Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 9657 – 9674

Research Article

Facial Expression Recognition Using VGG16 and LSTM

N. Krishnadas[1*], Nikhila T Bhuvan[2]

Abstract

Facial expression recognition has made significant progress using deep learning, which has received increasing attention across all fields. Mainly, conventional facial expression recognition systems require constrained datasets for optimal performance, making them unsuitable for use on real-time data. Such real-time sequences limit the efficiency and accuracy of the traditional system. An innovative deep learning framework is proposed in this work that combines dual VGG16 and long short-term memory (LSTM) cells to recognize facial expressions in real-time. Three main aspects of the novel framework are: (i) To enhance each image's edge detail and resolve illumination variances, edge enhancement pre-processing techniques are utilized; (ii) In order to extract spatial features from pre-processed images, the VGG16 model is used, which extracts them quite effectively; (iii) An LSTM layer is used in conjunction with VGG16 for extracting temporal relationships between successive frames along with the spatial feature maps. Comparing the experimental data to existing implementations, facial expression recognition has improved considerably in terms of robustness and accuracy.

Keywords: *Deep learning, edge enhancement, facial expression recognition, LSTM, transfer learning, VGG16.*

Introduction

Facial expression is the main non-verbal means of expecting goals in human communication. The work of Mehrabian [1] in 1974 says that 55% of messages relating to feelings and attitudes is in facial expression, 7% of which is in the words that are spoken, the rest of which is spoken communication that does not involve words. According to Mehrabian, facial expressions play a significant role in the exchange of information. Artificial intelligence has provided recent insight into how to discern facial expressions instinctively. Psychology, computer vision, and pattern recognition have all gained considerable interest in the study of Facial Expression Recognition (FER). FER has extensive applications in multiple domains, including human-computer interaction [2,3], augmented reality [4], virtual reality [5], advanced driver assistance systems [6,7], entertainment [8], and education [9].

[1*] Dept. of Information Technology, Rajagiri School of Engineering & Technology, Ernakulam, Kerala, India. E-mail: krishnadas.narayanapillai@gmail.com

[2] Dept. of Information Technology, Rajagiri School of Engineering & Technology, Ernakulam, Kerala, India. E-mail: nikhilatb@rajagiritech.edu.in

According to Ekman and Friesen [10], six universal emotions can be characterized by facial expressions (fear, happy, disgust, sad, anger, and surprise).

An early method of recognizing facial expressions used the Facial Action Coding System (FACS) [11, 12], developed to distinguish between changes in movement patterns of the muscles of the face using 44 action units. The use of handmade and engineered techniques such as facial landmarks have been proposed for many traditional methods. As its evolution evolved, geometric-based and appearance-based feature extraction approaches were used to automate the facial expression recognition system. In some research studies [13, 14, 15], geometric-based methods are used to extract geometric topographies of pre-defined facial landmarks. Similarly, the active appearance model (AAM) [16] and active shape model (ASM) [17] extract geometric topographies based on position and shape. These methods are more sensitive to noise and have trouble delineating the subtle changes in facial muscle movement. Surfaces and edges are investigated using appearance-based techniques, such as local binary pattern (LBP) [18], local directional ternary pattern [19], histograms of oriented gradients (HOG) [20], and Gabor wavelets [21]. Appearance-based techniques, like geometric methods, are robust to noise and can be used to extract distinct features. In some cases, hybrid-based topographies are used to combine both approaches to yield improved recognition accuracy [22, 23]. All the above-mentioned approaches are good for the images in constrained environments.

In real time, images are characterized by multiple backgrounds, occlusions, and illuminations, as well as spontaneous expressions. Expressions of these kinds can vary slightly from normal expressions depending on the culture and style of individuals [24]. Using traditional feature extraction methods in such a case results in increased training time and computational expense, as well as increased noise that adversely affects system performance. Furthermore, studying the complex image requires extra memory, and exploring discriminative facial topographies to recognize facial expressions and link them to a person's internal emotions is an open question.

In recent years, researchers have used deep learning to address the above issues and have made promising progress in emotion recognition from image sequences. Today, GPUs and large training databases are increasingly available, and deep learning has become quite common in computer vision. In fact, convolutional neural networks (CNNs) [25], have made several noteworthy developments in the recognition of emotional expression through facial expression [26–29]. The use of a large dataset lends itself to outperforming a hand-engineered method, as previously explained. For real-time videos and FER databases, the issue is how to handle temporal and spatial signals for improved emotion recognition. There are some 2D CNN approaches that are incapable of identifying temporal information. Researchers developed an integrated method that learns spatial and temporal features, i.e. by engaging both convolutional neural networks and long-short term memory cells (LSTMs) [30]. Inspired by various approaches [50,51], Further investigated the facial expression recognition framework by presenting a new technique for the sequence of images prepared from the CK+ [49] dataset.

This paper presents a novel integrated method for improving the performance of the facial expression recognition framework against challenges such as illumination, pose variation, etc. A dual VGG16 as a base model is designed to extract high-level features. The edge enhancement pre-processing technique is employed to handle different image sequences while training. A novel aspect of the paper is that instead of fusing a dual VGG16-base model with LSTM, the output of the dual VGG16 is fused to produce spatial feature maps. This feature maps then combined with LSTM to create a learnable

model for dynamic temporal features using integrating memory units. Predicting expressions is done using the softmax layer.

The significant contributions in this paper can be summarized:

1. First, instead of raw input, each image frame is treated with a pre-processing technique to correct illuminance and to preserve edge topographies. Proposed framework learns the subtle features of the expression very effectively with the help of pre-processing technique.
2. It is proposed to extract high-level topographies with dual VGG16, and then fuse the feature maps and integrate them with LSTM so they can be used to recognize facial expressions.
3. This proposed method is evaluated on the most common dataset (CK+) to display that this proposed model performs well to the state-of-the-art.

The paper is organized as follows. Section 2 explains the literature related to a few of the existing models. Section 3 explains the proposed methodology. Section 4 defines the experimental details with results. Finally, the conclusion is added in section 5.

