MDLFace: Memorability Augmented Deep Learning for Video Face Recognition

Gaurav Goswami, Romil Bhardwaj, Richa Singh, and Mayank Vatsa
IIIT-Delhi, India
{gauravgs,romil11092,rsingh,mayank}@iiitd.ac.in

Abstract

Videos have ample amount of information in the form of frames that can be utilized for feature extraction and matching. However, face images in not all of the frames are "memorable" and useful. Therefore, utilizing all the frames available in a video for recognition does not necessarily improve the performance but significantly increases the computation time. In this research, we present a memorability based frame selection algorithm that enables automatic selection of memorable frames for facial feature extraction and matching. A deep learning algorithm is then proposed that utilizes a stack of denoising autoencoders and deep Boltzmann machines to perform face recognition using the most memorable frames. The proposed algorithm, termed as MDLFace, is evaluated on two publicly available video face databases, Youtube Faces and Point and Shoot Challenge. The results show that the proposed algorithm achieves state-of-the-art performance at low false accept rates.

1. Introduction

Face recognition is a very important area of research in biometrics due to the non-intrusive nature of capture and possibility of applications. Face recognition with still face images has been extensively studied and several approaches have been proposed to perform still image face recognition in the presence of challenging covariates such as pose, illumination, aging, disguise, and plastic surgery [2, 6, 8, 9, 20]. Despite these advancements, face recognition using still face images has its limitations, especially in unconstrained scenarios. In the past few years, video based face recognition has emerged as an alternative to address the challenges involved in still face recognition. Video face recognition is also relevant due to the increasing number of CCTV cameras installations at key locations as well as the easy availability of video recordings.

Videos encompass a multitude of pose, illumination, and expression variations which can aid in extracting robust and resilient features for each subject. There are a few algorithms on video based face recognition that utilize varying approaches ranging from frame by frame matching to advanced deep learning architecture. Park *et al.* [29] demonstrated 99% accuracy on the CMU FIA [25] database using an adaptive fusion scheme with multiple matchers. Liu *et al.* [30] achieved 98.8% accuracy on the CMU MoBo [26] database by utilizing adaptive hidden Markov models. Thomas *et al.* [7] reported 99% accuracy on the UCSD database [18] by applying principal component analysis. Recent experiments in video face recognition have been performed on the Youtube Faces database [19], where Bhatt *et al.* [13] have applied clustering based re-ranking and fusion to obtain 80.7% accuracy. Hu *et al.* [15] have utilized deep metric learning to achieve 82.3% accuracy. Taigman *et al.* [33] have reported 91.4% on this database using a nine-layer deep neural network. Barr *et al.* [16] have provided a detailed review of the algorithms proposed for video based face recognition.

The current results on video face recognition show very high performance on databases including the YouTube Faces database; however, closer observation illustrates a major limitation. While, as per defined protocol, existing algorithms demonstrate verification accuracies of over 80% at Equal Error Rate (EER), their performance at lower False Accept Rates (FARs) is quite low. In applications where false accepts are associated with a significant cost, such as law enforcement and secure access, it is important that the false accept rate should be as low as possible. It is our assertion that, while existing algorithms yield good results at higher FARs, there is a significant scope of improving the genuine accept rates (GAR) at lower FARs.

Depending on the frame rate, videos can contain a large number of frames. However, all of these frames may not contribute equally towards face recognition. Some frames are almost identical to each other, whereas some frames can have a face image which is challenging to recognize due to extreme expression, illumination, and pose variations. Utilizing such frames for recognition can significantly affect the performance. Therefore, it is desirable to select the best set of frames from a video to achieve accurate recognition. Generally, existing algorithms either select a small number of frames from all the available frames or use all the frames. While using all frames increases the computational overhead to process one video, selecting random frames is not equivalent to selecting optimal frames.

To address (1) low performance at lower FAR and (2) high computational overhead in video based face recognition, this research presents a memorability based frame selection followed by deep learning architecture for feature extraction. As the first contribution, this paper presents a memorability based frame selection approach. Memorability can be defined as the quality of being easy to remember [21]. Recently, Isola *et al.* have proposed to automatically predict the memorability of scenes [21, 22]. In past, human cognition research has shown that memorability does impact face recognition [14, 17]. Faces that are more memorable are more likely to be accurately identified by human subjects. Therefore, it is our hypothesis that incorporating memorability in frame selection can improve the recognition performance.

