International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

MHA_VGG19: Multi-Head Attention with VGG19 Backbone Classifier-based Face Recognition for Real-Time Security Applications

Pallavaram Venkateswar Lal[1], Uppalapati Srilakshmi[2], D.Venkateswarlu[3]

Submitted: 06/06/2022 **Accepted:** 10/09/2022

Abstract: Face recognition remains a general biometric verification approach employed for assessing the face images and excerpting beneficial identification data out of them that is consistently named as a feature vector, which is employed for differentiating the biological features. The face recognition procedure starts with excerpting the coordinates of features like mouth's width, eyes' width, pupil, and correlating these with a saved face template. The objective of the proffered scheme remains to craft an independent security system, which executes face recognition-based surveillance alongside a hardware mechanism for locking up the protected area. Surveillance camera photographs of people tend to be of Low Resolution (LR), making it difficult to match them with High Resolution (HR) images. Super resolution, linked mappings, multidimensional scales, and convolutional neural networks are only moderately effective in practice.This study proffers Multi-Head Attention with VGG19 Backbone (MHA_VGG19) that is trained by face images of 3 remarkably disparate resolutions that are employed for excerpting distinctive features strong to the resolution alteration. This as well gives a quantization of the image specimens into a topological region in which inputs, which remain close in the original region as well as remain close in the output region; consequently, they give size decrement and invariability to petty alterations in the image specimen. The proffered methodology is widely analyzed employing LFW and Color FERET datasets by correlating with the advanced methodologies concerning different criteria. Subsequently, the proffered MHA_VGG19 attains 98.69% of accuracy, 99.06% of precision, 98.51% of recall, 98.75% of F1-score, and 100% of ROC for colour FERET database. By employing the LFW database, the proffered MHA_VGG19 attains 95.96% of accuracy, 96.41% of precision, 95.73% of recall, 96% of F1-score, and 99.83% of ROC.

Keywords**:** *Classification, face recognition, feature extraction, gamma correction, neural networks.*

I. Introduction

There are numerous practical applications of face recognition (FR), such as biometric identification, human-computer interaction, video surveillance and so on, in pattern recognition and computer vision. FR[1] is built around two key components: face recognition and verification.Face identification is the process of finding a previously unidentified face in an image, whereas face verification is the process of confirming the accuracy of a previously reported face. Traditional techniques for face identification, such as extracting features or landmarks from a picture, are among the many methods available. SVM (Support Vector Machine) [3] and PCA (Principal Component Analysis) [4] can be used to search for more photos with comparable features. Third-dimensional recognition uses sensors to acquire a lot of information about the face's shape and use this information to detect distinctive traits, such as the shape of the nose.Face recognition can be improved by using this data [5]. Analyzing skin texture is a third method used. Yet another new skin-enhancing design, this one employs skin-enhancing visual features. Face recognition performance was improved by including skin texture analysis [6]. Thermal cameras are the fourth method. A new kind of data is being used for facial recognition, one that is in an entirely different format. Only the shape of a person's head will be recognized using this method. In addition, the wearer's cosmetics, headwear, and sunglasses are not visible. There is a major problem when we use thermal images for face identification because of the small dataset.Traditionally speaking, face recognition is an issue

[1]Reasearch Scholar Of VFSTR Deemed To be University and Associate Professor Of CSE Department,Narayana Engineering College,Gudur , A.P.
[2] Assistant Professor Of CSE Department, VFSTR Deemed To be University, Vadlamudi,Guntur, A.P.
[3] Professor Of CSE Department, VFSTR Deemed To be University, Vadlamudi,Guntur, A.P.
E-Mail : 1venkateswarlal@gmail.com, 2druppalapati2019@gmail.com,

of recognizing more than one face at a time, which is known as a multi-sample problem. Recognition of individuals' faces from a single sample (SSPP). Due to the high accuracy of modern face recognition techniques like deep learning algorithms; these techniques were unable to attain good accuracy in the circumstances where there was very little information.In order to achieve high accuracy with deep learning techniques, more data is required [8]. Face identification with SSPP remains a challenging problem in real-world surveillance applications despite recent advances in machine learning and computer vision [9]. The visual field change between the Enrollment Domain (ED) and the Operational Domain (OD) is one key difficulty [10]. Gathering samples is another challenge for emerging face-recognition methods. It's easier to handle a small number of samples per person, which saves time and money. Sadly, most face recognition methods rely heavily on the size and representativeness of the training set and would suffer greatly or fail completely if each person had only one example to work with. One sample per person is known as an SSPP dilemma. This issue results in poor performance and makes it difficult to develop a reliable FR system [11]. A lack of training data makes this an uphill battle for the vast majority of existing algorithms.In recent years, deep learning algorithms such as CovNets, Recurrent Neural Networks has achieved great success in experiments. The convolution layer extracts visual features through connections and weight sharing. It has the advantage of feature extraction through dimension reduction of convolution layers. After nonlinear mapping, the network can automatically form feature extraction and classifiers adapted to the task from training samples. Therefore, with the improvement of image and video processing technology, the appearance of deep learning has made a great contribution to computer vision. This also makes it possible to achieve face recognition with better algorithms and models. This motivation tends to make following contributions

- To make use of EfficientDet as the Object Detection and MultiHeadAttention with VGG19 Backbone is used as both feature extractor and classifier
- Specifically, VGG19 is used as the backbone for feature extractor and the output is evaluated with the required performance metrics Accuracy ,Precision, Recall,F Measure, AUC.

The rest of this chapter is ordered as follows - Section 2 mentions a few existing research works, Section 3 shows the proposed approach and methodologies, Section 4 exhibits the experimental outcomes and discussion, and, finally, Section 5 ends up with conclusion and future work.