Related Work

With access to large datasets and high-performance computing systems, CNN has seen spectacular success in related fields such as object recognition, image classification, semantic segmentation, and other computer vision requirements. In the last few years, research has proposed a number of deep learning models based on CNN and the fusion of different deep learning models to handle the issue of FER in different areas. The main differences between each model are the representation of CNN's architecture and how the spatial-temporal approaches for managing image sequences.

In 1988, Fukushima [31] defined the CNN structure, but the project failed due to hardware limitations. After sometime, in 1989 LeCun et al. [32] used supervised back-propagation networks and presented an excellent result for digit recognition. In face analysis, A method is proposed by Fasel [33] that uses CNNs to handle pose variations in face analysis. Matsugu et al. [34] also proposed FER based on subject independence. Computer vision and image classification were revolutionized by CNN recently. AlexNet is a deep CNN architecture designed by Krizhevsky et al. [35] in 2012. In comparison to LeNet, AlexNet swept the ILSVRC-2012 image classification competition, where more than millions of images were classified across 1000 different classes. It consists of many convolutional layers, a max-pooling layer, and a fully connected layer with a ReLu layer which is an activation function tailed by a softmax layer. Similarly, other most popular CNN architectures such as VGGNet [36], Google Net [37], ResNet [38], and so on were also giving impressive results on well known benchmark databases over the existing state-of-the-art approaches.

CNN-based deep learning methods have been reported in several works as the most effective for FER and have achieved state-of-the-art results in expression categorization tasks. Liu et al. [39] used CNN and restricted Boltzmann machines combined deep learning models added with FACS and utilized the previous knowledge to identify facial expression. Liu et al. [40] proposed a maximum boosted deep belief network which is a unified structure for identifying facial expressions followed by previous work. This method does feature learning, feature selection, and classifier construction as a loopy procedure. The network robustness is improved by the model proposed by Lopes et al. [41] by fusing the CNN model with some pre-processing techniques. This is the reason for measuring pre-processing

steps as a data augmentation for training of the deep neural network. Burkert et al. [42] proposed DeXpression which is a novel CNN technique. This technique creates a unique structure which contains many convolutional and max-pooling layers to learn features automatically at various scales called named FeatEx. Mollahosseini et al. [43] proposed GoogLeNet inception layers and they utilized all subjects from various databases for the training, which results overfitting due to lack of data. A novel architecture was proposed by Fathalla et al. [44] that aimed to achieve higher classification accuracy by fine-tuning the parameters. The identity-enhanced network (IDEnNet) presented by Li et al. [45] mitigates negative identity impacts while learning the informative topographies. Because there is a lack of data for training, the authors used data augmentation and combining CNNs with other deep learning networks to achieve an improved performance. Two small deep networks DTGN and DTAN are proposed by Jung et al. [46], which are trained with image sequences and facial landmarks distinctly. They combined the networks using joint-fine tuning techniques in order to achieve a better result.

A number of the above tasks were performed under supervision. CNN learns spatial topographies and temporal information is ignored, which creates is not suitable for real-time video sequences. For such reasons, several automated deep learning networks can be developed which are proficient to learn both the spatial and temporal features simultaneously. Fan et al. [47] used CNN–RNN hybrid model for emotion recognition and the hybrid model feature map was combined with 3D-CNN to attain greater performance with the input of audio plus video. Donahue et al. [22] introduced a combined CNN and LSTM based model called long-term recurrent convolutional network (LRCN) to learn spatial and temporal features jointly for different object recognition tasks. Jaiswal and Valstar [48] presented a combination of CNN and BiLSTM deep models to obtain temporal information and this approach attained a greater recognition accuracy than the winner of FERA 2015. Saranya et al. [50] used maximum boosted CNN to learn the spatial features and it is integrated with LSTM to overcome long and short-term dependencies which resulted in an accuracy of 99.01% for CK+ dataset. They are using image pre-processing techniques weighted histogram equalization and edge enhancement for better output. Haiqiang et al. [51] used the concept of transfer learning to develop the model. Inception-v3 pre-trained model is used to extract features that provide an accuracy of 98.2% for CK+ dataset.

Methodology

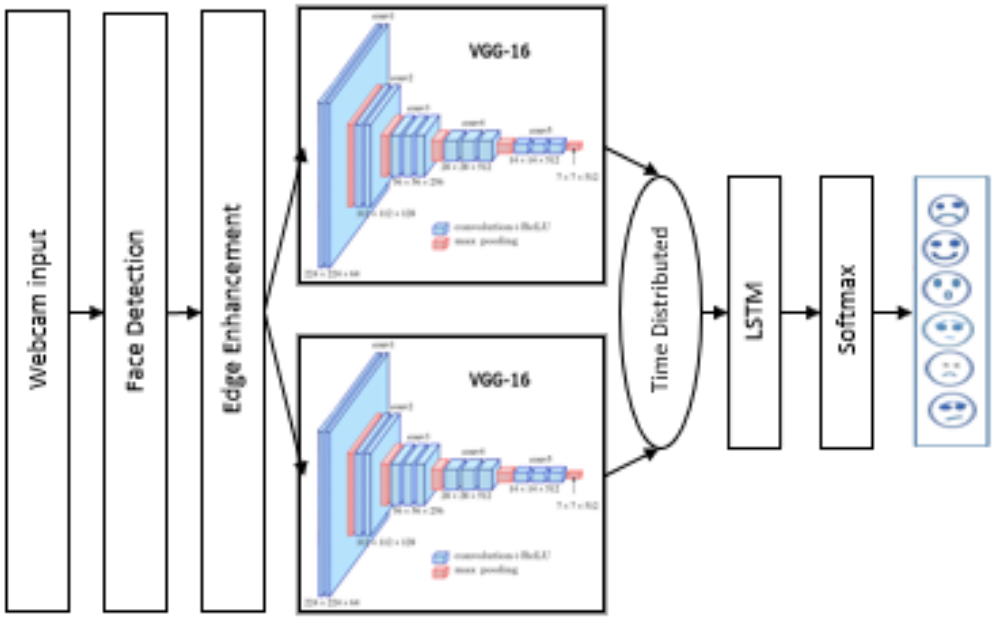

Figure 1. Architecture of the Proposed Model

In this section, the proposed method is conceptually explained with its components. This method consists of three stages: First, the preprocessing method is applied to enhance the edges of each image, which is shown in Fig. 2. Second, a dual VGG16 layer is proposed to create the feature maps. So fully connected layers of VGG16 are not used in this model. Edge enhanced input is fed to the dual VGG16. Finally, the features from dual VGG16 are fused and integrated with LSTM. The feature maps from dual VGG16 and the LSTM are connected to a softmax layer to predict the expression. Fig. 1 illustrates the proposed model of facial expression recognition.

