Face Recognition from Video using Generalized Mean Deep Learning Neural Network

Poonam Sharma
Department of Computer Science & Engineering,
Visvesvaraya National Institute of Technology, Nagpur,
India
e-mail: poonamsharma@cse.vnit.ac.in

R. N. Yadav
Department of Electronics & Communication Engineering,
Maulana Azad National Institute of Technology, Bhopal,
India
e-mail: rnyadav@gmail.com

K. V. Arya
Department of Information & Communication Technology
ABV-Indian Institute of Information Technology & Management
Gwalior, India
e-mail: kvarya@iiitm.ac.in

***Abstract*— Proposed algorithm is a face recognition algorithm from video using Generalized mean Deep Learning Neural Network. Generalized mean provides fast convergence of the feature set and Deep learning neural network is enhanced using wavelet transform as it improves the classification efficiency of the neural network. The performance of the proposed algorithm is evaluated on PaSC and Youtube dataset. The results proved that the proposed algorithm is better in terms of identification accuracy.**

Keywords- Encoder, face recognition, video, wavelet, deep belief network.

I. INTRODUCTION

Face recognition [1] [2] [3] [4] is the most challenging problem in the fields of biometrics and computer vision because of the variation in illumination, pose variation, aging, disguise, body weight etc. An extensive research has been done in these areas and many different approaches have been proposed.

Despite these developments in face recognition, identification accuracy is very low for images under unconstrained environment. Due to its advantages over other biometric techniques, face recognition is used in the areas of law enforcement and surveillance to fight terrorism, find fugitives at secured places like airports, defence areas, government buildings etc., smart cards like ATM cards, driving license, passport etc., entertainment in using human-computer interaction, information security using face image as the passwords and commercial areas like telemarketing for authentication.

Face recognition system is an application where a person is either automatically verified with the information stored in the system or a person is identified from a database of stored images or frames obtained from stored video. It includes two basic parts: face detection where a face is detected from a cluttered scene by positioning and detecting the edges by discarding the image background. Recognition methods are used to verify or identify one or more faces using stored database of face image. Face recognition system may be divided into four steps: The first step is image acquisition either using a camera or from a video. Second step is detecting or segmenting face image from the still image. Third step includes feature extraction from the face detected which act as the image signature. Fourth step includes the face recognition method which leads to verification or identification of the person as shown in Fig. 1.

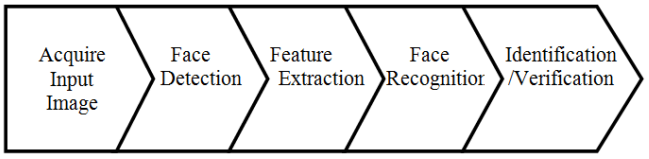

Figure 1. Steps of face recognition system

Face recognition may be defined as a biometric method that involves automated approach used for verification or recognition of the identity of a person based on his or her facial features and characteristics. Face recognition is a very high level image analysis task due to variation in pose, illumination, scaling, and unconstrained environment, etc. Due to the very complex pattern distribution and wide variations in pose, illumination, unconstrained background, expression and occlusion, face recognition has been interesting to researchers in the field of pattern recognition. Although numerous methods have been developed for different applications, the available methods may not produce efficient results due to wide variation in lighting pose and other factors.

Face recognition has proved to be challenging due to its intra-class variances, and there is a requirement of a robust face recognition technique which can overcome the effect of variation in pose, illumination, aging and expression, and will be used in a number of face recognition applications for security and authentication.

978-1-5090-3488-8/16/$31.00 ©2016 IEEE

In the last decade, face recognition from video have also been a very challenging research area. Face recognition is done using image sequences. It requires tracking of still images from video followed by recognition.

Tracking estimates faces location in the frame and returns as sub image. This sub image is used as the still image for recognition. The major problems associated with video based face recognition are head pose variation, partial occlusion. But it provides an advantage of availability of different poses of a person across the video. Goswami et al. [5] presented a memory based method for obtaining the still images from the video which can further be used for facial feature extraction and matching. A deep learning algorithm is then proposed that uses a stack of denoising autoencoders and deep Boltzmann machines for face recognition using the most memorable frames. The efficiency is improved at low false accept rates. In [6] authors proposed to explicitly learn enhanced face representations on a per-individual basis, and presented two methods enabling this approach. By learning and operating within person-specific representations, algorithm is able to significantly outperform the previous state-of-the-art on PubFig83, a challenging benchmark for familiar face recognition in the wild, using a novel method for learning representations in deep visual hierarchies.

