International J. of Engg. Research & Indu. Appls. (IJERIA).
ISSN 0974-1518, Vol.9, No. III (December 2016), pp.23-42

A NOVEL APPROACH OF FACE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS WITH AUTO ENCODER

T. SYED AKHEEL[1] AND DR. S. A. K JILANI[2]

[1] Research Scholar, Dept. of Electronics & Communication Engineering,
Rayalaseema University Kurnool, Andhra Pradesh.

[2] Research Supervisor, Professor, Dept. of Electronics & Communication Engineering,
Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh.

Abstract

In this paper we present a new technique for face recognition with Convolutional auto-encoder and Siamese neural network. The face recognition is the crucial step for many applications like surveillance, access control or human-computer interaction. Existing Convolution Neural Network (CNN) based methods for face recognition are not robust and reliable .We propose a novel face recognition approach based on the auto-encoder with Convolution neural network and Siamese network. The first step of the for this network is to create a near frontal face for each and every individual and then train the data using Siamese network. The method of learning Siamese neural networks which employ a unique structure to naturally rank similarity between inputs. Using convolution architecture in the Siamese neural network we are able to achieve strong results which exceed those of other deep learning models with near state-of-the-art performance. We also present an analysis and discussion in comparison with the existing models like VGG deep learning neural network for face recognition and CNN based face recognition with labelled faces in the wild database and also the YouTube faces database with suitable performance graphs.

Key words: CNN Convolution neural networks, Siamese network

1. INTRODUCTION

The automatic processing of images to extract semantic content is a task that has gained a lot of importance during the last years due to the constantly increasing number of digital photographs on the Internet or being stored on personal home computers. The need to organize them automatically in an intelligent way using indexing and image retrieval techniques requires effective and efficient image analysis and pattern recognition algorithms that are capable to extract relevant semantic information. Especially faces contain a great deal of valuable information compared to other objects or visual items in images. For example, recognizing a person on a photo graph, in general, tells a lot about the overall content of the picture. In the context of human-computer interaction (HCI), it might also be important to detect the position of specific facial characteristics or recognize facial expressions, in order to allow, for example, a more intuitive communication between the device and the user or to efficiently encode and transmit facial images coming from a camera. Thus, the automatic analysis of face images is crucial for many applications involving visual content retrieval or extraction The principal aim of facial analysis is to extract valuable information from face images, such as its position in the image, facial characteristics, facial expressions, the person's gender or identity.

We will outline the most important existing approaches to facial image analysis and present novel methods based on Convolutional Neural Networks (CNN) with auto-encoder to normalize the detected face and recognize faces and facial features. CNNs shown to be a powerful and flexible feature extraction and classification technique which has been successfully applied in other contexts, i.e. hand-written character recognition, and which is very appropriate for face analysis problems as we will experimentally show in this work .We will focus on the processing of two-dimensional gray-level images as this is the most widespread form of digital images and thus allows the proposed approaches to be applied in the most extensive and generic way. However, many techniques described in this work could also be extended to color images,3D data or multi-modal data.

There are numerous possible applications for facial image processing algorithms. The possible application for face authentication are mainly concerned with access control, e.g. restricting the physical access to a building, such as a corporate building, a secured zone of an airport etc. This principle could equally be applied to the access to systems, automatic

teller machines, mobile phones, Internet sites etc. Video surveillance is another application for face identification.

2. RELATED WORKS

It's a true challenge to build an automated system which equals human ability to detect faces and estimates human body dimensions from an image or a video. The conceptual and intellectual challenges of such a problem, because faces are non-rigid and have a high degree of variability in size, shape, color and texture. Numerous techniques have been developed and designed typically to deal with single images .Detect human appearance in image or track humans in videos become a very important task for many applications such as: auto focus in cameras ,visual surveillance, traffic safety monitoring and human computer interaction. From both neurological and computational point of view symmetry plays an important role in facial part dimensions as it has been demonstrated that an exceptional dimension reduction can be made by taking into account facial symmetry.

The present method usually returns the image location of a rectangular bounding box containing a face. is focusing to isolate the human body in a relatively clean background from input images or videos containing people roughly localized and applying a list of human-based measuring units on the human bodies. This proposed algorithm goes beyond the limits of all existing technologies as it obtains the unique functional features by enabling the proposed model to work with different skin color tone, applying it to low-quality images, detecting faces with eye glasses, determining the position of facial parts (e.g. eye pupils, nose, lips, etc.) and detect several faces on one image is typically designed to deal with single images.This method uses a deep CNN trained to classify faces using a dataset of 4 million examples spanning 4000 unique identities. It also uses a siamese network architecture, where the same CNN is applied to pairs of faces to obtain descriptors that are then compared using the Euclidean distance. The goal of training is to minimize the distance between congruous pairs of faces (i.e. portraying the same identity) and maxi mise the distance between incongruous pairs, a form of metric learning. In addition to using a very large amount of training data, Deep Face uses an ensemble of CNNs, as well as a pre-processing phase in which face images are aligned to a canonical pose using a 3D model. When introduced, Deep Face achieved the best performance on the Labelled Faces in the Wild (LFW; [8]) benchmark as well as the Youtube Faces in the Wild (YFW) benchmark.