II Related Works

Face recognition is made easier with CNNs that represent learning. Although promising findings have been produced, they are often restricted to accuracy.. Taigman et al. [12] for example, defined face representation learning as a CNN problem of facial recognition. Using deep metric learning approaches, [13] hoped to improve the discriminative capacity of previously learned face representations. CNN-based Facial recognition, on the other hand, has received only a few studies. To improve CNN features' ability to withstand position shifts and occlusion, the Trunk-Branch Ensemble (TBE-CNN) model described in [14] extracts additional information from holistic face images and patches clipped around facial components. When used on the trunk and branch networks, low and medium level convolutional layers are used to efficiently extract features in the TBE-CNN model. The representations trained by TBE-CNN have already been given a boost in their discriminative power, but we now suggest an even better triplet loss function. As part of the training and testing of a DCNN for FR applications, [15] devises a novel "Gabor DCNN ensemble" method that effectively uses many Gabor face representations.Both of these components comprise the GDCNN ensemble approach that we're proposing: Construction of a GDCNN ensemble and combining of a GDCNN ensemble are the first steps. First, a group of GDCNN members (basis models) will be assembled, each trained using a different sort of Gabor face representation.Human face recognition was improved by using back-propagation neural networks (BPNN) and correlation-based feature extraction in [16]. That new data set, called the T-Dataset, was created from the original training data set used to train the BPNN is a key addition of this paper. The connected T-Dataset aids the BPNN's convergence and accuracy by providing a high differentiation layer between the training images, and was constructed using correlation.In [17], we suggest an FRS that combines a powerful image enhancement technique with a new collection of hybrid features for face image pre-processing. Metaheuristic optimization is used in our image enhancement method to enhance facial images effectively regardless of environmental conditions. As a result, the facial image has more features, resulting in better recognition performance than the original image. Improve classification performance of convolutional neural network designs by introducing a new hybrid feature.Deep face and emotion recognition is possible thanks to a convolutional neural network (CNN) and an additional capsule network that employs dynamic routing to understand hierarchical connections between capsules, CapsField. Every long-wavelength image has CapsField extracting the spatial characteristics and learning about the angular part to whole relations for the selected 2D sub-aperture images that are formed from each subaperture.A

fresh constrained face dataset, taken from the same participants as the initial wild LF face dataset, has been captured and made available for analysis of the proposed solution's performance in the wild.Automatic extraction of effective features is proposed in [19] using a weighted mixture deep neural network (WMDNN). There are a number of pre-processing techniques used to limit the regions where FER can operate, including face detection, rotation rectification, and data enhancement. There are two channels of facial images analyzed by WMDNN, one for grayscale images and the other for local binary patterns (LBPs). By fine-tuning a portion of a VGG16 network, the parameters of which were trained on the ImageNet database, we extract expression-related features from facial grayscale photos. A shallow convolutional neural network (CNN) based on DeepID is used to extract facial features from LBP images. Both channels' outputs are combined and weighted. SoftMax classification is used to determine the final recognition result.Realism can be improved in face simulators by using unlabeled real faces, although the identification information can be preserved during the realism improvement, as demonstrated in [20]. The dual agents were created with the specific purpose of being able to discern between authentic and fictitious identities at the same time. A 3D face model can be used as a simulator to generate a variety of different profile images. To generate high-resolution images, DA-GAN uses a fully convolutional network as the generator and an auto-encoder as the discriminator. Using a convolutional neural network and appearance flow, [21] presents a method for fractalizing faces (A3F-CNN). A3F-CNN, in particular, learns to connect the non-frontal and frontal faces in a way that is unique. The frontal face is formed by "shifting" pixels from the non-frontal face. This method of employing synthetic frontal faces preserves the finer details of the face's texture. Training convergence can be achieved by using an appearance-flow-guided learning technique. Face mirroring and generative adversarial network loss are used to solve the problem of self-occlusion. Pose an idea for a multi-tasking exercise. A CNN that simultaneously learns pose-specific identifying features in a combined framework [22] suggests. To examine how CNN-based MTL works, we present an energy-based weight analysis method. Our results show that the side tasks are employed as regularizations to distinguish between the PIE variations and the learnt identity features of participants.In-depth testing on the complete PIE data set has shown that this new method is effective. To our knowledge, this is the first time that multi-PIE face recognition data has been used in this way.

To extract the face region from the face detection process, it is important to segregate the face from the background pattern, which serves as a basis for the following extraction of face difference characteristics. Face detection methods based on deep learning have recently risen in popularity, reducing the time it takes to identify a face and increasing its accuracy. Feature extraction and contrast identification of normalized face images are used to identify human faces in the images that have been separated.

III System Paradigm

In this segment, a proffered method is presented for feature extraction and classification for face identification employing important methodologies at each phase. At first, the databases like LF and Color FERET are embraced for facial images. These images are provided to the pre-processing phase in which Isotropic Smoothing Normalization (ISN) and Gradient Faces procedure could be executed. In the next phase, the feature extraction is performed employing Multi-Test Convolutional Neural Network that remains the amalgamation of P-Net, R-Net, and O-Net. Lastly, the excerpted features are sorted by employing the MHA_VGG19 network. The schematic portrayal of the proffered method is illustrated in figure 1.An elaborate explanation of each approach employed in every step is concisely described in the following segments.