Besides memorability based frame selection, this research also proposes a deep learning based algorithm for feature extraction and face verification. Deep learning can be defined as a set of algorithms in machine learning that attempt to model high-level abstractions in data by using architectures composed of multiple non-linear transformations [31]. Recently, deep learning based algorithms have been successfully applied in speech recognition [10], object recognition [3, 28], natural language processing [24], and face recognition [33]. In the proposed approach, a stack of denoising autoencoders and deep Boltzmann machines are utilized to extract face representation that, along with memorability based frame selection, helps to improve the recognition performance. The contributions of this research can be summarized as follows:

1. A novel algorithm to compute the memorability scores of face images is proposed and utilized to devise a frame selection approach for video based face recognition.

2. A deep learning architecture based video face recognition algorithm is proposed that achieves state-of-the-art results on two existing video databases: Point and Shoot Challenge (PaSC) [5] and YouTube Faces [19].

2. Proposed MDLFace Algorithm

Figure 1 illustrates the overview of the proposed memorability augmented deep learning algorithm for face recognition, termed as MDLFace. It comprises of two components: (a) memorability based frame selection and (b) deep learning architecture based face verification algorithm. These components are discussed in detail in the following subsections.

2.1. Memorability based Frame Selection

Past research in human cognition has shown that certain faces are more memorable than others [14, 17]. They have also shown that certain faces can be more accurately remembered by human subjects as seen/unseen compared to other faces. In context of human faces, this phenomenon is termed as memorability. Memorability is not the same as attractiveness, implying that attractive faces are not the same as memorable faces. In accordance with the observations made in these studies, it is our assertion that using memorable frames should facilitate face recognition by providing the images that contain discriminative/unique information for feature extraction. In addition to the inherent performance advantage in utilizing only a subset of frames instead of all the frames, the algorithm can also avoid inclusion of spurious feature information if it is able to filter redundant or non-informative frames.

Following these observations, the primary hypothesis of memorability based frame selection algorithm is that the memorability of a face is not static but may vary depending on the image in consideration. In case of a video, multiple images of a face are available with different illumination, pose, expression, viewpoint, and camera distance. The memorability of an individual's face changes with these variations and the objective of applying memorability to frame selection is to select the set of most memorable frames from the available pool.

The proposed algorithm computes memorability by quantifying the *feature richness* of the image and analyzing the image content locally at a fine level. By analyzing the image in both dense and fine manner, the image is represented as a sum of small parts which resembles the way visual data is processed by the human visual system in V1 and V2 cells [34]. In order to compute memorability score, first the input image I is preprocessed to be of a fixed size and is converted to the HSV (Hue, Saturation, Value) color mode. After conversion to HSV color mode, only Hue is considered for further processing in order to de-couple memorability computation from brightness and attain resilience towards illumination variations. Thereafter, the image is divided into equally sized overlapping windows of size 2×2 and visual entropy [4] of each window is computed. Visual entropy can provide an estimate of the feature-richness of an image region. For an image region, visual entropy signifies the variation in pixel intensity values. An image region with constant pixel values has the lowest visual entropy value, whereas it is high for a region with high variations. The visual entropy, $H(\mathbf{x})$ of an image region $\mathbf{x}$ is computed according to Eq. 1:

$$H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \qquad (1)$$

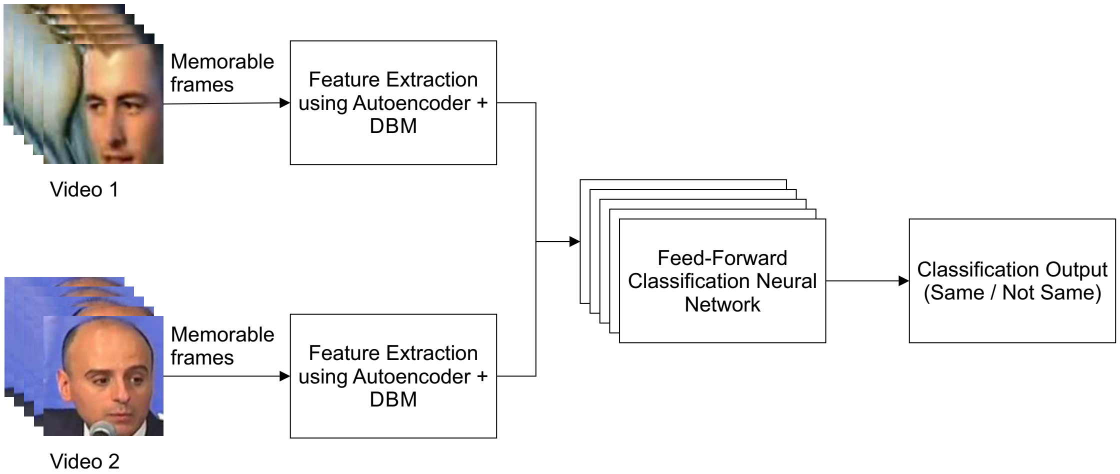

Memorable frames
Feature Extraction using Autoencoder + DBM
Video 1
Memorable frames
Feature Extraction using Autoencoder + DBM
Video 2
Feed-Forward Classification Neural Network
Classification Output (Same / Not Same)

Figure 1. An overview of the steps involved in the proposed MDLFace algorithm.

where, $p(x_i)$ is the value of the probability mass function for x_i. In case of images, $p(x_i)$ signifies the probability that the pixel value x_i appears in the neighborhood and n is the total number of possible pixel values. If the size of image region $\mathbf{x}$ is $M_H \times N_H$ then

$$p(x_i) = \frac{n_{x_i}}{M_H \times N_H} \tag{2}$$

Here, n_{x_i} denotes the number of pixels in the region with value x_i, and $M_H \times N_H$ is the total number of pixels. The visual entropy value of each window is combined to obtain the memorability of image I according to Eq. 3:

$$\mathbf{H} = \sum_{i=1}^{n}(|H(\mathbf{x_i})|) \tag{3}$$

where, $\mathbf{H}$ denotes the memorability score of the image, n is the number of windows in I and $H(\mathbf{x_i})$ denotes the entropy of the i^{th} window in I.

The memorability score of an image is thus the sum total of variations in visual entropy of fine local region of the image. Higher the value, higher is the memorability of the input image I. Therefore, in order to perform frame selection for a video, memorability of each frame is computed using the proposed algorithm and top frames are selected according to high memorability scores.

2.2. Deep Learning Approach for Face Recognition

Once the memorable frames are obtained, feature extraction and matching are performed using a deep learning architecture that comprises of stacked denoising autoencoders (SDAE) and Deep Boltzmann Machine (DBM) for feature/representation learning and neural network classification. Figure 2 shows the steps involved in the deep learning architecture. First, we briefly present an overview of SDAE and DBM followed by the proposed architecture.

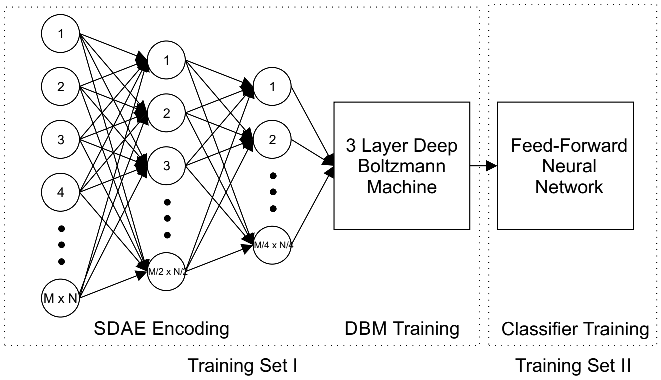


Figure 2. Deep learning architecture for face recognition.

2.2.1 Stacked Denoising Auto-Encoder and Deep Boltzmann Machines

Let $\mathbf{x} \in R^a$ be the input data; an autoencoder [32] maps the data into a feature (latent representation) $\mathbf{f}$ using a deterministic (encoder) function g_p

$$g_p(x) = s(\mathbf{w} \cdot \mathbf{x} + \mathbf{b}) \tag{4}$$

where, $p = \{\mathbf{w}, \mathbf{b}\}$ is the parameter set, s represents sigmoid, $\mathbf{w}$ is the $a' \times a$ weight matrix, and $\mathbf{b}$ is the offset vector of size a'. Feature $\mathbf{f}$ can be mapped to the feature vector $\hat{\mathbf{x}}$ of dimensionality a using a decoder function $g'_{p'}$ such that

