Key Frames Based Video Face Verification

Dasari Subbarao, G.Sai Ram

Abstract*: *The availability and abundance of video capture devices like mobile phones and surveillance cameras have instigated research in video face recognition, which is highly related to the law enforcement applications. And the proposed approaches are giving high accuracy in error rates, the performance at lower false accept rates requires significant improvement. Now here we are proposed face verification algorithm where Wavelet Transform and Entropy are used for frames selection from the video followed by feature extraction using deep learning where we will combine Deep Boltzmann Machine (DBM) and Stacked Denoising Sparse Auto-Encoder (SDSAE) with learnt representations for Face verification. After completion of all these steps finally we obtained verification details by multilayer neural network system (MNNS). The proposed feature richness-based frame selection shows fair performance compared to the other methods namely Random frames or frame selection based on no visual reference image quality measures. The proposed method in this paper shows good performance in face verification. Face verification accuracy of proposed method is about 97% and 95% with false 1% accept rate on point and Shoot(PaS), YouTube Video(YTVF) face databases respectively.

Index Terms*: *Face Verification, Face Recognition, Feature Frames Selection, Deep learning, Auto Encoder, Deep Boltzmann Machine.

I. INTRODUCTION

Face Recognition from Video was turned out to be highly in observation situations. For instance, in excess of Eighty thousand individuals checked using Face Recognition in Beijing Olympics-2008.

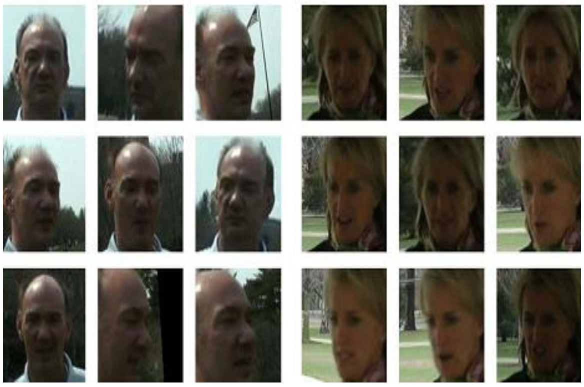


Figure 1: A Frames subset of a video showing the information quantity.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author

Dr.Dasari Subbarao*, Professor, Department of ECE, Siddhartha Institute of Engineering and Technology, Ibrahimpatnam, Hyderabad, India.

G.Sai Ram, Assistant Professor, Department of ECE, Siddhartha Institute of Engineering and Technology, Ibrahimpatnam, Hyderabad, India.

The Video consisting of more frames that means sequence of frames represents the video that contains different poses, expressions. Some frames in the video are used face recognition [2]. In some situations, videos which are recorded by the gadgets are used by Law organizations for face verifications gives more inspiration. Figure 1 shows video cuts having different faces. Face Recognition from Video has been broadly contemplated with few proposed algorithms.

The video face recognition mainly classified in two set-based and sequence based. A video as an arrangement of pictures (outlines) which are then demonstrated and coordinated utilizing an assortment of approaches in set and sequence-based which are not uses the worldly data present in the video. Then again, grouping based methodologies are particularly intended to use transient data in video shows grouping of pictures and use arrangement characterization systems to get face recognition.

For correlation, outcomes are for the most part given an account of benchmark databases, for example, the UCSD Honda [7], YTVF [3] and PaS Challenge databases [2]. The existing calculations have achieved superior on YTVF dataset [3]. These database requires outcomes at meet Equal Error Rate(EER) [2]. In view of execution point, the calculations need to restrict False Accept Rate(FAR) or False Reject Rate (FRR).

Low EER will never give lower FRR. Figure 2 summarizes performance of algorithms existed for database YouTube Faces [3] having more accuracies with low FAR at same error rate. As shown in figure 2, EigenPEP gives 85% EER with 49% acceptance rate at 1% false accept rate (FAR).

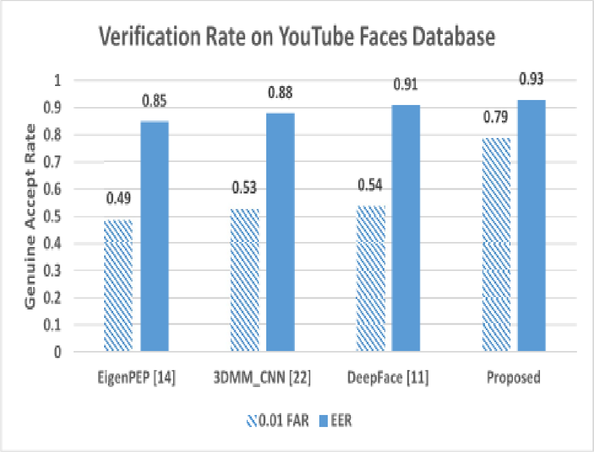


Figure 2: Summary of performance of algorithms existed.