Gao et al. [7] proposed a supervised autoencoder, which is a new type of building block for deep architectures. There are two features distinct supervised autoencoder have from standard autoencoder. First, it enforces the faces with variants to be mapped with the canonical face of the person, for example, frontal face with neutral expression and normal illumination; Secondly, it enforces features corresponding to the same person to be similar. As a result, our supervised autoencoder extracts the features which are robust to variances in illumination, expression, occlusion, and pose, and facilitates the face recognition. Ding et al. [8] proposes a comprehensive deep learning framework to jointly learn face representation using multimodal information. It includes a set of elaborately designed convolutional neural networks (CNNs) and a three-layer stacked auto-encoder (SAE). The set of CNNs extracts complementary facial features from multimodal data and concatenates to form a high-dimensional feature vector, whose dimension is compressed by SAE.

In [9] authors used the ability of a CNN to learn local patterns from data is used for facial recognition. The symmetry of facial information is exploited to improve the performance of the system by considering the horizontal reflections of the facial images. Nagpal et al. [10] proposed a regularizer-based approach to learn weight invariant facial representations using two different deep learning architectures, namely, sparse-stacked denoising autoencoders and deep Boltzmann machines. A body-weight aware regularization parameter has been included in the loss function of these architectures to help learn weight-aware features. Zhang et al [11] presented a face recognition method based on deep neural network. The sparse coding neural network and the softmax classifiers were used in this paper to build and train the deep hierarchical network after the face image preprocessing.

The state of art algorithms shows high verification accuracies of 80 % and above under constrained environment but the verification accuracy reduces as the environment changes to unconstrained environment. Deep learning is especially applicable to large training sets, and has been recently used in different fields of image processing such as vision, speech and language modeling. In the proposed algorithm generalized mean auto encoder is used for feature extraction and deep belief network is used for classification with wavelet as the activation function. This improves the recognition accuracy of the algorithm. The paper is structured as follows: Section 2 describes the deep neural network for face recognition in more detail. The concept of generalized neural network is discussed in Section 3. The proposed face recognition method is described in Section 4. Section 5 will show the experimental results, and in section 6, conclusions and future works are discussed.

II. Deep Neural Network for Face Recognition

Frames are obtained by tracking the face images in the video. After obtaining the frames, feature extraction and verification are performed by deep learning architecture including generalized auto encoder for feature extraction and deep belief network with wavelet as the activation function for classification.

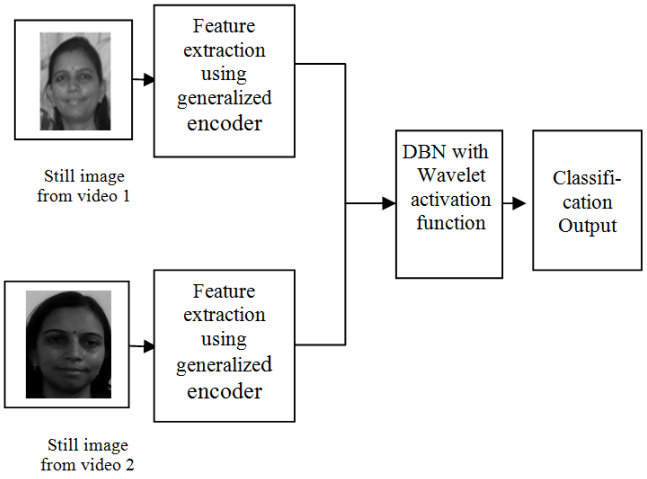

Figure 2. An overview of the steps involved in the proposed algorithm

Especially for faces, the success of the deep learning neural network in faces detection in a robust manner but needs further improvement. Once face is properly aligned, the location of all facial features is fixed at the pixel level. It is therefore possible to learn from the raw image using a single layer only. Thus the proposed method is useful for face recognition on very large databases, once the deep learning network is trained.