The authors later extended this work, by increasingthe size of the dataset by two orders of magnitude, including 10 million identities and 50 images per identity.

3. CONVOLUTION NEURAL NETWORK

Convolution Neural Networks (CNN) are an approach that tries to alleviate the above mentioned problems. That is, they automatically learn local feature extractors, they are invariant to small translations and distortions in the input pattern, and they implement the principle of weight sharing which drastically reduces the number of free parameters and thus increases their generalization capacity compared to NN architectures without this property.Siamese Neural Networks have first been presented by Bromley et al. using Time Delay Neural Networks (TDNN) and applying them to the problem of signature verification, i.e. to verify the authenticity of signatures. This idea was then adopted by Chopra et al who used Siamese CNNs and employed them in the context of face verification. More precisely, the system receives two face images and has to decide if they belong to the same person or not. The architecture of the CNN proposed by LeCun et al.

3.1 ILLUSTRATION OF A SIAMESE CONVOLUTIONAL NEURAL NETWORK.

Siamese NNs learn [21,22,24] and chopara et.al a non-linear similarity metric by repeatedly presenting pairs of positive and negative examples, i.e. pairs of examples belonging to the same class or not. The principal idea is to train the NN to map the input vectors into a non-linear subspace such that a simple distance, e.g. the Euclidean distance, in this subspace approximates the "semantic" distance in the input space. That means, two images of the same category are supposed to yield a small distance in this subspace and two images of a different category a large distance. Let us call this mapping GW(X) and its parameters (i.e. weights) W. Thus, the goal is to learn the parameters W of the function GW(X) such that the similarity metric is small if X1 and X2 belong to the same class and large if they belong to different classes. The choice of GW(X) is arbitrary and, in this case, is a CNN.

$$E_W(X_1,X_2) = \|G_W(X_1) - G_W(X_2)\| \text{ ----------(1)}$$

The Convolutional Auto-encoder is used to get the frontal face approximated image and then we need to train the system using the Siamese network for classification purposes. The architecture and the training of the system is discussed with results and the comparisons.

4. NETWORK ARCHITECTURE AND TRAINING

Here there are two approaches which are used for training for the face recognition. The first approach will be training the data base for getting a near frontal face for each and every individual of the person. Then the second approach is used to recognize the person with Siamese algorithm.

A. CONVOLUTIONAL AUTO ENCODER (CAE)

The CAE is an acronym for convolutional auto encoder. The CAE is a back-propagation approach where the encoder and decoder are embedded together. The network for CAE will have only 4 layers. The first layer is the input layer with the input as 112X112 gray scale image data. The image data face image is cropped from the original image after face detection with a bounding box where the face is centred where the nose, mouth and eyes are at the approximate positions while maintaining the aspect ratio. Then the gray scale values are normalized between 0 and 1. These image data base are given as an input to the first layer with 5X5 filtering. Then the second layer will be the pooling layer where the images are down-sampled by 2. Then the un-pooling of the images are done instantly to get the original image. Then it is followed by a decoder and encoder which are fully connected where we try to reconstruct the whole gray image again. Here the batch normalization and near and Frontal images is done to get the near frontal face for each and every person. The architecture of the Convolution Auto-Encoder is as shown in the figure below.

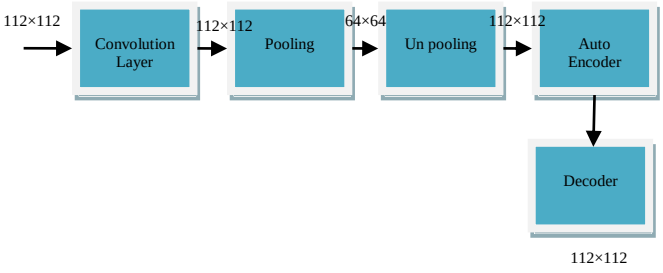


Figure 1 Architecture for Convolution Auto- Encoder.

B. SIAMESE NETWORKS

Siamese nets were first introduced in the early 1990s by Bromley and LeCun to solve signature verification as an image matching problem. A Siamese neural network consists of

twin networks which accept distinct inputs but are joined by an energy function at the top. This function computes some metric between the highest-level feature representation on each side (Figure 6). The parameters between the twin networks are tied. This strategy has two key properties:

- It ensures the consistency of its predictions. Weight tying guarantees that two extremely similar images could not possibly be mapped by their respective networks to very different locations in feature space because each network computes the same function.
- The network is symmetric: if we present two distinct images to the twin networks, the top conjoining layer will compute the same metric as if we were to we present the same two images but to the opposite twins:
- The model is presented to show the procedure how linear weights can carry the sequence and the output is presented.
- The relation between inputs and outputs can categorized in the block diagram to analyze the flow of weights and parameters variations.

