Face Recognition using Deep Neural Network Technique

[1] Eman Zakaria, [2] Wael Abdel Rahman, [3] Abeer Twakol, [4] Ashraf Shawky
[1][2][3][4] Faculty of engineering/ Benha University, Benha, Egypt
[1] Emanzakaria988@gmail.com, [2] Wael.ahmed@bhit.bu.edu.eg, [3]Abeer.Twakol@bhit.bu.edu.eg,
[4] amohra@bhit.bu.edu.eg

Abstract: **In recent years, the use of Convolution Neural Network (CNN) with a huge amount of images in databases, has made the deep learning technique very beneficial. Our objective is to improve the Face Recognition system using Deep Neural Network because of the importance of this system in many applications such as security systems, mobile authentication, access control and banking using ATM. We will use a convolution neural network to make the face recognition performance being analogous to humans. CNN technique learns features discriminatively not handcrafted to improve recognition accuracy. The learned face representations are very valuable for face recognition and are also capable of reconstructing face images in their frontal views. We propose a deep neural network model which is 15-layer to learn discriminative representation, obtain and outperform the state-of-the art methods on ORL (Olivetti Research Laboratory face database) and YTF (YouTube Faces database). The comparison will be done to CNN with Fuzzy Hidden Markov Models (FHMM) and Principle Component Analysis (PCA). For our presented CNN method, we have obtained the best recognition accuracy of 99.69 %. The presented system based on deep neural network transcends the state of the art methods in the field of face recognition.**

Index terms: **Convolution Neural Network (CNN), Face recognition, ANN, Database, Layers**

1. INTRODUCTION

A face recognition system can be defined as a computer application which is capable of verifying or recognizing a person either from a video frame or from a digital image. One of the techniques to accomplish this is making a comparison between facial features that is selected from the image and a face database. Face recognition is employed in different applications and purposes such as identification of persons, psychology, imagefilm processing, security system, computer interaction, surveillance, entertainment system, law enforcement, smart card and so on. The system of face recognition [1] in general involves two phases:

- Face detection: in this phase, the input image is checked to discover any face and then for easier recognition, image processing cleans up the facial image.
- Face recognition: in this phase, the detected and processed face is contrasted with the dataset of known faces in order to determine who that person is.

The three basic approaches for existing face recognition can be summarized as follows:

- Holistic Matching Method and is also called Appearance-Based method.
- Feature-Based Approach
- Hybrid Approach

We can distinguish between these approaches based on the feature extraction method.

CNN falls under the second category which is the feature-based approach, but the values and position of the filters are automatically decided by the CNN throughout the training process. LeNet-5 [2] is the ancient CNN model. There are several works which use CNN to fit the face recognition problem. Among the first works on the face recognition victimization, the CNN which rumored by Lawrence et al. [3]. In their work, learning was performed utilizing a standard backpropagation method. Their approach had excessive complexity due to the fact that two dissimilar neural networks were consolidated to carry out the recognition functions. In 2007, Duffner and Garcia, showed an extraordinary methodology contrasted with other CNN-based work [4]. Their work is to train the system for converting the input image into a reference image which is predefined for each subject. Extra recently, Khalajzadeh et al. [5] made various level structures based CNN that was tested on Yale, JAFFE, and ORL databases. It has consisted of 4 layers and standard backpropagation are performed as the learning method. Unfortunately, both of these approaches have achieved unsatisfying recognition rate.

In 1986, the backpropagation (BP) algorithm was proposed as a common method of training deep neural networks and many years passed to reach the stage of enabling the deep architecture to work for real applications. Deep learning technique's history in computer vision is:

- In 1998, Yann Lecun et al. proposed LeNet-5 [2] which is used for classifying digits, handwritten and machine-printed character recognition.

- In 2006, Hinton proposed an architecture called Deep Belief Nets (DBN) that has a strong capability to learn high-level features.
- In 2012, Alex Krizhevsky [6] proposed an architecture called AlexNet which is used in image classification and have the ability to classify images into 1000 object categories. AlexNet consists of 8 layers deep and it drops the Top-5 error from ~26% to 15.3%.
- In 2014 Simonyan proposed the VGG Net (Visual Geometry Group Network) which consists of 19 layers with Top-5 error rate of 7.3%. Also in 2014 Google proposed the Google Net which consists of 22 layers with Top-5 error rate of 6.67%.