A. Generalized Auto Encoder

Auto encoder is a neural network for learning to map between input and output layer. It can have multiple layers stacked together the mapping between the input layer and hidden layer is modified by taking generalized mean [12] [13][14].

$$z = f(x) = S[Wx^r + w_0]^{\frac{1}{r}} \quad (1)$$

where, f is an activation function
x is input data
w_0 is bias
S is sigmoidal function
r is generalization parameter

The mapping between hidden layer and output layer is given by

$$\hat{x} = g(x) = S\left[\widehat{W}z^r + \widehat{w_0}\right]^{\frac{1}{r}} \quad (2)$$

where, $\widehat{W}$ is the weight applied between hidden layer and output layer
$\hat{x}$ is recovered input data
$\widehat{w_0}$ is bias at output layer
S is sigmoidal function
r is generalization parameter

The reconstruction error is given by [5]

$$\underset{W,\widehat{W}}{argmin} \|x - \hat{x}\|^2 \quad (3)$$

B. Deep Belief Network (DBN)

Deep belief network is a stack of multiple restricted Boltzmann machine (RBM). Each layer of RBM is trained and stack as a directed graph. This method is used iteratively for efficient feature extraction. So as to improve the convergence wavelet activation function is used in the network. The advantage of using RBM is that the weights between two layers are local and have less chances of getting trapped in local minima
Gabor filter $\psi_{\mu,v}(z)$, is defined as follows [12]

$$G_{\psi f}(x, y, \mu, v) = f(x, y) * \psi_{\mu,v}(z) \quad (4)$$

where, * denotes the convolution operator. Here Gabor filters use five frequencies $v \in \{0, \dots, 5\}$ and eight orientations $\mu \in \{0, \dots, 8\}$.

This function is used as the activation function.

The output of the k-th unit in the output layer is given by

$$Y_k^n = S\left[\sum_{i=1}^{L} W_k \; G_{\psi f}(x, y, \mu, v) + W_{ok}\right]^{1/r} \quad (5)$$

where $S(x) = \frac{1}{1+e^{-x}}$ is a sigmoid function.

III. Proposed Method

The proposed system is divided into two major phases.

A. Pre-Processing
B. Feature Extraction
C. Verification

A. P re-Processing

Image pre-processing is done by denoising, cropping, resize and edge detection after obtaining the image from the video.

B. Feature Extraction

The selection of a powerful set of feature is essential in face recognition. The features that are extracted in this part are used to create a feature vector. A feature vector of dimension has been used to uniquely distinguish face. These features are extracting as follows:

1. An image of size A×B is converted to 1×AB and taken as input to the encoder. Weights are initialized with random value.
2. Layer by layer greedy approach is used using equation (1)(2) and (3).
3. Initially the network is trained on train images until error becomes zero.

C. Verification

1. Feature extracted as 1×256 is obtained from the encoder as is taken as input to DBF.
2. Wavelet activation function is applied to all layers of the DBF for fast convergence.

IV. Experimental Setup and Analysis

For direct comparison of the experimental results, both YouTube Faces [16] and PaSC [15] databases are tested on standard experimental protocols so that results can be directly compared. For the YouTube Faces database, it consists of 10 splits, with 250 genuine and impostor pairs. Random number of splits varying from 250 to 10 has been taken for training and the remaining instances have been taken for testing.

The PaSC database have an advantage that it contains videos of low resolution captured from a handheld camera and high resolution video from a control camera. The results are tested on handheld-to-handheld videos with low resolution and compared with other existing algorithms in the PaSC protocol. In PaSC database, random number of splits varying from 500 to 10 has been taken for training and the remaining instances have been taken for testing.

