networks. As a result, dropout strengthened the robustness.

3.3.1 Network Optimization:

Adam and SGD are usually employed as optimisers in the deep learning process. The convergence pace of SGD is modest in the early stages of the training network. Meanwhile, with SGD, determining an adequate learning rate for the goal of facial emotion recognition is difficult. SGD has the potential to give poor outcomes and slow training speeds if the training dataset is insufficient. SGD is also receptive to convergence to local optima and may become stuck at the saddle point. Gradient updating can also be made more flexible using the optimization methods given by Nesterov and Momentum [37]. Regardless of this, an artificially established learning rate is more difficult for operating than adaptive learning rates. Adam has become the default strategy for several deep learning systems [38] because of its fastest training speed. Due to the intricate and constantly changing shooting environment of face images and the substantial variance in diseases in different phases, the learning rate of Adam's adaptive technique is inconsistent.

Adam is a popular deep learning optimization algorithm that offers a faster convergence rate than stochastic gradient descent (SGD). Adam's learning rate in the later training period, on the other hand, is low, which has an impact on effective convergence [39]. Furthermore, the Adam method has the potential to overfit early-stage features, making it harder for later-stage features to compensate for the previous fitting effect. Luo et al. [40] introduced the Adabound optimization technique to address these restrictions. The Adabound uses a dynamic learning rate range to enable a gradual, seamless transition from Adam to the SGD, avoiding the consequences of extreme learning rates. In addition, the Adabound has a fast-learning rate at the initiation of training and an excellent convergence outcome at the conclusion. As a result, the Adabound optimization technique is used in this study to learn the recommended model parameters.

To overcome SGD's slow convergence and Adam's weak generalisation capabilities, the Adabound optimizer is chosen based on the properties of the face image. Adabound is an optimiser that starts as Adam and advances through training to become SGD. When it comes to dealing with non-linear goals and sparse data processing, Adabound easily surpasses the competition. Furthermore, it necessitates less memory. For each parameter, the self-adaptive learning rate is established, which is beneficial for high-dimensional space, non-convex optimisation, and large datasets. The following is the Alexnet model that has been proposed.

An input layer and the Batch Normalisation layer make up the first layer. The input image is processed by the BN layer, which improves the generalisation capability of the model and speeds up the convergence speed of the network.

Convolution module 1 (Conv1), which has a kernel size of 11×11 pixels and 64 filters, makes up the second layer. To increase the feature extraction capabilities of face expressions, Conv1 is extended by employing dilated convolution. Furthermore, the ReLu activation function is used to deal with the problem of disappearing gradients. Pooling layer 1 is the maximum pooling layer, with a kernel of 3×3 and a stride of 2. The BN processing is done after the pooling.

Convolution module 2 (Conv2) is the third layer, and it is made up of kernel sizes of 5×5 pixels and 192 filters that use the Relu operation. The highest pooling layer is pool layer 2, with a 3×3 kernel and a stride of 2. BN is carried out after pooling. Convolution module 3 (Conv3), which consists of a kernel size of 3×3 pixels and 384 filters and is supported by the Relu operation, makes up the fourth layer. The maximum pooling layer is pooling layer 3, with a kernel of 3×3 and a stride of 2. After pooling, BN is carried out.

Convolution module 4 (Conv4), which consists of a kernel size of 3×3 pixels and 256 filters and is supported by the Relu operation, makes up the fifth layer. Batch Normalization is conducted after pooling.

The sixth layer is made up of a multi-scale convolution module with 96 and 16 filters, respectively, in the first layer of multi-scale convolution and 1×1 kernel size. Relu is the activation function. From upper to lower layer, the second layer sum of multi-scale convolution is 64, 128, 128, and 128, with kernel sizes of 1×1, 3×3, 5×5, and 1×1, respectively. The concatenate layer processes it after it has been collected.

Convolution module 5 (Conv5) is the seventh layer, and it consists of a kernel size of 3×3 pixels and 256 filters. The Relu operation is applied, and the BN layer is processed for the collection, as mentioned before.

A ReLU and a dropout operation are used to process the first fully connected layer, which has 200 neurons. A ReLU operation and a dropout operation are used to process the second fully connected layer, which has 100 neurons.

The final entirely connected layer, which represents the number of face expression types, is made up of seven neurons. After that, the output of the last completely connected layer is passed to the output layer, which determines the categorization of the input image. At last, a softmax activation function will be utilised, resulting in a total output value of 1.0 and a single output value that is confined to a range of 0-1. The softmax function is an excellent fit for Alexnet since it compensates for the comparative magnitude of all outputs. The Fig.1 gives the structure of improved Alexnet.