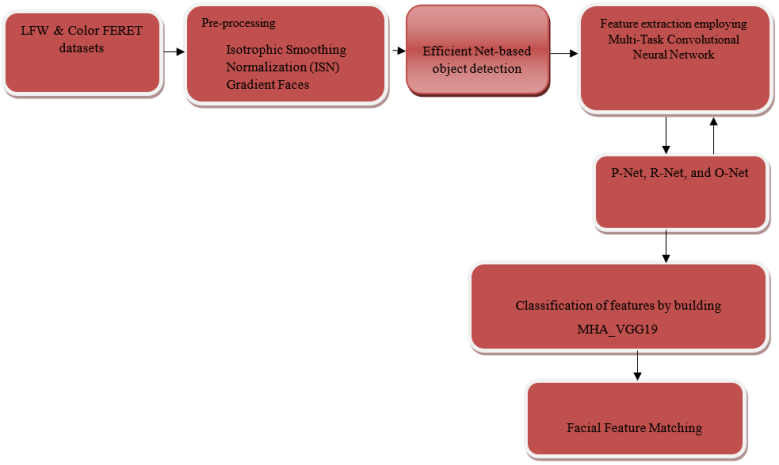

Figure1- Blockschematic illustration for the proffered face recognition methodology

3.1Database explanation

For analyzing the proffered technique, LFW and Color FERET datasets are embraced. The images are obtained out of the front position having dissimilarities in lighting and, especially, in the constitution of the backdrop of every image. The setting is compiled of several items for confirming the strength of this methodology. The FERET dataset comprises 11.338 images gathered out of 994 individuals out of different angles and disparate circumstances and accuracy of various methods. Lastly, the HPI dataset comprises 2 sequences of 93 images per sequence having 15 disparate individuals giving few differences in complexion, hairdo, and adornments (spectacles). Labeled Faces in the Wild (LFW) remains a dataset of face images crafted for learning the issue of unrestricted face identification. This dataset was developed and controlled by analysts at the University of Massachusetts, Amherst (particular references are provided in the Acknowledgment segment). 13,233 images of 5,749 individuals are discerned and formulated by the Viola-Jones face detector and gathered out of the web. 1,680 individuals' images contain 2 or more unique images in the database. The original dataset comprises 4 disparate sets of LFW images and as well 3 disparate kinds of "aligned" images. As per the analysts, deep-funneled images generate excellent outcomes for major face authentication algorithms when correlated with the rest of the image kinds. Thus, the database uploaded herein remains the deep-funneled iteration.

3.2 Pre-processing

Before the multi-focus face detection step, the problem of lighting inconsistency is required to be discussed via using particular contemporarily presented optimization methodologies.

- Isotropic smoothing normalization (ISN) – in the method, the problem of face verification via brightness by isotropic smoothening normalization is used (a transmission step, which basically updates each pel using the average of its adjoining intensity values disregarding the image data neighbouring the region below consideration).
- Gradient faces (GFs) – it remains improperly an enhancement technique however a lightning unaffected measure attained out of image gradient that remains adequately robust for lightning differences encompassing in the uncontrolled natural lighting environment. In this study, gradient faces resembling the pre-processing approach are used for indicating an image in the gradient region.

3.3 Efficient Net-based object recognition

This segment addresses the network framework and a novel compound scaling methodology for Efficient Net. ImageNet-pretrained The network's backbone is made up of Efficient Nets. We propose that a feature network called BiFPN collects features of levels 3-7 from its backbone network and continually performs bidirectional feature fusion from the top to bottom (P3, P4, P5, P6, P7). Based on this knowledge, the class and box networks creates object class and bounding box prognostications.

Backbone network – we re-employed the similar width/depth scaling coefficients of EggicientNet-B0 to B6 intending to effortlessly re-employ their ImageNet-pretrained checkpoints.

Bi-directional feature pyramid network (BiFPN) – BiFPN depth Dbifpn (#layers) are directly raised as the depth requires to be rounded to little integers. For BiFPN width W_{bifpn} (#channels), there remains aggressive growth of BiFPN width W_{bifpn} (#channels). Especially, grid search is executed on a listing of values {1.2, 1.25, 1.3, 1.35, 1.4, 1.45}, and select the finest value 1.35 as the BiFPN width scaling factor. Conventionally, BiFPN width and depth are measured with the ensuing equation:

$$\text{Wbifpn} = 64.(1.35)\text{degrees}$$

Box/class prediction network – The width is set to be consistently similar toBiFPN (i.e., Wpred = Wbifpn), yet directly improve the depth (#layers) employing the following equation:

$$D_{box} = D_{class}\text{=3+(degree/3)}$$

As feature levels 3-7 are employed in BiFPN, the input resolution should be divisible by $2\times 7 = 128$; thus, the resolutions are directly improved employing the following equation:

$$R_{input} = 512 + degree.128$$

3.4 Feature extraction employing Multi-Task Convolutional Neural Network (MTCNN)

The human face incorporates several features – a few of these alter with the surroundings whereas the rest possess robust anti-interference to ecological modifications. For explaining these exclusives more finely, the facial features should be found out. This segment addresses feature extraction out of an individual's face employing Multi-Task Convolutional Neural Network (MTCNN), which processes frames out of a camera live or out of a recorded video document having following entry of the recognized individual into the dataset. The framework of MTCNN, as illustrated in figure 3, contains 3 convolutional networks (P-Net, R-Net, and O-Net) and remains competent in surpassing several face recognition tests when

administering live execution. An image is obtained and rescaled to disparate measurements for constructing the images' pyramid that remains the input to the following three-phase cascading network.

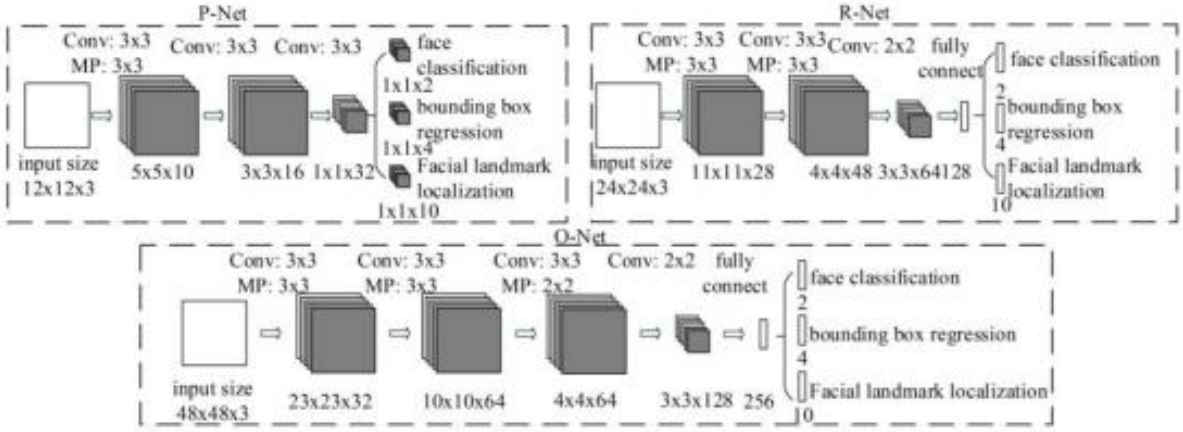

Figure 3- Framework of Multitasking Cascade Convolutional Neural Network (MTCNN)