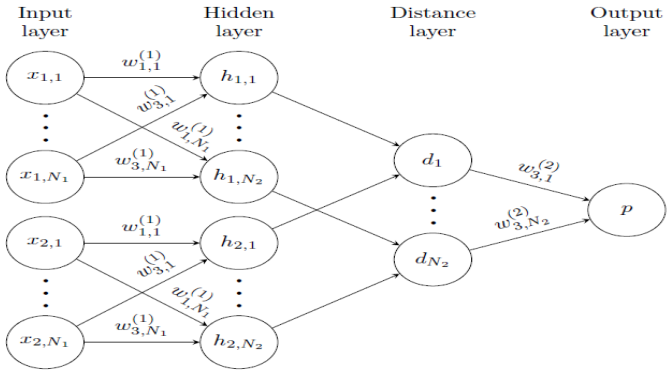


Figure 2: A simple 2 hidden layer Siamese network for binary classification with logistic prediction p. The structure of the network is replicated across the top and bottom sections to form twin networks, with shared weight matrices at each layer.

Vijay Badrinarayanan, et.al [25] Deep Convolution Encoder-Decoder Architecture has presented segNet in a more efficient way , However, less attention has been paid to smaller and more efficient models for real-time applications of segmentation .we experiment our analysis with facial database of normalization and back propagation algorithm.

5. PROPOSED METHOD

INPUT: Given a training set X ={X^i m} consisting of N face images from 'C' classes, with N_i face images per class specified with their class label, and X ={Xi^m} denoting the m^{th} sample from the i^{th} class.

OUTPUT: A classification function that assign the testing sample 'x' correctly to a class label r ε C, where C = {1, ….c} denotes a label set with c classes.

Step 1: FACE NORMALIZATION:

INPUT: Face database with slight pose variations of same identity.

OUTPUT**:** Normalized Frontal Face image.

The frontal face normalization is done for all the faces of the same identity. This is done by taking all the face the same person identity and creating a normalized face which is a frontal face.

Learning:

1. Select a representative face for each identity
    a. This is done manually
    b. Automatically it can be done using following steps :

i. Facial feature detection

ii. Classifying frontal face using simple classifier

1. Choosing the representative face to the mean face image.

The face image of the individual 'i' that is closest to mean image of I is chosen:

$$r_i = \underset{im^{+1_{ij}}}{argmin} \left\| im_{ij}^{+} - \overline{im_i^{+}} \right\| \text{----------(2)}$$

$$\forall i,k \in 1,.....N,\ j \in 1,...M, l \in 1,....M_k, k \neq i$$

2 .Choosing the most distant image.

The face image of individual 'i' that has greatest distance to the face images of all the other individuals is chosen

Give all other sample of the same person as input to signet and expect representative face.

Construct the learning input using the procedure

To construct the training data to the neural network the face images are normalized in the following way. First each image is cropped in such a way that the face is centered and that

the eyes and the mouth are roughly at predefined positions while keeping the aspect ratio. Then the gray values of the face image are normalized between 0 and 1 and the face image is resized to a predefined size.

Frontal face database creation.

1. Frontal faces will be created by sending all the faces through the SEGNET.
2. Synthesis slightly different offsets from the frontal faces to take care of variations possible on the unseen data
3. Synthesis noise variations in the appearance of the data

a. Training the data using the SEGNET

b. Training is performed using the Back propagation algorithm which has been slightly adapted to account for the shared weights.

For a given example im_{ij} the objective function is as shown in the following:

$$E_i = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(d_i(x,y) - y_1(x,y)\right)^2 \text{-----}\rightarrow (3)$$

Which is the mean squared error (MSE) between the computed outputs for training example and the desired outputs , where represents the respective reference image normalized between +1 and -1. Before the actual training, the weights are initialized at random. Then, they are updated after each iteration Note that by training the Neural Network, i.e. by minimizing the objective function, all parameters are learned conjointly: the convolution filters, the projection and the reconstruction parameters. In other words, the proposed architecture optimizes the filters and, at the same time, the projection parameters in order to reconstruct best the respective reference image.

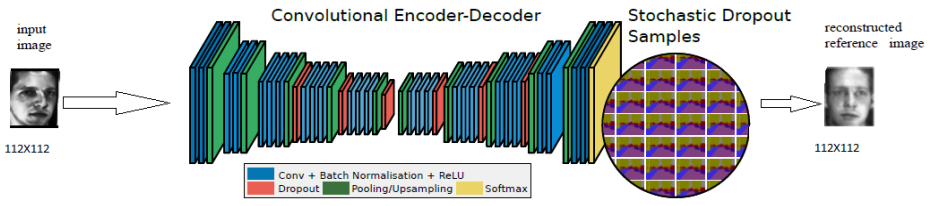


Figure: 3 Senet [25] courtesy Architecture Demonstrated Using Single Face Image

STEP 2: FACE RECOGNITION:

INPUT: Face image data base of different persons.

OUTPUT:

1) Learning: Prepare a batch

The batch preparation must be done in such a way that for each batch we need to have some faces of the same person identity and other faces need to be in such a way that at least one face needs to be included for each person of identity. These faces are from the SEGNET which are face normalized and frontal faces.

Compute the triplet loss and extract features from the triplets for each batch.

For each image $\left(X_i^a\right)$ (anchor) of a specific person is closer to all the other images $\left(X_i^p\right)$ (positive) of the same person than it is to any image $\left(X_i^n\right)$(negative) of any other person. This is known as the triplet loss.