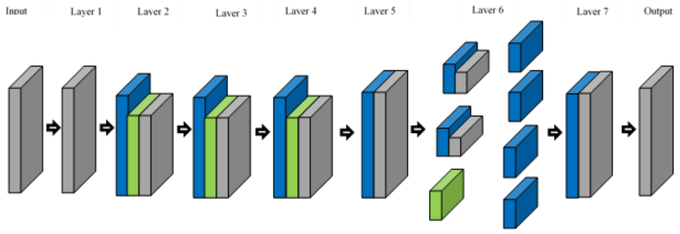


Fig.1. The Structure of Improved AlexNet

4. EXPERIMENTAL EVALUATION

The yale and face recognition datasets are used to test and evaluate the proposed facial expression recognition system. It uses crucial procedures including pre-processing, feature extraction, and classification to recognise expressions. Higher accuracy, faster retrieval, and less calculation are all advantages of the suggested approach. The suggested system input can be taken from the real-time images or the dataset. The image is pre-processed after it has been loaded from the dataset. Using a pre-trained model, the position of landmark points on the facial image will be mapped with coordinates. Based on these landmark points, the face is scaled, cropped, and aligned for further processing. Alexnet is used to extract features from the images using convolution layers and filters, resulting in dimensional feature vectors. By selecting the best features from the retrieved features, the feature selection strategy decreases the feature vector. The experimental process is carried out in the Matlab platform.

4.1 DATASET DESCRIPTION

The Table.1 show the description of the dataset utilised for the suggested facial expression recognition method. Two types of datasets are utilised in the proposed system namely the Yale faces and the Face expression recognition dataset.

Table.1. Dataset Description

Dataset name	Number of classes	Total number of images
Yale’s faces	11	165
Face expression recognition	7	288821

4.2 PERFORMANCE EVALUATION

To properly report the performance of our model, we use 10-fold cross-validations for determining the generalisation performance of our approach. Cross-validation is performed on the entire dataset, including both the train and test sets. The information is randomly jumbled and divided into ten subsets with a 90:10 percent split. This subdivision is used to train the model on 9 of the 10 subgroups. After the model has been trained, it is estimated on the remaining data set. The ultimate accuracy is calculated as the average of all achieved accuracies. The Fig.2 and Fig.3 demonstrate the input image imported from the dataset for the suggested facial expression recognition process. It is the outcome of image processing that finds every similar image that matches the input images. The Fig.2 gives the sample images of the yale faces dataset and Fig.3 gives the sample images of the facial expression recognition dataset.

Fig.2. Sample images of Yale face dataset

Fig.3. Sample images of the Face expression recognition dataset

Our Generic Model is the proposed model with batch normalisation and feature extraction. In this method, the regularising factor is set to 0.0005 and the learning value is set as 0.001. The batch size might range from 16 to 512 items. The maximum number of training iterations is 1000. The trained and tested datasets are mirrored, and any misclassified trained dataset must be deleted before they are used while testing samples with prediction confidence of less than 90% should not be considered when adjusting the model with the proposed method. The Table.2 gives the accuracy for different batch sizes in a different database. In addition, the suggested algorithm is compared to other existing methods such as ResNet and VGG16 to assess its efficiency.

Table.2. Accuracy for different batch sizes in different database

Dataset	Batch size	Proposed	ResNet	VGG16
YALE	16	0.7143	0.5325	0.4026
FER		0.6883	0.4286	0.3117
YALE	32	0.7273	0.5584	0.4026
FER		0.7143	0.5974	0.4156
YALE	48	0.7403	0.5455	0.4286
FER		0.7273	0.5325	0.3377
YALE	64	0.7792	0.5844	0.4675
FER		0.7922	0.6234	0.4805
YALE	80	0.8052	0.6753	0.5455
FER		0.8182	0.6494	0.5195
YALE	96	0.8571	0.6753	0.5584
FER		0.8442	0.6883	0.6364
YALE	112	0.8961	0.7922	0.6623
FER		0.8701	0.7273	0.6364
YALE	128	0.8831	0.7792	0.6753
FER		0.8961	0.7922	0.6494
YALE	144	0.9091	0.8052	0.7273
FER		0.9221	0.8182	0.7143
YALE	160	0.9351	0.8442	0.7922
FER		0.7143	0.5325	0.4026
YALE	176	0.6883	0.4286	0.3117
FER		0.7273	0.5584	0.4026
YALE	192	0.7143	0.5974	0.4156
FER		0.7403	0.5455	0.4286
YALE	208	0.7273	0.5325	0.3377
FER		0.7792	0.5844	0.4675
YALE	224	0.7922	0.6234	0.4805
FER		0.8052	0.6753	0.5455
YALE	240	0.8182	0.6494	0.5195
FER		0.8571	0.6753	0.5584

YALE	256	0.8442	0.6883	0.6364
FER		0.8961	0.7922	0.6623

The experimental results in the table showed the improved recognition accuracy for different datasets. For all batch sizes, the suggested algorithm has greater recognition accuracy than other current algorithms. When the batch size in the datasets is raised, the accuracy increases, as seen in the table. The suggested technique minimises the distance between the feature of the testing sample and the centroid of the most relevant category, preferably the most important feature of the trained dataset. As a result, individual prejudice can be significantly reduced.