Three phases of the MTCNN:

- The Proposal Network (P-Net) – FCN is the beginning phase (Fully Convolutional Network). An FCN differs from a CNN in that the thick layer is not used in any way as a framework. The bounding box regression vectors and candidate windows are obtained using P-Nets.

- The Refine Network (R-Net) – All candidates from P-Net have been transferred to R-Net. It remains a CNN rather than an FCN since there is still a thick layer in the structure of the network in the final phase. R-Net generates a four-component vector that represents the face's bounding box and a ten-component vector that identifies the picture's facial landmarks regardless of whether the image is still recognized as a face or not.

- The Output Network (O-Net) – Unlike R-Net, the output network of this phase concentrates on the face and outputs the locations of the five facial landmarks for eyes, nose, and mouth.

Feature extraction, bounding box coordinates, and the location of face landmarks are the three functions that the MTCNN must infer. The ability to discern what images are similar and different can be learned in this way. The input to every network is a single or a pair of photos. The final outputs of each network's final layers are then sent to an action that determines the images. The extent between the two outputs is computed. Using the network, high-quality facial features may be extracted and represented as a 128-component vector. During face recognition, the MTCNN flips a bidimensional set of the bounding boxes' coordinates of all the detected faces. A trained classifier is used to identify the faces recovered from the frame once the dimensions have been leveled. In deciding whether or not a person fits in one of the classes, the maximum credibility score is used, and if that score is higher than the predetermined threshold, the person is added to the registry of identifiable faces.As long as the predetermined cutoff point is not reached, the existence of an unknown individual is officially declared. The attendance accounting department receives a copy of the registry containing the names of those who have been identified. The names of the persons from the dataset are included in the attendance document that is generated by this unit.

3.5 Features classification

The classification layers are entirely joined with their neighboring layer and its aim remains to obtain the last attributes computed by the remaining network and generate scores or probabilities correlation to the trained classes that were chosen for clustering the features employing their eigenvalues as attributes. A disadvantage of employing this methodology remains that the clusters quantity should be predetermined. This technique is employed for addressing this issue by executing many K-Means performances with different K and for computing the Dunn's Index per clusters set. The Dunn's Index calculates the conciseness and division of the clusters acquired for every K. A greater Dunn's Index indicates a little intra-cluster difference and a great inter-cluster distance, that is, the features incorporated in every cluster remain extra identical to every another, and extra disparate out of the features appertaining to rest of the clusters. Hence, the quantity of clusters per feature is chosen as K, which optimizes the Dunn's Index.

3.6 Modeling of MHA_VGG19

The vision transformer paradigm remains an image classification paradigm and does not require a decoder in the transformer architecture, which includes an array of encoder components and an array of decoder elements. As

a result, there is just one encoder in the transformer architecture of the vision transformer. The encoder consists of six identical stacks of encoders. LayerNorm and residual connection architectures are used in the construction of every encoder's multi-head attention and feed-forward layers.The framework for MHA_VGG19 is illustrated in figure 4.

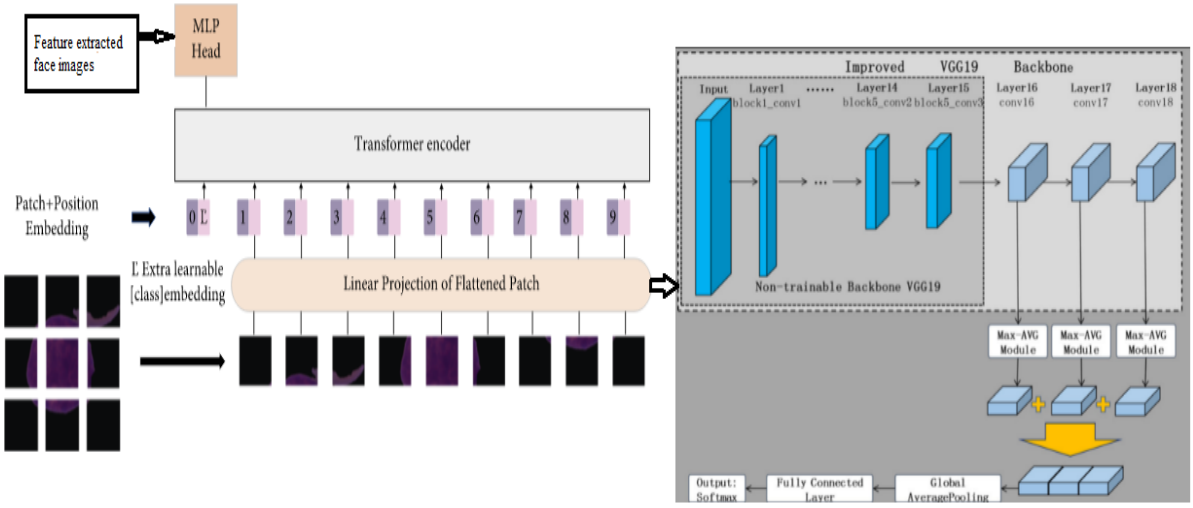

Figure 4- Framework for MHA_VGG19 classification

The multi-head attention remains a self-attention architecture type and this permits the paradigm to focus on disparate qualities of data as illustrated in equation (1) toequation (3) of multi-head attention.