Image Pre-processing

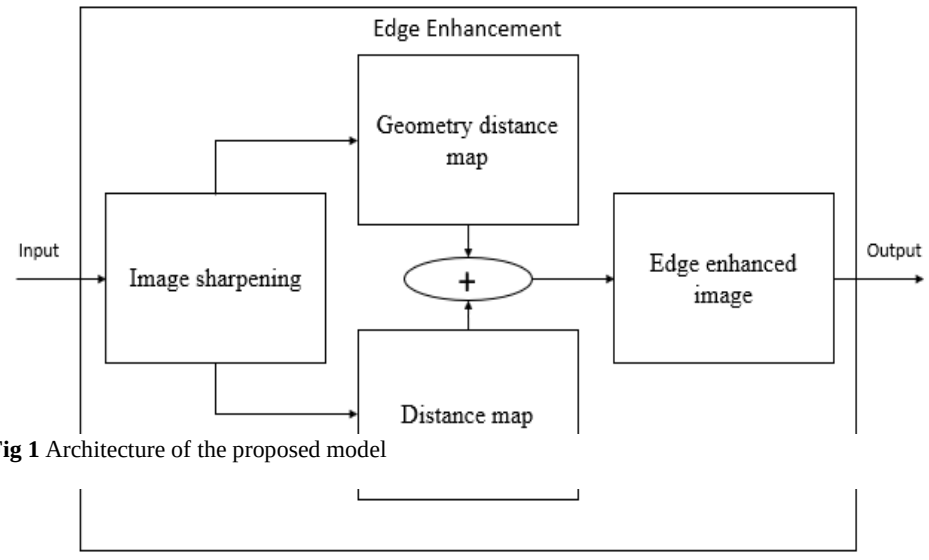

Fig 1 Architecture of the proposed model

Figure 2. Edge Enhancement

For formulating input data, the first face region is detected using OpenCV. Then, crop and resize the detected face into 128 × 128-pixel sizes. The probability density function of the resized image is calculated as

$$P(G_k) = \frac{N_k}{N} \tag{1}$$

where G_k is the greyscale image, N_k is the number of times the occurrence of G_k, and N is the total number of pixels in the image. Since the convolution operation in CNN recognizes the edges using filters, an edge enhancement technique is used to enrich the latent topographies to identify subtle expression variations to get accurate results. The distance map [52] and geometric distance map [53] techniques are added to improve the edges based on pixel distance. The distance map supports labeling each pixel in the image with a distance value, which is nearest to the active neighborhood pixel.

Using the Euclidean distance metric, the distance between two prominent edge points is calculated. Equation (2) gives the mathematical representation of the Euclidean metric

$$D_{\text{euc}}(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{2}$$

where M = mi, N = ni, and p is the number of points in M. The distance map and geometry distance map techniques is superimposed to get better edge information as a resultant image that assists in detecting expressions well. Fig. 2 illustrates the flow of image pre-processing.

VGG16

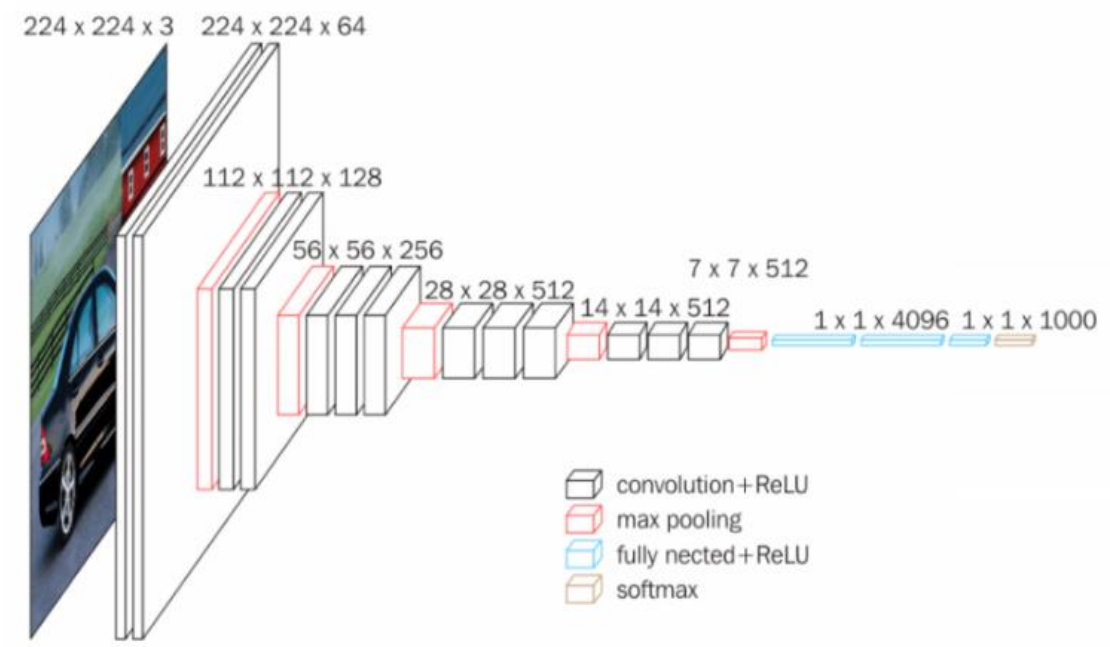

Figure 3. VGG16 Architecture

VGG16 [54] is the pre-trained model used for feature extraction. VGG16 is a convolutional neural network model developed by K. Simonyan and A. Zisserman from the Oxford University in their paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset with 14 million images belonging to 1000 classes. It was one of the well-known models submitted to ILSVRC-2014. It enhances AlexNet by

replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another.