$$\hat{\mathbf{x}} = g'_{p'}(\mathbf{f}) = s(\mathbf{w}' \cdot \mathbf{f} + \mathbf{b}') \tag{5}$$

where $p' = \{\mathbf{w}', \mathbf{b}'\}$ is the approximately sized parameter set. Note that $\hat{\mathbf{x}}$ is not the exact representation of $\mathbf{x}$ but is a probabilistic approximation. The parameters are optimized by utilizing the unsupervised training data. Denoising autoencoders [23], a variant of autoencoders, operates on the noisy input data $\mathbf{x}_n$ and attempts to reconstruct $\hat{\mathbf{x}}$. It is observed that these autoencoders are robust to noisy data and have good generalizability. If the autoencoders are stacked in a layered manner, they are called stacked autoencoders and form a deep learning architecture.

Deep Boltzmann Machine is an undirected graphical model, a deep network architecture, with symmetrically coupled binary units [27]. It is designed by layer-wise training of Restricted Boltzmann Machine and stacking them together in an undirected manner. In this research, a three layer DBM is utilized with a greedy learning approach [12].

2.2.2 Stacking SDAE and DBM for Face Recognition

SDAE and DBM both individually learn the "useful" (intermediate) representation of input data. In the MDLFace architecture, we propose to stack SDAE and DBM in a sequential manner. While SDAE is robust to noise in the input data, DBM learns the internal complex representations probabilistically. Let I be a face image of size $M \times N$ (80×100). It is first converted into a vector form $[1 \times MN]$ and provided as input to a two layer SDAE. In each layer of SDAE, the number of units in the hidden layer is half of the size of previous layer. Layer-by-layer greedy approach [11] with stochastic (mini batch = 100) gradient descent is utilized to train the SDAE followed by fine-tuning with back-propagation method. The output of SDAE is used as input to the three layer DBM (500, 500, 1000 units) and pre-training approach [12] combined with generative fine-tuning is followed to train the DBM.

By using face images in an unsupervised manner, the stack of SDAE and DBM provides a feature representation that can be utilized for verification task. Let I_g and I_p be the two face images to be matched. Trained SDAE-DBM is used to extract the features ϕ from I_g and I_p. These extracted features are concatenated and given to a five layer neural network (one input layer - 3 hidden layers - one output layer) for classification (verification). In order to use this feature extraction and classification architecture for video based face recognition, as shown in Figure 1, frame by frame matching is performed. Once the neural network classifier is trained to verify a pair of input images, during testing on videos, undecimated output is utilized to combine information from multiple frames. In the proposed algorithm, from each of the gallery and probe videos, 25 memorable frames are selected. Using these frames, top n ($n = 15$ in our experiments) possible match pairs are obtained via neural network classification. For these n pairs, their undecimated network (classifier) outputs are combined using normalized sum rule [1]. Using this combined information (score), the verification threshold is used to take a decision of accept or reject.

3. Experimental Results

In order to evaluate the efficacy of the proposed MDLFace algorithm, face verification experiments are performed on two popular video face recognition benchmark datasets: YouTube Faces dataset [19] and Point and Shoot Challenge dataset [5].

3.1. Database and Experimental Protocol

Both YouTube Faces [19] and PaSC [5] datasets operate with specific experimental protocols. The respective protocols provided by both datasets are followed in order to enable a direct comparison with existing results. For the YouTube Faces dataset, the restricted protocol is followed which consists of 10 splits, each containing 250 genuine and 250 impostor pairs. Further, no information outside of these splits is to be used during any stage of evaluation. 10 fold cross validation is performed using these splits, utilizing 9 splits for training and the remaining one split for testing.

The PaSC dataset contains videos from a handheld camera of low resolution and a control camera of high resolution. The handheld-to-handheld experiment evaluates the accuracy of an algorithm when matching videos of low resolution, whereas the control-to-control evaluates the accuracy for high resolution videos. The experiments are performed for both handheld-to-handheld and control-to-control challenges in the PaSC protocol. The database distribution includes the source code and resources to perform face detection. The signature sets provided with the database are utilized to select the testing pairs while training is performed on separate training videos provided with the dataset.

For both datasets, we have divided the given training data into two sets. Training Set I consists of 60% of the training data and is utilized for training the deep learning algorithm. Training Set II consists of the remaining 40% training data and it is utilized for training the neural network classifier. After training, the proposed algorithm is evaluated on the entirety of the testing data. The results are reported in terms of Receiver Operating Characteristic (ROC) curves and the verification accuracies at different false accept rates.

