Key frame based Face in Video Recognition using Multi-Artificial Neural Network

DR. S. WILSON

Assistant Professor, Department of Computer Science,CSIJayarajAnnapackiam College,Nallur - 627 853 Tamilnadu, India

Abstract

Face in video recognition has several challenges. Although deep learning approaches have achieved performance that surpasses people for still image- face recognition, video- face recognition remains a challenging task due to the large volume of data to be processed and intra / intervideo variations in pose, lighting, occlusion, scene, blur, video quality, etc. In this paper, deep convolutional neural network is used for feature extraction and artificial neural network is used for face recognition. The computation overhead of these deep learning approaches is reduced by introducing keyframe based face recognition in video. The low-quality frames are removed by extracting keyframes. The proposed method is tested on YTF dataset and the results are compared with recent methods. The experimental results substantially proved that the proposed method achieves a higher accuracy rate of 98.36% when compared with other recent methods.

Keywords: Artificial Neural Network, Keyframe, Convolutional Neural 8Network

1. Introduction

Face recognition is one of computer vision and biometrics issues most actively studied. Video in Face Recognition (VIFR) is an active research due to wide necessity.Compared to still image-based face recognition, video-based face recognition is more difficult due to a much greater amount of data to be processed, and major intra-/inter-class variations induced by motion blur, poor video quality, occlusion, frequent scene changes, and unconstrained acquisition conditions.

In [1], Histograms of Oriented Gradients (HOG) is used for feature extraction and feed-forward neural network is used for classification.A face image is processed by several Deep Convolutional Neural Networks (DCNN) which generate multiple deep features. 3D rendering isused to generate multiple face poses from the input image [2].In [3], PCA is employed to learn multistage filter banks which are followed bybinary hashing and block histograms.

Acomprehensivedeep learning framework [4]is introduced to jointly learn face representationusing multimodal information. This framework consists of a set of elaborately designed CNNs and a three-layer Stacked Auto-Encoder(SAE). This set extracts facial featuresfrom multimodal data which are concatenatedto form a high-dimensional feature vector.

A new face detection method is developed by extending the state-of-the-art Faster R-CNN algorithm [5].This method improves the existing faster RCNN scheme by combining several importantstrategies, including feature concatenation, hard negative mining, and multi-scale training, etc [6]. A structured ordinalmeasure method is usedfor video-based face recognition that simultaneously learns ordinal filters and structured ordinal features[7].

In [8], the design details of a deep learning system are presented for unconstrainedface recognition, which includes face detection,association, alignment and face verification. Zheng et al.is developed a face recognition system [9]for unconstrained video, which is composed of face detection, association and face recognition. Initially, multi-scale single-shot face detectors are used to localizefaces in videos. The detected faces are then grouped through face association methods, especially formulti-shot videos. Finally, the faces are recognized by the face matcher based on an unsupervised subspace learningapproach and a subspace-to-subspace similarity metric.

In [10], many methods are presented for all steps of a face recognition system. In the step of face detection, a hybrid model combining AdaBoost and Artificial Neural Network (ABANN) is introduced to solve the process efficiently.

Inspired by the work in [10], this research makes use of ANN for face detection. It also includes keyframe extraction in video for improving the computational complexity and to reduce poor quality frames.