At low false accepts rates there is a lack of performance compared the performance near EER.

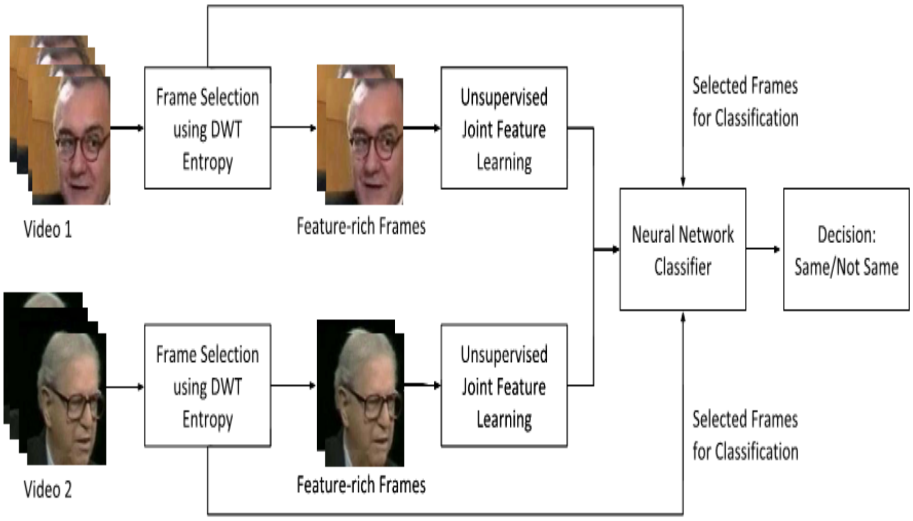

Video 1
Frame Selection using DWT Entropy
Feature-rich Frames
Unsupervised Joint Feature Learning
Selected Frames for Classification
Neural Network Classifier
Decision: Same/Not Same
Video 2
Frame Selection using DWT Entropy
Feature-rich Frames
Unsupervised Joint Feature Learning
Selected Frames for Classification

Figure 3: Steps involved in Face Recognition Algorithm: Proposed

A. Research Advancements

As a rule, Face recognition from video includes coordinating utilizing every one of the casings introduces in two recordings. However, not all edges are similarly instructive and a few casings may experience the ill effects of low picture quality or outrageous varieties because of illusion, expression and pose. Because of face recognition covariates of nearness. A few edges may influence the inter and intra classes varieties. As it were, it is exceedingly likely that highlights take out from frame might leads to results incorrect.

Subsequently, choose and use of high data content of video is vital and effective and it gives video information all the more difficult and additionally compensating recognition of face. To specify some of these restrictions, increases general execution, we stated a novel video face recognition algorithm with casing determination process for feature extraction and matching along with deep learning architecture as shown in Fig. 3.The first contribution is quantifying entropy based feature-richness in wavelet domain by feature-richness based frame selection algorithm for no-reference feature-richness [9] in compared with traditional biometric quality measures with no-reference [3].

The second commitment is planning a novel joint element understanding structure that is used for consolidate middle of the road highlights processed in a profound system. Profound learning designs by and large process a progression of middle of the road highlights from information input, use last highlights layer just for order and portrayal. We consolidate middle of road portrayals registered by an auto encoder utilizing a combine portrayal layer in the stated deep architecture,. This combined portrayal is used in contribution to a Deep Boltzmann Machine (DBM).

II. LITERATURE SURVEY

For the motivations behind this synopsis, the primary discoveries and suggestions of the report are separated into five general classifications: execution, assessment, activity, approach, moral and political contemplations. Here discoveries as well as proposals utilize specific specialized ideas. In particular, in any case, this report suggests that good and political contemplations be viewed as on a standard with practical execution contemplations, impacting the plan of innovation and establishment and additionally operational approaches all through the procedure of improvement and arrangement and not just attached toward the end.