Thus we want:

$$L = \sum_i \left[f\left(X_i^a\right) - f\left(X_i^b\right)_2^2 - f\left(X_i^a\right) - f\left(X_i^n\right)_2^2 \right] + a \dashrightarrow (3)$$

It Is crucial to select hard triplets, that are active and can therefore contribute to improving the model.

$$f\left(X_i^a\right) - f\left(X_i^b\right)_2^2 \dashrightarrow (5)$$

Hard positive

$$f\left(X_i^a\right) - f\left(X_i^n\right)_2^2 \dashrightarrow (6)$$

Hard negative

The lowest triplet loss needs to be selected as the best one. The triplet will extract the features of M dimensions for each face image.

2. Train these features using the Convolutional neural network. The Convolutional neural network has a convolution layer with different filter sizes, there is a pooling layer which down samples by a down sampling factor and the final layer which is present in the CNN is the RELU layer which makes all the negative to the positive. At the end of this convolutional

layer there is a flly connected layer which connects to each and every node. The typical CNN architecture as shown in the figure below.

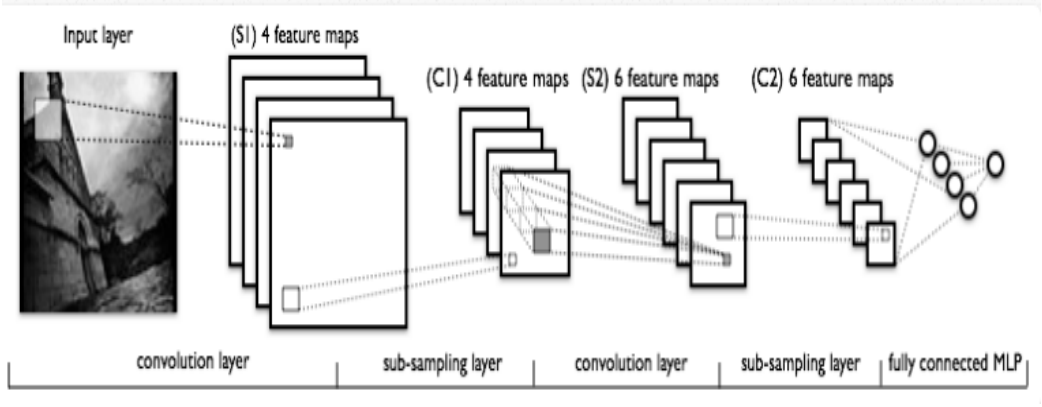


Figure 4. : CNN architecture for a sample image. Image courtesy-deep learning methods

3. Suppose that we have some N X N square neuron layer which is followed by our convolutional layer. If we use an mXm filter w, our convolution layer output will be of size (N-m +1)X(N-M+1). In order to compute the pre-non linearity input to some unit $\left(X_{ij}^{l}\right)$ in our layer, we need to sum up the contributions (weighted by the filter components) from the previous layer cells:

$$X_{ij}^{l} = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} W_{ab} y_{(i+a)(j+b)}^{(l-1)} \text{ ------} \rightarrow (7)$$

4. The max-pooling layers are quite simple, and do no learning themselves. They simply take some kXk region and output a single value, which is the maximum in that region. For instance, if their input layer is a NXN layer, they will then output $\frac{N}{K} x \frac{N}{K}$ layer, as each k X k block is reduced to jus a single value via the max function.

STEP 3: TESTING

In the testing part for each face image get the near frontal face using the SEGNET trained parameters.

II. COMPARISON AND RESULTS

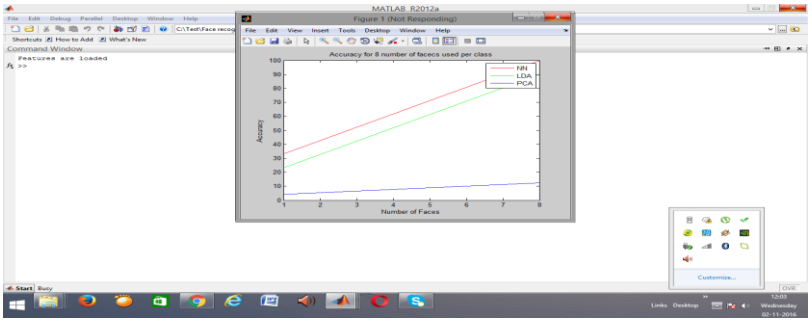


Figure 5 Showing The Accuracy with Most Challenging LFW Data Base.

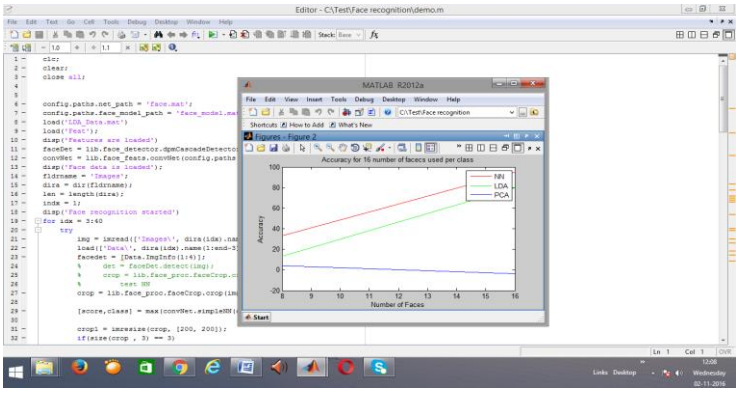


Figure 6: Comparsion Of LDA,PCA And Siamese NN-Siamese Neural networks

The present system is compared with other systems like Face recognition with CNN and VGG face recognition which is a deep learning technique. Here we present the analysis on two types of database:

1. Labeled faces in the wild.
2. YouTube face database.

For the calculation of loss of the system we use two types of loss functions. Soft max regression based method uses the L2 distance comparison. The distance between the actual face and the near frontal face is derived using the metric L2 distance. Another type of method used for the loss is Embedding loss as the loss function. The following are the

accuracies given by two different types of databases in comparison with the present technique and having two different types of loss functions.