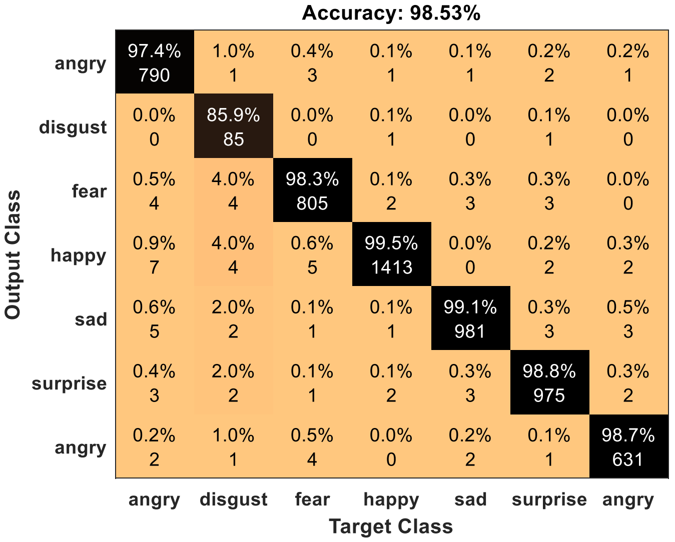


Fig.4. Confusion matrix of Face expression recognition dataset for the proposed method

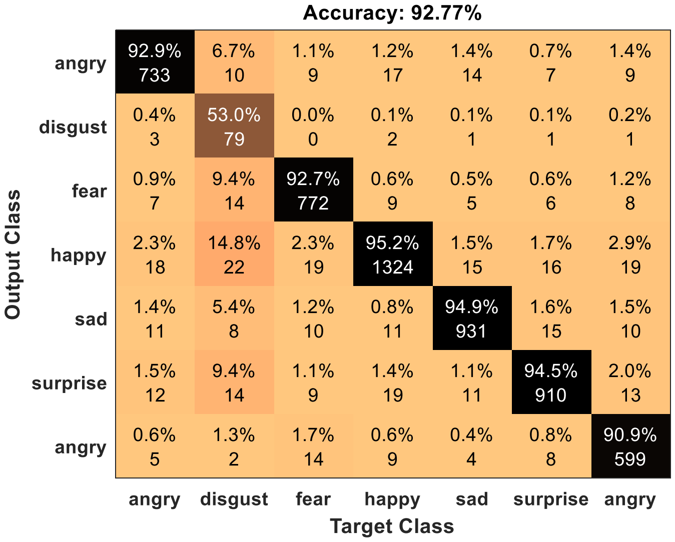


Fig.5. Confusion matrix of Face expression recognition dataset for ResNet method

The confusion matrix of the proposed method and the other existing methods like ResNet and VGG16 method for facial expression recognition dataset is shown in Fig.4 - Fig.6. Facial expressions such as anger, disgust, fear, happiness, sadness and surprise are considered for the experiment. The results reveal that the category related to happiness is predicted correctly each time in these three methods since its feature is obvious and its training data is substantial. On the other hand, the category related to disgust is the least predicted category in all the three methods. Generally, the mouth is the most important component of facial emotions. Here, several disgust expressions are misclassified in the proposed method because they do not depict the character of the mouth. Also, the proposed method has gained higher accuracy of 98.53% compared to other existing methods. The ResNet method has occurred with 92.77% accuracy and the VGG16 method has gained 90.98% accuracy.