In recent years, all these CNN models lead to great attention in the domain of face recognition. In the Facebook AI group, researchers trained a CNN model of 8 layers called DeepFace [7] that reaches an accuracy of 97.35% on LFW. DeepFace work was extended by model series: DeepID (2014) [8] [9], DeepID2 (2014) [10], DeepID2+ (2015) [7] and DeepID3 (2015) [11]. The DeepID reaches an accuracy of 97.45%. DeepID2 is an approach of learning representations of the deep face by joint face identification-verification with an accuracy of 99.15% and was further improved by DeepID2+ network that reaches 99.47% accuracy. DeepID2 [10] an extension of DeepID, uses both identification and verification information to train a CNN objecting for maximizing the inter-class variations and maximizing the intra-class difference. DeepID3 is deeper than DeepID2+ and has accuracy of 99.53%. FaceNet [12] model was proposed by a Google research group and it achieves an accuracy of 99.6% on labeled faces in the wild (LFW) database, 95.12% on YouTube faces database.

2. THEORETICAL BACKGROUND

2.1 Artificial Neural Networks (ANN)

Artificial Neural Network is a combination of interrelated arithmetic nodes, which forms a computing system. ANN model has three rules of simple sets which are multiplication, summation, and activation as shown in figure (1).

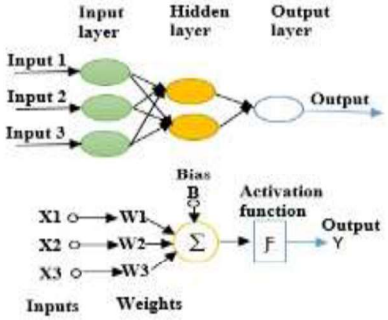

Figure 1. Simple Artificial Neural Network

In the first section, the process of weighting the inputs is executed (i.e.) multiplication between every input value and individual weight. In the second part, the process of summing bias and all weighted inputs. In the third part, the process of passing the resultant summation from the previous layer through activation function. The mathematical model of ANN is:

$$y = F\left(\sum_{i=0}^{n}(w_i x_i) - b\right)$$

Where: X: input value, b: bias, w: weight value, F: activation function, and y: output value.

2.2 Convolution Neural Network (CNN)

CNN operates on images by convolving them with banks of filters and passing the output of the convolution through the nonlinear projection to obtain the identity classification and weights of the filters are learned to minimize the loss in the identity classification. Every layer takes its input from the output of the preceding layer and also uses it to detect top-level features [13].
In CNN, number of neurons in each layer is deduced from the parameters of the layer. A CNN Model Structure example [14] is shown in figure (2).

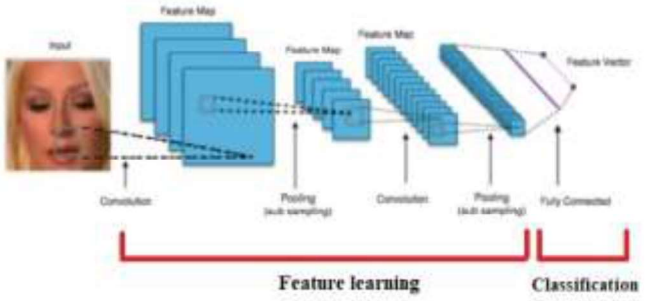

Figure 2. CNN Model Structure [14]

As we see the CNN model consists of many layers and the input of each layer is a multi-dimensional array of numbers. The main layers which build CNN architectures [15]:
Input: contains the raw pixel values of the input image.
Convolution Layer: is a feature identifier that automatically learns to filter out not needed information from an input.
Batch Normalization layer: this layer is used between convolution layers and nonlinear layers like RELU layers to reduce the sensitivity to network initialization and speeds up network training.
RELU (Rectified Linear Unit): this activation function is a = max (0,) and it prevents the gradient vanishing problem.
Pooling Layer: used for downsizing input images.
Fully Connected Layer: is essentially neural layer which is connected to all neurons from the previous layer.