.

Training method	**Loss function**	**Labeled faces in the wild**	**YouTube faces dataset**
CNN	Softmax (L2 distance comparision)	95.4	91.7
VGG	Softmax (L2 distance comparision)	97.27	92.8
CAE with Siamese network	Softmax (L2 distance comparision)	99.3	94.6
CNN	Embedding Loss	96.5	94.3
CGG	Embedding Loss	99.13	97.4
CAE with Siamese network	Embedding Loss	99.45	98.7

Table 1: Accuracies Compared To CAE with Siamese Network and CNN, VGG Deep Learning Network For Different Types of Datasets

The Above shows the accuracies compared to CAE with Siamese network and CNN, VGG deep learning network for different types of datasets. It is observed that the CNN is giving good accuracy results but not as good as the deep learning networks. The Deep learning networks are giving good results when compared to CNN based face recognition. The CNN based face recognition will create a face image which is near to the mean image of a

particular person according to the distance metric and then the training of the algorithm is done which is same as the Present CAE with Siamese network. But the only difference is that the training of the algorithm is taken care by the Siamese algorithm and the CNN with auto-encoder will create the near frontal face for each individual. The CAE with Siamese network has the highest accuracy for two different types of loss functions and using the two datasets when compared to the other networks.

Here we display the FAR and FRR for face data with respect to PCA and LDA

For CAE ::