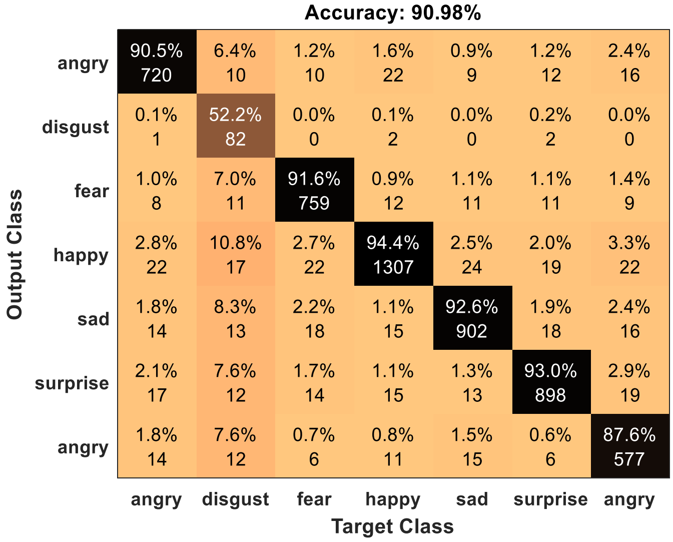


Fig.6. Confusion matrix of Face expression recognition dataset for VGG16 method

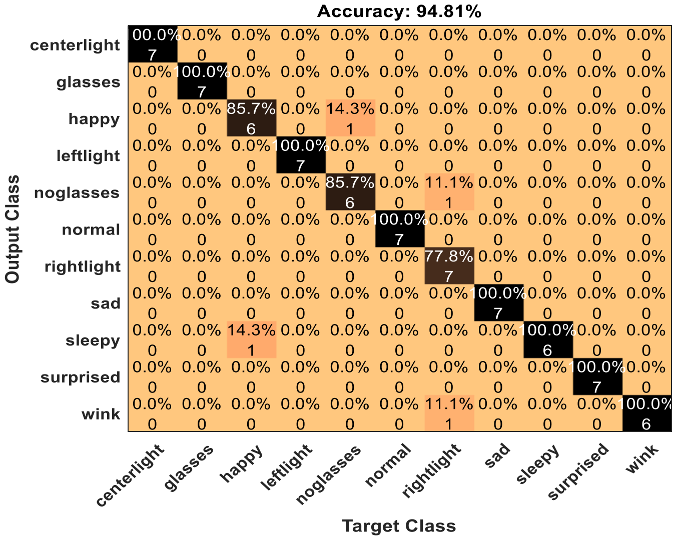
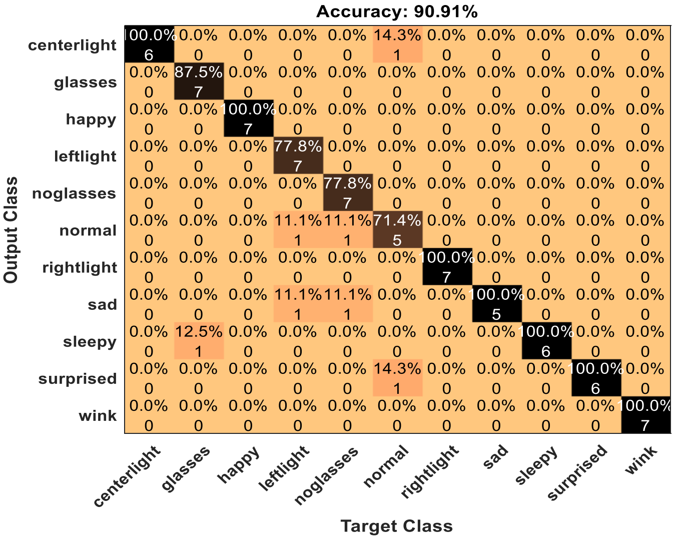


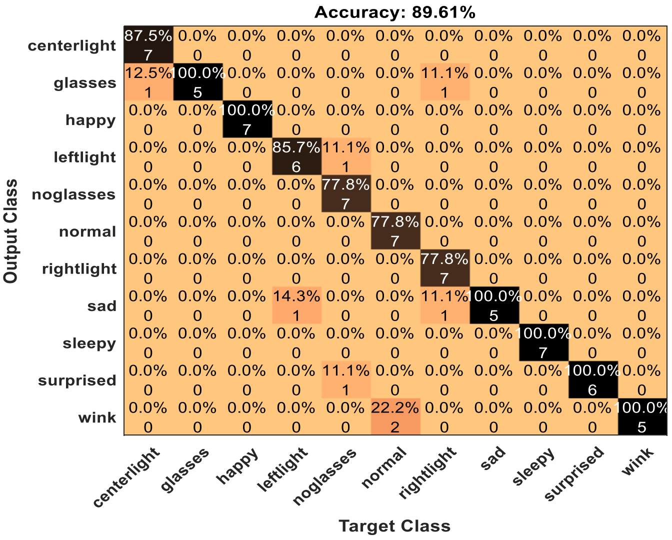
Fig.7. Confusion matrix of Yale faces dataset for the proposed method

The confusion matrix of the proposed method and the other existing methods like ResNet and VGG16 method for the Yale dataset is shown in Fig.7 - Fig.9. The facial expressions such as centre light, glasses, happy, left light, no glasses, normal, the right light, sad, sleepy, surprised and wink are considered for the experiment. Here, the proposed method has gained higher accuracy of 94.81% compared to other existing methods. The ResNet method has occurred with 90.91% accuracy and the VGG16 method has gained 89.61% accuracy. According to the experiments, all three types of comparative algorithms can assist in improving the performance of a model in specific cases. For the overall recognition accuracy, the proposed algorithm works the best.



	centerlight	glasses	happy	leftlight	noglasses	normal	rightlight	sad	sleepy	surprised	wink
centerlight	100.0% 6	0.0% 0	0.0% 0	0.0% 0	0.0% 0	14.3% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
glasses	0.0% 0	87.5% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
happy	0.0% 0	0.0% 0	100.0% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
leftlight	0.0% 0	0.0% 0	0.0% 0	77.8% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
noglasses	0.0% 0	0.0% 0	0.0% 0	0.0% 0	77.8% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
normal	0.0% 0	0.0% 0	0.0% 0	11.1% 1	11.1% 1	71.4% 5	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
rightlight	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0
sad	0.0% 0	0.0% 0	0.0% 0	11.1% 1	11.1% 1	0.0% 0	0.0% 0	100.0% 5	0.0% 0	0.0% 0	0.0% 0
sleepy	0.0% 0	12.5% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 6	0.0% 0	0.0% 0
surprised	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	14.3% 1	0.0% 0	0.0% 0	0.0% 0	100.0% 6	0.0% 0
wink	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 7

Fig.8. Confusion matrix of Yale faces dataset for ResNet method



	centerlight	glasses	happy	leftlight	noglasses	normal	rightlight	sad	sleepy	surprised	wink
centerlight	87.5% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
glasses	12.5% 1	100.0% 5	0.0% 0	0.0% 0	0.0% 0	0.0% 0	11.1% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0
happy	0.0% 0	0.0% 0	100.0% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
leftlight	0.0% 0	0.0% 0	0.0% 0	85.7% 6	11.1% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
noglasses	0.0% 0	0.0% 0	0.0% 0	0.0% 0	77.8% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
normal	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	77.8% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
rightlight	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	77.8% 7	0.0% 0	0.0% 0	0.0% 0	0.0% 0
sad	0.0% 0	0.0% 0	0.0% 0	14.3% 1	0.0% 0	0.0% 0	11.1% 1	100.0% 5	0.0% 0	0.0% 0	0.0% 0
sleepy	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 7	0.0% 0	0.0% 0
surprised	0.0% 0	0.0% 0	0.0% 0	0.0% 0	11.1% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 6	0.0% 0
wink	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	22.2% 2	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 5

