Deep Learning Face Detection and Recognition

Samar S. Mohamed, Wael A. Mohamed, A. T. Khalil, and A. S. Mohra

***Abstract*— Due to the development in the digital image processing, its wide use in many applications such as medical, security, and others, the need for more accurate techniques that are reliable, fast and robust is vehemently demanded. Face recognition technology is very important for the security that provides intelligence services. Recently, a new trend has emerged to raise the efficiency of the facial recognition systems by using neural networks. Furthermore, using Convolution Neural Networks (CNN) with a huge number of images in databases has made the deep learning technique very beneficial. The use of deep learning networks supports to learn much more complicated and high-level abstracted features automatically not handcrafted to improve recognition accuracy. Our objective is to enhance 2D face recognition accuracies based on convolution neural network which consists of 15 layers to learn discriminative representation. We use CNN training on differently aligned face images and use stochastic gradient descent algorithm to train the feature extractor and the classifier, which can extract the facial features and classify them automatically. The experiments on the Face96 database show that our proposed method achieves 99.67% accuracy.**

***Keywords*— Face Recognition, Deep Learning, CNN, Image Processing.**

I. INTRODUCTION

NOWADAYS, many application systems use biometric recognition systems because they are authentication techniques, have high recognition accuracy, and are convenient for users. Recognition techniques depend on the difference in specific physical or behavioral characteristics among people. Face recognition technology is one of the biometric identifications that is widely and mainly used to know someone among others. Face recognition has great development for various potential applications in security and emergency [1-3], Law enforcement and video surveillance [4,5], access control, smartphone login, and identification cards. Face recognition is cost effective and high reliability. Face recognition technology is using for verifying and identifying faces. Face verification system only classifies whether two faces image are the same or not, while the face identification system gives a list of matches people after it compares the presented person with all people in the database. The facial recognition system [6,7] commonly involves two stages:

- Face detection where the image of input is searched to discover all the faces in the image for recognition.
- Face recognition where compares faces detected and equipped with the known faces database to determine who this person is.

Electrical Engineering Department, Benha Faculty of Engineering, Banha University, Benha.
engsamar56@gmail.com, {wael.ahmed, abeer.twakol, amohra} @bhit.bu.edu.eg.

In the past, many methods were proposed to begin face recognition such as the eigenface analysis [8], Fisher face analysis [9], independent component analysis (ICA) [10], tensor face analysis [11], and their extensions. These approaches commonly assume that face images are well aligned and have a similar pose to the registered images of a face in the gallery. However, these assumptions are invalid in wild environments and practical applications. Face recognition system has several challenges in wild environments such as different illumination, facial poses, facial expressions, background, angle and the distance from the camera. So, there are many types of research in face recognition to defeat those obstacles and increase accuracy.

II. RELATED WORK

In this sector, due to a multiplication of the number of related papers, we briefly review some relevant face recognition methods drawn from two main categories: (A) conventional face recognition classifications and (B) deep neural network face recognition classifications.

A. The Conventional Face Recognition Classifications

Those methods have two networks training network and test network. In training network extracting features from an input image and creating the information base. In the test grid, use the information base to extract facial expressions and action units. For example, face recognition system using PCA and Euclidian Distance Classifier [12]. (Riddhi A. Vyas 2016). Different facial expressions are recognized as good face recognition rate through this method by using the PCA feature extraction technique. For Happy facial expression, it gives 93.33%, Sad facial expression gives 60 %, sleepy facial expression gives 73.33%, surprised facial expression gives 80%, Wink facial expression gives 78.57% and normal facial expression gives 86.66%.

Face recognition with Local Binary Patterns and Local Ternary Patterns (LBP + LTP) [13]. The recognition performances increase due to the increase in face images in the training set. It achieved the best performance with accuracy 98.75% on ORL dataset when the relation between training/test images 80% to 20%.