$$Q_i = QW_i^Q$$
$$K_i = KW_i^k$$
$$V_i = VW_i^v, i = 1,2,3 \ldots .8$$
$$head_i = Attention(Q_i K_i V_i), \ i = 1,2, \ldots . .8$$

The query vector Q, the key vector K, the value vector V, and the weight matrix W are all represented in the following formulas. In the vision transformer paradigm, the linear embedding layer remains important. These patches are then flattened into a one-dimensional tensor by the following layer. Position embedding and class embedding are added to the transformer encoder once the patch embedding operation is complete. An MLP head architecture with a completely attached layer and activation function would follow the output of the transformer encoder. Gaussian error linear unit activation function is used in this, and its equation is shown as follows:

$$\mathrm{GELU(x)} = \mathrm{o.5x}(1 + \tanh(\frac{2}{\pi}(\mathrm{x} + 0.44715\mathrm{x}^3)$$

A classification operation is run after output is routed via the MLP head design. As a result of this investigation, the final output class of the vision transformer paradigm has been changed from 3 to 2 classes. It is still possible to reduce the size of the excerpted features while still preserving their important data by using the pooling layer.The convolutional layer remains the central architecture of CNN as depicted in the ensuing equation.

$$y(t) = \int_{-\infty}^{\infty} x(p)h(t-p)dp = x(t) * h(t)$$

The VGG19 network serves as the backbone of this new network. Network's first fifteen layers are non-trainable frozen layers; whereas the network's lower four layers are used in this study to calibrate the training database.Furthermore, the loss function SoftMax cross-entropy classification is joined. Lastly, complementing the output classes, the strip steel defect-recognition transfer learning network cantered upon VGG19 is built. The first learning rate is fixed at lr = 10^{-6}, the attention rate is fixed at decay = 10^{-6}, the attenuation momentum moment is fixed as 0.9, and the training phase is fixed as 500 phases. The paradigm is crafted employing Python 3.6 programming language alongside Keras and TensorFlow libraries. The experiential device in this study remains a desktop personal computer having a 15-8500 processor, NVIDIA GTX 1050 graphics card, 32G RAM, and hard disk capacity of 1T. For the less learning rate issue, this study proffers a stratified variation learning rate methodology. This signifies that the lower layers' learning rates are fixed less as the edges and the rest of the good geometry features could be meticulously learnt and replied to; when the layers' learning rates are fixed greater for quickly assuring the network learnt images' top-level features and for resolving the issue of tardy convergence on the network. The 4 network layers are substituted by 3 convolution network layers. Thus, the enhanced VGG19 backbone neural network contains eighteen layers on the whole; the network layers fix their learning rates at 10^{-6}, 10^{-4}, and 10^{-2} for disparate layer regions by the 2:2:6 rule.

Branch 1: The original features are sent in a sequence via a convolution layer having dimension 1 × 1 and a

convolution layer having dimension 2 × 2. The branch remains in no way particularly processed thereby this branch could store the features in the original image as far as feasible.

Branch 2: The original features are sent in a sequence via a convolution layer having dimension 1 × 1 and a mean pooling layer having dimension 2 × 2, and, lastly, a ReLU activation layer is joined. This branch employs the averaging pooling layer chiefly for refining the intrusion data in the original features.

Branch 3: The original features are sent in a sequence via a convolution layer having dimension 1 × 1 and a maximal pooling layer having dimension 2 × 2, and lastly, a ReLu activation layer is joined. This branch embraces the maximal pooling layer chiefly for excerpting the features having greater illumination out of the original features for finely seeking the deficit region.

For permitting the enhanced VGG19 backbone neural network for excerpting multi-level feature data, the seventeenth, eighteenth, and nineteenth layers are joined having a maximal and mean feature extraction unit and a convolutional layer accordingly. Next, the multi-level features are excerpted out of the 3 branch layers that are joined for combining an international pooling layer and an entire connection layer. Lastly, the network is joined to the softmax classifier.

For facial image databases, minute noise remains inevitable. The morphological relatedness of face identification leads to a few noises upon the label. Overfitting can be avoided by using a symmetric cross-entropy loss function. The following is an illustration of the equation for the symmetric cross-entropy loss function:

$$I_{rce} = I_{ce} + I_{rce}$$

Here, l_{ce} represents the cross-entropy loss function, and l_{rce} remains the reverse cross-entropy function. The equation for the same is depicted as ensues:

$$I_{rce} = -\sum_{k=1}^{k} \mathrm{p(k)} \log \mathrm{q(k)}$$

Hence, the symmetric cross-entropy loss function is employed in this study as the paradigm's loss function for lessening the impact of noise upon the generalization capacity of the paradigm.

3.7Algorithm for Multi head Attention with VGG19 Backbone-adam optimizer procedure

Set initial values for all N networks.
Randomly generate the input layer parameters $(a_j^i, b_j^i), j = 1, \dots, L$ *(is the number of nodes in hidden layer), of the th Enetwork.*
The ith networks' hidden layer output matrix should be calculated.
Determine the weights of the ith output matrix (i.e., the target output weights).
Assume that k networks are left after dissimilarity elimination of N networks.
Base $=$ *Target Vector* $+$ *Random Integer*
for number of iterations do
TempVector $1 =$ *Target* $+$ *Random Deviation*
TempVector 2 = Target-Random Deviation
Weighted = Weighted Difference (TempVector 1 , Temp Vector 2)
$\{(yj', xj)\}j =$, *where* $xj = 0{,}1$
For $r =, \dots, R$ *do*
standardize the weights $\omega r, ,$ *in amount to* $\sum \omega r, jnj = 1 = 1$
For all feature, train a weak classifier hi.
For error ϵi *of a classifier* hi *is designed matching to the weight* $\omega r_n. \dots, \omega r, n$

$$\epsilon i = \sum \omega r' j | hi(yj') - xj | nj = \mathbf{1}$$

Base = Base+Weighted
Trial = Aggregate(Target, Base)
if MF(Trial) $<$ *MF(Base) then*
Base = Trial
else
Base = Target
end if
end if
end for