[2] J. Beveridge et al published his "point and shoot" camera innovation given inspiration to utilize confront acknowledgment innovation. Notwithstanding the obvious straightforwardness of the issue, confront acknowledgment in this setting is hard. Generally, disappointment rates in the 4 to 8 out of 10 territory are normal. Conversely, blunder rates fall about one out of one thousand and is very much symbolism. This paper presents the Point and Shoot Face Recognition(PaSFR) Challenge. The test incorporates 9,376 still pictures of 293 individuals adjusted as for separation to the camera, elective sensors, frontal as opposed to not-frontal perspectives, and shifting area. There are additionally 2,802 recordings for 265 individuals: a subset of the 293. Check comes about are introduced for open standard calculations and a business calculation for three cases: contrasting still pictures with still pictures, recordings to recordings, and still pictures to recordings.

[3] L. Wolf, T. Hassner, and I. Maoz, Recognizing faces from video recordings is an errand of mounting significance. While clearly identified with confront acknowledgment in still pictures, it has its own particular interesting attributes and algorithmic prerequisites. Throughout the years a few strategies have been proposed for this issue, and a couple of benchmark informational collections have been amassed to encourage its examination.

III. FACE RECOGNITION ALGORITHM PROPOSED

The Algorithm proposed in this paper is mainly carried out in three steps namely frame selection, feature extraction based on deep learning and verification of face based on learnt representations. The steps involved in the algorithm are shown clearly in Figure 3.

A. Frame Selection Based On Entropy

Four to five seconds video clip will have 100 to 200 frames depending upon its frame rate and time. Liu et al. [3] has given in his Principal Component Analysis (PCA) about how to partition the video into cluster frames without redundant frames. Convert the detected face image I into gray scale by preprocessing to get feature richness. By taking only the facial region frames face detection is performed here.

Mean and Standard Deviations are used for image normalization. Then DWT is applied on the image I with the following equation (1).

$$\left[I_{APP}, I_{Hor}, I_{Ver}, I_{Diag}\right] = DWT(I) \qquad (1)$$

Here, I_{App} represents image the approximation coefficients, $[I_{Hor}, I_{Ver}, I_{Diag}]$ represents horizontal, vertical, and diagonal sub-bands detailed coefficients respectively.

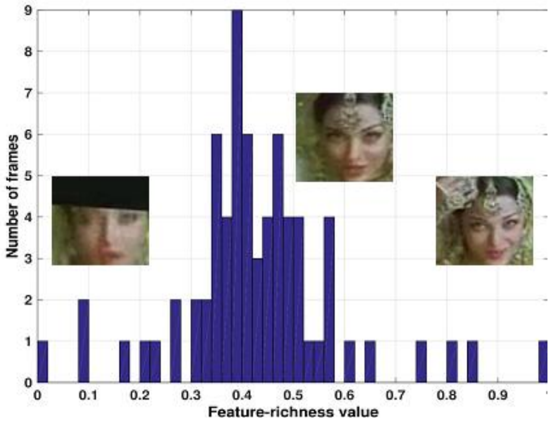


Figure 4: Distribution of Feature Richness values of video.

Values one as well as zero represents the most and least feature-rich frames having score with high and poor fidelity frames respectively.

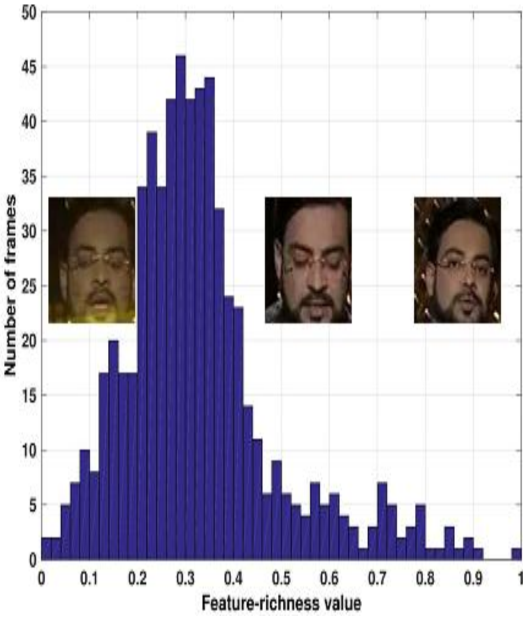


Figure 5: Distribution of Feature Richness values of video.