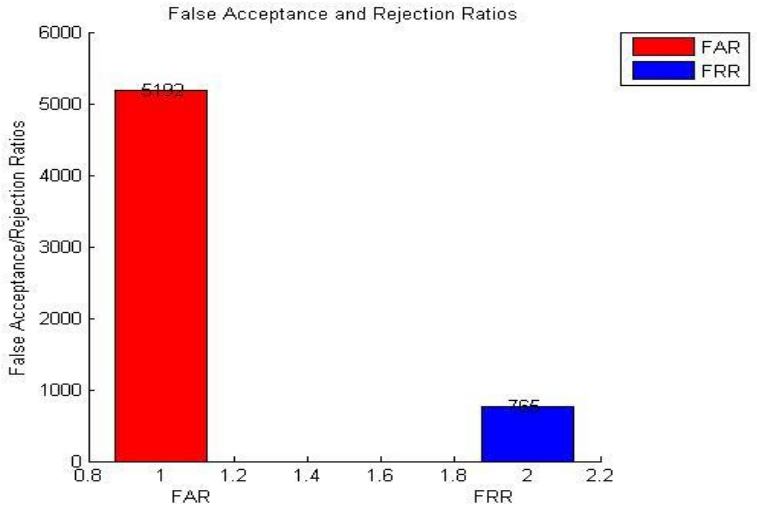


Figure 7: FAR and FRR using CAE

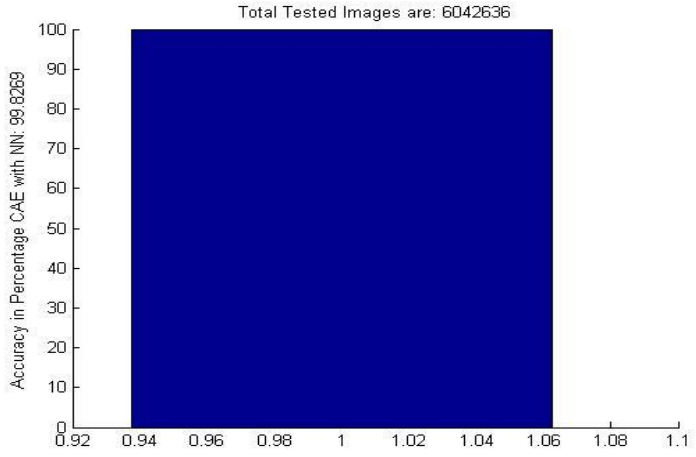


Figure 8: Accuracy Rate using CAE

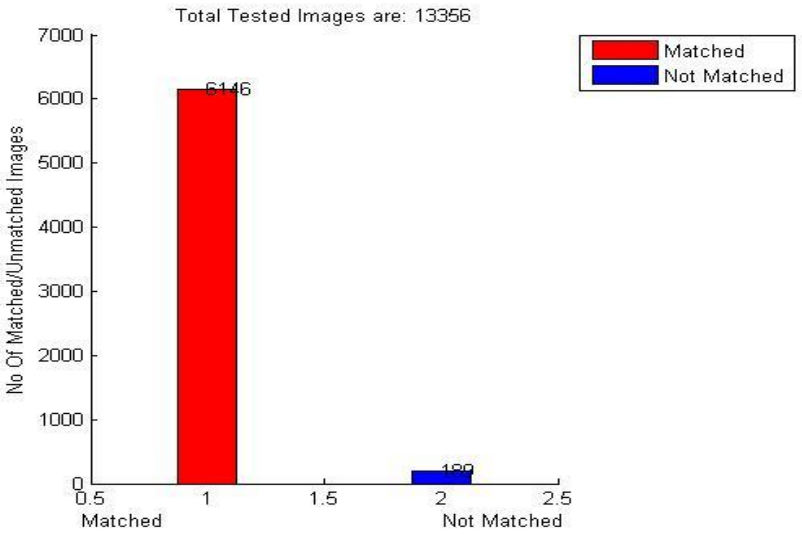


Figure 9: Matched and Unmatched face images using CAE

For PCA

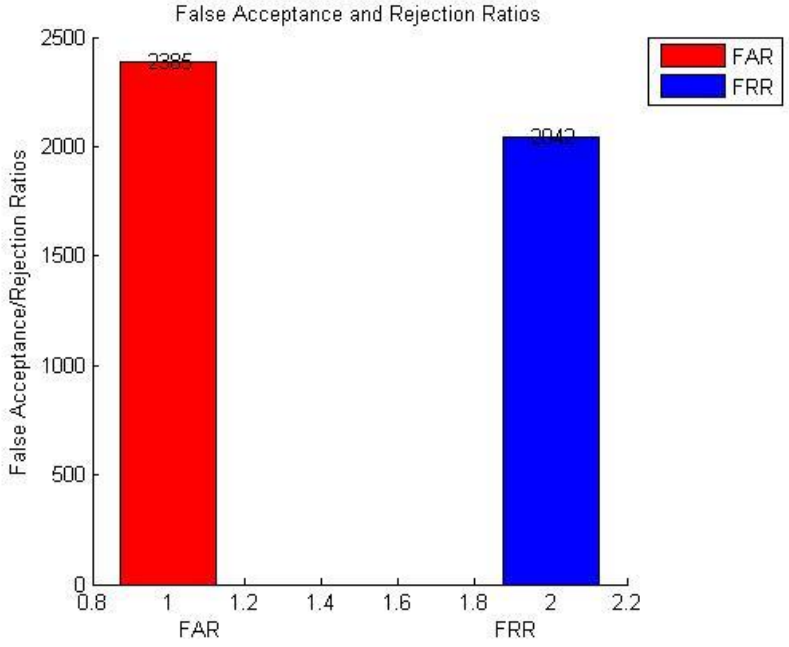


Figure 10: FAR and FRR using PCA

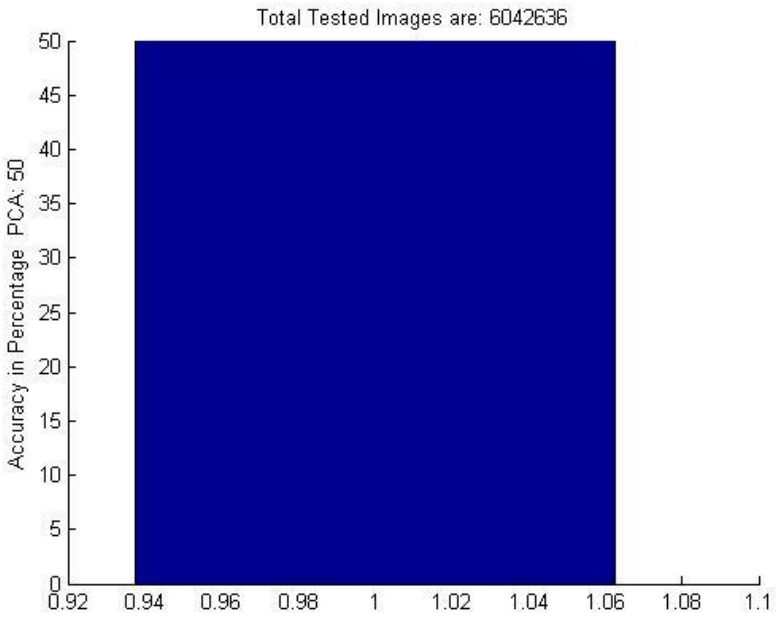


Figure11: Accuracy Rate using PCA

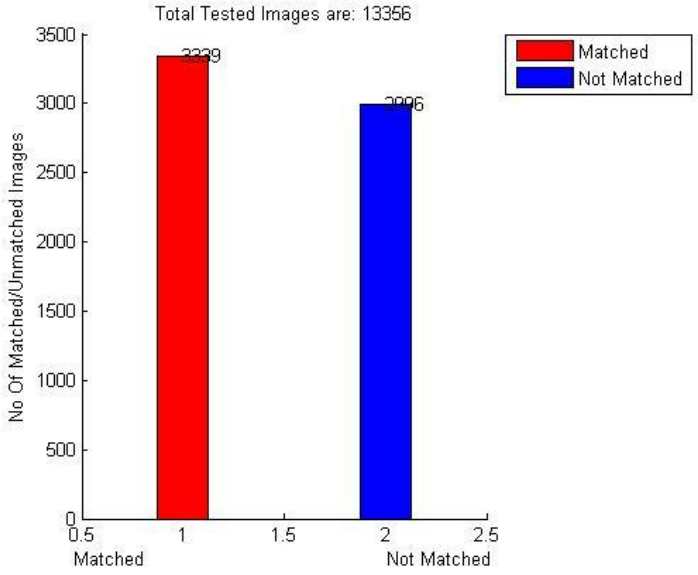


Figure 12: Matched and Unmatched face images using PCA

For LDA

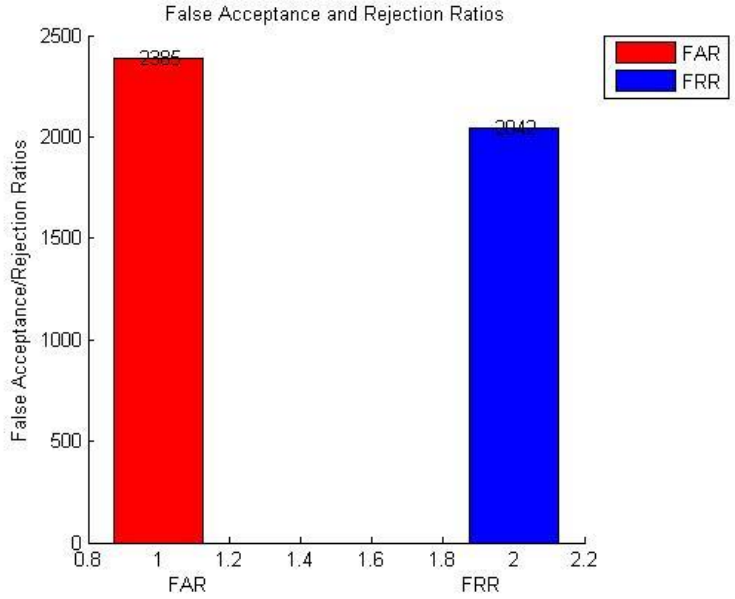


Figure 13: FAR and FRR using LDA

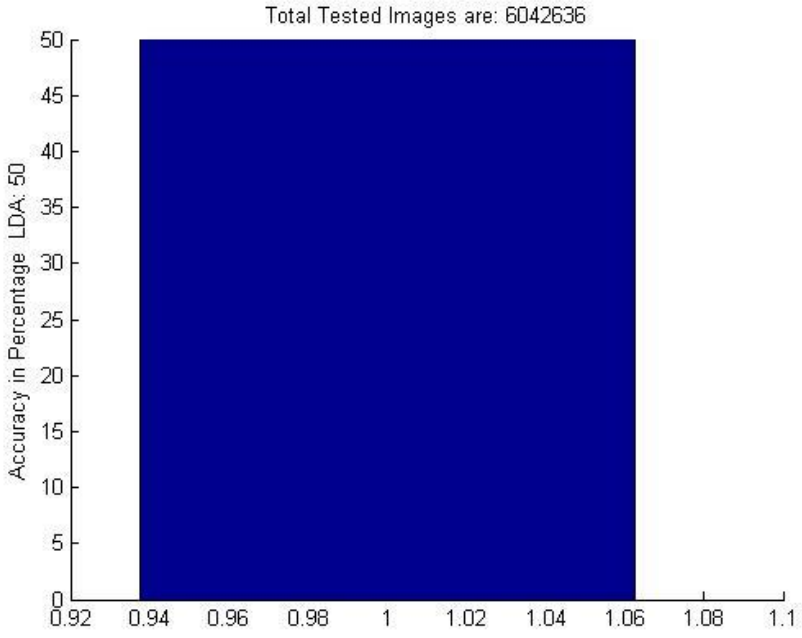


Figure 14: Accuracy Rate using LDA