A pre-trained model is a saved network that was formerly trained on a large dataset, typically on a large-scale image classification task. The pre-trained model is used as is or use transfer learning to modify this model to a given task. Instead of developing a model from scratch to solve a similar problem, the model trained on other problems is used as a starting point. The insight behind transfer learning [55] for image classification is that if a model is trained on a large and general enough dataset, this model will efficiently serve as a generic model of the visual world. Model can then take advantage of these learned feature maps without starting from scratch by training a large model on a large dataset.

VGG16 was trained for weeks using the IMAGENET dataset. ImageNet is a dataset of over 15 million labeled high-resolution images belonging to almost 22,000 categories. The images were identified from the web and labeled by human labelers using Amazon's Mechanical Turk crowd-sourcing tool. Starting in 2010, as part of the Pascal Visual Object Challenge, a yearly competition called the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) has been held. ILSVRC uses a subset of ImageNet with almost 1000 images in each of 1000 categories. At all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images. ImageNet consists of variable-resolution images. Therefore, the images have been down-sampled to a 256×256 resolution. Given a rectangular image, the image is rescaled and cropped out of the central 256×256 patch from the final image.

The input to conv1 layer is a fixed size 224 x 224 RGB image. The image is moved through a stack of convolutional (Conv) layers, where the filters were used with a tiny receptive field: 3×3 (which is the smallest size to capture the concept of left/right, up/down, center). In one of the configurations, it also utilizes 1×1 convolution filters, which can be viewed as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed as 1 pixel; the spatial padding of Conv layer input is such that the spatial resolution is conserved after convolution, i.e. the padding is 1-pixel for 3×3 Conv layers. Spatial pooling is carried out by five max-pooling layers, which are followed by some of the Conv layers (not all the Conv layers are followed by max-pooling). Max-pooling is performed in a 2×2 pixel window, with stride 2.

Three Fully-Connected (FC) layers tail a stack of convolutional layers (which have a different depth in different architectures): the first two have 4096 channels each, the third one performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the softmax layer. The configuration of the fully connected layers is identical in all networks.

LSTM Cell

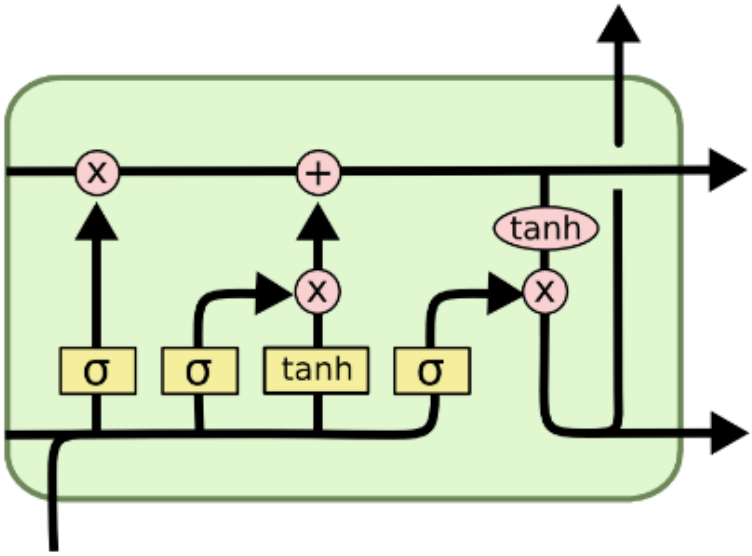

Figure 4. LSTM basic Architecture

LSTM is a form of recurrent neural network (RNN) [22] introduced to overcome long-term dependencies. Traditional (RNN) architectures have shown promising outcomes in short-term image sequences, but show poor performance in long-term sequences due to the vanishing/exploding gradient problems. However, LSTM solves such issues by providing memory for remembering and forgetting the previous information for a long period. In this work, the advantages of the LSTM cell [56] are adopted.

LSTM structure has three gates: forget gate, input gate, output gate, and includes memory cell state. Implementation of these gates supports retaining and update the valuable information for every time step t to the successive network layers, respectively. The basic structure of the LSTM unit is added in Fig. 3. It illustrates the working of three gates and the memory cell state. The forget gate (ft), which chooses either to forget or remember the past information based on the dependencies of the network. The input gate (it) decides to save and update new information on the present state. The output gate (ot) produces the output. Finally, the memory cell state Ct is used to remember long-term historical and future information. LSTM unit contains xt and ht, which indicate the input and output information, respectively. The sigmoid function and the dot product operation govern the information transformation in the LSTM unit. The sigmoid function ranges to 0 and 1. Suppose the dot product of sigmoid operation results in value 1 then all the information is transmitted and if results 0 then information is not transmitted. The equations to calculate the units of LSTM are given as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o \cdot [\, h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * \tanh(c_t) \tag{8}$$

Where σ is the sigmoid activation function, W is the weight matrix, and b is a bias vector. Equation (3) scales the forget gate memory cell value at a time t − 1 and based on the sigmoid range, decides whether the information is retained or thrown. The input gate in (4) is similar to forget gate, and (5) scales the value of the memory cell state at a time t. Equation (6) adds the past and present memories. Finally, (7) shows the output gate, and (8) gives the output of the cell state.

Dual VGG16 and LSTM Integrated Model

The spatial features from dual VGG16 structures are fused and combined with LSTM for learning the deterministic feature information of each image sequence for the FER task. As shown in Fig. 1, the proposed model is a composition of two deep learning structures. The proposed VGG16 pre-trained model without fully connected layers can extract features, and on the other hand, LSTM models the contextual information of arbitrary sequences at each time step. Besides, the edge enhancement pre-processing technique forms two input layers and fed into both VGG16 pre-trained models. VGG16 pre-trained model learns the spatial features and LSTM learns the temporal features of the image. As a combination of both, model is able a provide a very high accuracy compared to existing implementations.

