IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

SPECIAL SECTION ON DATA MINING FOR INTERNET OF THINGS

Received April 9, 2019, accepted May 6, 2019, date of publication May 23, 2019, date of current version June 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918275

Deep Unified Model For Face Recognition Based on Convolution Neural Network and Edge Computing

MUHAMMAD ZEESHAN KHAN[1], SAAD HAROUS[2], SALEET UL HASSAN[1], MUHAMMAD USMAN GHANI KHAN[1], RAZI IQBAL[3], (Senior Member, IEEE), AND SHAHID MUMTAZ[4]

[1]Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology Lahore, Lahore 54000, Pakistan
[2]College of Information Technology, United Arab Emirates University, Abu Dhabi 15551, United Arab Emirates
[3]College of IT, American University in the Emirates, Dubai 503000, United Arab Emirates
[4]Instituto de Telecomunicaces, 4099 Lisboa, Portugal

Corresponding author: Muhammad Zeeshan Khan (zeeshan.khan@kics.edu.pk)

This work is partially supported by grant 31T102-UPAR-1-2017 from UAE University.

ABSTRACT Currently, data generated by smart devices connected through the Internet is increasing relentlessly. An effective and efficient paradigm is needed to deal with the bulk amount of data produced by the Internet of Things (IoT). Deep learning and edge computing are the emerging technologies, which are used for efficient processing of huge amount of data with distinct accuracy. In this world of advanced information systems, one of the major issues is authentication. Several techniques have been employed to solve this problem. Face recognition is considered as one of the most reliable solutions. Usually, for face recognition, scale-invariant feature transforms (SIFT) and speeded up robust features (SURF) have been used by the research community. This paper proposes an algorithm for face detection and recognition based on convolution neural networks (CNN), which outperform the traditional techniques. In order to validate the efficiency of the proposed algorithm, a smart classroom for the student's attendance using face recognition has been proposed. The face recognition system is trained on publically available labeled faces in the wild (LFW) dataset. The system can detect approximately 35 faces and recognizes 30 out of them from the single image of 40 students. The proposed system achieved 97.9% accuracy on the testing data. Moreover, generated data by smart classrooms is computed and transmitted through an IoT-based architecture using edge computing. A comparative performance study shows that our architecture outperforms in terms of data latency and real-time response.

INDEX TERMS CNN, face, attendance, RCNN, anchors, RPN, edge computing.

I. INTRODUCTION

The human face is an important entity which plays a crucial role in our daily social interaction, like conveying individual's identity. Face recognition is a biometric technology that extracts the facial features mathematically and then stores those features as a face print to identify the individual. Biometric face recognition technology gained a lot of attention during the past few years due to its wide range of applicability in both law enforcement and other civilian areas, institutes and organizations.

Face recognition technology has a slight edge on other biometric systems like finger-print, palm-print and iris due to its non-contact process. Face recognition system is also able to recognize the person from a distance without touching or any interaction with the person. Moreover, the face recognition system also helps in crime deterrent purpose, because the captured image can be stored in a repository and later can be helpful in many ways like to identify a person. Currently, face recognition applications are deployed in social media websites like Facebook, in the entrance of Airports, Railways

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

Stations, Bus Stop, highly secured areas, advertisement, and health care. The purpose of these applications is to minimize criminal activities, fake authentication, tracking addictive gamblers in casinos, whereas Facebook is using face recognition system for automatic tagging purpose. For face recognition purpose, there is a need for large data sets and complex features to uniquely identify the different subjects by manipulating different obstacles like illumination, pose and aging. During the recent few years, a good improvement has been made in facial recognition systems. In comparison to the last decade, one can observe an enormous development in the world of face recognition. Currently, most of the facial recognition systems perform well with limited faces in the frame. Moreover, these methodologies have been tested under controlled lighting conditions, proper face poses and non-blurry images.

Machine learning on edge computing nodes is also gaining popularity and likely to rise up even more with the passage of time. Edge computing is defined as the technology that enable the data processing at the edges of the network. The edge computing is helpful in many ways such as; to reduce the traffic, cloud computing, and storage resources over the network. It is also helpful in reducing the response time and data latency and making the data transmission more safe and secure [1], [2]. The concept of smart homes and buildings equipped with smart devices has becomes popular now days. These devices are capable of automating the human activities. Face recognition has also its implications in making the class rooms smart. In currently available smart architecture deployed at homes, cities and some specific buildings, the data produced by the nodes are passed to the cloud for further processing. The amount of data produced at the edge of the network is very large, so there is a need to process the data at the edges of the network to reduce data latency. Face recognition requires a large amount of computation and processing power with the large amount of database with whom the encodings of the input image is compared. With the availability of cheap bandwidth and fast internet speed, the computational data of the face recognizer is transfer to the edge device to get the faster results.

The accuracy of the recognition task has remarkably increased due to the availability of high computational power, required for the deep learning algorithms. To achieve better results, proposed algorithm utilizes the Convolution Neural Network, which is a deep learning approach and state-of-the-art in computer vision. The proposed methodology is able to recognize the people even when frame has multiple faces. This system is capable of recognizing the people from different positions and under different lighting conditions, as light does not have much effect on the system. Moreover, to improve the data latency and response time, edge computing have been utilized for implementing the smart class rooms in real time. Below are the major contributions of this paper:

- Propose a deep unified model for Face Recognition based on Faster Region Convolution Neural Network.
- Design a group-based face attendance system based on the proposed deep unified model.
- The proposed model is able to recognize 30 faces out of the 35 detected faces.
- Achieve an accuracy of 97.9% by implementing the proposed algorithm under different conditions.
- Edge Computing have been utilized for processing the data at the edges of the nodes to reduce the data latency and increase the real time response.

The rest of the paper is organized in the following manner. Section II and III review literature and explain methodology respectively. Section IV discusses the practical implementations of the proposed system, whereas experiments and results are explained in section V. Finally we concluded the paper in section VI.

II. LITERATURE REVIEW

Face recognition is considered as a substantial challenging task due to its complicated nature. In the literature, two distinguished methods have been utilized for solving the problem of the face recognition, one is shallow learning and the other one is the deep learning. Since both approaches are based on supervised learning, a good training dataset is required to achieve the efficient and acceptable results.

Currently most of the databases for face recognition are based on the data collected in a controlled environment with some specific parameters. Huang *et al.*, [3] proposed Labelled Faces in the Wild dataset which covers all constrains like pose, lighting, accessories, occlusions, and background. The dataset contains 13233 images of unique 5749 persons of different ages.

Most of the methodologies of face recognition have used this Labeled Faces in the Wild (LFW) dataset for the training of their proposed solution. Shallow methods are based on the hand crafted features decided by humans on a local face image descriptor such as scale-invariant feature transform (SIFT), local binary patterns (LBP), histogram of oriented gradients (HOG) [4]–[7] etc. Then, they aggregate these local face descriptor features on the overall face descriptor through pooling mechanism such as Fisher Vector [8], [9]. Sivic *et al.* [10] modeled two faces representation jointly along with the prior face representation. Instead of using the classical Bayesian method which is based on the difference of the features between two faces [11]. The system is trained on LFW dataset with 92.4% accuracy. Sift features have also been used for the face recognition problem.