B. The Deep Neural Network Face Recognition Classifications

In recent studies, the Deep Neural Networks (DNNs) [14] has made remarkable achievement in the field of image processing. Deep Learning is presenting significant discoveries in solving the problems that have faced many struggles of machine learning and artificial intelligence community in the past. A representative approach among the recent works based on DNNs in the trend of face recognition is convolution neural networks (CNNs) which are used to reduce the distance between intra-personal of faces and enlarge the distance between extra-personal of faces.

CNNs have feed-forward networks with the capability of extracting properties from the untreated input image without any pre-processing required. Thus, the features extractions are emerged into the path of training by these networks. Furthermore, CNNs can identify models with the greatest variability, with robustness to distortions and simple geometric alterations like translation, scaling, rotation, squeezing, stroke width and noise. CNNs have several advantages such as provide the best performance in pattern recognition problems and even exceed humans in certain cases, ruggedness to shifts and distortion in the image, unique superiority in image processing because of its special structure of local weight sharing, and fewer memory requirements. The researchers of computer vision community developed the methods of deep learning to Deep Face [15], DeepID [16], DeepID2 [17], and DeepID3 [18].

DeepFace is used to distinguish personal faces images. It uses a deep neural network of a nine-layer. It employed a private database of 4 million training face images applying to more 4,000 unique identities. It achieved the best performance with accuracy 97.35% on LFW dataset.

The Deep Identification (DeepID) has been learned by designing deep convolutional networks. This the deep neural network consists of four convolution layers, three max-pooling layers, one fully connection layer to extract the features, one softmax layer for prediction, and activation layer (RELU). It achieved the best performance with accuracy on LFW dataset 97.45%.

The DeepID2 has been learned by designing deep convolutional networks. The task of identifying the face increases the inter-personal variations by using DeepID2 features extracted from different individualities apart. It achieved the best performance with accuracy on LFW dataset 91.15%.

III. Theoretical Background

A more recent trend is deep learning, which arose as a promising framework that provides the latest image processing performance. Furthermore, deep neural networks showed important conclusions to solve problems of recognition in the domain of image processing.

The idea of deep learning which is searching for the main features to be capable of distinguishing between all the images it's given and figure out the unique features that through low levels features and high-level features that are extracted by the convolution layer without using feature algorithms as presented in figure 1.

The algorithms of machine learning are separated into three levels:

- Supervised Learning which learning with a labeled data.
- Unsupervised Learning where identifying patterns in unlabeled data.
- Reinforcement learning where learning based on feedback or reward.

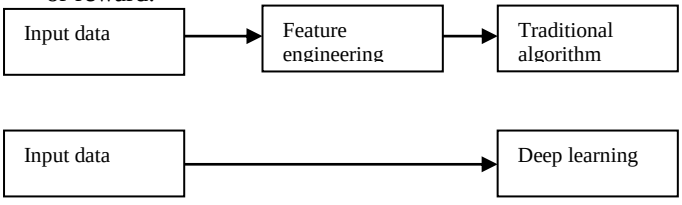

Fig. 1: Traditional learning algorithm versus deep learning algorithm

.

Deep Learning [19] is providing major discoveries in solving the issues that have withstood several tries of machine learning and Artificial Intelligence Community in the past. Deep learning has overcome many of the traditional neural network problems such as vanishing gradient problem, overfitting, and local optima. As a result, it is currently used to decipher hard scientific problems at an unusual scale, e.g. in the reconstruction of brain circuits, analysis of modifications in DNA, prediction of structure-activity of potential drug molecules, and recognize traffic sign. Deep neural networks have additionally become the well-liked option to solve several difficult tasks in speech recognition and language understanding.

DNNs [20] have produced robust machine learning models which show differences from conventional approaches for the image classification. 2012 was the first year that neural networks have grown to prominence as Alex Krizhevsky used them to win that year's Image Net competition, dropping the classification error record from 26% to 15%, associate degree astounding improvement at the time. DNNs with deep structures depend on learning complicated models and allow for learning robust object representations without the wanted to work designed features. This has been practically demonstrated on the ImageNet classification responsibility for thousands of classes. Due to the varied valuable contributions that the computer vision has created in the deep learning field, was to produce solutions for the issues encountered in medical science to mobile applications.