One level DWT of image is obtained from equation(1) and another level of DWT is related to I_{APP}.

$$\left[I_{AP}, I_{Hor}\ I_{Ver}, I_{Diag}\right] = DWT(I_{APP}) \qquad (2)$$

Here, I_{APP} and $[I_{Hor}, I_{Ver}, I_{Diag}]$ are DWT second level of estimation and detail . Entropy is calculated for these DWT Levels. Each DWT band is sub divided into 3x3 windows and Local entropy of it is computed, for an image window k Entropy is given by the equation(3).

$$H(k) = -\sum_{i=1}^{n} P(K_i)\log_2 P(K_i) \qquad (3)$$

where, n represents aggregate of pixel esteems, and $P(K_i)$ indicates estimation of the likelihood mass capacity for K_i . On the off chance that the span of the window κ is Mk × Nkat that point

$$P(Ki) = \frac{n(Ki)}{M_k \times N_k} \qquad (4)$$

Where $n(K_i)$ indicates pixels quantity in K_i window. The score of feature richness of a DWT of a window number w with entropy H(i) is given by the equation (5).

$$H(F) = \sum_{i=1}^{\omega}(|H(i)|) \qquad (5)$$

The image I final score is the aggregate of the individual bands feature-richness values given by the equation (6).

$$HF(I) = HF(I'_{APP}) + HF(I'_{Hor}) + HF(I'_{Ver}) + HF\left(I'_{Diag}\right) + HF(I_{Hor}) + HF(I_{Ver}) + HF\left(I_{Diag}\right) \qquad (6)$$

Give m_i a chance to speak to the element abundance esteem comparing I^{th} outline f_i, acquired utilizing min-max standardization.

$$mi = \frac{HF(fi) - \min(HF)}{\max(HF) - \min(HF)} \qquad (7)$$

Here, HF signifies component extravagance scores of video V, base esteems min(HF) and max(HF) greatest esteems in HF. More component rich edge will have higher value of m_i. Figure 4 and 5 demonstrates element abundance circulation for two recordings of various people from the YouTube Faces database [3] alongside test edges of high, normal, and low element wealth esteems. Once the score of each casing is registered, versatile edge determination is performed to decide the ideal arrangement of edges to speak to a video clip. Assume σ_m indicate the standard deviation(SD) and μ_m signify the mean relating to the arrangement of highlight lavishness estimations of the video V. With a specific end goal to choose which outlines are chosen for check, ϕ_i is figured for each edge.

$$\varphi i = \begin{cases} 1 & \text{if } mi \geq \mu m + \frac{\sigma m}{2} \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

B. Feature Extraction using Deep Learning Framework

After component rich edges acquired, following stage includes highlight extraction and coordinating.

A few condition of-theart calculations in late writing use convolutional neural network(CNN) systems.

i. Stacked Denoising Auto Encoder(SDAE) and Deep Boltzmann Machines(DBM):

Gaussian-Bernoulli Restricted Boltzmann Machines(RBMs) are used here, vitality is characterized as:

$$\mathrm{E(v,h;\theta)} = -\sum_{\mathrm{i=1}}^{\mathrm{D}} \frac{\mathrm{vi}}{\mathrm{\sigma i}} \sum_{\mathrm{j=1}}^{\mathrm{F}} \mathrm{Wij\ \ hj} - \sum_{\mathrm{i=1}}^{\mathrm{D}} \frac{\mathrm{(vi-bi)^2}}{2\sigma^2} - \sum_{\mathrm{j=1}}^{\mathrm{F}} \mathrm{aj\ hj} \qquad (9)$$

Here, v ∈ RD signifies the genuine esteemed obvious vector and model parameters θ = {a, b, W, σ}.

The vitality of this DBM with three layers can be characterized as:

$$\begin{aligned}\mathrm{E(v,h;\theta)} = &-\sum_{\mathrm{i=1}}^{\mathrm{D}}\sum_{\mathrm{j=1}}^{\mathrm{F1}} \mathrm{Wij}^{(1)}\frac{\mathrm{vi}}{\mathrm{\sigma i}}\ \ \mathrm{hj}^{(1)} \\ &-\sum_{\mathrm{j=1}}^{\mathrm{F1}}\sum_{\mathrm{l=1}}^{\mathrm{F2}} \mathrm{Wij}^{(2)}\mathrm{hj}^{(1)}\mathrm{hl}^{(2)} \\ &-\sum_{\mathrm{l=1}}^{\mathrm{F2}}\sum_{\mathrm{m=1}}^{\mathrm{F3}} \mathrm{Wim}^{(3)}\mathrm{hl}^{(2)}\mathrm{hm}^{(3)} \\ &-\sum_{\mathrm{i=1}}^{\mathrm{D}}\frac{\mathrm{(vi-bi)}^2}{2\sigma^2} - \sum_{\mathrm{j=1}}^{\mathrm{F1}} \mathrm{aj}^{(1)}\mathrm{hj}^{(1)} \\ &-\sum_{\mathrm{l=1}}^{\mathrm{F2}} \mathrm{al}^{(2)}\mathrm{hl}^{(2)} \\ &-\sum_{\mathrm{m=1}}^{\mathrm{F3}} \mathrm{am}^{(3)}\mathrm{hm}^{(3)} \qquad (10)\end{aligned}$$