Cinbis *et al.* [4] first detect the faces from the frame by applying the Viola Jones face detector and then apply the sift algorithm to extract the features of the detected faces which are further used for classification purpose. They have used the dataset of the TV series Buffy the vampire slayer and claimed to have achieved the accuracy of 86% on the testing data. However, the limitation of the sift features are that they do not perform well when there is illumination, rotation and blurriness in the image. To overcome this problem some researchers utilized the fisher vectors along with the sift features. For example, Lu and Tang, [12] extract the sift features from the detected faces, and then apply the fisher vector on these sift features. Since fisher vectors are very sparse so they used the compact descriptor which is learnt by discriminative metric learning. They have used the LFW dataset and achieved 93.1% accuracy. Parkhi *et al.*, [8] have used the same fisher vectors to extract the features of the detected faces. They have reduced scarcity by using the binarization. Due to the compact size of the vectors, they have computed the large scale repository quickly. Similarly, Simonyan *et al.*, [9] applied fisher vectors on the sift features and reduced dimensionality to achieve 93.03% accuracy on the LFW dataset. Face recognition have also been utilized in video retrieval systems. One of the flaws of the fisher vector is that it is very sparse. Although the dimensionality of the features have been reduced using different reduction techniques, but by doing so there is also a probability of the removal of features having important information. Moreover, Sivic *et al.*, [6] retrieved the videos based on the faces. The face recognition has been done by using the spatial orientation field. The matching of the sets has been done using the tubes of the spatial temporal volume. They have achieved 90.7% accuracy on the Hollywood Movies dataset. Whereas, Wolf *et al.*, [7] proposed a novel set to create similarity measure as well as matched background similarity to recognize the face from the videos. They have trained their proposed system using the You-tube Faces.

All these approaches used the traditional computer vision techniques which are based on the human decided features. The strength of the features is decided after the results have been generated by the system. However, the work proposed in this paper is primarily focused on the face recognition using the deep learning approaches. In deep learning, algorithm automatically picks the best features specific to the problem. Convolution Neural Network have taken the computer vision community by storm. The accuracies related to the recognition task have intensely increased when convolution neural network is used. One of the major reasons is the availability of the large-scale dataset. Parkhi *et al.*, [13] passed the training data to the convolution neural network. The network contains a total of 18 layers which is the combination of convolution, RELU, and fully connected layers. They composed the large scale dataset which contains 2.6M images, of 2.6k people and achieved 97.3% accuracy. Similarly, Schroff *et al.*, [14] proposed the deep convolution neural network named as Face Net. Firstly it extracts features from the input image and stores them into euclidean space, then for the test image it extracts the features and compare them with the features present in the euclidean space using euclidean distance. They have utilized the LFW, YouTube Faces DB dataset and achieved 95.12% accuracy.

But it did not perform well when there are different variations in face attributes. Predicting face attributes in the wild is challenging due to complex face variations. Liu *et al.*, [15] propose a novel deep learning framework for attribute prediction in the wild. It cascades two CNNs, LNet and ANet, which are fine-tuned jointly with attribute tags, but pre-trained differently. LNet is pre-trained by massive general object categories for face localization, while ANet is pre-trained by massive face identities for attribute prediction. Their proposed framework not only outperforms the state-of-the-art with a large margin, but also reveals valuable facts on learning face representation. They have evaluated their methodology on CelebA as well as LFWA dataset with accuracy of 83%.

Wu *et al.*, [16] used the facial landmark for face recognition. They extracted the features present in the intermediate layer of convolution neural network to revert the facial landmark attributes. So, they proposed the novel CNN architecture, which separates the facial landmark attributes of specific poses and appearances. They have used Annotated-wider-face (AFW) and Annotated-facial-landmarks in the wild (AFLW) dataset and were able to achieve 81.7% accuracy. The combination of two neural network architectures has also been explored for efficient face recognition.

Sun *et al.*, [17] proposed the deep convolution neural network named as DeepID3. Their convolution neural network is the combination of VGGNet and GoogLeNet. Their architecture is based on the convolution as well as inception layers. They achieved 96% accuracy on the test data using the LFW dataset. The 3D face modeling is also used to extract features by applying the piecewise affine transformation. Taigman *et al.*, [18] proposed the system based on the 3D face modeling. The nine layer convolution neural network have been applied for this purpose. They achieved 97.35% on LFW dataset. While the proposed method of Sun *et al.*, [19] used the convolution neural network for face recognition. They took the features from the last hidden layer of convolution neural network. These features are composed of various part of the face to form complementary and over complementary representations. They have achieved 97.45% on LFW dataset. The above proposed architecture by Sun *et al.*, [19] creates a bottleneck. Taigman *et al.*, [20] proposed the method to overcome the problem of the bottleneck in convolution neural network by replacing the naive random sub-sampling of the training set with the bootstrapping process. They have achieved the approximately 96% accuracy on LFW dataset.

A. DEEP LEARNING IN EDGE COMPUTING

As artificial intelligence impacts every field of life, so it has also changed the learning perception of the students, lectures and the instructors. Research community have shown great interest in work related to the smart class rooms. The perception about the smart class room is very broad and has changed with respect to time. Some researchers utilized the interactive white boards and other multimedia devices to setup the smart class rooms [21]. The voice commands and gestures of the instructors have also been utilized for implementing the smart class rooms. Deep learning approaches are also used for making the class rooms smart [22]. Dash *et al.*, [23] detect the hand gestures and recognize the hand movements of the instructors for the video stream coming from the class using