Figure 1 shows that the Deep Neural Network [21] has more hidden layers compared with the simple neural network which contains one hidden layer so; the deep network is more efficient than simple neural network. Each layer reconstructs the input data into more complex representations (e.g. edge → nose → face). Every layer is designed as a series of neurons and gradually extracts higher and higher-level features of the input which helps CNNs to use the context of the whole picture, not only the local parts as shown in figure 3. Those features are combined in the last layer essentially which decides about whatever the input shows. numerals.

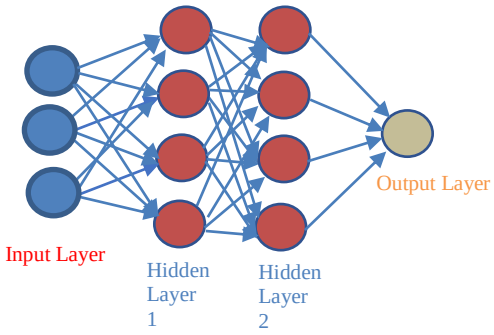

Fig. 2: architecture of deep learning network.

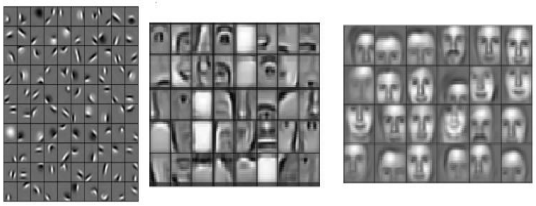

Fig. 3: Facial features extracted by a CNN [22].

IV. PROPOSED METHOD

This section presents the technique that was used in the development of a face recognition system. It was used a deep learning approach where the CNN model was developed and trained by using the different number of images.

A. Pre-Processing

Before feeding the face images into the CNN architecture, the raw input images are subjected to a set of pre-processing transformations. In this paper, the pre-processing procedures applied to the input images are as follows:

Face Detection: The first step in any automatic face recognition system. It is used to detect the face area from the background of an input image. We used the vision Cascade Object Detector to detect the location of a face in an input image. The cascade object detector uses the Viola-Jones detection algorithm.

Cropped image: After the detection step, the face area is cropped from an input image.

Image resizing: The input images were all different sizes, varying from 196x196 to 100x75 pixels. Thus, to reduce the computational cost and the complexity of the problem, all images of the database were resized to a constant value of 112x 92.

Image channels reduction: For some experiments, the input RGB images were converted to grayscale images, reducing the depth of the images from 3 to 1.

*Image normalization***:** The normalization process is made by applying the histogram equalizer on the input face image. In image processing, this technique is commonly used to enhance the contrast of images.

B. CNN Architecture

The network consists of three convolution layers; three batches normalize (BN) layers, three rectifiers linear unit (RELU) layers, two Max-pooling layers, fully connected layer, and one Softmax regression Each connection layer represents a linear mapping of different types of data. Figure 4 shows the architecture of this network. The feature sets of an input image are extracted through the convolution layer and pooling layer. Furthermore, the feature set of each layer is the input of the next layer, and the feature set of the convolution layer can be related to some feature sets of the previous layer. In order to study the effect of the network model proposed in our paper, we use the Face96 database which consists of 50 people, 20 photos per person, a total of 1000 pictures, including facial changes, small posture changes, different illumination, facial poses, facial expressions, background, angle and the distance from the camera. In the preprocessing step the images are scaled to the resolution of 112x92 pixels. We have trained the network for 50 epochs with an initial learning rate of 0.0001 and used CPU as hardware.