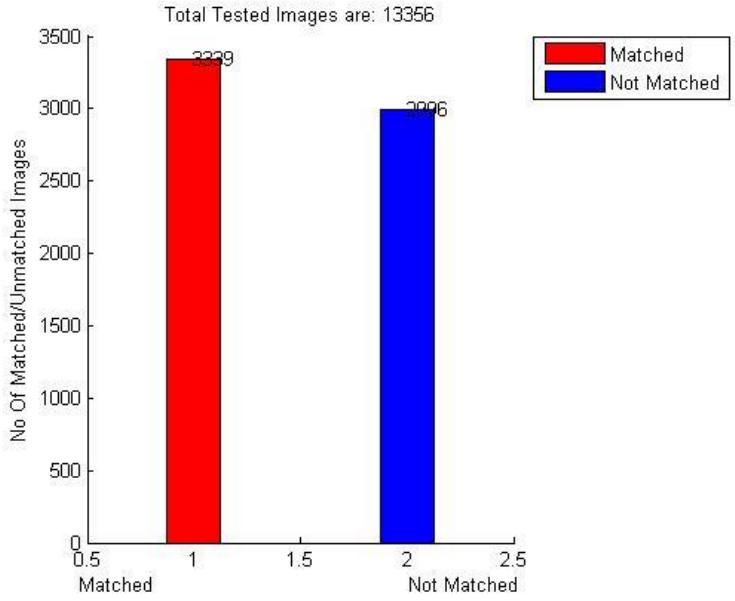


Figure 15: Matched and Unmatched face images using LDA

The Above figures shows the FAR and FRR rates for the face data with respect to PCA and LDA

For the calculation of loss of the system we use two types of loss functions. Softmax regression based method uses the L2 distance comparison. The distance between the actual face and the near frontal face is derived using the metric L2 distance. Another type of method used for the loss is Embedding loss as the loss function. The following are the accuracies given by two different types of databases in comparison with the present technique and having two different types of loss functions.