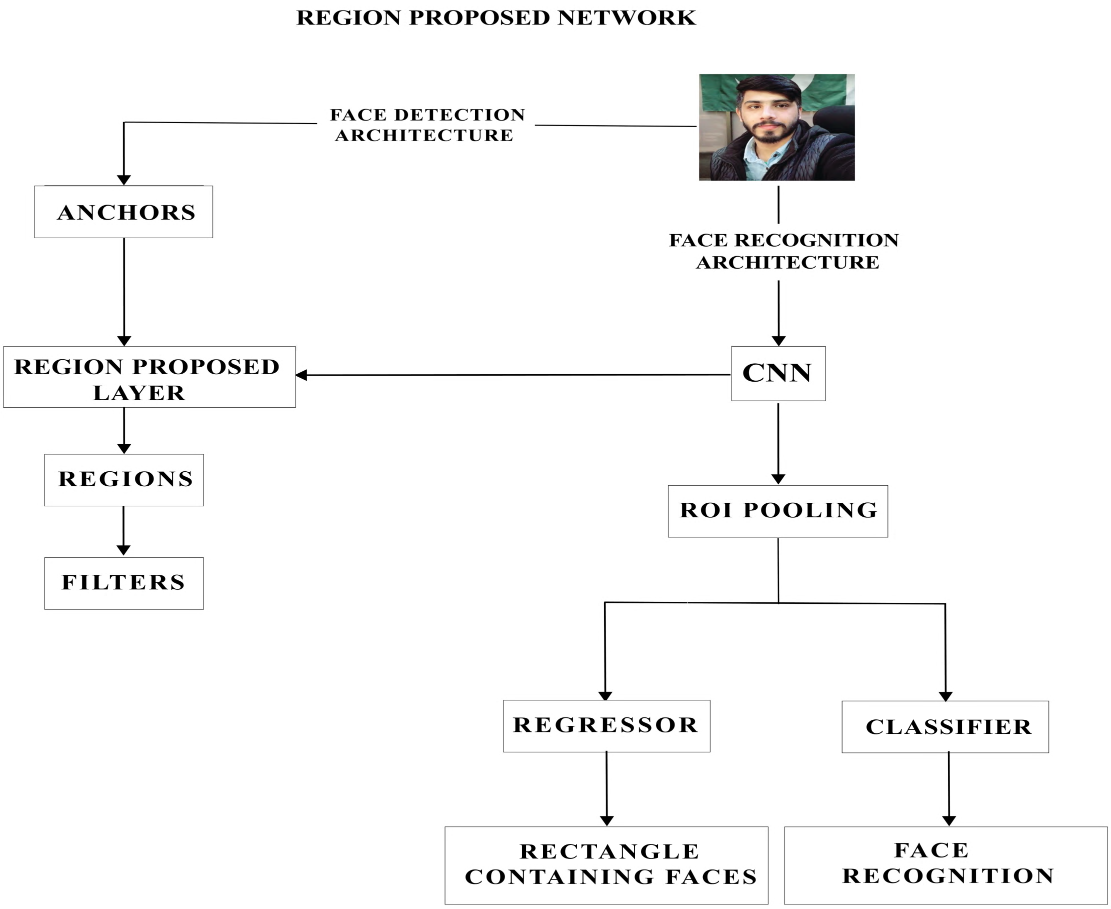


FIGURE 1. Network flow diagram.

the Internet of Things (IoT) architectures. Kim *et al.*, [24] used the pre-trained deep learning models to detect the emotions of the students present in the class. All above described approaches for face recognition performed well when there are limited number of faces in a frame. One of the main objectives and the focus of the proposed algorithm in this paper is face recognition from the images as well as from the videos.

Currently, most of the techniques in the literature for face recognition, first detect the face using the image processing techniques and then pass the detected face to CNN for recognition, however, the proposed architecture used the Convolution Neural Network for both detection and recognition purpose. So, our proposed system not only detects the appropriate number of faces from the frame, but also recognizes the detected faces with distinct accuracy as compared to existing solutions in the literature using the deep learning approaches. As, per our best knowledge there is no work in the literature that used the deep learning approaches for face recognition in order to implement the smart class room based on face recognition attendance system using the edge computing.

III. METHODOLOGY

The proposed methodology in this paper is inspired by Ren *et al.* [25] method for face detection and recognition purpose. The proposed model has two streams; one is for Region Proposal for detected faces and other one is used to recognize the detected faces. The architecture of the proposed algorithm is shown in Figure 2. The previous versions of the Region Convolution Neural Network is in selective search methods to detect the objects boundaries with in the image. However, this technique is very costly in terms of time and computation [26]. The time cost of the Region Proposal Network is less as compared to the selective search as it shares most of the computation time with detected object recognition network. Network flow of the proposed system is illustrated in Figure 1.

A. FACE DETECTOR

For face detection purpose, Region Proposal Network (RPN) draws anchors and outputs the one which most likely contains the objects. An anchor is box which is drawn overall on the image as shown in Figure 3. Top 2000 anchors based on scores on the image from the multiple anchors are selected.

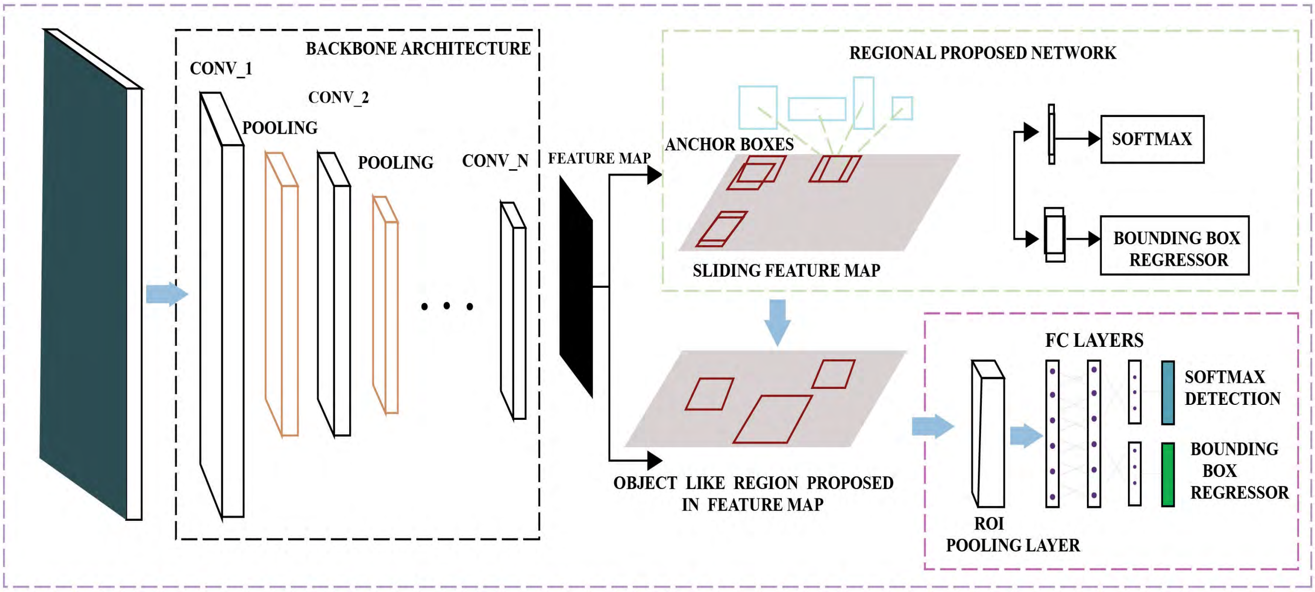


FIGURE 2. Network architecture.

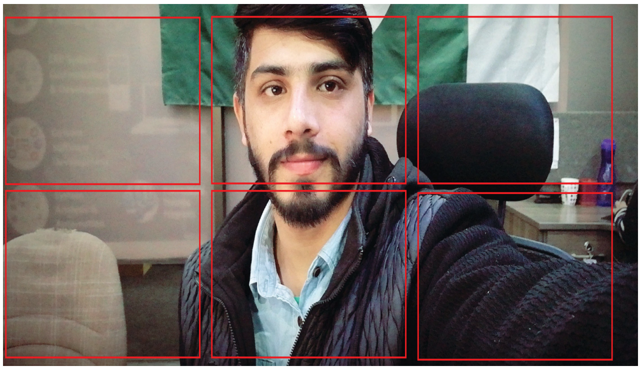


FIGURE 3. Anchors on the image.