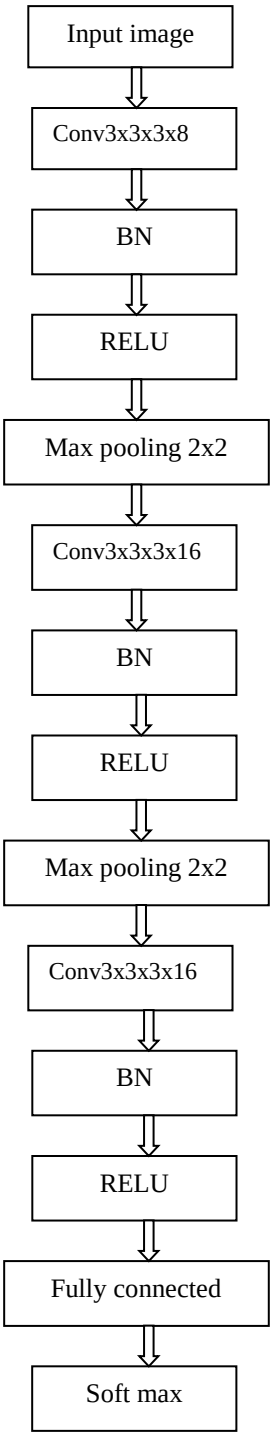

Fig. 4: The architecture of our network.

Convolution Layer

This is the distinguishing element of a CNN as compared to other neural networks. The convolution layer extracts the features automatically by using the convolution filter without external intervention. Additionally, these filters help obtain spatial features. The convolution filters are applied over the width and the length of the image by calculating the multiplication and summation between the filter and the input at any location. Additionally, the convolution filters are the contents to be learned by the convolution layer, including the weight matrix w and the bias b. In this paper, the size of the convolution kernel is 3x3 and have respectively 8, 16, 32 filters. The mathematical equation of this layer [23] is:

$$x_j^l = f\left(\sum_{i \in M_j^{l-1}} x_i^{l-1} k_{ij}^l + b_j^l\right) \quad (1)$$

Pooling layer

Pooling layers are placed among convolution layers. Pooling layers measure the max or average value of a feature across a region of the input data (downsizing of input images). Furthermore, aids to detect objectives in some unusual positions and decreases memory size. Figure 5 shows how max pooling operates. In the network, each feature map that has been put into the pooling layer is sampled, and the number of output feature maps is unchanged, but the size of each feature map will be smaller. Thus, the purpose of using the pooling layer to minimize the amount of calculation and resisting the change of microdisplacement is achieved with keeping the most important data for the following layer. In our paper, we are using the maximum pooling layer which has size 2x2 with step size 2.

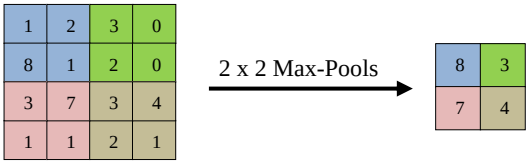

Fig. 5: Examples of how max pooling operates [24].

Batch Normalization layer (BN)

It is a technique to present any layer in a Neural Network with inputs that are zero mean/unit variance. Batch normalization layers are constructed between convolutional layers and nonlinearities such as ReLU layers to fast network training and reduce the sensitivity to network initialization.

Input: Values of x over a mini-batch: β = {x1….m}; parameters to be learned: β, ɣ
Output: {BN $_{\beta, \gamma}$ (xi)}

$$\mu\,\beta \leftarrow \frac{1}{m}\sum_{i=1}^{m} xi \quad \text{Mimi-batch mean} \quad (2)$$

$$\sigma^2\,\beta \leftarrow \frac{1}{m}\sum_{i=1}^{m}(xi - \mu\,\beta)^2 \quad \text{Mimi-batch variance} \quad (3)$$

$$xi^{\sim} \leftarrow \frac{xi - \mu\,\beta}{\sqrt{\sigma 2\,\beta + \epsilon}} \quad \text{normalize} \quad (4)$$

$$yi \leftarrow \gamma\, xi^{\sim} + \beta \equiv BN_{\beta, \gamma}(xi) \quad \text{scale and shift} \quad (5)$$

Rectifier Linear Unit layer (ReLU)

Nowadays, most of the deep networks use non-linear activation function ReLU–max (0, x) for hidden layers, since it trains much faster, is more significant than logistic function and overcomes the gradient vanishing problem.