3.2. Results on PaSC Dataset

Figure 3 compares the performance of the proposed MDLFace algorithm with a commercial matcher PittPatt[1] on the PaSC dataset. The results show that MDLFace achieves significantly better performance than PittPatt on both control and handheld protocols. At 1% FAR, MDLFace yields the verification accuracy of 93.4% on the control protocol and 87.4% on the handheld protocol. To the best of our knowledge, these are the highest accuracies reported on the PaSC database.

[1] The results of PittPatt are provided along with the database in [5].

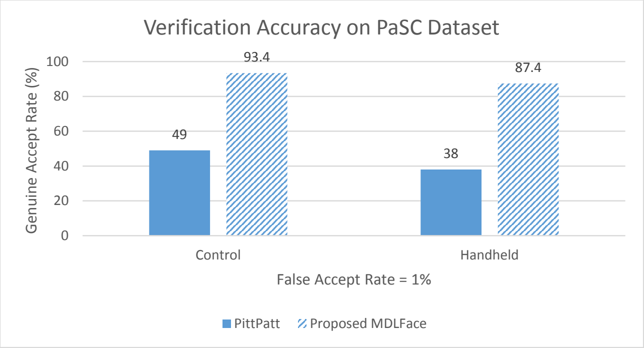


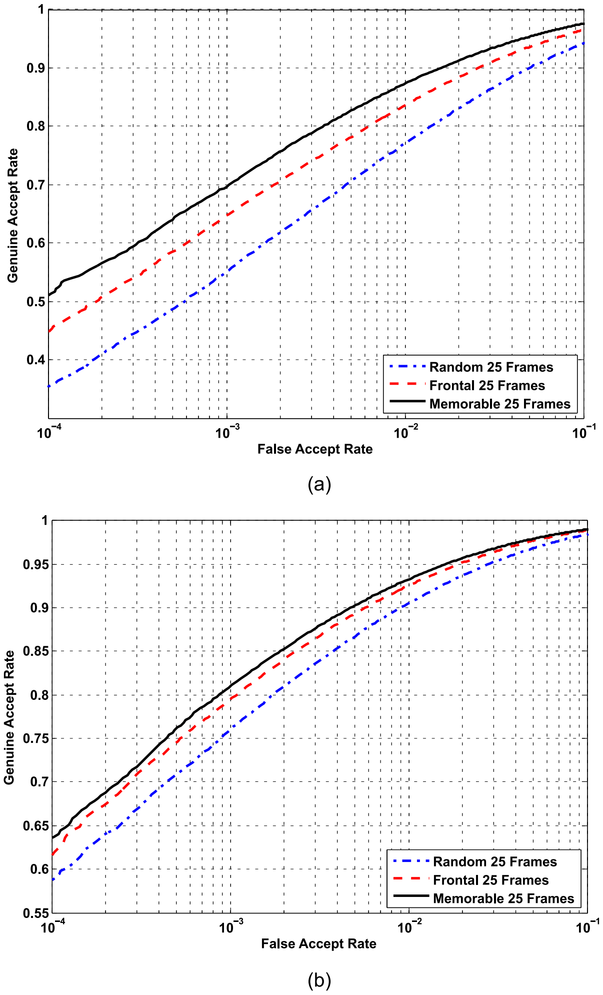
Figure 3. Comparing the performance of the proposed algorithm with PittPatt on the PaSC dataset at 1% FAR on both handheld and control video sets.

To understand the effectiveness of using memorable frames for recognition, another experiment is performed which compares the performance of MDLFace with verification accuracies obtained using 25 randomly selected frames and 25 frontal frames. Figure 4 show the ROCs of this experiment for both handheld and control protocols. Figure 5 presents some sample memorable frames from the PaSC dataset. The results indicate that memorable frames are much better for recognition compared to randomly selected frames. Also, the comparison with frontal frames indicates that memorable frames are not necessarily frontal frames.