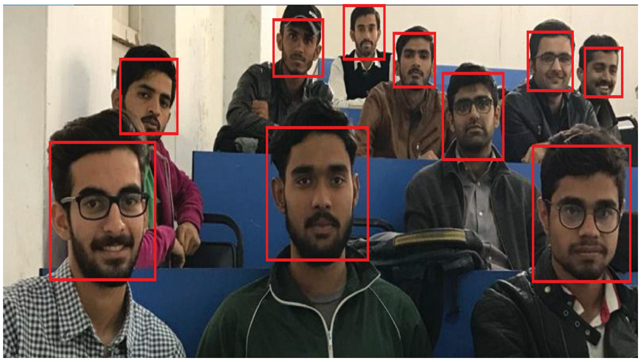


FIGURE 4. Detected faces from class room image.

RPN is a light weight network that scans the images with the help of the sliding windows over the anchors. After that it returns the anchors that have the maximum probability of containing the objects as depicted in Figure 4. To check the similarity of the particular bounding box with the other bounding boxes the intersection over union overlap is used. It can be calculated using equation 1.

$$\text{IOU:} \frac{Area\ of\ overlap\ of\ compare\ boxes}{Area\ of\ Union\ of\ compare\ boxes} \tag{1}$$

For calculating the loss of RPN's during training time the algorithm assigns the binary label to each anchor drawn on the image. The positive value is assigned to the anchor who has the value of intersection over union overlap greater than 0.7. Whereas negative value is assigned to the anchors who have intersection over union value less than 0.3. The anchors that are neither negative nor positive did not take part during training time. From these definitions, loss function is described by equation 2.

$$Loss(P_k, T_k) = \frac{1}{n_{class}} \sum_k M_{class}(P_k, P_k^*) + \lambda \frac{1}{n_{reg}} \sum_k P_k^* M_r eg(T_k, T_k^*) \tag{2}$$

Here k represents the number of objects. P_k represents the probability predicted by the network of the anchor having the object, whereas P_k^* is the ground truth probability which is 1 if it is positive and 0 in the negative case. T_k represents 4 coordinates of image where object is present based on the network prediction and T_k^* associated with the gorund–truth value of the bounding boxes. $P_k^* M_r eg(T_k, T_k^*)$ depicts that regression loss is activated when the anchor value is positive otherwise it remains deactivated. The sliding window of the RPN scans all the regions in parallel due to convolution nature of the RPN using GPU. RPN does not scan the overall image directly, instead, it uses the features that were extracted from the other network with in the same architecture which is named as the backbone feature map.

This will help the RPN to extract the features efficiently and avoid duplication of features. After getting the Region of Interest the proposed model predicts two outputs. One for the classification of the detected object and other for the bounding box size and location. But for the classification of the detected object, there is a challenge. Due to variable sizes of the Region of Interest (ROI) proposed by the RPN, classification does not handle well. Because classification requires fixed size of the image. This problem is solved with the help of the ROI pooling, which crops the particular part of the image containing the object and then resizes it. The back–bone architecture for convolution features for classification of detected objects is described in the sub–section B of the methodology section.

B. PROPOSED CNN FOR FACE RECOGNITION

Primarily, deep convolution neural network architecture is developed for the recognition of 2622 distinctive entities, structured as to solve the problem of the N–ways classification. The CNN applied to each training frame $I_x, x = 1 \ldots .X$, a score is computed by using $y_x = w\pi(I_x) + B\varepsilon\mathbb{R}^n$ through final fully–connected layer having N linear predictors $w\varepsilon \mathbb{R}^{n\times D}, B\varepsilon\mathbb{R}^n$ each for one identity. These scores are then compared to the ground truth label against each class to compute the loss value. After training, the classification layer has been removed and the score vectors are used for face verification using the Euclidean distance.

TABLE 1. Parameters of CNN architecture.

No	Layer	Filter Size	Kernel	Stride	Padding
1	Conv	11x11	96	4	0
2	Conv	7x7	128	1	2
3	Conv	5x5	256	1	1
4	Conv	3x3	256	1	1
5	Conv	3x3	384	1	1
6	FC	1	4096	1	0
7	FC	1	2622	1	0
8	FC	1	2622	1	0

The input to our proposed architecture is the cropped face image of 224 × 224 × 3. It comprises a total of 8 blocks, out of which five blocks are convolution and three are fully connected layers. The parameters of the convolution architecture have been depicted in Table 1. Each convolution layer is followed by the non–linearities such as rectification layer (ReLU) and Max Pooling. As there are three fully connected layers, the output of the first two fully connected layers are 4096 dimensional and the last fully connected layer has N = 2622 dimensional output depending upon the number of classes of the used dataset. Softmax layer has been placed after the second fully connected layer to normalize the un–normalized vectors and make the predictions into the probabilistic form. This can be achieved using equation 3

$$\text{Probabailistic score: } s_y = \frac{e^{s_y}}{\sum_x e^{s_x}} \quad for\ ALL\ y\ in\ \{1, 2..n\} \tag{3}$$

C. TRAINING PARAMETERS OF THE PROPOSED NETWORK

The proposed algorithm utilized the Labeled Faces in the Wild dataset. It is the database designed for the recognition of the unconstrained face recognition. Some frames of the dataset are shown in Figure 7. It contains more than 13000 images collected from the web. Label of the each image is the name of that particular person.The database contains images of 5749 unique individuals. More than 1620 people have two or more distinctive photos in the dataset [3].

During training of the convolution neural network, the primary task is to find the best parameters that minimize the average loss of the training data after softmax layer. The weights of the proposed architecture are initialized using the random sampling of the Gaussian Mixture with zero mean and standard deviation of 10^{-2}. Initially the biasness is set to be 0. All the training images are rescaled in such a way that the minimum width and height of the frame is 256.

During training, initially the network is fed with 224 × 224 pixels patches which are cropped from the training images. Data augmentation of the images has been performed flipping left to right having 50% probability, but did not perform the color augmentation [9], [27]. The weights have been optimized through stochastic gradient descent, by using the batch size of the 64 with the coefficient momentum of 0.9 [28]. Weight decay and momentum are utilized for the regularization of the model. Later the coefficient was set to 5×10^{-4} and drop out layer is used after the fully connected layer with the ratio of 0.5. Initially learning rate is 10^{-2}, after that it will decrease with the multiple of 10^{-1} if the accuracy on validation data stops increasing.

D. LOSS FUNCTION

In most of the machine learning and deep learning algorithms, the error is calculated by comparing the actual label with the predicted label. The function used to compute the error is known as the loss function. Learning of a machine is estimated with the help of the loss function. Basically, it is the technique to measure how well the model is trained on the given data. If the predictions on the training data deviate too much from the actual results, the value of the loss function will be high.

This value can be reduced gradually with the help of some optimization parameters. Loss function has significant influence on the performance of the model. Mean Square Error loss function is utilized for such calculations. Mean Square Error is one of the most commonly used loss function. It is measured by the taking the difference of the average of the actual labels with the predictions. This loss function is only apprehensive with the average magnitude of the error regardless of their direction. Moreover, due to square, the value of the predictions which are far away from the actual predictions are penalized heavily as compare to the less deviated predictions. The computation power is very important, because of the large amount of the data set and the complexity of the deep neural architectures.

$$\text{Mse:}\frac{1}{N}\sum_{n=1}^{N}(\hat{P}_n - P_n)^2 \tag{4}$$

Here in equation 4, N is the number of the data points, while $\hat{P}$ is the predicted values and P is the actual values of particular sample.