The Above analysis shows the accuracies compared to CAE with Siamese network and CNN, VGG deep learning network for different types of datasets. When we observe that the CNN is giving good accuracy results but not as good as the deep learning networks. The Deep learning networks are giving good results when compared to CNN based face recognition. The CNN based face recognition will create a face image which is near to the mean image of a particular person according to the distance metric and then the training of the algorithm is done which is same as the Present CAE with Siamese network. But the only difference is that the training of the algorithm is taken care by the Siamese algorithm and the

CNN with auto-encoder will create the near frontal face for each individual. The CAE with Siamese network has the highest accuracy for two different types of loss functions and using the two datasets when compared to the other networks.

CONCLUSION

Experimental results proved that the Siamese architecture with Convolution auto encoder can perform far better than the LDA , PCA, and neural networks , when it comes to handle large and most challenging data base, Siamese architecture with Convolution auto encoder performs remarkable results because cumulative result.

REFERENCES

[1] L. B. Almeida, T. Langlois, J. D. Amaral, and A. Plakhov. Parameter adaptation in stochastic optimization. In D. Saad, editor, On-Line Learn-ing in Neural Networks, chapter 6. Cambridge University Press, 1999.

[2] S.-I. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind source separation. In D. S. Touretzky, M. C. Moser, and M. E. Hasselmo ,editors, Advances in Neural Information Processing Systems, volume 8,pages 757–763. MIT Press, Cambridge, MA, 1996.

[3] G. Arulampalam and A. Bouzerdoum. A generalized feed forward neural network architecture for classification and regression. Neural Networks,16:561–568, 2003.

[4] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland. Visually controlled graphics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(6):602–605, June 1993.

[5] S. O. Ba and J. M. Odobez. A probabilistic framework for joint head tracking and pose estimation. In Proceedings of the 17th International Confer-ence on Pattern Recognition, volume 4, pages 264–267, August 2004.

[6] F. R. Bach and M. I. Jordan. Kernel independent component analysis .Journal of Machine Learning Research, 3:1–48, 2002.

[7] E. Bailly-Bailli`ere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mari´ethoz,J. Matas, K. Messer, V. Popovici, F. Por´ee, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In proceedings of the Fourth International Conference on Audio- and Video-Based Biometric Person Authentication, pages 625–638, 2003.

[8] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In Computer Vision and Pattern Recognition, volume 1, pages1090–1097, 2001.

[9] P. Baldi and K. Hornik. Neural networks and principal component analysis :Learning from examples without local minima. IEEE Transactions on Neural Networks, 2:53–58, 1988.

[10] S. Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. In Neural Information Processing Systems, 1998

[11] S. Baluja and H. A. Rowley. Boosting sex identification performance .International Journal of Computer Vision, 71(1):111–119, 2007.

[12] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition byin dependent component analysis. IEEE Transactions on Neural Networks,13(6):1450–1464, 2002.

[13] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. In International Conference on Computer Vision, volume 2, pages 383–390,2001.R. Battiti. Accelerated back propagation learning: Two optimization methods. Complex Systems, 3:331–342, 1989.

[14] P. Belhumeur, J. Hespanha, and D. Kriegman. What is the set of images of an object under all possible lighting conditions. International Journal of Computer Vision, 28(3):245–260, 1998.

[15] P. Belhumeur, J. Hespanha, and D. Kriegman. What is the set of images of an object under all possible lighting conditions. International Journal of Computer Vision, 28(3):245–260, 1998.

[16] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegmann. Eigen faces vs Fisher faces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(7):711–720, 1997.

[17] A. Bell and T. Sejnowski. An information maximization approach to blind separation and blind deconvolution. Neural Computation, 7:1129–1159,1995.

[18] C. Ben AbdelKader and P. Griffin. A local region-based approach to gender classification from face images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[19] Y. Bengio, Y. LeCun, and D. Henderson. Globally trained hand written word recognizer using spatial representation, space displacement neural networks and hidden markov models. In Advances in Neural Information Processing Systems, volume 6, pages 937–944. Morgan Kaufmann, 1994.

[20] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh,E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, volume 2, pages 848–854, 2004.

[21] A Convolution Neural Networks by Kalil hani and liew shang suang convolution neural network approach for face recogntion

[22] Dimensionality Reduction by Learning an Invariant Mapping

Raia Hadsell, Sumit Chopra, Yann LeCunThe Courant Institute of Mathematical Sciences New York University, 719 Broadway, New York, NY 1,USA

[23] Learning Algorithms For Classification A Comparison N Hand Writteen Digital Recogntion, yabb lecun ,L.D jackel,leon bottou,corinna cortes.

[24] Signature Verification using a "Siamese" Time Delay Neural Network Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Sickinger and Roopak Shah AT&T Bell Laboratories.

[25] SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, Senior Member, IEEE.