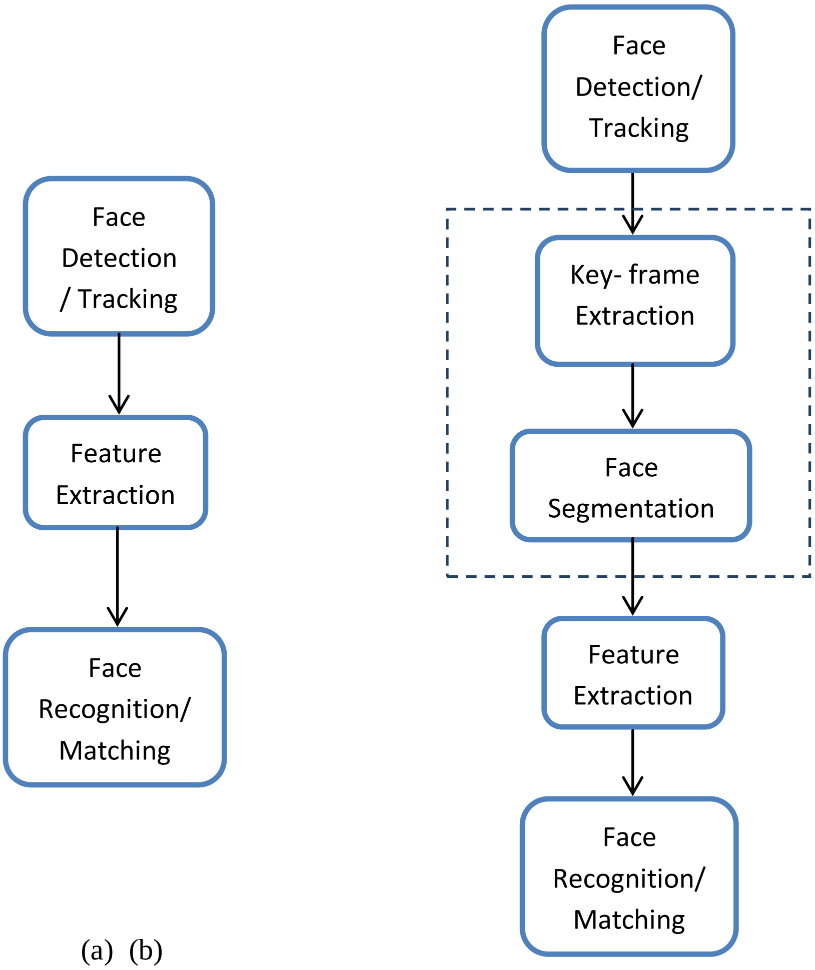
The rest of the paper is organized as follows: Section 2 describes the overall architecture of the proposed method. Section 3 elaborates the workflow of the proposed method followed by experimental analysis in Section 4. Finally, this research is concluded in Section 5.

2. System Architecture

The overall architecture of the general Video in Face Recognition (ViFR) system and the proposed system is shown in Fig. 1. The conventional ViFR consists of 3 phases: Face Detection/Tracking, Feature Extraction and Face matching/ Recognition. The proposed method includes two more phases: Keyframe Extraction and Face Segmentation. These two phases reduces the overhead of the Face recognition system.

Face detection segments the facial regions from the background.In the case of video, the detected facets mayneed to be tracked using a *face tracking* method. After facedetection, *featureextraction*is used to extract featuresthatare useful for differentiating faces and other geometrical features. In*face matching*, the extracted feature vector ofthe input face is matched with enrolled faces inthe database. The outputs of this phase are the identity of the face when a matchis found or indicates an unknownface otherwise.

In the *Keyframe Extraction* phase, some representative frames are chosen from the video sequence. This phase removes the poor quality frames or frames with unconditional lightings. In the *Face Segmentation* phase, only the face is cropped or segmented from the background. This phase reduces the extraneous time required for extracting features in the background.



(a) (b)

Fig. 1 (a) Conventional Face Recognition System (b) Proposed Face Recognition System

The algorithm of the proposed method is given in Algorithm 1. The inputs of the face recognition algorithm are the video dataset and the query image for which the face should be identified. The output is the Face ID which is declared in the video dataset which is matched with the query image. Initially the faces are tracked to identify only faces and leave non-faces using face tracking method.

Algorithm 1: Face Recognition Algorithm

```
Input: Video Dataset $\mathcal{H}$, Query face $\boldsymbol{\Omega}$
Output: Face ID
    Steps:
    1.  For each Video V in $\mathcal{H}$
            1.1 Track faces using Face Tracking Method
            1.2 Partition V into frames $f_1, f_2, \dots ,f_n$
            1.3 Select Keyframes $k_1,k_2,\dots k_m$ using SCS Algorithm where m<n
            1.4  Crop faces fromkeyframes using Background Subtraction method.
            1.5 For each keyframes $k_1$ to $k_m$
                  1.5.1 Extract Features using CNN
1.6  End
    2. End
    3. Generate $\mathcal{G}$ as feature vector set.
    4. Extract Feature for $\boldsymbol{\Omega}$
    5. Match $\boldsymbol{\Gamma}$ in $\mathcal{G}$ using ANN.
    6. Generate Face ID
    7. End
```

3. Proposed Method

In this section, the methods used in all the phases are described. The first phase uses Face Tracking Method to identify faces and non-faces in the video. In the second phase, the keyframes are extracted using Scene Change Segmentation (SCS) algorithm. For cropping only face, Background Subtraction Method is used. Finally, ANN is used for face detection.