E. SMART CLASS ROOMS VIA EDGE COMPUTING AND DEEP LEARNING

The amount of the data produced by the (Internet of Things) IoT devices have increased tremendously on the clouds. So, the efficient way to process this data is at the edges of the specific network, rather than to pass it to the cloud. Several techniques like micro –data centres [29], fog computing [30] and cloudlet [31] have been utilized for IoT based architectures, because cloud computing is not so good for processing of the data generated at the edges of the network with good response time. Placing all the processing data into the cloud is a good approach, because it has the needed memory capability and processing power by comparison to the edges of the node. But when it comes to the processing and mining from the huge amount of the data, the processing time and data latency become high. Due to the huge data at edges, the speed of the transmitted data creates the bottleneck for the cloud computing phenomena. Consider, the example of the autonomous vehicle, it produces 1 Gigabyte data after each second and requires the real time processing to make correct decisions. In case, we sent the data to the cloud, the processing time and data latency will be increased. So, if there are large autonomous vehicles in an area it will be challenging to process the data with current bandwidth and data latency. So, in this case the data needs to be processed at the edges, with less response time, shorter and efficient processing with an insignificant network pressure.

The edge computing processes the data at the edges of the nodes, here edge is a computing device and network resource along with the dedicated path of generated data sources and cloud data centres. For example, smart device like mobile phone is the edge between the holding body and the cloud data centres, a gateway is the edge between the home appliances and the cloud. A microdata center and the cloudlet are also examples of edges between the mobile devices and the cloud computing. The cloud computing is also defined as the process in which the processing of data takes place at the proximity of the data generation sources. Often, edge computing and fog computing are used interchangeably. But there is a slight difference between these two paradigms, edge computing is more focused on the things, whereas the fog computing is more concerned towards the infrastructure side. In edge computing, the things are not only the data producers but they also work as consumer of the data. At the edge, nodes not only gets the content and services from the clouds, but they can also perform processing on the data. Cloud computing is not suitable for the tasks where the processing of the visual data is involved due to long data latency and late response. Searching from the bulk amount of data present in cloud and then processing it, requires huge time. So the possible solution is, to process the generated data at the edges of the network to enhance the data latency and response time.

In our proposed system, we have a number of class rooms of a specific institute in which we set–up our face recognition system for making a smart class rooms. Several images from different smart class room Buffys are being sent simultaneously for processing, in order to take the attendance. All class rooms are connected to the gateway device which is placed in some central place of the institute. At the time of the registration of the user, all the data is passed to the clouds, because at that time data latency rate and response time does not matter. The necessary cloud data is synchronized into the gateway device after the particular time stamp. But at the time when teacher upload the attendance quick response time really matters. To getting the names of the recognized persons, at the time of attendance the captured image from the device is passed to the gateway which is edge node in our case. The predicted names are passed back to the user interface, from which the image is captured. After taking the decisions based on the prediction the names are passed to the cloud for generating reports of the attendance of the specific day. The system flow of this process is shown in Figure 5. Based on the prediction the system generates, the teacher or smart class room will have to take actions, like door lock, turn on the electronic white boards etc, so this process must be a performed in an efficient manner. In our proposed edge computing architecture, at the time of attendance, the uploaded image is passed to the gateway device, it returns the predicted information in a shorter time per comparison to the time it would take if the data sent to the cloud. The architecture includes the User Interface and Backend services developed in Csharp. Net, face recognition model, whereas HTTP M2M protocol is used for data transfer between the devices and micro server implementation.

IV. PRACTICAL IMPLEMENTATION OF THE PROPOSED SYSTEM

A. PROPOSED SYSTEM

In order to measure the validity of the proposed algorithm, a web application of a group–based face attendance system is developed. The input to the proposed system is in the form of the image. The image can be uploaded either from the directory or by capturing through device's camera. Two types of users faces have been used; First one is the student, who can add and update his record. The second one is the teacher. Teacher has privilege to add, update as well as mark attendance functionality based on the images of the group

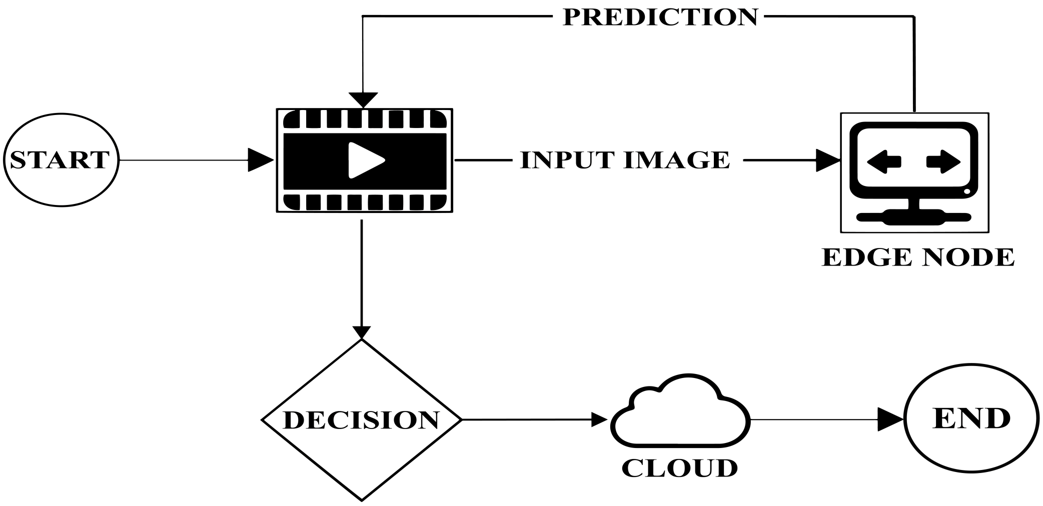


FIGURE 5. Data transmission architecture based on Edge Computing.

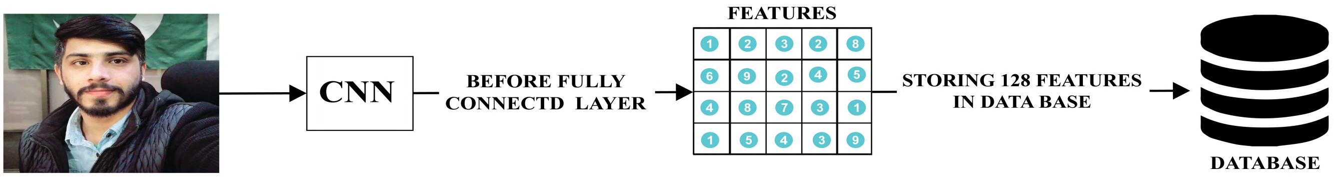


FIGURE 6. Extract encodings and save into database.

of students. The teacher can also view the reports of class attendance based on the specific subject id, and date.