Both the databases are divided into two sets, one for training and other for testing. Training Set consists of random percentage of the training data and is utilized for training. After training, the proposed algorithm is evaluated on the entirety of the testing data. The deep learning

architecture comprises 11 blocks; each block contains a linear operator followed by ReLU. The sequence of eight blocks each of which is said to be convolutional because it have a linear operator work as a bank of linear filters. The last three blocks are fully connected with the size of the filters same as the size of the input data, such that each filter "senses" data from the entire image. All the convolution layers are followed by a rectification layer (ReLU). The input to all networks is a preprocessed face image of size 224×224. Our mini-batch size is 64, and learning rate for all trainable layers to 0.05, which was manually decreased, each time by an order of magnitude to a final rate of 0.001. The weights were initialized with non zero-mean Gaussian distribution with $\sigma = 0.06$, and bias of 0.4.The results are reported in terms of Receiver Operating Characteristic (ROC) curves and the verification accuracies at different false accept rates.

This method is having an advantage above the existing region of interest (ROI) method in terms of recognition accuracy and recognition time or latency time. Methods that include finding ROI for recognition miss some of the crucial features as is discussed in [17]. Thus, reducing the recognition accuracy. This method improves the recognition accuracy due to increased number of layers and reduces the time of recognition once the network has been trained.

To visualize the effectiveness of the proposed method the receiver operating curve (ROC), is plotted in Fig. 3 for PaSC dataset. It was observed that the identification accuracy of the proposed method is higher than most of the existing methods. It also explains that as the false acceptance rate increases false rejection rate decreases. It is due to the fact that if the size of the training set is more, then there are more chances that an errant image may be accepted as the image of the subject for which it is trained and vice versa. This results in reduced accuracy also.

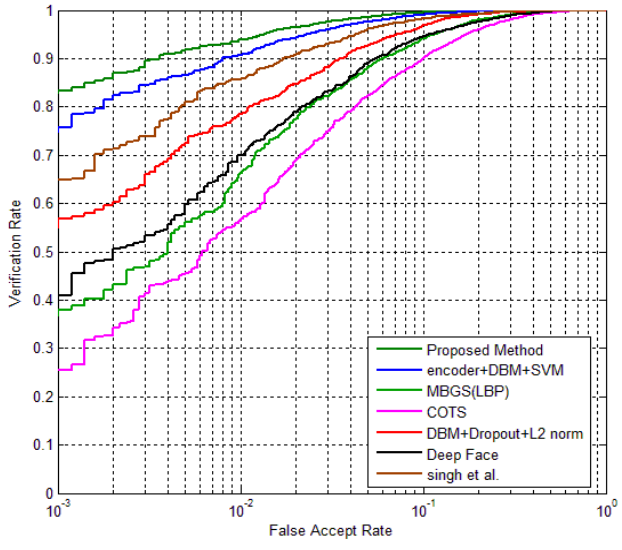

Figure 3. ROC curve for PaSC Dataset.

The performance of the proposed method is evaluated by the cumulative match characteristic (CMC) curve which is computed by plotting Rank against the values of varying Identification accuracy from 0 to 80. The MC curve for PaSC databases is given in Fig. 4.

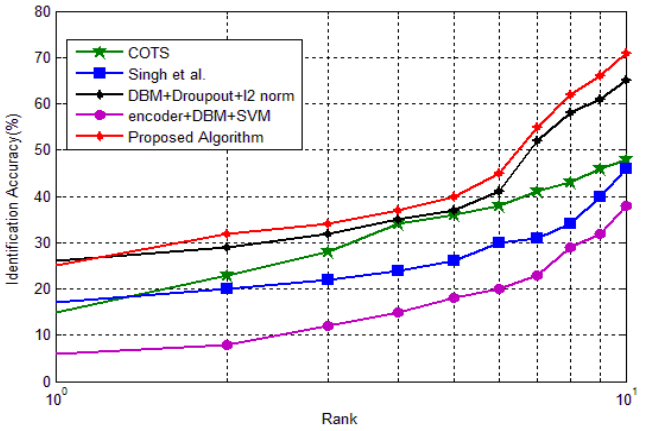

Figure 4. CMC curve for PaSC Database

The results for Youtube database is given in Table 1. The table shows that the proposed algorithm outperform out existing algorithms

TABLE 1. IDENTIFICATION ACCURACIES (5) OF DIFFERENT STATE OF ART ALGORITHMS

Algorithm	Rank 1	Rank 10
COTS[5]	14.3	47.0
SDAE+Dropout+l2 norm[5]	24.6	64.6
DBM+ Dropout+l2 norm[5]	26.0	65.8
Proposed Algorithm	31.0	71.8

V. CONCLUSION

The performance of the proposed algorithm is tested on two most commonly used databases, i.e. PaSC and Youtube databases. The results proved that the proposed algorithm is better in terms of identification accuracy. This improvement is due to the use of Generalized mean encoder and modified Deep Belief Network.