Fully Connected Layer (FC)

The last layers of a CNN have used fully connected layers which, all the parameters of all the features of the previous layer get applied in the estimation of each parameter of each output feature. The objective of using fully connected layers to achieve the classification.

Softmax regression layer

The softmax classifier which has a strong non-linear classify ability is used at the last layer of the network because the facial characteristics are more complicated, and the face category is more and there is no uniform template. Softmax classifier is a multi-classifier, which can not only complete the dichotomy task but also can complete the multiples (greater than 2) task. Given sample vector input x and weight vectors {wi}, the predicted probability of y = j

$$P(y=j \mid x) = \frac{e^{x^T wj}}{\sum_{k=1}^{k} e^{x^T wk}} \quad (6)$$

Definition of hyperparameters

For the learning process, some parameters must be carefully considered in order to achieve the best possible performance. Thus, the parameters to be defined before the training algorithm are the following:

Batch size: The batch size is the number of samples fed to the network in one training iteration, in order to make one update to the model parameters. Since the entire dataset cannot be propagated into the neural network at once for memory limitations, it is divided into batches, which makes the overall training procedure require less memory and become faster. It should be highlighted that the higher the batch size is, the more memory will be needed and the slower is the training procedure. We used mini batch=40 in the proposed method.

Epochs: The number of epochs denotes how many times the entire dataset has passed forward and backward through the neural network, i.e., one epoch is when every image has been seen once during training. Nevertheless, this concept should not be confused with iterations. The number of iterations corresponds to the total number of forward and backward passes, with each pass using a batch and depends on the batch size, the number of epochs and number of training images. It is computed as follows:

$$\neq Iterations = \frac{\neq \text{epochs x} \neq \text{training images}}{\text{batch size}}$$

We used the number of epochs in our method Max Epochs= 50.

Learning rate: the learning rate parameter controls the step size for which the weights of a model are updated regarding the loss gradient. The lower its value is, the slower the convergence is but it is ensured that it is not missed any local minimum.

V. Result and Discussion

A. Face Database

Database Face96 [25] contains the images of 152 individuals, whereas a sequence of 20 images is taken per individual. The images have large head scale variation along with expression variation. The images were acquired under varying lighting condition and the background of images is complex.

B. Implementation Details:

MATLAB R2017b is used to execute the framework described in this paper. The decision to use MATLAB was made because Vanderlande uses the program internally for all its applications. The framework itself is meant to run on a desktop computer. The CPU used is an Intel Core i5- 5200U processor and memory 6 GB. A dataset is made of 1000 images, with 20 images for each 50 individual, each image in JPEG format. All images have resolution 112 by 92. Those images have been divided into training and validation database.

C. Performance evaluation

The performance of the face recognition system in this paper is evaluated by using different numbers of face images. The accuracies of the proposed face recognition system which is based on Convolutional Neural Network can be viewed in figure 6 where the increase in the number of images leads to increase in the accuracy of the system. But we can do that up to a certain extent then, accuracy is decreasing. So, the network tends to 'overfit' the data. Overfitting can lead to errors in some of the other form like false positives. Our system achieved high accuracy of 99.67% at used dataset consists of 1000 images. Divide the database into 70% training and 30% validation database.