B. REGISTRATION

Our system works in the following manner, firstly an admin assigns a username and password to the user (user can be either a teacher or a student). Username is the registration number in case of a student, while in case of a teacher, the username is teacher id. After assigning the user name and password to the teacher, Admin will register a subject for that teacher as well. Once a user is assigned a username and password, they can go to the website and log in with the help of username and password. After logging in a user can update his/her password, and he/she will have to add his/her personal information along with six photos of the face from different angles. Encodings against each image is performed and saved into the database as shown in Figure 6.

C. FACE BASED LOGIN

After Registration Step, Students can have the facility of face based login. The system matches the encoding of the image that user uploads with all the encodings of the faces stored in a database. After logging in, the user can update his/her images as well. During updating, user does not have to upload all the six images, it's all up to the user which image to update. Validations for the images are set in case a user uploads a blurry picture or a picture without a face the system will generate an alert for a user to upload a picture with a properly aligned face.

D. MARK ATTENDANCE

Once all the students of a class get registered, Teacher will be able to mark the attendance by capturing a photo of students. Teacher will capture the image of students and upload it on our server and gets a list of students present in that image. The technique of recognition is depicted in Figure 8. Now the question arises, how he will be able to cover all the students in a single picture. He is not bound to capture only a single picture. The number of pictures depends on the strength of the class and in how many pictures he can easily cover the whole class. And if somehow all the faces are not recognized, teacher will have the privilege to mark the attendance of those students manually on the list and updates it in the data base. Teacher can also generate the report of a subject on the specific date, that report will be downloadable in PDF format.

V. EXPERIMENTS AND RESULTS

A. IMPLEMENTATION DETAILS

For the purpose of training of the proposed architecture tensor flow deep learning frame work is used. Nvidia CUDDN libraries have been utilized to accelerate the training process with the help of GPU. Nvidia 1080 ti GPU is used for the training of the dataset.

FIGURE 7. Some sample frames from dataset.

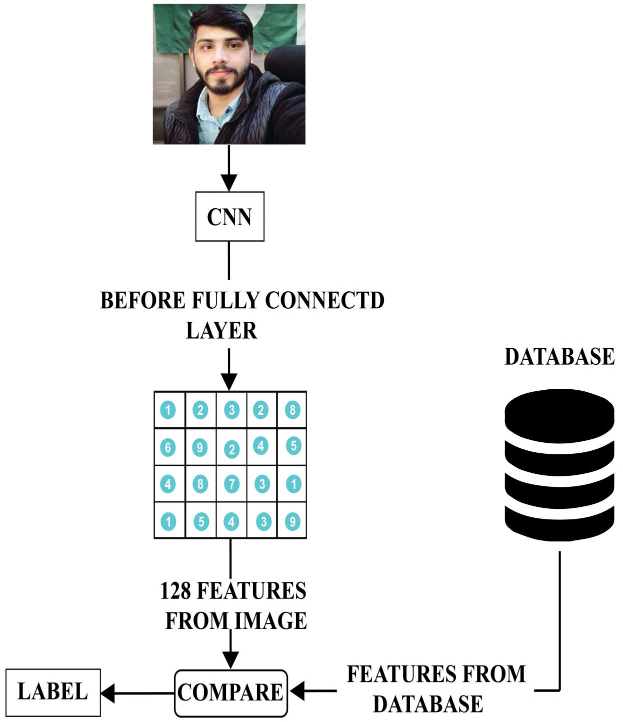


FIGURE 8. Flow of face recognition process.

TABLE 2. Quantitative evaluation.

	Total students	Detect faces	Recognize faces
class1	35	35	30
class2	31	28	25
class3	33	30	29
class4	26	26	24
class5	34	32	28

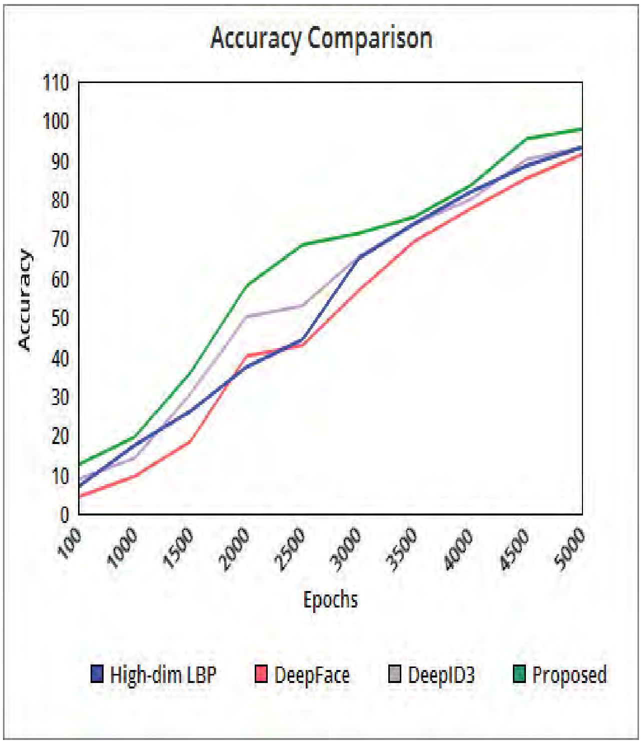


FIGURE 9. Accuracy comparison.

B. EVALUATION

Proposed algorithm has been evaluated both qualitatively and quantitatively. To evaluate the system quantitatively, survey are done by 5 different subjects. 35 students were registered in a particular class. The face encodings of the students have been stored into the database. These face encodings include the 128 convolution features which are obtained after the last convolution layer of the architecture. The teacher took the image of the class by making sure that all faces of the students are incorporated. The statistics of this activity is illustrated in Table 2. The first column of Table 2 represents the specific class in which teacher took the images for attendance. From the results shown in Table 2, it is concluded that our system performed efficiently in all the different classes. However, some of the challenges which proposed system faced are; like increasing beard, glasses, and tilted face. So, if the accuracy of the proposed system is measured in terms of percentage, the system achieved 94.6% accuracy in face detection and 85.5% in recognition concluded from the Table 2. For qualitative evaluation, 3000 given face pairs were fed to the proposed system to verify whether they are of the same person or not. These images are from the validation data which was split on the LFW dataset. The system achieved 99.64% mean accuracy. In figure 9, the proposed system is compared with the previous state of art algorithms of face recognition. It is

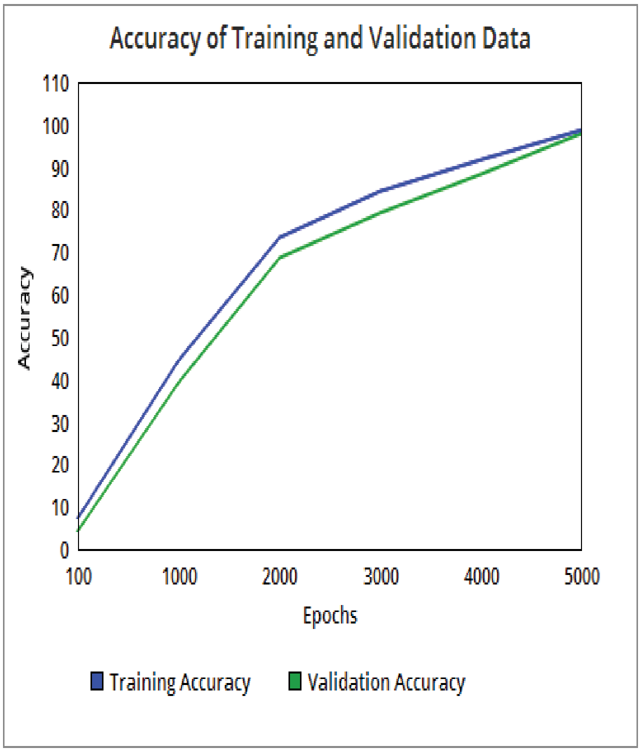


FIGURE 10. Training and validation accuracy.

clearly depicted that proposed algorithm has performed better at each epoch in terms of mean accuracy.