Initially, the video sequence is divided into frames. The face and non-face are identified and faces are tracked using Face Tracking system. Then Keyframes are identified by neglecting poor quality frames. From the keyframes, faces alone are cropped using background subtraction method. Deep features are extracted from the cropped face using DCNN which is then given to ANN to classify the faces. All the identified faces are given Face ID which is used for query image classification. When the query image is given, it is matched with the already identified faces and Face ID is given as output. The steps involved in the proposed method is shown in Fig. 2.

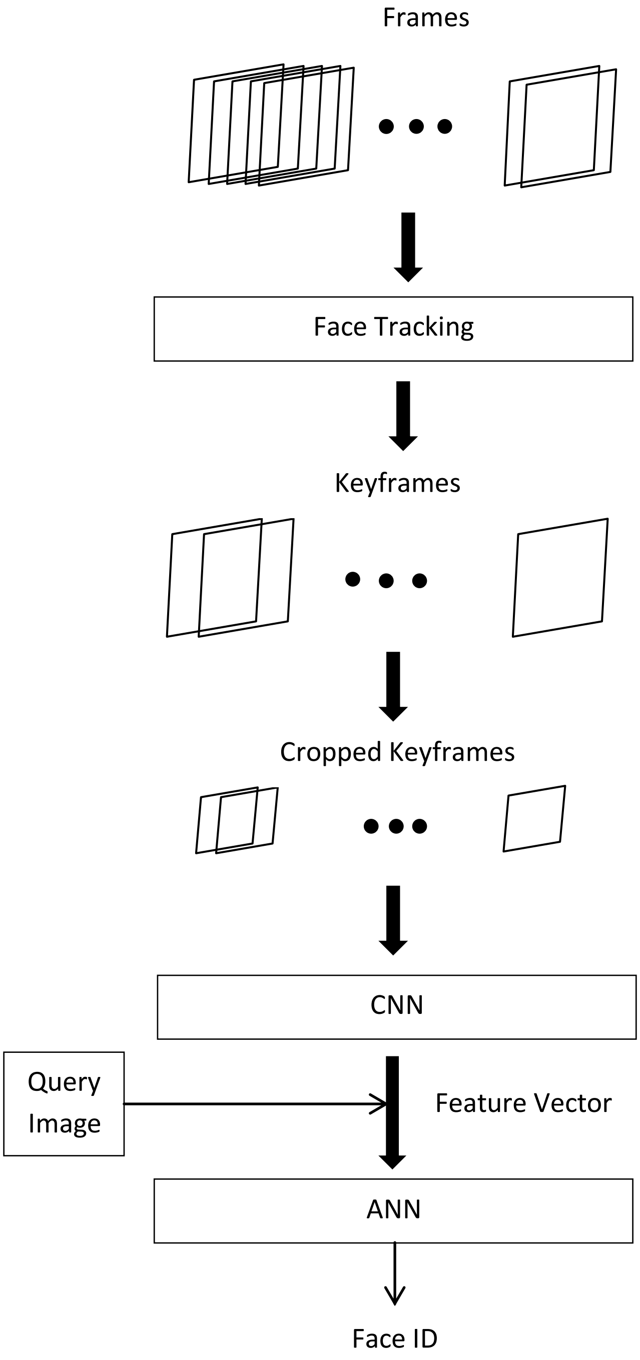


Fig. 2 Steps in Proposed Method

For extracting keyframes, the most popular similarity measure such as Pearson Correlation Coefficient (PCC) is used [11]. The value of PCC can fall between 0(no correlation) and 1(perfect correlation). In this research, PCC less than 0.80 are considered as next scene. The PCCis expressed as follows.

$$PCC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(f(i,j)-f^{m})(f_p(i,j)-f_p^{m})}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(f(i,j)-f^{m})^2(f_p(i,j)-f_p^{m})^2}} \quad (1)$$

InBackground Subtraction, for each pixel, a set of values taken in the past at the same location or in the neighborhood is stored. It then compares this set to the current pixel to decide if that pixel belongs to the background, and adapts the model by choosing the values from the background to be replaced. This methodvaries from others based on the common assumption that the oldest values should be replaced first. Finally, if the pixel is found to be part of the background, its value is propagated into the background model of a neighboring pixel [12].

In [13], SOM is combinedwith deeply learnedfeaturesofCNNtoenhance coding stability. Benefiting from CNN's deep architectureand supervised learning approach [14], CNN's caneffectively manage with large volumes of data and produceahierarchical and discriminative representation of features. Theuse of deeply learned features makes the learned ordinalfeatures not only include the prior data structure butalso the hierarchical structure of local image patches.Alex4's CNN network is used as a deep-architecture. This CNN first feeds two convolution layers of gray scale images, each followed by a normalization layer and a max-pooling layer.Then, two locallyconnected layers are connected to the output of the secondmax-pooling layer, and finally to a C-way soft-max regressionlayer (C is the number of classes) which generates a distributionover class labels. The inputs to this network are images of the cropped gray-scale face without pre-processing.The last soft-max regression layer on the C-way provides supervised information on how to learn face representations. The outputs of the last locally connected layers are used as deep feature representations.