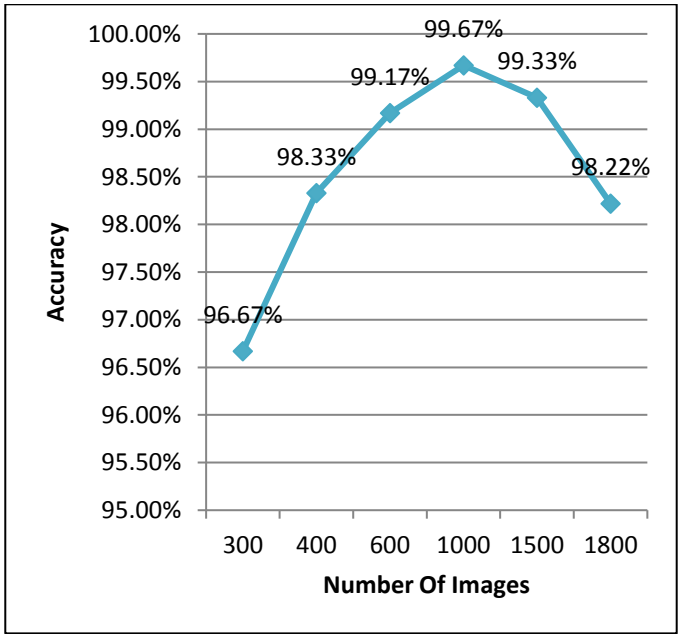

Fig. 6: Accuracy of different number of images from database Face96.

TABLE I
COMPARISON BETWEEN THE RATIO OF TRAIN TO TEST AT 1000 IMAGES

The ratio of train to test	Accuracy
(50% to50%)	98.60%
(55% to45%)	98.89%
(65% to35%)	99.43%
(70% to30%)	99.67%

Table I shows that increasing the number of training images help to improve the performance of the network.

TABLE II
THE RELATED WORK IN FACE RECOGNITION

Related work	Accuracy
PCA + Euclidian Distance [12]. (Riddhi A. Vyas 2016)	93.33%
LBP + LTP [13]. (Chi Kien Tran, Tsair Fwu Lee 2014)	98.75%
DeepFace [15]. [Taigman et al., 2014]	97.53%
DeepID [16]. [Yi Sun 2014]	97.45%
DeepID2 [17]. [Yi Sun, Xiaoou Tang 2014]	99.15%
DeepID3 [18]. [Yi Sun, Xiaoou Tang 2015]	99.53%
Proposed method	99.67%

Table II shows the validation of the proposed CNN superiority method over state-of-the-art methods.

TABLE III
THE COMPARISON BETWEEN PROPOSED METHOD AND CONVENTIONAL METHOD ON ORL DATABASE

Proposed method	Conventional method
CNN Accuracy 97.89%	Fuzzy Hidden Markov Models (FHMM) classifier accuracy 95%

Table III shows the comparison between traditional methods such as Fuzzy Hidden Markov Models (FHMM) and the Proposed CNN Method using the ORL database which consists of 400 images of size 112 x 92. There are 40 persons, 20 images per each person. We can say that using CNN can give higher accuracy than the conventional method.

CONCLUSION

This paper introduces an efficient convolution neural network architecture which has 15 - layers for accurate face recognition in wild environments. In our model, we use a different number of images of the Face96 Database. The Images have a resolution of 112x92 pixels. The best accuracy achieved is 99.67% which outperforms the state-of the art methods. And we perform a comparison between our proposed CNN method and a traditional method such as Fuzzy Hidden Markov Models (FHMM) method. Also, we found that our Proposed CNN method outperforms the conventional method.

REFERENCES

[1] Z Xu, C Hu, L Mei, "Video structured description technology-based intelligence analysis of surveillance videos for public security applications". *Multimed. Tools Appl*.75(19), 1–18 (2015).F. Author, H. Author, and I. Author, "Journal style," *Journal*, vol. 1, Jan. 1999, pp. 140–151 [*Conference*, 2016, pp. 300-307].

[2] Z Xu, Y Liu, H Zhang et al. "Building the multi-modal storytelling of urban emergency events based on crowdsensing of social media analytics". *Mob. Netw. Appl*.22(2), 218–227 (2017).J. A. Author, "Periodical style," *Periodical*, vol. 1, no. 1, pp. 30–38, Jan. 1999. DOI: 01XYZ.