The proposed model is trained on the LFW wild dataset [3]. The data is split into 80% training and 20% validation. The proposed model is trained till the 5000 epochs. The learning rate is set depending upon the loss value of training. Initially the learning rate is fixed at 10^{-2} and then it decreased with the factor of 10 if the loss value and validation accuracy is stops improving. Early stopped function is utilized to stop the training if the loss value and validation accuracy stop improving. A maximum accuracy of 97.9% on validation data is achieved. Training and validation accuracy against each epoch of the trained model is illustrated in figure 10. Each experiment during training of the proposed architecture carried out till minimum loss value is achieved. The main parameter to determine how accurate the model is trained is loss value. If the value of the loss function is high, it means most of the prediction that model makes on the validation set are wrong. Whereas, if the loss value is quite low it shows that the model is accurately trained on the dataset. If there is a large gap between training and validation accuracy and loss value, it means that model is going towards the over fitting. Over fitting is also a major problem while training the deep neural network.

Over fitting occurs when the model learns the features as well as noise from the training dataset in such a way that the model performs inaccurate on the unseen data. Mostly, this happens when the noise and fluctuation of the training set is also learned by the model. The main problem of the over fitted model is that, it is does not perform good on new data and adversely affect the model's capability to generalize. To prevent from over fitting drop out was used. Initially, the drop out value of 0.7 was set, it can be increased depending upon the fixed threshold of the difference between the training and validation loss and accuracy.

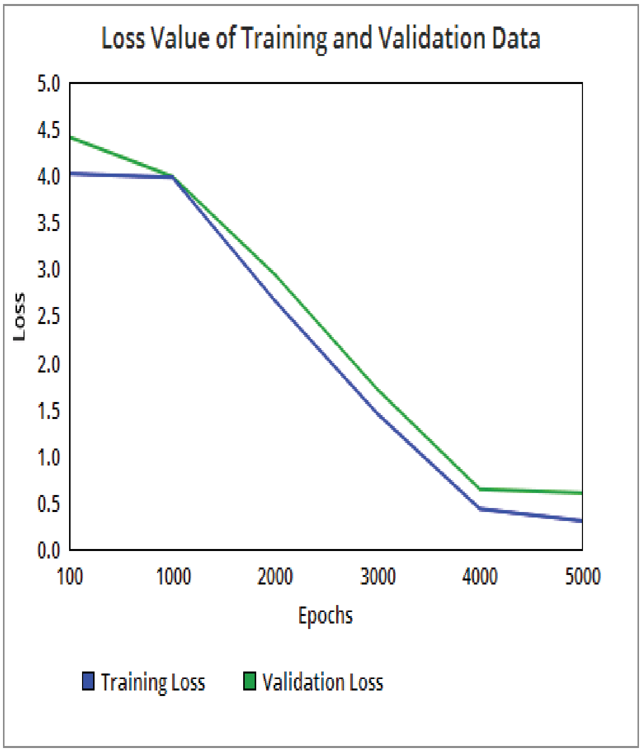


FIGURE 11. Loss on training and validation data.

TABLE 3. Accuracy comparison.

Algorithms	Accuracy	Parameters
Deep Face [20]	97.35%	CNN
Video Fisher Vector Faces [27]	93.10%	Fisher Vectors
Proposed	97.9%	Faster RCNN

Training and validation loss of the trained model is shown in Figure 11 after each epoch. The accuracy of proposed model on validation data is compared with the best available algorithms trained on LFW wild dataset. It can be observed from the Table 3, that the results achieved by the proposed model are better than the state of art algorithms till now by using simple neural network architecture.

To evaluate the strength of the edge computing in terms of the data latency and response time, we have pass the five images each to our edge node device and the cloud for getting the names of the students present in it. Figure 12 shows the comparison and it clearly depicts that the computing and processing which is performed in the edge node outperforms the computing done at the cloud in terms of response time and data latency. The images processed by

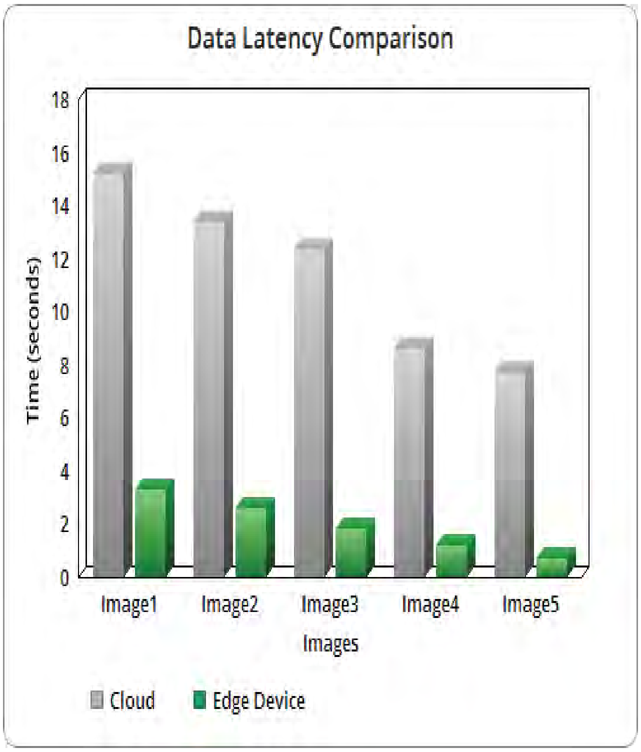


FIGURE 12. Data latency comparison.

both architecture contain the number of persons in a following manner; Image1 contains 30 faces, Image2 contains 25 images, Image3 contains 20 faces whereas Image4 and Image5 contain 15 and 10 faces respectively. The reason of doing so is to evaluate the architectures, when the system is performing the processing with different number of faces in the image. Figure 12 shows the data latency rate for both architectures tested on five images.