Fig.9. Confusion matrix of Yale faces dataset for VGG16 method

The Fig.10 gives the accuracy of the Yale faces database in terms of batch size. The results indicate that the proposed algorithm has higher accuracy than the other existing methods used for comparison. Also, in each algorithm, the accuracy has been increased when the batch size of the algorithm is increased.

The Fig.11 gives the accuracy of the facial expression recognition database in terms of batch size. The results indicate that the proposed algorithm has higher accuracy than the other existing methods used for comparison. Also, in each algorithm, the accuracy has been increased when the batch size of the algorithm is increased.

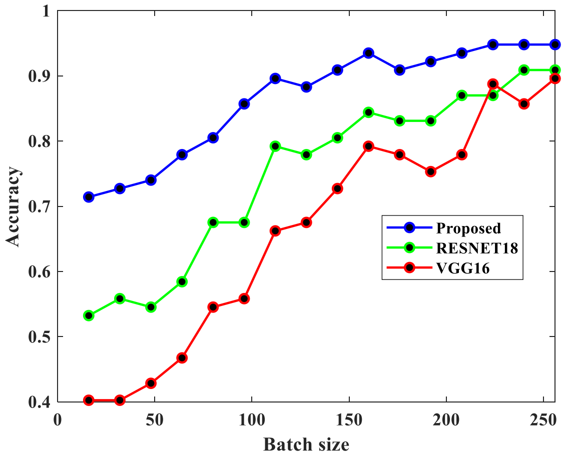


Fig.10. Yale Faces Database Accuracy vs. Batch Size Graph

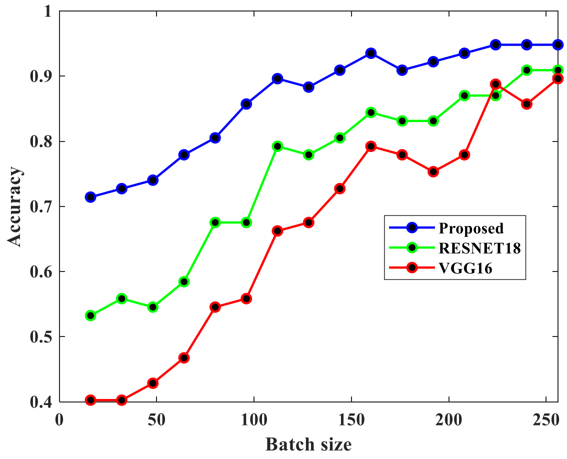


Fig.11. Face expression recognition database accuracy vs batch size graph

The Fig.12 gives the training accuracy for different algorithms and Fig.13 gives the training loss of different algorithms. After the learning procedure, the approaches mentioned are all static. They can perform well in general circumstances if adequate data is used for training, but their performance in specialised testing, such as the experiment findings, will be relatively low.

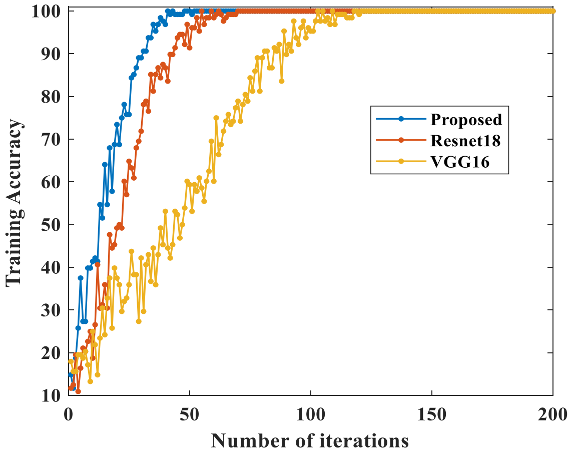


Fig.12. Training accuracy for different algorithms

The proposed strategy in this study transfers the feature space of testing samples as closely as feasible to that of training data. To equalise the image count across all categories, the category with fewer images is enhanced before merging the training data by duplicating the images selected randomly. This ensures that each class has the same total amount of images. The extracted features of training samples must be kept in the feature database after the Alexnet model has been trained.