3.3. Results on YouTube Faces Dataset

Figure 6 shows the ROC curves on the restricted protocol of the YouTube Faces dataset. The comparisons are performed with the two currently best performing algorithms: DeepFace [33] and Discriminative Deep Metric Learning [15]. It shows that the proposed algorithm outperforms all existing algorithms and is comparable to DeepFace. In order to further analyze the performance of the proposed MDLFace algorithm, the results are compared at three different false accept rates, 0.1%, 1% and 10%. Figure 7 shows that while DeepFace performs better at higher FARs, MDLFace performs much better at lower FARs of 0.1% and 1%. Specifically, at 10% FAR, DeepFace outperforms the proposed algorithm by 3.4% whereas at 0.1% FAR, the proposed MDLFace outperforms DeepFace by around 33%.

Similar to PaSC database experiments, the performance of the memorability based frame selection algorithm is also evaluated if 25 random or frontal frames are selected instead of memorable frames. The results are presented in Figure 8. It is observed that random and frontal frames do not perform as well as memorable frames, especially at very low false accept rates where memorable frames outperform by a large margin of around 17%. It is also evident that while frontal frames perform better than randomly selected frames, memorable frames achieve the best performance consistently across all FARs. Computationally, the algorithm requires less than a second to match two input videos.



(a)

(b)

Figure 4. Evaluating the effectiveness of memorable frames by comparing with randomly selected frames and frontal frames on the (a) handheld and (b) control video sets of the PaSC database.

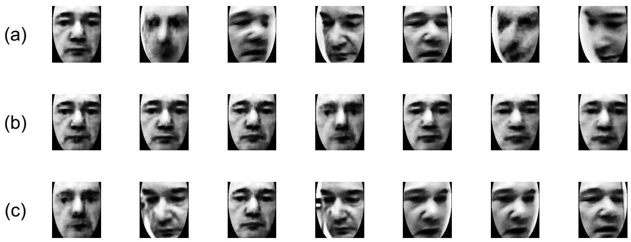


Figure 5. Sample (a) random, (b) frontal, and (c) memorable frames from a video in the PaSC dataset. It can be observed that while some frontal frames are memorable, not all memorable frames are frontal. Images shown here are processed and cropped frames.

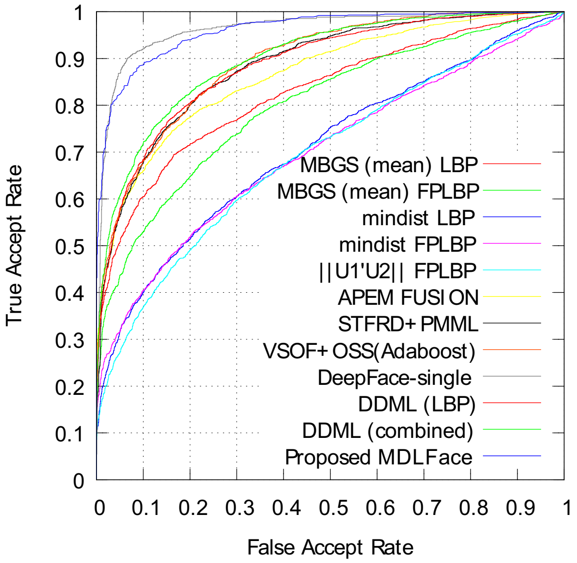


Figure 6. ROC curves on the restricted protocol of the YouTube Faces dataset [19]. The proposed algorithm outperforms almost all existing algorithms and is second only to DeepFace.

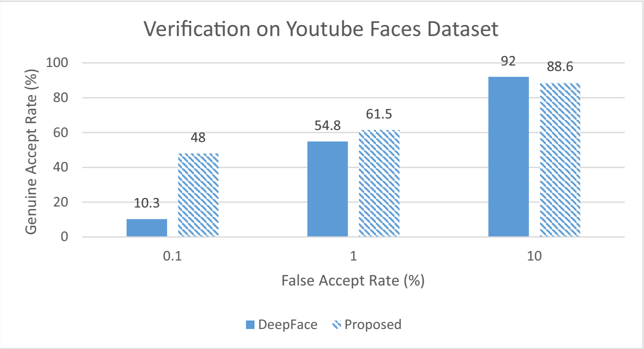


Figure 7. Comparing the performance of the proposed algorithm on the YouTube Faces database with DeepFace on three different FARs. The bar charts show that at lower FARs, the proposed algorithm yields the best results.