2.3 Image pre-processing stages

Images are pre-processed to improve its quality and to be accepted by the architecture. The basic image pre-processing stages are shown in figure (3).

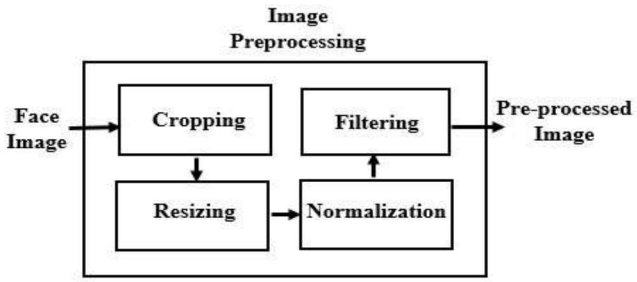

Figure 3. Block diagram of image pre-processing

- Image Cropping is an important action to reach high recognition rate.
- Image Resizing: Image resulted from different face detection techniques can be resized using nearest neighbour interpolation technique.
- Image Normalization: Distribution of intensity levels in images with uncontrolled light conditions is not alike. So many techniques such as histogram equalization can be used to make the levels of intensity equal or roughly equal.
- Image filtering and de-noising: By default, images have Gaussian noise because of the variations in illumination. Pixel based filtering such as low pass filter can be used to de-noise images.

2.4 Convolution Neural Network optimization

For image classification, the most excessively used technique is CNN. The performance of CNN relies on many hyper-parameters such as convolution layer number, size, and number of filters, depth and number of epochs. Optimization of the hyperparameters means choosing the optimal value of these hyperparameters in order to optimally control the learning process. Many techniques can be used for optimizing the CNN such as Genetic Algorithms (GA). GA technique starts with a number of chromosomes, which generated randomly. Then the process of recombination and selection based on each chromosome fitness is executed. Parent genetic materials are recombined for generating child chromosomes in order to produce the next generation. Iterating this process until some stopping criterion is achieved.

3. METHODOLOGY

A simple neural network is a sequence of layers. The Convolutional Neural Networks are extremely like common Neural Networks. As shown in figure 4 (a, b) the proposed face recognition system consists of four parts of the Neural Network.

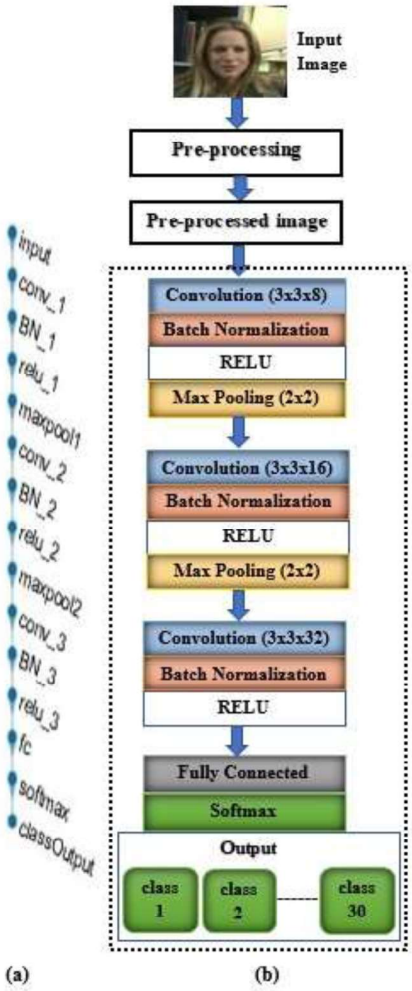

Figure 4. (a, b): (a) Layer graph of the network & (b) Flow Chart of proposed CNN system

The three basic types of layers which used to build our CNN architecture are Convolution (CONV) Layer, Pooling Layer, and Fully-Connected Layer. These layers are stacked to create a full CNN architecture which is a 15*1 Layer array:

- INPUT [227*227*3] carries the crude pixel values of the image, with zero centre normalization. The image size is 227-pixel height, 227-pixel width, and 3 channel sizes because the images are color.
- Conv_1 which is a convolution layer that evaluates the neuron's output, which is attached to local areas in the input, each making a dot multiplication between a small area they are connected to in the input volume and their weights. And in our case, they result in a volume of [227*227*8] in the first layer because we used 8 filters in this layer so convolution2dlayer (3, 8, padding, same) means creating a 2-D convolutional layer with 8 filters of size [3 3] and same padding. And results in a volume of [227*227*16] in the fifth layer

because we used 16 filters in this layer. And results in a volume of [227*227*32] in the ninth layer because we used 32 filters in this layer. At training time, the software computes and sets the zero padding so the output of the layer has the same size as the input.

- Batch Normalization is used to accelerate network training and minimize the sensitivity to network initialization.
- RELU layer: is a nonlinear activation function and is used to prevent the gradient vanishing problem.
- Max Pooling layer makes the down-sampling function along the spatial volumes (width, height) to remove redundant spatial information so maxPooling2dLayer (2,'Stride', 2) produces a max pooling layer with pool size [2 2] and stride [2 2] where a stride represents the number of pixels which are shifted on from the current position.
- Fully-Connected layer: neurons in this layer are attached to all the neurons of the preceding layer. This layer joins all the features learned by the preceding layers in order to recognize the larger patterns to be able to easily classify the images. So in the last fully connected layer, the Output Size parameter is equal to the number of classes in the target data. In our system, the number of classes is 30 corresponding to 30 outputs.
- Softmax Layer: softmax activation function is one of the most important output functions which used to normalize the fully connected layer output. The resultant of this layer composed of positive numbers where the total sum of these outputs is equal to 1 to be used by the classification layer as classification probabilities.
- Classification layer: in this layer, the softmax activation function for each input is used to evaluate the loss and specify the input to one of the mutually exclusive classes. The final output was a classified ranking relatively to 30 mutually exclusive classes.

As shown in figure (4-b) the proposed face recognition system consists of four parts of the Neural Network [16] [17].

4. RESULTS AND DISCUSSION

All our experiments were carried on a desktop computer with specifications of Intel ® Core ™ I7-5500U CPU @ 2.40GHZ 2.40 GHZ processor, 8GB Ram, 64-bit Operating System, x64-based processor, GeForce 820M NVIDIA, 96 CUDA Cores and 16.02GB/s NVIDIA memory bandwidth.
The proposed system was trained on YouTube Faces database which mainly contains 621126 face images of 1595 identities taken from 3,425 YouTube videos and we choose from it 30 subjects then making some variations on the number of images to follow the resultant accuracy in order to know the relation between the number of images and resultant accuracy. And this image database contains images with different poses, lightening, and background. We trained the proposed network on matlab with learning rate = 0.01.

4.1 Preprocessing Operation

Input face images are preprocessed to get the best image size and resolution, which can be accepted by the proposed CNN architecture. The images are cropped to the same aspect ratio to reduce the problem complexity. We train the proposed system on many different sizes of images in order to be able to choose the best image size to use it in the following stages as shown in figure (5).

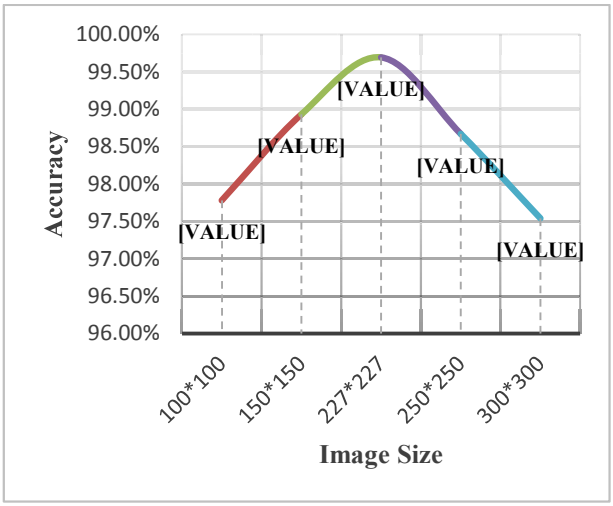

Fig (5) Accuracy based on different image sizes

So from figure (5) we deduce that accuracy increases gradually by increasing the size of images until reaching the aspect ratio of 227*227 from the raw face input image, and after that the accuracy decreases. So we will choose the 227*227 volume to be used in the next stages.