Here, D, F1, F2, F3 gives quantity of units and obvious and concealed layers, and θ = {W(1), W(2), W(3), b, a(1), a(2), a(3), σ}. The likelihood doled out by above model, V as obvious vector is given by the Boltzmann appropriation:

$$P(V;\theta) = \frac{1}{Z(\theta)}\sum_h \exp\left(-E(V,h^{(1)},h^{(2)},h^{(3)};\theta)\right) \qquad (11)$$

Here, normalizing constant is Z(θ). On the off chance that exclusive W(1) is viewed as, the subsidiary of the log-probability as for the model parameters is:

$$\frac{\delta \log P(V;\theta)}{\delta W(1)} = Epdata\left[Vh^{(1)^T}\right] - Epmodel\left[Vh^{(1)^T}\right] \quad (12)$$

Where EPdata[•] and EPmodel[•] indicates the expectation with respect to the data distribution and data distribution by the DBM respectively.

ii. Unsupervised Joint Feature Learning

We declaration our proposed engineering ought to have the capacity to create a strong portrayal contrasted with utilizing SDAE or DBM in disengagement. Later DBM can decipher the highlights from joint portrayal and consolidate every one of its segments as needed to get improved larger amount discriminative portrayal, particularly later adjusting. Give the measure of the information a chance be M × N; in the proposed engineering method, each one of SDAE layer is one-fourth the span of its past layer. A joint portrayal as delineated in Figure 6.

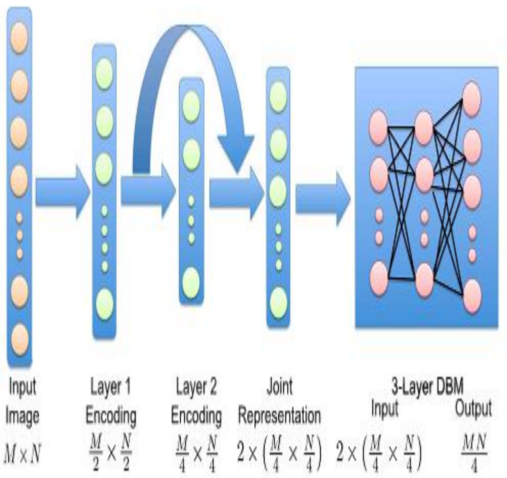

Input
Image
$M \times N$
Layer 1
Encoding
$\frac{M}{2} \times \frac{N}{2}$
Layer 2
Encoding
$\frac{M}{4} \times \frac{N}{4}$
Joint
Representation
$2 \times \left(\frac{M}{4} \times \frac{N}{4}\right)$
3-Layer DBM
Input
$2 \times \left(\frac{M}{4} \times \frac{N}{4}\right)$
Output
$\frac{MN}{4}$

Figure 6: Proposed Facial representation Deep Learning Architecture.

A joint portrayal is obtained by consolidating the data from two SDAE encoding layers.

C. Feature Richness and Deep Learning Based Representations for Face Verification

The videos to be coordinated may have critical varieties in quality and highlight abundance. It has been appeared in writing that if the pictures are of altogether different quality, at that point the coordinating execution may break down. In this manner, we play out a post-handling advance to choose frame pairs with comparable component wealth and dispose of the rest of the sets. Give V1 and V2 a chance to be the two recordings to be coordinated, a couple astute element lavishness esteem is figured for every conceivable casing pair utilizing the calculation clarified in Section II-A.

$$\begin{bmatrix}\mathrm{m1{,}1m1{,}2;\, m2{,}1m2{,}2;\, \dots,\, mi,1mj,2;} \\ \mathrm{\dots,\, mN1{,}1mN2{,}2}\end{bmatrix} \qquad (13)$$

Here mi,1mj, 2 means the result of highlight wealth esteem related with the combine shaped by the I th outline from V1 and the j th outline from V2. N1 and N2 indicate the aggregate number of chose outlines from V1 and V2 individually. Give σm a chance to standard deviation and μm is mean relating to arrangement of match savvy highlight wealth esteems for all sets conceivable amongst V1 and V2. To at last select the sets for basic leadership, following condition is used:

$$\gamma_{\mathrm{i,j}} = \begin{Bmatrix}1 & \mathrm{if\ mi,1mj,2} \geq \mu'\mathrm{m} + \frac{\sigma'\mathrm{m}}{2} \\ 0 & \mathrm{otherwise}\end{Bmatrix} \qquad (14)$$

In the event that the joined score of a couple fi,1 f j,2 is more than the limit, i.e., if Yi, j = 1, at that point this combine is considered for figuring the match score. While sets with Yi, j < 1 are not considered for confirmation, other chose outline sets are weighted by the joint element extravagance esteem.