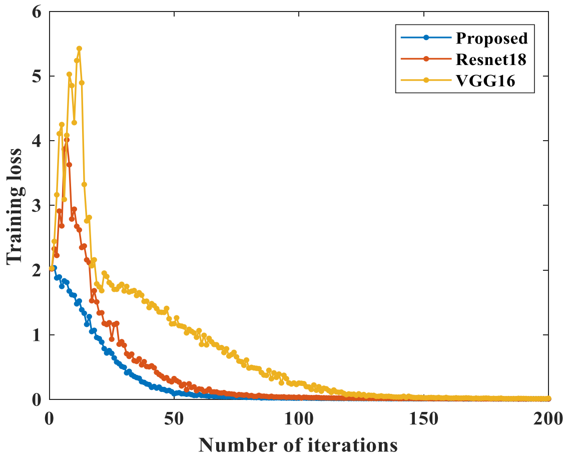


Fig.13. Training Loss of Different Algorithms

5. EVALUATION PARAMETERS

For the evaluation process, the parameters like F1 score, precision, sensitivity, specificity, and accuracy, are considered. The number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are used to calculate the assessment parameters (FP). The right values will be forecasted as correct in true positive (TP). The proper values will be predicted as wrong values in the true negative (TN). The false values will be predicted as the right value in false positive (FP), while the false values will be forecasted as the wrong value in false-negative (FN).

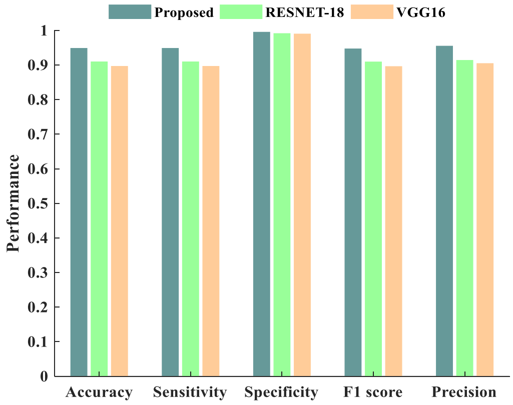


Fig.14. Performance Measure of Yale Faces Dataset

The parameters utilised for evaluating the performance of the proposed method are specificity, accuracy, sensitivity, precision and F1 score. The evaluation process is done for the proposed method and other existing algorithms like ResNet-18 and VGG16. The Fig.14 gives the performance measure of the Yale faces dataset for different face expression recognition algorithms. The proposed method has a better performance measure than the other existing algorithms employed for comparison, according to the experimental data. The Fig.15 shows the performance of different face expression recognition methods on a face recognition dataset. The proposed method has a better performance measure than the other existing algorithms employed for comparison, according to the experimental data.

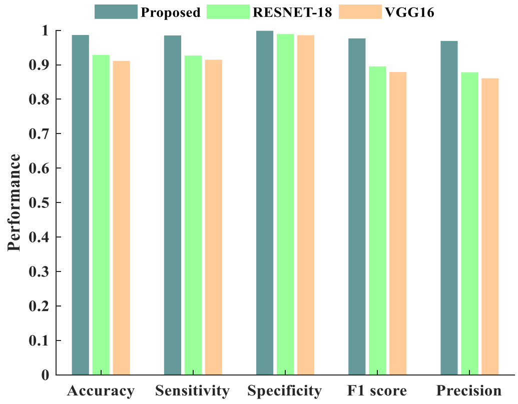


Fig.15. Performance Measure of Face Recognition Dataset

The false-positive rate is the probability of incorrectly discarding the null hypothesis for a given test. The ratio of false positives (negative measures wrongly classified as positive) to the total number of negative measures, irrespective of classification, is used to compute the false positive rate. The Fig.16 gives the false positive rate of proposed algorithms and other existing algorithms for the yale face dataset and face recognition dataset. For the algorithm to perform well, the false positive rate must be kept to minimal. As a result, the experimental data show that the suggested method outperforms other current methods.

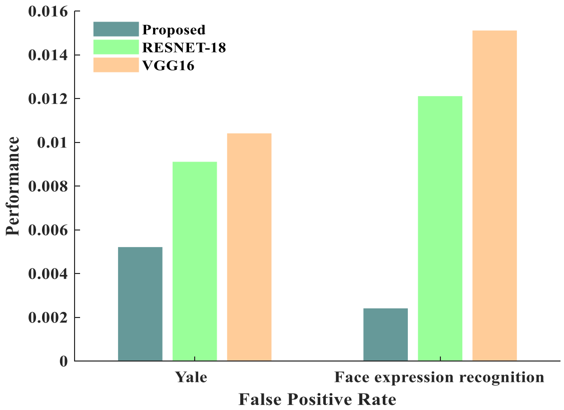


Fig.16. False-Positive Rate of Different Datasets

Our model gradCAM saliency maps are shown in Fig.17. The human face has non-symmetrical facial expressions, indicating that the proposed method is learning to distinguish true anatomic traits rather than sampling artefacts. They demonstrate that the model concentrates on facial expressions such as anger, surprise, fear, sad, happiness and normal. These findings show that the method employs the same areas for gender classification that forensics professionals do, indicating that the method learned which sections are critical for gender categorization and the process of interpreting them for accurate classification. The model was not provided with any prior information about those areas.