3.8 Facial Feature Matching

Experimentations on LFW and Color FERET datasets comprising 441 labelled faces of twenty-six individuals classified by MHA_VGG19 were performed. The dataset is split into 2 portions – a training set comprising 120 images and a testing set comprising 321 images. The experimentations were reiterated for ten haphazard splits of the dataset in order that each image of the subject could be employed for testing. The outcomes were stated upon the mean execution. A labeled face remains a pattern, which is portrayed by three vectors $(y_{face}, y_{eyes}, y_{face\_no\_mouth})$ containing a hundred sizes with k=100. Face images require to be classified within one among the twenty-six individuals. Hence, the MHA_VGG19 paradigm, in the present instance, contains 3 sigmoid functions of 26 CNNs. An image's i[th] vector would be processed for generating the dimensional output

vector. For combining the entire k-th, the components of the output vectors obtain the collective vector. The collective vectors remain the CNNs' input. Merely a single CNN's output node remains MHA_VGG19's output node. This application employs backbone VGG19 consisting of three layers having the transfer function, which remains the sigmoid function for CNN. The quantity of hidden nodes of and is experientially discerned from ten to hundred hidden nodes. Each contains (feature vector dimensions) input nodes and (classes' quantity) output nodes. The k-th output of the remains the probability measure that indicates in any case the input image remains in the class.

IV. Performance Analysis

The experimentation result is executed in Python software and the parameters used for analysis remain accuracy, precision, recall, f1-score, and AUC curve. These parameters are compared with 2 kinds of databases like LFW and Color FERET for the proffered MHA_VGG19.

Accuracy: This determines the quantity of rightly estimated values to the entire quantity of estimations.

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

Recall: This is described as the rightly estimated value to the entire estimation value.

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

Precision: This gives the proportion of true positive values to the entire estimated values.

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP}$$

F1-Score: This gives the proportion betwixt the average mean of precision and recall.

$$F1 - Score = 2 * \frac{\mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$$

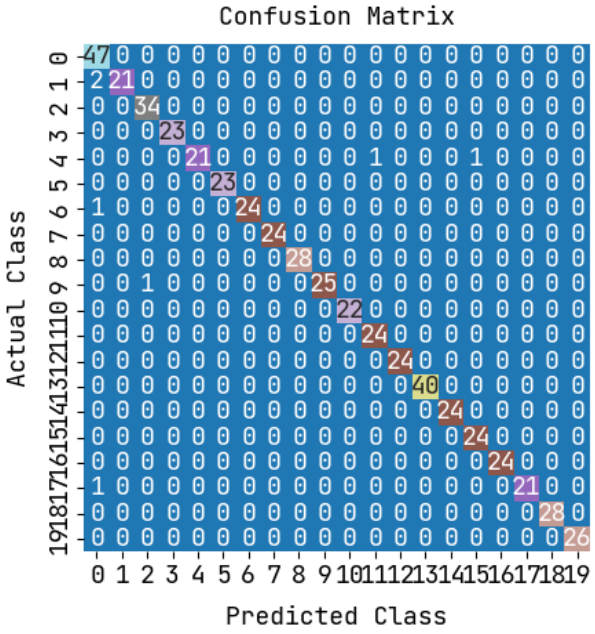

Figure 5- Confusion matrix for Color FERET database

The atop figure 5 exhibits the confusion matrix for the Color FERET database where the rows portray the estimated class (output class) and columns portray the real class (target class) for classification. The crosswise cells portray the tested networks, which are rightly and wrongly classified. The column on the right denotes each estimated class whereas the row below denotes the execution of each real class.

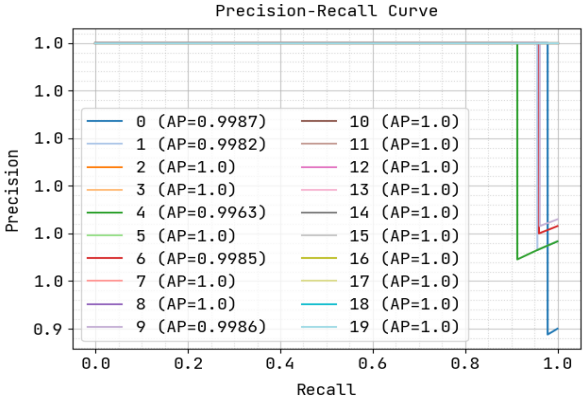

Figure 6- Precision-Recall curve for Color FERET database

On the x-axis is the recall value, and on the y-axis is the precision value, as shown in the precision-recall curve in figure 6.While doing this procedure, the AP range differs betwixt 1.0 and 0.998 exhibiting that the proffered MHA_VGG19 possesses finer precision and recall.

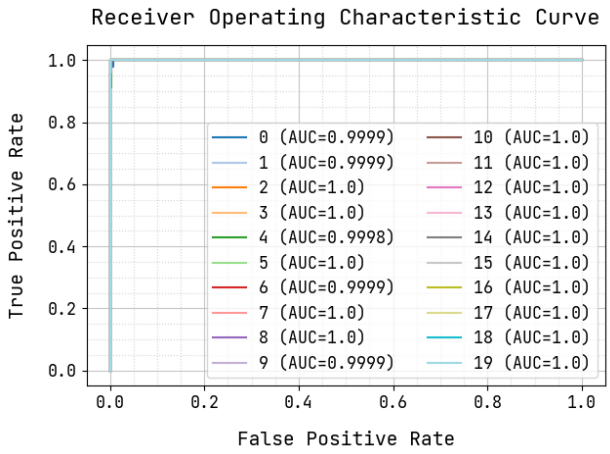

Figure 7- ROC curve for Color FERET database

The atop figure 7 illustrates the ROC curve for the Color FERET database in which the x-axis denotes the false positive rate, and the y-axis denotes the true positive rate. While doing this procedure, the AUC range differs betwixt 1.0 and 0.999 exhibiting that the proffered MHA_VGG19 possesses finer ROC.