4.2 Comparison of accuracy when using different number of Images

We trained the proposed system on many different numbers of images, and the resulting accuracy is illustrated in table (1) given that the total number of subjects is 30.

Table (1) Accuracies of Face Recognition system

No of Images	No of Iterations per epoch	Maximum iterations	Accuracy (%)
300	3	150	96.67
600	6	300	99.33
990	10	500	99.44
1500	16	800	99.69
2550	27	1350	99.36
3000	32	1600	98.56

So as we see the accuracy of the proposed face recognition system which is based on Convolutional Neural Network increases whenever the number of images increases until reaching an extent of input data which is 1500 image in our system, after that the accuracy decreases with increasing the number of input images. The main cause of this is the over-fitting or over training of the system. In our system, the data are divided into 70% training and 30% validation. We can deduce that the highest accuracy achieved is 99.69% at 1500 input face images.

4.3 Comparison of accuracy when using different number of Layers

We next evaluate the performance of the facial recognition system when changing the number of layers and using the same dataset which contains 1500 images for each case as shown figure (6).

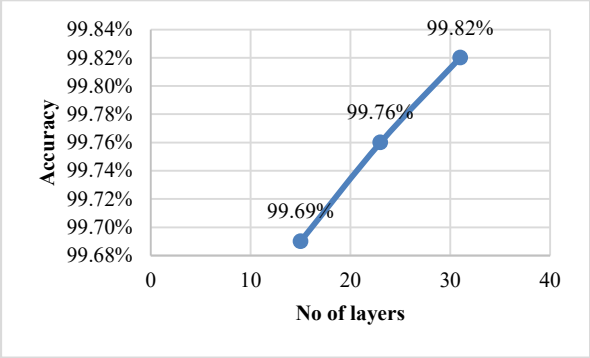

Fig (6) Accuracy based on different number of layers

So from figure (6), we deduce that adding more layers help to extract more features and consequently, the accuracy improved and making the performance better, but the time taken is very large and also more resources consumed.

4.4 Experiments on a multi-epochs for optimization

For achieving the optimization of the proposed CNN architecture, we trained our system on many different numbers of epochs to know how the number of epochs affects the results and also to determine the best number of epochs. Table (2) shows the obtained results.

Table (2) No of epochs versus accuracy

No of epochs	Accuracy (%)
20	97.43
30	97.56
40	97.78
50	99.69
60	98
70	97.56
80	97.52

So from table (2), we deduce that the accuracy increases whenever the number of images increases until reaching an extent of number of epochs which is 50 epoch, after that the accuracy decreases with increasing the number of epochs. So the best number of epochs used in our CNN system is 50 epoch.

4.5 Comparison with State of the Art

We compare our method with state of the art methods as shown from table (3).

Table (3) Comparison with State of the Art

Method	Accuracy (%)
Deep Face [Taigman et al., 2014] [7]	97.35
DeepID [Sun Yi, Wang and Tang] [8]	97.45
DeepID2 [10]	99.15
DeepID2+ [7]	99.47
DeepID3 [11]	99.53
FaceNet[Schroff, Kalenichenko, and Phinlbin] [12]	99.63
Proposed CNN system	99.69

So from the table (3) we deduce that our proposed CNN outperforms the state of the art methods.