VI. CONCLUSION

This paper proposes an algorithm for face detection and recognition based on Convolution Neural Networks (CNN), that outperforms the traditional techniques. Automatic attendance system has been anticipated for the purpose of minimizing the human errors which take place in the conventional attendance taking system to validate the efficiency of the proposed algorithm. The basic aim is to automate the system and implement the smart class room which is useful for educational organizations. Faster Region Convolution Neural Network along with the Edge Computing techniques are utilized to achieve the state–of–the–art results. The system managed to recognize 30 faces out of 35 detected faces, the achieved accuracy can be more enhanced by taking clearer image of students. Although the system is achieving higher accuracy, but the main limitation of the system is distance, naturally as a distance increases, the picture becomes blurry, so the system produces false results on the blurry faces in some cases. The system works well if pictures are taken from around 20–25 feet. However, the outcome so far is very encouraging and promising. To increase the data latency and response time between the devices edge computing techniques have been utilized. The proposed method is secure, reliable and easy to use. No additional hardware and software are required for the utilization of the proposed system. Proposed system is currently working in collaboration with LMS (Learning Management System) of different educational institutes. The future of this work is to enhance the robustness of system by overcoming the following challenges : tilted face, moustache and growing beard. Furthermore we are planning to work on observing introvert and extrovert behavior based on our proposed face recognition algorithm.

REFERENCES

[1] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the Internet of Things with edge computing," *IEEE Netw.*, vol. 32, no. 1, pp. 96–101, Jan./Feb. 2018.

[2] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[3] G. B. Huang, M. Mattar, T. Berg, and E. Learned–Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, Oct. 2008, pp. 1–11.

[4] R. G. Cinbis, J. J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in TV video," in *Proc. ICCV*, Nov. 2011, pp. 1559–1566.

[5] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with gaussianface," in *Proc. AAAI*, 2015, pp. 2307–2319.

[6] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: Video shot retrieval for face sets," in *Proc. CIVR*, 2005, pp. 226–236.

[7] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, Jun. 2011, pp. 529–534.

[8] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman, "A compact and discriminative face track descriptor," in *Proc. CVPR*, Jun. 2014, pp. 1693–1700.

[9] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. BMVC*, 2013, p. 4.

[10] J. Sivic, M. Everingham, and A. Zisserman, "'Who are you?'—Learning person specific classifiers from video," in *Proc. CVPR*, Jun. 2009, pp. 1145–1152.

[11] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer, Oct. 2012, pp. 566–579.

[12] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with Gaussian face," in *Proc. AAAI*, Mar. 2015, pp. 3811–3819.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, Sep. 2015 vol. 1, no. 3, p. 6.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.

[16] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3067–3074, Dec. 2018.

[17] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," Feb. 2015, *arXiv:1502.0087*. [Online]. Available: https://arxiv.org/abs/1502.00873

[18] 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.*, Jun. 2014, pp. 1701–1708.

[19] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2746–2754.

[21] M. Alderton. (2016). *Smart Classrooms Give Tech Boost to Learning*. Accessed: Jan. 30, 2018. [Online]. Available: https://goo.gl/6RK8nB

[22] L. R. Winer and J. Cooperstock, "The 'intelligent classroom': Changing teaching and learning with an evolving technological environment," *Comput. Educ.*, vol. 38, nos. 1–3, pp. 253–266, Jan./Feb. 2002.

[23] A. Dash, A. Sahu, R. Shringi, J. Gamboa, M. Z. Afzal, M. I. Malik, A. Dengel, and S. Ahmed, "AirScript—Creating documents in air," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 908–913.

[24] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards emotionally aware AI smart classroom: Current issues and directions for engineering and education," *IEEE Access*, vol. 6, pp. 5308–5331, 2018.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[26] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

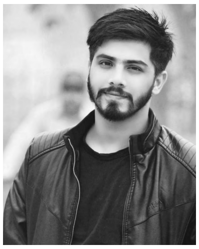
[29] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 68–73, Dec. 2009.

[30] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Comput.*, vol. 1, no. 4, pp. 14–23, Oct. 2009.

[31] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st Ed. MCC Workshop Mobile Cloud Comput.*, Helsinki, Finland, 2012, pp. 13–16.

MUHAMMAD ZEESHAN KHAN is currently pursuing the M.S. degree in computer science with the University of Engineering and Technology Lahore, Pakistan, where he is currently a Research Officer with the Computer Vision and Machine Learning Laboratory, Al–Khawarizmi Institute of Computer Science. His areas of specialization are computer vision, machine learning, deep learning, and block chain.

SAAD HAROUS received the Ph.D. degree in computer science from Case Western Reserve University, Cleveland, OH, USA, in 1991. He has more than 30 years of experience in teaching and research in three different countries including USA, Oman, and UAE. He is currently a Professor with the College of Information Technology, United Arab Emirates University. His teaching interests include programming, data structures, design and analysis of algorithms, operating systems, and networks. He has published more than 150 journal and conference papers. His research interests include parallel and distributed computing, P2P delivery architectures, wireless networks, and the use of computers in education and processing Arabic language.



SALEET UL HASSAN is currently pursuing the M.S. degree in computer science with the University of Engineering and Technology Lahore, Pakistan, where he is currently a Research Assistant with the Computer Vision and Machine Learning Laboratory, Al–Khawarizmi Institute of Computer Science. His areas of specialization are computer vision, machine learning, deep learning, and block chain.

MUHAMMAD USMAN GHANI KHAN received the Ph.D. degree from Sheffield University, U.K., concerned with statistical modeling for machine vision signals, specifically language descriptions of video streams. He is currently an Associate Professor with the Department of Computer Science, University of Engineering and Technology, Lahore. He has been studying on spoken language processing using statistical approaches with applications, such as information extraction from speech and speech summarization. His recent works are concerned with multimedia, incorporating text, and audio and visual processing into one frame work.

RAZI IQBAL (M'12–SM'18) received the master's and Ph.D. degrees in computer science and engineering from Akita University, Akita, Japan. He is currently an Associate Professor with the College of Computer Information Technology, American University in the Emirates (AUE), Dubai, UAE. He served as the Chairman of the Department of Computer Science and IT, Director of the Office of Research, Innovation and Commercialization and a Research Scientist. His current research interests include short range wireless technologies in precision agriculture, transportation, and education. He serves as an Editor and a Reviewer for several peer-reviewed journals. He is a member of the IEEE Computer and Computational Society.

SHAHID MUMTAZ received the master's degree in electrical and electronic engineering from the Blekinge Institute of Technology, Sweden, in 2006, and the Ph.D. degree in electrical and electronic engineering from the University of Aveiro, Portugal, in 2011. He has been with the Instituto de Telecomunicaces, since 2011, where he is currently an Auxiliary Researcher and adjunct positions with several universities across the Europe-Asian Region. He is also a Visiting Researcher at Nokia Bell labs. He has more than 12 years of wireless industry/academic experience. He has authored four technical books, 12 book chapters, more than 150 technical papers (100+ Journal/Transaction, 60+ conference, and two IEEE Best Paper Award in the area of mobile communications. He is an ACM Distinguished speaker, the IEEE Senior member, Editor-in-Chief of the *IET journal of Quantum communication*, and the Vice-Chair of the IEEE standard on P1932.1: Standard for Licensed/Unlicensed Spectrum Interoperability in Wireless Mobile Networks.