The CNN structures of VGG16 are integrated, and it is given as

$$F(x) = \{ f(I_m) w_i \oplus f(I_n) w_j \tag{9}$$

Where F(x) is the output of the fusion layer, ⊕ is the operation of matrix addition, f(Im) is the mth equalized image features, and f(In) is the nth edge enhanced topographies obtained from the original image sequence after pre-processing. More specifically, the knowledge of each expression feature is handed over from VGG16 to the LSTM model. LSTM learns the long-term feature information linked to previous and subsequent states and processing it for a long period. Now the LSTM feature learning layer creates a new representative feature vector. Finally, the softmax layer classifies and predicts the likelihood of each expression labels based on the input and learned feature maps. In multi-class facial expression classification, the softmax can classify a non-linear function easily. However, it increases the generalization of this model. The softmax layer is given as follows:

$$P(x)_j = \frac{e^{x_j}}{\sum_{i=1}^{N} e^{x_i}} \tag{10}$$

Results and Discussion

Dataset

To demonstrate the effectiveness of this proposed architecture, the experiment is done on the extended Cohn–Kanade database downloaded from Kaggle [57] website. The CK+ database is one of the popular benchmark databases for facial expression recognition. It contains 1562 image sequences are annotated with six basic expression labels (anger, disgust, fear, happy, sad, and surprise) except contempt. These face images are 128*128 in size.

Implementation Details

The proposed model is implemented using KERAS deep learning [58] library, along with OpenCV [59] and Pillow [60] libraries for face detection and edge enhancement pre-processing techniques. With the help of the pre-processed image, the facial expression recognition time is reduced and shows better performance in recognizing transformed and occluded images. The model is trained with 160 epochs which provided a testing accuracy of 99.45%. High-performance F-Series v2 (Up to 2X performance boost for vector processing workloads) virtual machine in Azure cloud portal [61] is used for training the model. This virtual machine is configured with 4 virtual CPUs and 8GB RAM. It took 24 hours for the training with a sequence of 16 images.

Experimental Results

The model is trained with 1562 images and 20% of the dataset is used for testing. 1250 images are utilized for training and 312 images are used for testing. It is observed that the model is attaining >90% accuracy within 10 epochs in training and testing. Model attained an accuracy of 99.45% in 160 epochs which are ahead of the accuracy achieved in the MBCNN-LSTM Model [50] and Inception V3 Model [51]. As per Table. 1, it is clear that the proposed model is providing improved accuracy compared to the existing implementations. Hence the proposed model with pre-processing forms the whole framework to yield better performance compared to the other base models.

Table 1
Accuracy Comparison Table

	VGG16-LSTM Model	MBCNN-LSTM Model	Inception V3 Model
Accuracy	99.45%	99.01%	98.2%

Accuracy

Classification Accuracy [62] is what generally mean when the term accuracy is used. It is the ratio of sum of correct predictions to the total sum of input samples. Accuracy can be calculated as,

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample} \tag{11}$$

Accuracy graph for the proposed model is added below.

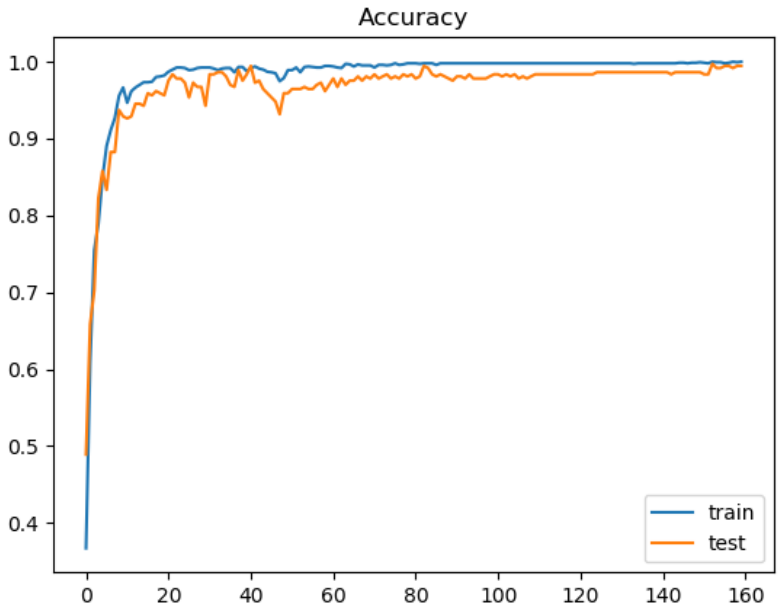

Figure 5. Accuracy Graph

AUC Curve

Area Under Curve (AUC) [62] is one of the most widely used metrics for evaluation. It is used for binary classification problems. AUC of a classifier is equal to the probability that the classifier will rank an arbitrarily chosen positive example higher than a randomly chosen negative example. Before defining AUC, let's understand two basic terms:

1. True Positive Rate (Sensitivity): True Positive Rate corresponds to the proportion of positive data points that are properly considered as positive, to all positive data points.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive} \tag{12}$$

2. False Positive Rate: False Positive Rate corresponds to the proportion of negative data points that are incorrectly considered as positive, to all negative data points.

$$FalsePositiveRate = \frac{FalsePositive}{TrueNegative + FalsePositive} \tag{13}$$

AUC curve for the proposed model is added below.

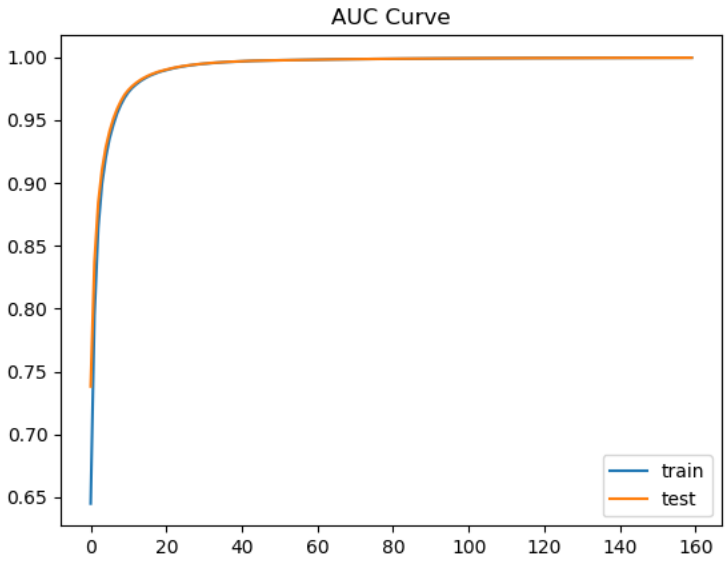

Figure 6. AUC Curve

F1 Score

F1 Score [62] is the Harmonic Mean between precision and recall. F1 Score ranges in [0, 1]. The better the F1 Score, the better is the performance of the model. Mathematically, it can be stated as:

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \tag{14}$$

F1 Score is the most recommended method for evaluating the multilabel classification algorithm. This model provides an F1 score of 98.49% in 160 epochs.