IV. RESULTS

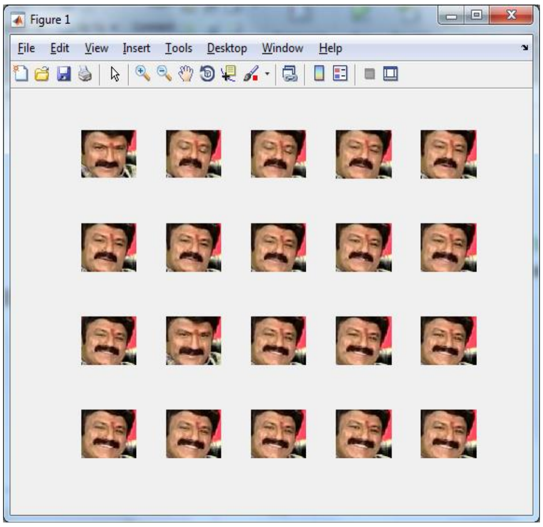


Figure 7: Video face images for training the data.

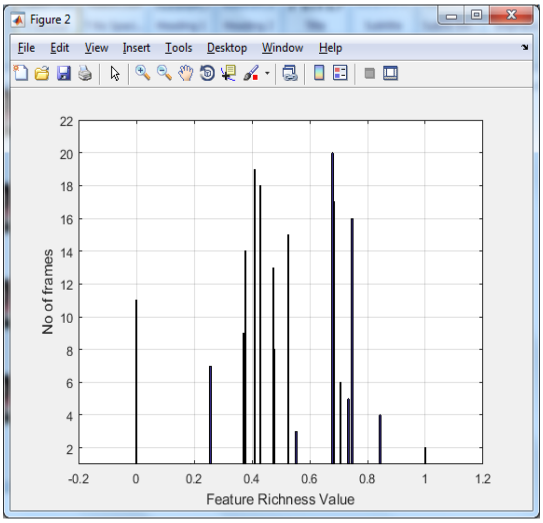
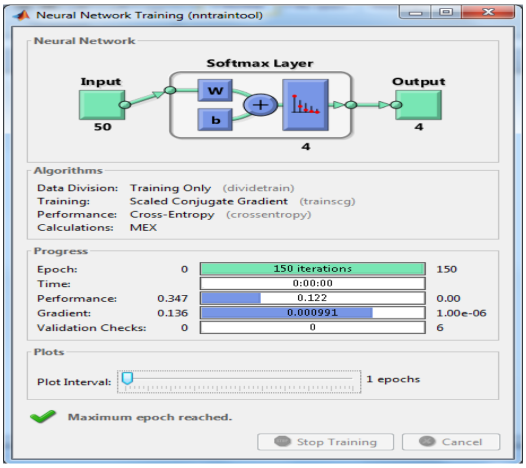


Figure 8: Feature reachness value.



Algorithms
Data Division: Training Only (dividetrain)
Training: Scaled Conjugate Gradient (trainscg)
Performance: Cross-Entropy (crossentropy)
Calculations: MEX

Progress
Epoch: 0 150 iterations 150
Time: 0:00:00
Performance: 0.347 0.122 0.00
Gradient: 0.136 0.000991 1.00e-06
Validation Checks: 0 0 6

Plots
Plot Interval: 1 epochs

Maximum epoch reached.

Figure 9: NN training.

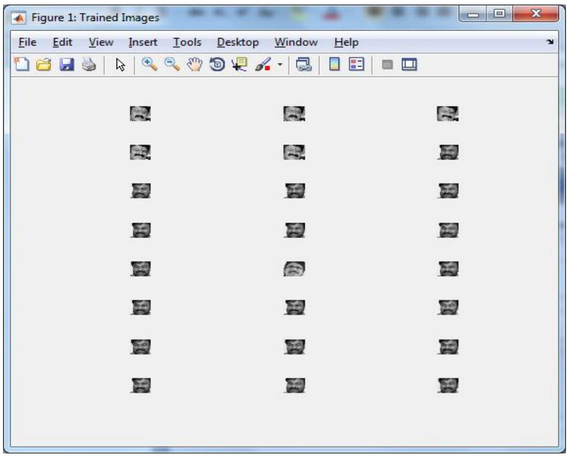


Figure 10:Training images for classifier.