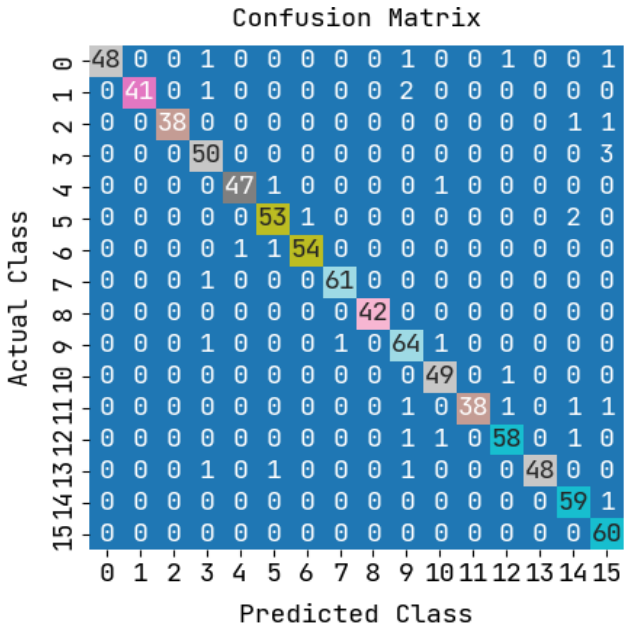

Figure 8- Confusion matrix for LFWdatabase

The atop figure 8 exhibits the confusion matrix for the LFW database where the rows portray the estimated class (output class) and columns portray the real class (target class) for classification. The crosswise cells portray the tested networks, which are rightly and wrongly classified. The column on the right denotes each estimated class whereas the row below denotes the execution of each real class.

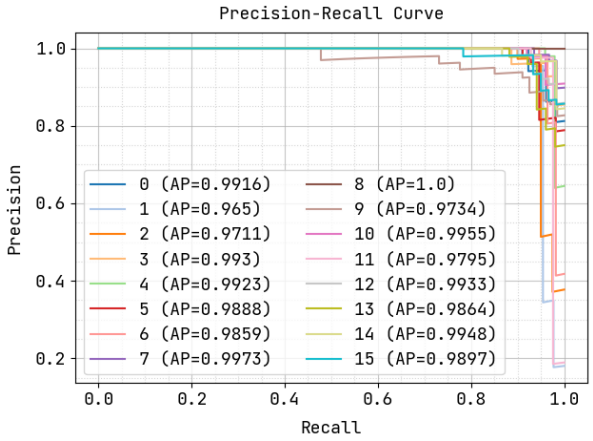

Figure 9- Precision-Recall curve for LFW database

As depicted in the top figure 9, the LFW database's precision-recall curve is depicted by x-axis (recall value) and y-axia (precision).While doing this procedure, the AP range differs betwixt 1.0 and 0.965 exhibiting that the proffered MHA_VGG19 possesses finer precision and recall.

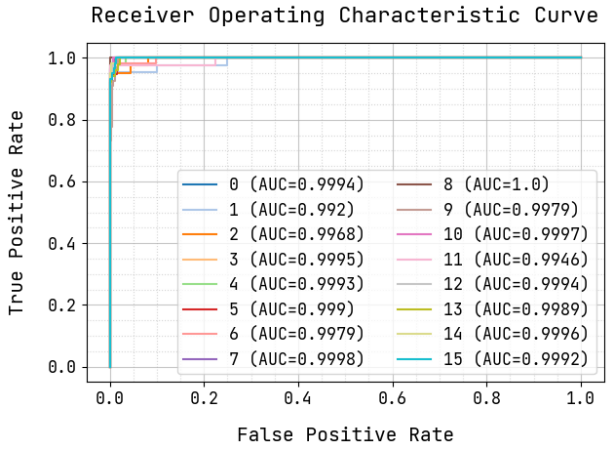

Figure 10- ROC curve for LFW database

The atop figure 10 illustrates the ROC curve for the LFW database in which the x-axis denotes the false positive rate, and the y-axis denotes the true positive rate. While doing this procedure, the AUC range differs betwixt 1.0 and 0.997 exhibiting that the proffered MHA_VGG19 possesses finer ROC.

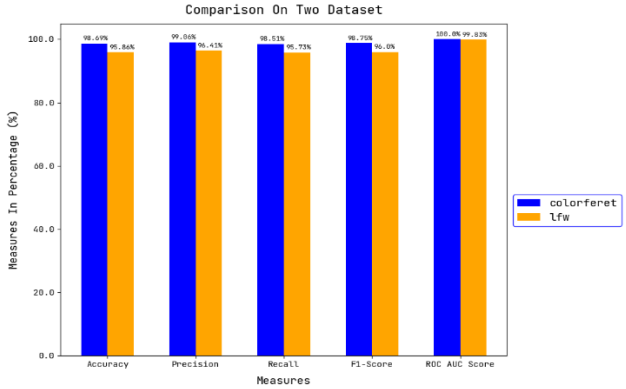

Figure 11- Comprehensive correlation betwixt the two databases

Figure 11 illustrates the comprehensive correlation betwixt LFW and Color FERET databases concerning the accuracy, precision, and recall. It is found out that the function of the Color FERET database in this proffered MHA_VGG19 attains 98.69% o accuracy, 99.06% of precision, 98.51% of recall, 98.75% of F1-score, and 100% of ROC. By employing the LFW database, the proffered MHA_VGG19 attains 95.96% of accuracy, 96.41% of precision, 95.73% of recall, 96% of F1-score, and 99.83% of ROC