1. Precision: It is the sum of correct positive results divided by the sum of positive results predicted by the classifier.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{15}$$

2. Recall: It is the sum of correct positive results divided by the sum of all relevant samples (all samples that should have been identified as positive). False positives in the precision equation will be replaced by False negatives for Recall.

Precision and recall graphs for the proposed model are added below.

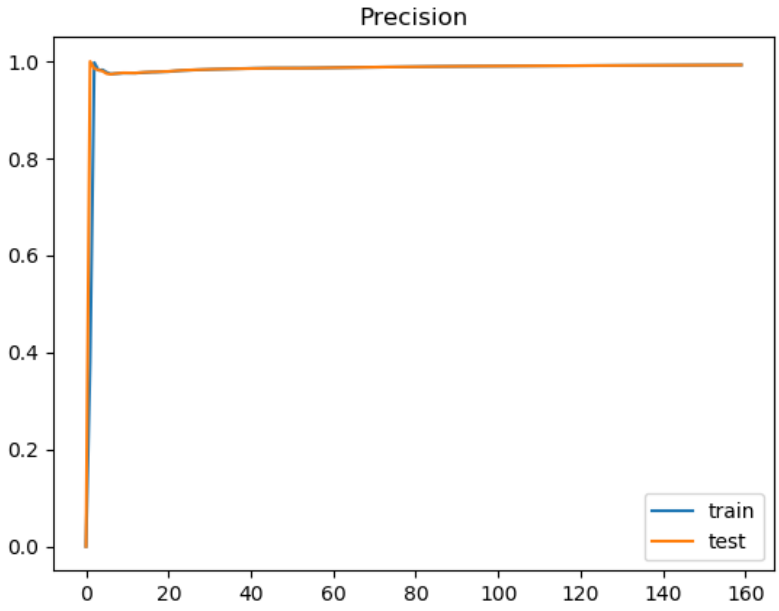

Figure 7. Precision Graph

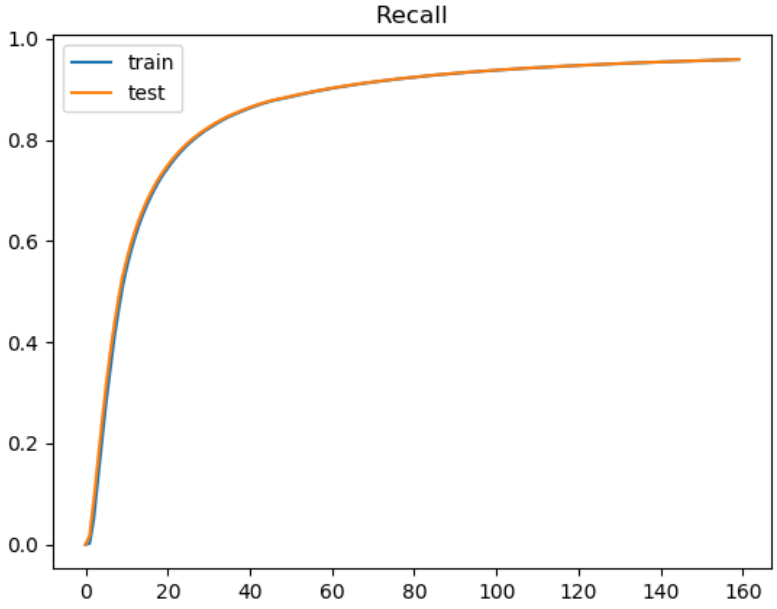

Figure 8. Recall Graph

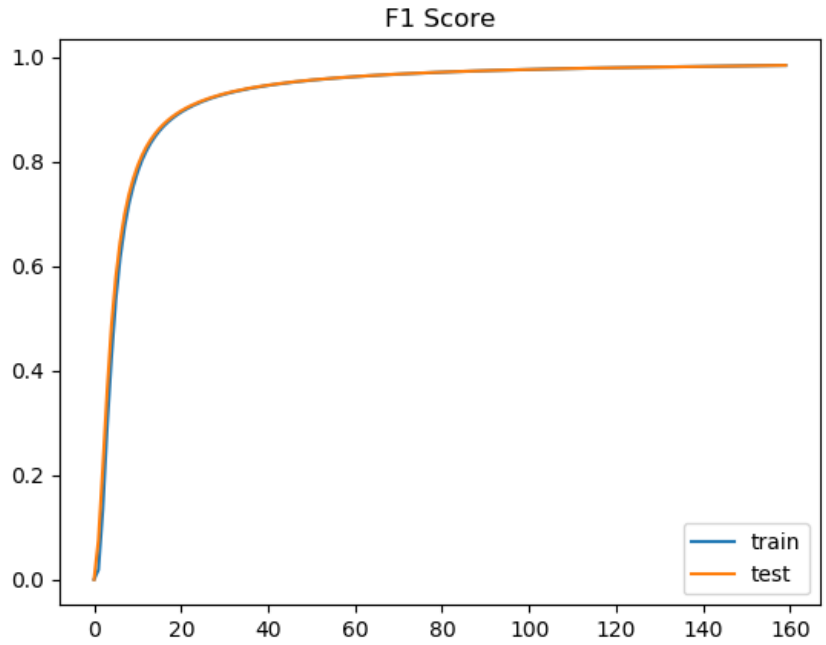

Figure 9. F1 Score Graph

Conclusion

In this paper, a novel technique is proposed to analyze facial expression recognition. This method slightly differs from the literature work, which performs FER task through high-level feature learning by fusing the dual VGG16 pretrained model. Before feature learning, preprocessing techniques are applied to increase the feature extraction ability through edge enhancement and to handle various illumination around the distinct environment. More specifically, the convolutional layers of VGG16 are used to learn the spatial features and it is integrated with LSTM to overcome long and short-term dependencies. The proposed method is trained by using the concept of transfer learning and moreover, the data augmentation technique and batch normalization increase the efficiency of this model during training. These experimental results show that the combination of VGG16 and LSTM produces a highly separable feature map that helps in achieving superior performance in FER. Future scope is the enhancement of current model by replacing VGG16 with GoogLeNet or ResNet.