4.6 Difference between traditional and Convolutional Neural Network in terms of Accuracy

Table (4) Comparison between traditional and CNN

Method	Accuracy (%)
FHMM	96.5
PCA	97.50
Proposed system using CNN	99.14

Comparison between traditional methods such as Fuzzy Hidden Markov Models (FHMM), Principle Component Analysis (PCA) and the Proposed CNN Method uses the same database which is an ORL face database which consists of 400 pictures of size 112 x 92 for 40 persons, 10 pictures for each person. As known, Hidden Markov Model follows the Feature-Based Approach. The Fuzzy Hidden Markov Models (FHMM) involve fuzzy integral theory and Hidden Markov Model. Using fuzzy expectation-maximization (FEM) Method in the Hidden Markov Model (HMM) is to evaluate the relative parameters of faces that are close to real values in a better condition and the weights are designed by using the fuzzy c-means (FCM) function to achieve a better result. With respect to the Principle Component Analysis (PCA) which lies under the

Appearance-Based method. PCA is a powerful tool used for analyzing data and in face recognition, it is a statistical method used for reducing the number of variables. In the training set, every image is represented as a linear combination of weighted eigenvectors named eigenfaces. The covariance matrix of a training image set consists these eigenvectors. Recognition is achieved by projecting a test image onto the subspace scanned by the eigenfaces and then classification is achieved by evaluating the minimum Euclidean distance. From table (4) we see that the CNN method outperform FHMM and PCA methods.

5. CONCLUSION

In this paper, we presented a face recognition method based on a convolution neural network (CNN). And the used network has fifteen layers. Because of very huge training databases lead to high memory usage and high computation load, which then requires high processing power to be capable of working beneficially. In our case, we use the YouTube Face Database and using different number of images. The largest number of images used are 3000 colored images with a resolution of 227x227 pixel for 30 different persons. The best accuracy achieved is 99.69%, which outperforms the state of the art methods. And we make a comparison between our proposed CNN method and two traditional methods such as FHMM method and PCA method using the ORL face database which composed of 400 grayscale images of size 112 x 92 pixels. And these 400 images for 40 people (i.e) ten images per each person. And also we found that our Proposed CNN approach outperforms these traditional approaches.

REFERENCES

[1] Meena, D. and R. Sharan. An approach to face detection and recognition. in 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE). 2016. IEEE.

[2] LeCun, Y., et al., Gradient-based learning applied to document recognition. 1998. 86(11): p. 2278-2324.

[3] Lawrence, S., et al., Face recognition: A convolutional neural-network approach. 1997. 8(1): p. 98-113.

[4] Duffner, S. and C. Garcia. Face recognition using non-linear image reconstruction. in 2007 IEEE Conference on Advanced Video and Signal Based Surveillance. 2007. IEEE.

[5] Khalajzadeh, H., et al., Hierarchical structure based convolutional neural network for face recognition. 2013. 12(03): p. 1350018.

[6] Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.

[7] Sun, Y., X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[8] Taigman, Y., et al. Deepface: Closing the gap to human-level performance in face verification. in Proceedings of the IEEE conference on computer vision and pattern recognition.2014.

[9] Sun, Y., et al. Deep learning face representation by joint Identification-verification in Advances in neural information processing systems. 2014.

[10] Wang, G., et al. Deep Embedding for Face Recognition in Public Video Surveillance. in Chinese Conference on Biometric Recognition. 2017. Springer.

[11] Sun, Y., et al., Deepid3: Face recognition with very deep neural networks. 2015.

[12] Schroff, F., D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Goodfellow, I., Y. Bengio, and A. Courville, Deep learning. 2016: MIT press.

[14] Phạm, D. V. Online handwriting recognition using multi convolution neural networks. In Asia-Pacific Conference on Simulated Evolution And Learning (2012, December). (pp. 310-319). Springer, Berlin, Heidelberg.

[15] Kamencay, P., et al., A new method for face recognition using convolutional neural network. 2017.

[16] Lee, Honglak et al. “Unsupervised learning of hierarchical representations with convolutional deep belief networks”. In: Communications of the ACM 54.10,pp.95–103. 2011.

[17] Lee, Honglak et al. “Unsupervised learning of hierarchical representations with convolutional deep belief networks”. In: Communications of the ACM 54.10,pp.95–103. 2011.