Table-1 Correlation of two databases concerning different criteria

Database	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC (%)
Color FERET database	98.69	99.06	98.51	98.75	100
LFW database	95.96	96.41	95.73	96	99.83

V. Conclusion

The constant progress and broadening of deep learning resulted in a base for the continual augmentation of classification networks that has steadily enhanced accuracy. For enhancing the face recognition degree of security monitoring scenes having varied scales in dense face images, this study proffers the MHA_VGG19 classifier appropriate for intricate scenes. The image super-resolution reconstruction technology is incorporated into the network architecture of the target detection algorithm. The proffered methodology attains advanced execution concerning the two – visual quality and quantitative performance metrics. For instance, the recognition images that were generated appeared extra natural and contain sharper edges and higher resolution. The proffered methodology is widely analyzed employing LFW and Color FERET datasets by correlating with the advanced methodologies concerning different criteria. Consequently, the proffered MHA_VGG19 attains 98.69% of accuracy, 99.06% of precision, 98.51% of recall, 98.75% of F1-score, and100% of ROC for the Color FERET database. By employing the LFW database, the proffered MHA_VGG19 attains 95.96% of accuracy, 96.41% of precision, 95.73% of recall, 96% of F1-score, and 99.83% of ROC. The prospective studies focus upon incorporating facial countenance and facial maturity-based identification methodology in a video frame.

References

[1]. Y. Sun, X.Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," Proc. IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), pp. 2892-2900, 2015.

[2]. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet, A unified embedding for face recognition and clustering," Proc. IEEE Conf. Comput. Vision and Pattern Recognition, pp. 815-823, 2015.

[3]. O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," Proc. British Mach. Vision Conf., vol. 1, no. 3, p. 6, 2015.

[4]. F. Matta and J.-L. Dugelay, "Person recognition using facial video information: A state of the art," Journal of Visual Languages & Computing, vol. 20, no. 3, pp. 180–187, 2009.

[5]. S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Dynamic ensembles of exemplar-svms for still-to-video face recognition," Pattern recognition, vol. 69, pp. 61–81, 2017.

[6]. W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng, "Discriminant analysis of principal components for face recognition," in Face Recognition. Springer, 1998, pp. 73–85.

[7]. J. A. Cook, V. Chandran, and C. B. Fookes, "3d face recognition using log-gabor templates," 2006.

[8]. J.-S. Pierrard and T. Vetter, "Skin detail analysis for face recognition," in 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007, pp. 1–8.

[9]. D. A. Socolinsky, A. Selinger, and J. D. Neuheisel, "Face recognition with visible and thermal infrared imagery," Computer vision and image understanding, vol. 91, no. 1-2, pp. 72–114, 2003.

[10]. D. Maturana, D. Mery, and A. Soto, "Learning Discriminative Local Binary Patterns for Face Recognition," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition and Workshops (FG), pp. 470-475, March, 2011.

[11]. Krishnaveni, S. ., A. . Lakkireddy, S. . Vasavi, and A. . Gokhale. "Multi-Objective Virtual Machine Placement Using Order Exchange and Migration Ant Colony System Algorithm". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 6, June 2022, pp. 01-09, doi:10.17762/ijritcc.v10i6.5618.

[12]. N.-S. Vu, and A. Caplier, "Enhanced Patterns of Oriented Edge Magnitudes for Face Recognition and Image Matching," IEEE Trans. Image Processing, vol. 21, no. 3, pp. 1352-1365, March, 2012

[13]. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1701–1708.

[14]. J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1875–1882

[15]. Chaudhary, D. S. . (2022). Analysis of Concept of Big Data Process, Strategies, Adoption and Implementation. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 05–08. https://doi.org/10.17762/ijfrcsce.v8i1.2065

[16]. Ding, C., & Tao, D. (2017). Trunk-branch ensemble convolutional neural networks for video-based face

recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 1002-1014.

[17]. Choi, J. Y., & Lee, B. (2019). Ensemble of deep convolutional neural networks with Gabor face representations for face recognition. *IEEE Transactions on Image Processing*, *29*, 3270-3281.

[18]. Abuzneid, M. A., & Mahmood, A. (2018). Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network. *IEEE access*, *6*, 20641-20651.

[19]. André Sanches Fonseca Sobrinho. (2020). An Embedded Systems Remote Course. Journal of Online Engineering Education, 11(2), 01–07. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/39

[20]. Oloyede, M. O., Hancke, G. P., & Myburgh, H. C. (2018). Improving face recognition systems using a new image enhancement technique, hybrid features and the convolutional neural network. *Ieee Access*, *6*, 75181-75191.

[21]. Kiran, M. S., & Yunusova, P. (2022). Tree-Seed Programming for Modelling of Turkey Electricity Energy Demand. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 142–152. https://doi.org/10.18201/ijisae.2022.278

[22]. Sepas-Moghaddam, A., Etemad, A., Pereira, F., & Correia, P. L. (2021). Capsfield: Light field-based face and expression recognition in the wild using capsule routing. *IEEE Transactions on Image Processing*, *30*, 2627-2642.

[23]. Yang, B., Cao, J., Ni, R., & Zhang, Y. (2017). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access*, *6*, 4630-4640.

[24]. Zhao, J., Xiong, L., Li, J., Xing, J., Yan, S., & Feng, J. (2018). 3d-aided dual-agent gans for unconstrained face recognition. *IEEE transactions on pattern analysis and machine intelligence*, *41*(10), 2380-2394.

[25]. Zhang, Z., Chen, X., Wang, B., Hu, G., Zuo, W., & Hancock, E. R. (2018). Face frontalization using an appearance-flow-based convolutional neural network. *IEEE Transactions on Image Processing*, *28*(5), 2187-2199.

[26]. Yin, X., & Liu, X. (2017). Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, *27*(2), 964-975.