References

1. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology; The MIT Press:* Cambridge, MA, USA, 1974.
2. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.*, 2001, 18, 32–80.
3. Bartlett, M.S.; Littlewort, G.; Fasel, I.; Movellan, J.R. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. *In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, Madison, WI, USA, 16–22 June 2003; Volume 5, p. 53.

4. Chen, C.H.; Lee, I.J.; Lin, L.Y. Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Res. Dev. Disabil.*, 2015, 36, 396–403.

5. Bekele, E.; Zheng, Z.; Swanson, A.; Crittendon, J.; Warren, Z.; Sarkar, N. Understanding how adolescents with autism respond to facial expressions in virtual reality environments. *IEEE Trans. Vis. Comput. Graph.* 2013, 19, 711–720.

6. Assari, M.A.; Rahmati, M. Driver drowsiness detection using face expression recognition. *In Proceedings of the IEEE International Conference on Signal and Image Processing Applications,* Kuala Lumpur, Malaysia, 16–18 November 2011; pp. 337–341.

7. Jabon, M.; Bailenson, J.; Pontikakis, E.; Takayama, L.; Nass, C. Facial expression analysis for predicting unsafe driving behavior. *IEEE Perv. Comput.*, 2011, 10, 84–95.

8. Lankes, M.; Riegler, S.; Weiss, A.; Mirlacher, T.; Pirker, M.; Tscheligi, M. Facial expressions as game input with different emotional feedback conditions. *In Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology,* Yokohama, Japan, 3–5 December 2008; pp. 253–256.

9. Kapoor, A.; Burleson, W.; Picard, R.W. Automatic prediction of frustration. *Int. J. Hum. Comput. Stud.*, 2007, 65, 724–736.

10. Ekman, P, Friesen, W.: 'Constants across cultures in the face and emotion', *J. Pers. Soc. Psychol.*, 1971, 17, pp. 124–129.

11. Ekman, P, Friesen, W.: 'The facial action coding system: a technique for the measurement of facial movement' (Consulting Psychologists Press, San Francisco, 1978).

12. Tian, Y., Kanade, T., Cohn, J.F.: 'Recognizing action units for facial expression analysis', *IEEE Trans. Pattern. Anal. Mach. Intell.*, 2001, 23, (2), pp. 97–115.

13. Li, S., Deng, W.: 'Deep facial expression recognition: a survey', 2018, arXiv: 1804.08348.

14. Ko, B.C.: 'A brief review of facial emotion recognition based on visual information', *Sensors*, 2018, 18, (2), p. 401.

15. Saranya, R., Poongodi, C., Somasundaram, D., et al.: 'Facial expression recognition techniques: a comprehensive survey', *IET Image Process.*, 2019, 13, (7), 1031–1040.

16. Zhao-yi, P., Yan-hui, Z., Yu, Z.: 'Real-time facial expression recognition based on adaptive canny operator edge detection'. *Proc. Int. Conf. Multimedia and Information Technology (MMIT),* Kaifeng, April 2010, pp. 154–157.

17. Changbo, H., Chang, Y., Feris, R., et al.: 'Mainfold based analysis of facial expression'. *Proc. Int. Conf. Computer Vision and Pattern Recognition Workshop,* Washington, DC, USA, 2004, pp. 1–7.

18. Shan, C., Gong, S., McOwan, P.W.: 'Facial expression recognition based on local binary patterns: a comprehensive survey', *Image Vis. Comput.*, 2009, 27, pp. 803–816.

19. Ryu, B., Rivera, A.R., Kim, J., et al.: 'Local directional ternary pattern for facial expression recognition', *IEEE Trans Image Process*, 2017, 26, (12), pp. 6006–6018.

20. Hu, Y., Zeng, Z., Yin, L., et al.: 'Multi-view facial expression recognition'. *Proc. Int. Conf. Automatic Face Gesture Recognition,* Amsterdam, the Netherlands, 17–19 September 2008, pp. 1–6.

21. Liu, W., Song, C., Wang, Y., et al.: 'Facial expression recognition based on Gabor features and sparse representation'. *Proc. Int. Conf. Control Automation, Robotics and Vision (ICARCV),* Guangzhou, 2012, pp. 1402– 1406.

22. Chen, J., Chen, D., Gong, Y., et al.: 'Facial expression recognition using geometric and appearance features. *Proc. Int. Conf. Internet Multimedia Computing and Service - ICIMCS,* Wuhan, China, September 2012.

23. Yu, H., Liu, H.: 'Combining appearance and geometric features for facial expression recognition'. *Proc. Int. Conf. on Graphics and Image Processing (ICGIP),* Singapore, 2015, 9443.

24. Anastasios, K., Dimitrios, I.F.: 'Image processing and machine learning techniques for facial expression recognition', *In Exarchos, T.P., Papadopoulos, A., Fotiadis, D.I. (Eds.): 'Hand-book of research on advanced techniques in diagnostics imaging and biomedical applications' (IGI Global, 2009),* pp. 247–262.

25. LeCun, Y., Boser, B.E., Denker, J.S., et al.: 'Handwritten digit recognition with a back-propagation network'. *Advances in Neural Information Processing Systems, San Mateo, CA, USA.,* 1990, pp. 396–404.

26. Byeon, Y.H., Kwak, K.C.: 'Facial expression recognition using 3D convolutional neural network', *Int. J. Adv. Comput. Sci. Appl.,* 2014, 5, (12), pp. 107–112.

27. Lopes, A.T., Aguiar de, E., Oliveira-Santos, T.: 'A facial expression recognition system using convolutional network'. *Proc. Int. Conf. SIBGRAPI Graphics, Patterns and Images, Salva-dor,* 2015, pp. 273–280.