4. Conclusion and Future Work

In this research, a novel MDLFace algorithm is proposed for improving the state of art in video face recognition. The contributions of this is two fold: (1) algorithm for selecting the most memorable frames from a video and (2) a deep learning approach for feature extraction, coupled with neural network architecture for verification. The performance of the proposed algorithm is evaluated on the Point and Shoot Challenge and YouTube Faces databases. The comparison with state-of-the-art results on both the databases show that the proposed MDLFace provides best results on both the databases at 0.1% and 1% FAR. Currently, we are extending the proposed algorithm to further improve the verification accuracy at lower FARs.

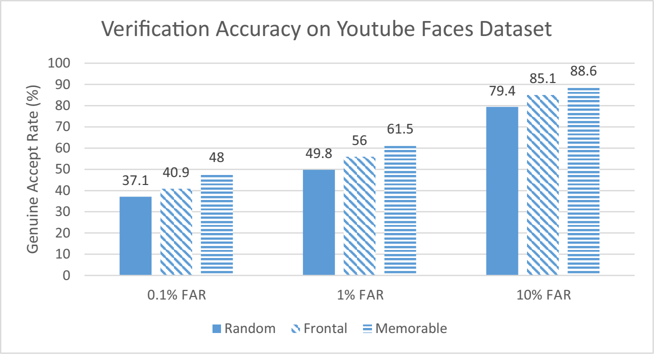


Figure 8. Comparing the performance of the proposed algorithm with random, frontal and memorable frames at different FARs.

Acknowledgement

The authors thank the CVRL lab for providing access to the PaSC database and Lior Wolf for providing access to the Youtube database. This research is partly supported through a research grant from DeitY, Government of India.

References

[1] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. 2006.

[2] A. K. Jain and S. Z. Li. *Handbook of face recognition*. Springer, 2005.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[4] A. Rrnyi. On measures of entropy and information. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961.

[5] J. Beveridge, P. Phillips, D. Bolme, B. Draper, G. Given, Y. M. Lui, M. Teli, H. Zhang, W. Scruggs, K. Bowyer, P. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE Conference on Biometrics: Theory, Applications and Systems*, 2013.

[6] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Recognizing surgically altered face images using multiobjective evolutionary algorithm. *IEEE Transactions on Information Forensics and Security*, 8(1):89–100, 2013.

[7] D. Thomas, K. W. Bowyer, and P. J. Flynn. Multi-frame approaches to improve face recognition. In *IEEE Workshop on Motion and Video Computing*, 2007.

[8] D. Yadav, M. Vatsa, R. Singh, and M. Tistarelli. Bacteria foraging fusion for face recognition across age progression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 173–179, 2013.

[9] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PLoS ONE*, 9(7):e99212, 2014.

[10] F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech*, pages 437–440, 2011.

[11] G. E. Hinton, and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[12] G. E. Hinton and R. Salakhutdinov. A better way to pretrain deep boltzmann machines. In *Advances in Neural Information Processing Systems*, volume 25, pages 2447–2455. 2012.

[13] H. S. Bhatt, R. Singh, and M. Vatsa. On recognizing faces in videos using clustering based re-ranking and fusion. *IEEE Transactions on Information Forensics and Security*, 10.1109/TIFS.2014.2318433, 2014.

[14] J. F. Cross, J. Cross, and J. Daly. Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics*, 10(6):393–396, 1971.

[15] J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[16] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(05), 2012.

[17] J. R. Vokey and J. D. Read. Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, 20(3):291–302, 1992.

[18] K. Lee, J. Ho, M. Yang and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99(3):303–331, 2005.

[19] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011.

[20] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. On recognizing face images with weight and age variations. IEEE Access, 2014.

[21] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.

[22] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–152, 2011.

[23] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[25] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In *Analysis and Modelling of Faces and Gestures*, volume 3723 of *Lecture Notes in Computer Science*, pages 255–263. 2005.

[26] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, June 2001.

[27] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[28] S. Rifai, Y. Dauphin, P. Vincent, Y. Bengio, and X. Muller. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pages 2294–2302, 2011.

[29] U. Park, A. K. Jain, and A. Ross. Face recognition in video: Adaptive fusion of multiple matchers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[30] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–345, 2003.

[31] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[32] Y. Bengio, L. Pascal, P. Dan, and H. Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, volume 19, pages 153–160. 2007.

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[34] D. Yoshor, W. H. Bosking, G. M. Ghose, and J. H. R. Maunsell. Receptive fields in human visual cortex mapped with surface electrodes. *Cerebral Cortex*, 17(10):2293–2302, 2007.