[3] Y Yang, Z Xu et al. "A security carving approach for AVI video based on frame size and index". *Multimedia Tools Appl*. 76(3), 3293–3312 (2017).L. A. Author and M. Author, "Presented Conference Paper style," presented at Meeting (Conference), City, Jan. 2–7, 2015.

[4] Yan, Zhiguo, Zheng Xu, and Jie Dai. "The Big Data Analysis on the Camera-based Face Image in Surveillance Cameras." *Intelligent Automation & Soft Computing* (2017): 1-9.

[5] Z Xu, et al., "The big data analytics and applications of the surveillance system using video structured description technology". *Clust. Comput*.19(3), 1283–1292 (2016).D. C. Author, "Write this article," *Periodical*, vol. 1, no. 1,pp. 5-16, 2017. DOI: 78XYZ.

[6] MEENA, D. and R. SHARAN. "An approach to face detection and recognition". In: *International Conference on Recent Advances and Innovations in Engineering (ICRAIE). Jaipur: IEEE*, 2016, pp. 1–6. ISBN 978-1-5090-2807-8. DOI: 10.1109/ICRAIE.2016.7939462.

[7] REKHA, E. and P. RAMAPRASAD. "An efficient automated attendance management system based on Eigen Face recognition". In: *7th International Conference on Cloud Computing, Data Science & Engineering – Confluence. Noida: IEEE*, 2017, pp. 605–608. ISBN 978-1-5090-3519-9. DOI: 10.1109/ CONFLUENCE 2017 7943223.

[8] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[9] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[10] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, Nov 2002.

[11] H.-S. Lee and D. Kim, "Tensor-based aam with continuous variation estimation: Application to variation-robust face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 6, pp. 1102–1116, 2009. 15.

[12] Riddhi A. Vyas. "Feature Extraction Technique of PCA for Face Recognition with Accuracy Enhancement". *International Journal on Recent and Innovation Trends in Computing and Communication*, 2016.

[13] Chi Kien Tra; Tsair Fwu Le; Liyun Chang; Pei Ju Chao. "Face Description with Local Binary Patterns and Local Ternary Patterns: Improving Face Recognition Performance Using Similarity Feature-Based Selection and Classification Algorithm". *IEEE*.2014.

[14] Shangfei Wang, Longfei Hao, and Qiang Ji. "Facial Action Unit Recognition and Intensity Estimation Enhanced through Label Dependencies". *IEEE Transactions on Image Processing*, 2018.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," *in 2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.

[16] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," *in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, ser. CVPR '14*. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1891–1898.

[17] Sun, Y., Chen, Y., Wang, X., et al.: "Deep learning face representation by joint identification-verification. In: *Advances in neural information processing systems*," pp 1988–1996 (2014).

[18] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," *ArXiv e-prints*, Feb. 2015.

[19] Kang, Bong-Nam, Yonghyun Kim, and Daijin Kim. "Deep convolutional neural network using triplets of faces, deep ensemble, and score-level fusion for face recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.

[20] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." *IEEE Access* 6 (2018): 14410-14430.

[21] Yang, Xulei, et al. "Deep Learning for Practical Image Recognition: Case Study on Kaggle Competitions." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018.

[22] Lee, Honglak, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. "Unsupervised learning of hierarchical representations with convolutional deep belief networks." *Communications of the ACM* 54, no. 10 (2011): 95-103.

[23] Syaffeza A R, Khalil-Hani M, Liew S S, et al. "Convolutional neural network for face recognition with pose and Illumination Variation[J]". *International Journal of Engineering & Technology*, 2014, 6(1): 44 - 57.

[24] Ziming Z, Song F. "Profiling and analysis of power consumption for virtualized systems and applications[C]". *Proceedings of IEEE 29th International Performance Computing and Communications ConferenceIPCCC*,2010: 329-330.

[25] The Database of face94, face95 and face96, Spacek DL (2012) Face recognition data, *University of Essex. UK. Computer Vision Science Research Projects*.