28. Yijun, G.: 'Facial expression recognition using convolutional neural network'. *Proc. Int. Conf. Vision, Image and Signal Processing (ICVISP),* Las Vegas, NV, USA, August 2018.

29. Yang, H., Yin, L.: 'CNN based 3D facial expression recognition using masking and landmark features. *Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII),* San Anto-nio, TX, 2017, pp. 556–560.

30. Donahue, J., Hendricks, L., Guadarrama, S., et al.: 'Long-term recurrent convolutional networks for visual recognition and description', *IEEE Trans. Pattern Anal. Mach. Intell.,* 2017, 39, pp. 677–691.

31. Fukushima, K.: 'Neocognitron: a hierarchical neural network capable of visual pattern recognition', *Neural Netw.,* 1988, 1, (2), pp. 119–130.

32. LeCun, Y., Jackel, L.D., Boser, B., et al.: 'Handwritten digit recognition applications of neural network chips and automatic learning', *IEEE Commun. Mag.,* 1989, 27(11), pp. 41–46.

33. Fasel, B.: 'Robust face analysis using convolutional neural networks. *Proc. Int. Conf. Pattern Recognition,* Quebec, Canada, 2002, pp. 40–43.

34. Matsugu, M., Mori, K., Mitari, Y., et al.: 'Subject independent facial expression recognition with robust face detection using convolutional neural network', *Neural Netw.,* 2003, 16, (56), pp. 555–559.

35. Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems,* Red Hook, NY, USA., 2012, pp. 1097–1105.

36. Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for largescale image recognition'. *Proc. Int. Conf. Learning Representations (ICLR),* 2015.

37. Szegedy, C., Liu, W., Jia, Y., et al.: 'Going deeper with convolutions'. *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA., 2015, 1–9.

38. He, K., Zhang, X., Ren, S., et al.: 'Deep residual learning for image recognition'. *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA., 2016, pp. 770–778.

39. Liu, M., Li, S., Shan, S., et al.: 'AU-aware deep networks for facial expression recognition'. *Proc. Int. Conf. Workshops. Automatic Face and Gesture Recognition (FG), Shanghai,* 2013, pp. 1–6.

40. Liu, P., Han, S., Meng, Z., et al.: 'Facial expression recognition via a boosted deep belief network'. *Proc. Int. Conf. Computer Vision and Pattern Recognition, Columbus, OH,* 2014, pp. 1805–1812.

41. Lopes, A.T., De Aguiar, E., De Souza, A.F., et al.: 'Facial expression recognition with convolutional neural networks: coping with few data and the training sample order', *Pattern Recognition.,* 2017, 61, pp. 610–628

42. Burkert, P., Trier, F., Afzal, M.Z., et al.: 'Dexpression: deep convolutional neural network for expression recognition', 2015. https:// arxiv.org/abs/1509.05371

43. Mollahosseini, A., Chan, D., Mahoor, M.H.: 'Going deeper in facial expression recognition using deep neural networks'. *Proc. Int. Conf. Applications of Computer Vision (WACV),* Lake Placid, NY, 2016, pp. 1–10

44. Fathallah, A., Abdi, L., Douik, A.: 'Facial expression recognition via deep learning'. *Proc. Int. Conf. Computer Systems and Applications (AICCSA),* Hammamet, 2017, pp. 745–750

45. Li, Y., Wang, X., Zhang, S., et al.: 'Identity-enhanced network for facial expression recognition'. 2018. Available at http://arxiv.org/abs/1812.04207

46. Jung, H., Lee, S., Yim, J., et al.: 'Join fine-tuning in deep neural networks for facial expression recognition'. *Proc. Int. Conf. Computer Vision (ICCV),* Santiago, 2015, pp. 2983–2991.

47. Fan, Y., Lu, X., Li, D., et al.: 'Video-based emotion recognition using CNNRNN and C3D hybrid networks'. *Proc. Int. Conf. Multi-modal Interaction (ICMI 2016), New York,* USA, 2016, pp. 445–450.

48. Jaiswal, S., Valstar, M.: 'Deep learning the dynamic appearance and shape of facial action units'. *Proc. IEEE Int. Winter Conf. Applications of Computer Vision (WACV), Lake Placid, NY, USA.,* 2016, pp. 1–8.

49. Lucey, P., Cohn, J.F., Kanade, T., et al.: 'The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion specified expression'. *Proc. Int. Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA,* 13–18 June 2010, pp. 97–101.

50. Saranya Rajan, Poongodi Chenniappan, Somasundaram Devaraj, Nirmala Madian, 'Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM', doi: 10.1049/iet-ipr.2019.1188

51. Haiqiang Feng, Jingfeng Shao, 'Facial Expression Recognition Based on Local Features of Transfer Learning'. *Proc. Int. Conf. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2020).*

52. Urbanova, P.: 'Performance of distance-based matching algorithms in 3D facial identification', *Egypt J. Forensic Sci.,* 2016, 6, (2), pp. 135–151

53. Ahdid, R., Taifi, K., Said, S., et al.: 'Euclidean & geodesic distance between a facial points in two-dimensional face recognition system', *Hum. Comput. Interact.,* 2017, 1, p.5

54. VGG16, Accessed Date: 30/05/2021, [Online], Available: https://neurohive.io/en/popular-networks/vgg16/

55. Transfer Learning, Accessed Date: 30/05/2021, [Online], Available: https://www.tensorflow.org/tutorials/images/transfer_learning

56. Hochreiter, S., Schmidhuber, J.: 'Long short-term memory', Neural Comput.,1997, 9, (8), pp. 1735–1780.

57. CK+ DataSet from Kaggle, Accessed Date: 30/05/2021, [Online], Avaliable: https://www.kaggle.com/xwdcrab/ckplus-ocface

58. Keras, Accessed Date: 30/05/2021, [Online], Avaliable: https://keras.io/

59. OpenCV, Accessed Date: 30/05/2021, [Online], Available: https://opencv.org/

60. Pillow Library, Accessed Date: 30/05/2021, [Online], Available: https://pillow.readthedocs.io/en/stable/

61. Azure Portal, Accessed Date: 30/05/2021, [Online], Avaliable: https://portal.azure.com/

62. Deep Learning Metrics, Accessed Date: 30/05/2021, [Online], Avaliable: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234