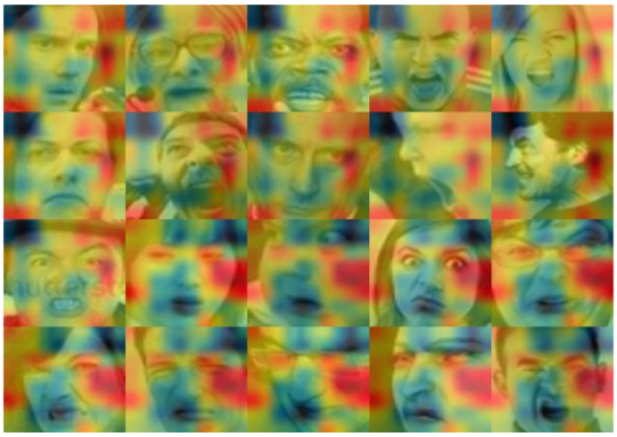


Fig.17. Grad Cam Results

Table.3. Accuracy Comparison with different optimizers

Optimizer	**Accuracy**
Adagrad	0.9712
RMSProp	0.9556
Adam	0.9783
Adabound	0.9853

The Table.3 gives an accurate comparison of different optimizers such as adagrad optimiser, RMSProp, Adam and Adabound optimizer. In the proposed model we utilised the Adabound optimizer for the optimization of the algorithm. From the table, it is known that the Adabound optimizer has gaithe maximum accuracy. Hence, the Adabound optimizer can be effectively utilised for the proposed approach. Meanwhile, the RMSProp optimizer has attained the least accuracy among the comparative algorithms. Apart from the Adabound optimizer and the Adam optimizer and the Adagrad optimizer have obtained the highest accuracy. However, since the Adabound optimizer has obtained the maximum accuracy, it is most suitable for the proposed method.

We present a facial emotion detection system in this research that combines traditional methods, such as improved Alexnet, with specialised pre-processing processes in images. Experiments have shown that combining the normalising techniques improves the accuracy of the proposed method greatly. As illustrated in the result section, our strategy obtains remarkable outcomes and gives a simple solution when compared to existing algorithms which utilise similar facial expression data and experimental processes. Furthermore, it takes less time to train and performs real-time recognition. It indicates that the proposed approach functions better even in unfamiliar situations, in which the image capture subjects and conditions differ from the training images, but there is still potential for development. Additionally, the improved AlexNet is intended to reduce the requirement for hand-coded functionalities. This occurs since the neural network model has the ability for learning the feature set that successfully develops the desired classification. Convolutional Neural Network requires a significant number of data to do such learning. Deep architectures are constrained by the high number of parameters that must be adjusted during training. To solve this issue, pre-processing techniques were conducted in the dataset for reducing the variation among images and choosing a subset of the characteristics to be learned, reducing the quantity of data required. Preliminary tests were carried out with deeper architectures that had been trained with a large number of datasets. A deep CNN termed as improved Alexnet has been suggested and briefly investigated in these trials to recognise faces. The already learned model was fed into a basic double-layered neural network trained with the Yale faces and facial recognition database as a pre-trained feature extraction method. A preprocessing operation was also used in this experiment. The experimental process indicates that the proposed method has improved the accuracy of facial expression recognition. Hence, it shows that the proposed method can be successfully utilised for developing a discriminative method for recognising facial expressions which can even perform in uncontrolled circumstances as it is the main issue recently in this field.

6. CONCLUSION

As machine vision technology and computer technology have improved, professionals and researchers both domestic and overseas have performed substantial research on image analysis technologies. Face expression detection depending on deep learning has had extensive application in the domain of face image recognition in recent years. To increase the accuracy and effectiveness of facial expression identification, this research proposes an approach depending on image enhancement and improved Alexnet for recognising facial expressions. This method exhibits the ability to recognise and distinguish between various facial expressions in various images. Initially, the features of facial images are improved. After that, an enhanced Alexnet is built for high-precision recognition and classification. The proposed strategy eliminates the necessity to choose specific traits, as evidenced by the experimental findings. As a result, it is thought to be a good fit for identifying facial expressions in images. Moreover, it can achieve the preferred effect and accuracy, and this method is simple to manage after training. More forms of facial expressions will be identified in the future, to make it easier to evaluate face expression information quickly and accurately.

REFERENCES

[1] F. Bourel, C.C. Chibelushi and A.A. Low, “Recognition of Facial Expressions in the Presence of Occlusion”, *Proceedings of British Conference on Machine Vision*, pp. 1-10, 2001.

[2] P. Ekman, “Facial Expressions of Emotion: An Old Controversy and New Findings”, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, Vol. 335, No. 1273, pp. 63-69, 1992.

[3] H. Sikkandar and R. Thiyagarajan, “Deep Learning based Facial Expression Recognition using Improved Cat Swarm Optimization”, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 2, pp. 3037-3053, 2021.

[4] H. Ali, M. Hariharan, S. Yaacob and A.H. Adom, “Facial Emotion Recognition using Empirical Mode Decomposition”, *Expert Systems with Applications*, Vol. 42, No. 3, pp. 1261-1277, 2015.

[5] H. Wu, Y. Liu, Y. Liu and S. Liu, "Efficient Facial Expression Recognition via Convolution Neural Network and Infrared Imaging Technology", *Infrared Physics and Technology*, Vol. 102, pp. 103031-103039, 2019.

[6] Y. Tang, X. M. Zhang and H. Wang, "Geometric-Convolutional Feature Fusion based on Learning Propagation for Facial Expression Recognition", *IEEE Access*, Vol. 6, pp. 42532-42540, 2018.