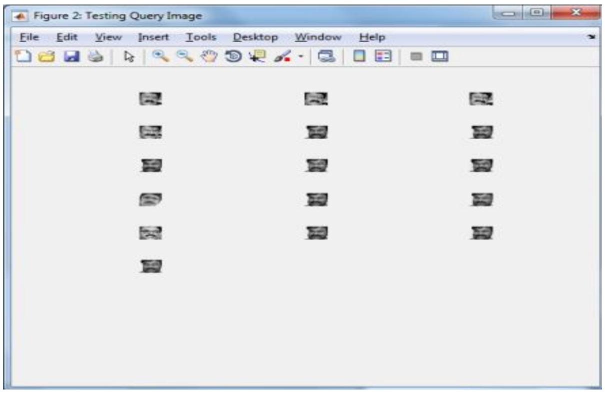


Figure 2: Testing Query Image
File Edit View Insert Tools Desktop Window Help

Figure 11:Testing images for classifier.

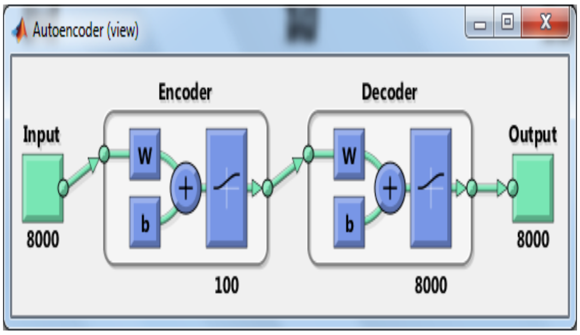


Figure 12: Autoencoder.

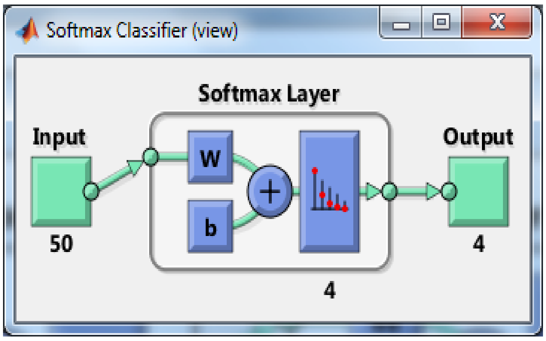


Figure 13: Softmax Classifier.

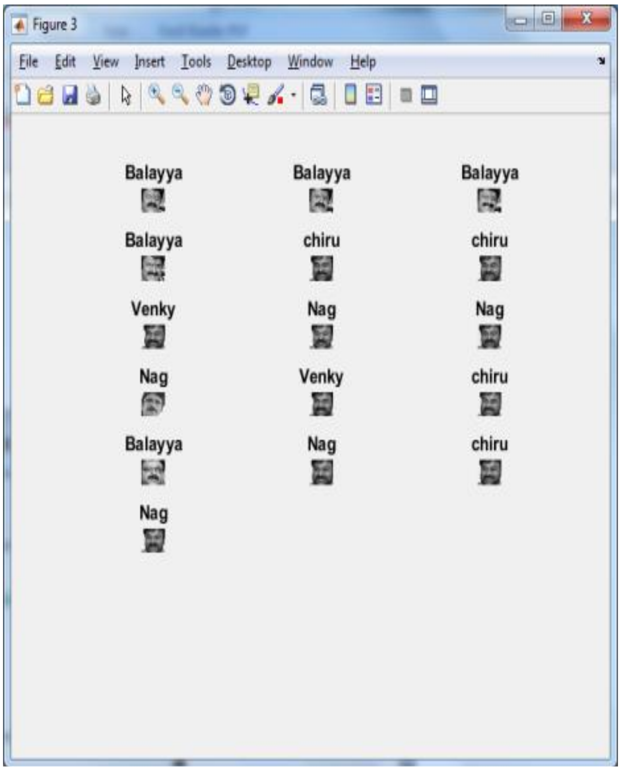


Figure 14: Face verified results.