REFERENCES

[1] S. Choi, C. Choi and N. Kwak, "Face recognition based on 2D Images under illumination and pose variations," Pattern Recognition Letters, vol. 32, pp. 561-571, 2011.

[2] J. Wang, J. You, Q. Li and Y. Xu, "Orthogonal discriminant vector for face recognition across pose," Pattern Recognition, vol. 45, no. 12, pp. 4069-4079, Dec., 2012.

[3] P. Moallem, B. S. Mousavi and S. A. Monadjemi, "A novel fuzzy rule base system for pose independent faces detection," Applied Soft Computing, vol. 11, no. 2, pp. 1801-1810, 2011.

[4] A. A. Mohammad, R. Minhas, Q. M. J. Wu, M. A. Sid-Ahmad, " Human face recognition based on multidimentional PCA and extreme learning machine," Pattern Recognition, vol. 44, no. 10-11, pp. 2588-2597, 2011.

[5] G. Goswami, R. Bhardwaj, R. Singh and M. Vatsa, "MDLFace: Memorability augmented deep learning for video face recognition," IEEE International Joint Conference on Biometrics (IJCB), 2014, Clearwater, FL, 2014, pp. 1-7.

[6] G. Chiachia, A. X. Falcão, N. Pinto, A. Rocha and D. Cox, "Learning Person-Specific Representations From Faces in the Wild," IEEE Transactions on Information Forensics and Security, vol. 9, no. 12, pp. 2089-2099, Dec. 2014.

[7] S. Gao, Y. Zhang, K. Jia, J. Lu and Y. Zhang, "Single Sample Face Recognition via Learning Deep Supervised Autoencoders," IEEE Transactions on Information Forensics and Security, vol. 10, no. 10, pp. 2108-2118, Oct. 2015.

[8] C. Ding and D. Tao, "Robust Face Recognition via Multimodal Deep Face Representation," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2049-2058, Nov. 2015.

[9] N. P. Ramaiah, E. P. Ijjina and C. K. Mohan, "Illumination invariant face recognition using convolutional neural networks," IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), *Kozhikode, 2015*, pp. 1-4.

[10] S. Nagpal, M. Singh, R. Singh and M. Vatsa, "Regularized Deep Learning for Face Recognition With Weight Variations," IEEE Access, vol. 3, no. 1, pp. 3010-3018, 2015.

[11] Z. Zhang, J. Li and R. Zhu, "Deep neural network for face recognition based on sparse autoencoder," 8th International Congress on Image and Signal Processing (CISP), *S*henyang, 2015, pp. 594-598.

[12] Poonam Sharma, K. V. Arya and R. N. Yadav, "Efficient face recognition using generalized mean wavelet neural network," Signal Processing, Vol. 93, no. 6, June 2013, pp. 1557–1565.

[13] Poonam Sharma, K. V. Arya and R. N. Yadav, "Pose-invariant face recognition using curvelet neural network," IET Biometrics, Vol. 3, no. 3, September 2014, pp. 128 – 138.

[14] R. N. Yadav, N. Kumar, P. K. Kalra, and J. John, "Learning with generalized-mean neuron model," NeuroComputing, vol. 69, no. 16, pp. 2026–2032, 2006.

[15] J. R. Beveridge, P. J. Phillips, D. S. Bolme, A. B. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn and S. Cheng, "The challenge of face recognition from digital point-and-shoot cameras," IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2013, Arlington, VA, 2013, pp. 1-8.

[16] L. Wolf, T. Hassner and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, Providence, RI, 2011, pp. 529-534.

[17] J. L. Raheja, K. Das, A. Chaudhary, "An Efficient Real Time Method of Fingertip Detection" Proc. of International Conference on "Trends in Industrial Measurements and Automation-TIMA–2011, Chennai.