[7] L. Ma, "Facial Expression Recognition using 2-D DCT of Binarized Edge Images and Constructive Feedforward Neural Networks", *Proceedings of IEEE International Joint Conference on Neural Networks*, pp. 4083-4088, 2008.

[8] S. Mohseni, H. M. Kordy and R. Ahmadi, "Facial Expression Recognition using DCT Features and Neural Network based Decision Tree", *Proceedings International Conference on Electronics in Marine*, pp. 361-364, 2013.

[9] H. Jung, S. Lee, J. Yim, S. Park and J. Kim, "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition", *Proceedings of IEEE International Conference on Computer Vision*, pp. 2983-2991, 2015.

[10] A. Khan, A. Sohail, U. Zahoora and A.S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks", *Artificial Intelligence Review*, Vol. 53 No. 8, pp. 5455-5516, 2020.

[11] P. Liu, S. Han, Z. Meng and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805-1812, 2014.

[12] P. Burkert, F. Trier, M.Z. Afzal, A. Dengel and M. Liwicki, "Dexpression: Deep Convolutional Neural Network for Expression Recognition", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1110-1117, 2015.

[13] M. Liu, S. Li, S. Shan and X. Chen, "Au-Inspired Deep Networks for Facial Expression Feature Learning", *Neurocomputing*, Vol. 159, pp. 126-136, 2015.

[14] P. Liu, S. Han, Z. Meng and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805-1812, 2014.

[15] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets", *Proceedings of 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200-205,1998.

[16] 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", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 94-101, 2010.

[17] I. Song, H.J. Kim and P.B. Jeon, "Deep Learning for Real-Time Robust Facial Expression Recognition on a Smartphone", *Proceedings of IEEE International Conference on Consumer Electronics*, pp. 564-567, 2014.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929-1958, 2014.

[19] L.B. Krithika and G.L. Priya, "Graph based Feature Extraction and Hybrid Classification Approach for Facial Expression Recognition", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 2, pp. 2131-2147, 2021.

[20] S.C. Tai and K.C. Chung, "Automatic Facial Expression Recognition System using Neural Networks", *Proceedings of International Conference on TENCON*, pp. 1-4, 2007.

[21] A.N. Sreevatsan, K.S. Kumar, S. Rakeshsharma and R. Mansoor, "Emotion Recognition from Facial Expressions: A Target Oriented Approach using Neural Network", *Proceedings of International Conference on Machine Learning*, pp. 497-502, 2004.

[22] A. Mollahosseini, D. Chan and M.H. Mahoor, "Going Deeper in Facial Expression Recognition using Deep Neural Networks", *Proceedings of International Conference on Applications of Computer Vision*, pp. 1-10, 2016.

[23] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller and M. Pantic, "Deep Structured Learning for Facial Expression Intensity Estimation", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 5709-5718, 2017.

[24] G. Yolcu, I. Oztel, S. Kazan, C. Oz and F. Bunyak, "Deep Learning-Based Face Analysis System for Monitoring Customer Interest", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 11, No. 1, pp. 237-248, 2020.

[25] K. Zhao, W. S. Chu and H. Zhang, "Deep Region and Multi-Label Learning for Facial Action Unit Detection", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3391-3399, 2016.

[26] H. Yang, U. Ciftci and L. Yin, "Facial Expression Recognition by De-Expression Residue Learning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2168-2177, 2018.

[27] S. Minaee, M. Minaei and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition using Attentional Convolutional Network", *Sensors*, Vol. 21, No. 9, pp. 3046-3056, 2021.

[28] K. Wang, X. Peng, J. Yang, D. Meng and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition", *IEEE Transactions on Image Processing*, Vol. 29, pp. 4057-4069, 2020.

[29] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition", *Proceedings of International Conference on computer Vision*, pp. 499-515, 2016.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *Proceedings of International Conference on Computer Vision*, pp. 1-14, 2014.

[31] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov and A. Rabinovich, "Going Deeper with Convolutions", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[33] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867-1874, 2014.

[34] R. Ranjan, V.M. Patel and R. Chellappa, "Hyperface: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation and Gender Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 1, pp. 121-135, 2017.

[35] J. Peters, D. Janzing and B. Scholkopf, "Identifying Cause and Effect on Discrete Data using Additive Noise Models", *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 597-604, 2010.

[36] A. Krizhevsky, I. Sutskever and G.E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1097-1105, 2012.

[37] A. Botev, G. Lever and D. Barber, "Nesterov's Accelerated Gradient and Momentum as Approximations to Regularised Update Descent", *Proceedings of International Conference on Neural Networks*, pp. 1899-1903, 2017.

[38] A.C. Wilson, R. Roelofs, M. Stern, N. Srebro and B. Recht, "The Marginal Value of Adaptive Gradient Methods in Machine Learning", *Proceedings of International Conference on Neural Networks*, pp. 1707-1714, 2017.

[39] N.S. Keskar and R. Socher, "Improving Generalization Performance by Switching from Adam to Sgd", *Proceedings of International Conference on Neural Networks*, pp. 1508-1514, 2017.

[40] L. Luo, Y. Xiong, Y. Liu and X. Sun, "Adaptive Gradient Methods with Dynamic Bound of Learning Rate", *Proceedings of International Conference on Neural Networks*, pp. 988-996, 2019.