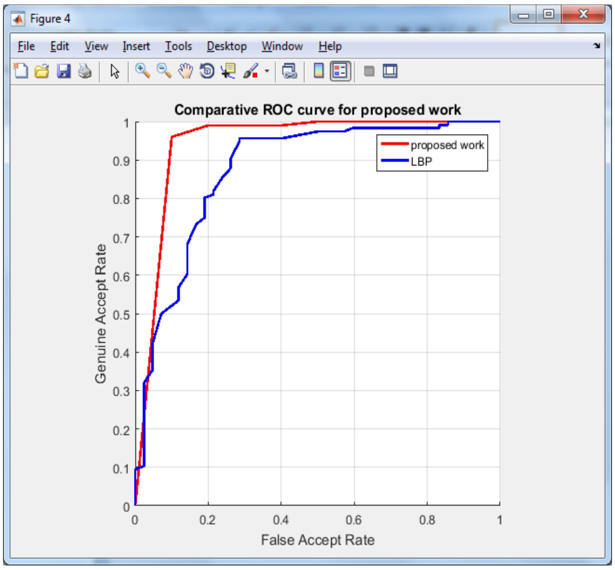


Figure 15: ROC graph for comparison.

V.CONCLUSION

The proposed profound learning design which consolidates SDAE portrayal in addition DBM is utilized to separate highlights from the chose outlines. The removed portrayals from two recordings are coordinated utilizing a forward neural bolster system. The outcomes exhibited here on the testing PaS Challenge and YTVF databases. Face verification from the two video databases is done with lower error rates. Face verification accuracy of proposed method is about 97% and 95% with false 1% accept rate on point and Shoot(PaS), YouTube Video(YTVF) face databases respectively are achieved finally.

REFERENCES

1. Beveridge J et al., “The challenge of face recognition from digital point and-shoot cameras,” in Proceedings of IEEE Conference on Biometrics Theory Application Systems, October. 2013, pp. 1–8.
2. K. Yadhul, Thusnavis Bella Mary I., Lakshmi P. S. and A. Saju, "Face detection and recognition with video database," 2014 International Conference on Electronics and Communication Systems (ICECS), Coimbatore, 2014, pp. 1-5.
3. L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” in Proceedings of IEEE Conference on Computer Visual Pattern Recognition, June(2011), pp. 529–534.
4. S. V. Tathe, A. S. Narote and S. P. Narote, "Face detection and recognition in videos," 2016 IEEE Annual India Conference (INDICON), Bangalore, 2016, pp. 1-6.
5. S. V. Tathe, A. S. Narote and S. P. Narote, "Human face detection and recognition in videos," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 2200-2205.
6. Tat-Jun Chin, U. James, K. Schindler and D. Suter, "Face Recognition from Video by Matching Image Sets," Digital Image Computing: Techniques and Applications (DICTA'05), Queensland, Australia, 2005, pp. 28-28.
7. Changbo Hu, J. Harguess and J. K. Aggarwal, "Patch-based face recognition from video," 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, 2009, pp. 3321-3324.
8. Y. Lee, C. Hsu, P. Lin, C. Chen and J. Wang, "Video summarization based on face recognition and speaker verification," 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), Auckland, 2015, pp. 1821-1824.
9. Gopatoti, A., Naik, M.C., Gopathoti, K.K.” Convolutional Neural Network based image denoising for better quality of images”, International Journal of Engineering and Technology (UAE), Vol.7, No.3.27, (2018), pp. 356-361.
10. H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, “Probabilistic elastic matching for pose variant face verification,” in Proceeding of IEEE Conference on Computer Visual Pattern Recognition, June(2013), pp. 3499–3506.

AUTHORS PROFILE

Dr.Dasari Subbarao is working in Siddhartha Institute of Engineering and Technology as a Professor in the Department of ECE located at Ibrahimpatnam, Hyderabad. He has done his B.Tech in ECE from JNTUH, Hyderabad in the year 2003 and M.Tech in Embedded Systems from SRM University in the year 2007 located at Chennai. He has done his Ph.D in ECE in 2014 from VBS Purvanchal University, Jaunpur. About 96 papers got published by him in International Journal, He is the life Time member of ISTE, Fellow of Institute of Electronics and Telecommunication Engineers (IETE) and also he is the member of International Association of Engineers (IAENG).His area of Interest is Wireless Communications and Embedded Systems.

Mr.G.SAI RAM is working as Assistant Professor in ECE Department in Siddhartha Institute of Engineering and Technology located at Ibrahimpatnam, Hyderabad. He has done his B.Tech in ECE from JNTUH, Hyderabad in the year 2013 and M.Tech in Digital Electronics and Communication Systems from JNTUH, Hyderabad in the year 2016 .His area of interest is Communication systems, Digital Signal Processing, Digital Image Processing.