For*face matching*, a model which combines many artificialneural networks is used. (ANNs) [15]. ANNs have been widely implemented over the last two decades to address problems over signal processing[16]. Researchers have suggested a variety of different artificial neural network models. One challenge is to find the most appropriate model of neural network that can work efficiently to solve practical problems. ANN is the term for the method of solving problems by simulating the activities of the neurons.In depth, ANNs can be best described as "computational models" with unique properties such as the ability to adapt or learn, generalize, or cluster or organize data, and which operation is based on parallel processing[15]. Many of the above-mentioned properties can however be attributed to non-neural models.The selected neural network here is three-layer feed-forwardneural network with back propagation algorithm.The number of input neurons T is equal to the length ofextracted feature vector, and the number of output neurons. Rowley's ANN model [15] is used in this research for detecting faces which is presented in Table 1.

Table 1 The ANN structure for detecting faces.

Name	Input Nodes	Hidden Nodes	Output Nodes	Learning Rate
ANN_FACE	25	25	1	0.3

The system is implemented by the three-layer feed-forwardANN with the Tanh activation function [18] and theback-propagation learning algorithm [16]. The Tanh activation function is given as

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}}, f(x) \in [-1, 1] \qquad (2)$$

4. Experimental Results

The efficiency of the proposed method is evaluated on the YouTube Faces (YTF) dataset [17]. Each YTF dataset identifier contains many sequences of videos. Unless the faces for each frame are identified separately, the features may not be compatible. Keyframes are also derived from the video sequence and features are only extracted from them.Accuracy, sensitivity and specificity are the metrics used to analyse the performance of the proposed method.

Accuracy is the most important factor of any recognition system. Sensitivity rate is the rate of correctly classified faces to the sum of faces that are correctly and incorrectly classified. Specificity is the rate of faces that are accurately classified as negative to the sum of faces that are correctly classified as negative and the images that are wrongly classified as positive. The formulae for computing accuracy, sensitivity and specificity rates are given below.

$$acc_r = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100 \qquad (3)$$

$$Sen_r = \frac{T_p}{T_p + F_n} \times 100 \ (4)$$

$$Sp_r = \frac{T_n}{F_p + T_n} \times 100 \ \ (5)$$

whereacc_r is the accuracy rate, Sen_r is the sensitivity rate and Sp_r is the specificity rate. $T_p, T_n, F_p, F_n$are the True Positive, True Negative, False Positive and False Negative rates. Table 2 shows the experimental results of the proposed method.

Table 2 Results of the Proposed Method

Measure	Value (%)
Sensitivity	98.2
Specificity	96.5
Accuracy	98.36

The proposed method is compared with methods such asWebFace [19], DeepFace [20], Lookup- based CNN (LCNN) [21], VGG [22], Cloud Based Face Recognition (CBFR) [23], DeepID2+ [24], FaceNet [25], CASIA WebFace [26], SphereFace [27], CosFace [28], SeqFace [29], Iqbal et al. [30] method, Hasnat et al. [31] method and Wen et al. [32] method. Table 3 shows the comparison of accuracy of the proposed method with the recent methods. Fig. 3 shows the bar chart of the same comparison.

Table 3 Accuracy Comparison of Proposed Method with Recent Methods

Method	Accuracy (%)
WebFace(2014) [19]	90.6
DeepFace (2014) [20]	91.4
LCNN (2015) [21]	91.6
VGG (2015) [22]	92.8
CBFR(2018) [23]	93.0
DeepID2+ (2015) [24]	93.2
FaceNet(2015) [25]	95.1
CASIA WebFace (2014) [26]	90.60
SphereFace (2017) [27]	95
CosFace (2018) [28]	97.6
SeqFace (2018) [29]	98.12
Iqbalet al. (2019) [30]	96.4
Hasnat et al. (2017) [31]	96.24
Wen et al. (2016) [32]	94.9
Proposed Method	98.36

From the Table 3, it is observed that SeqFace [29] method achieves a maximum accuracy of 98.12% which is higher than all other recent methods. The proposed method achieves 98.36% accuracy which overwhelm seqFace method.

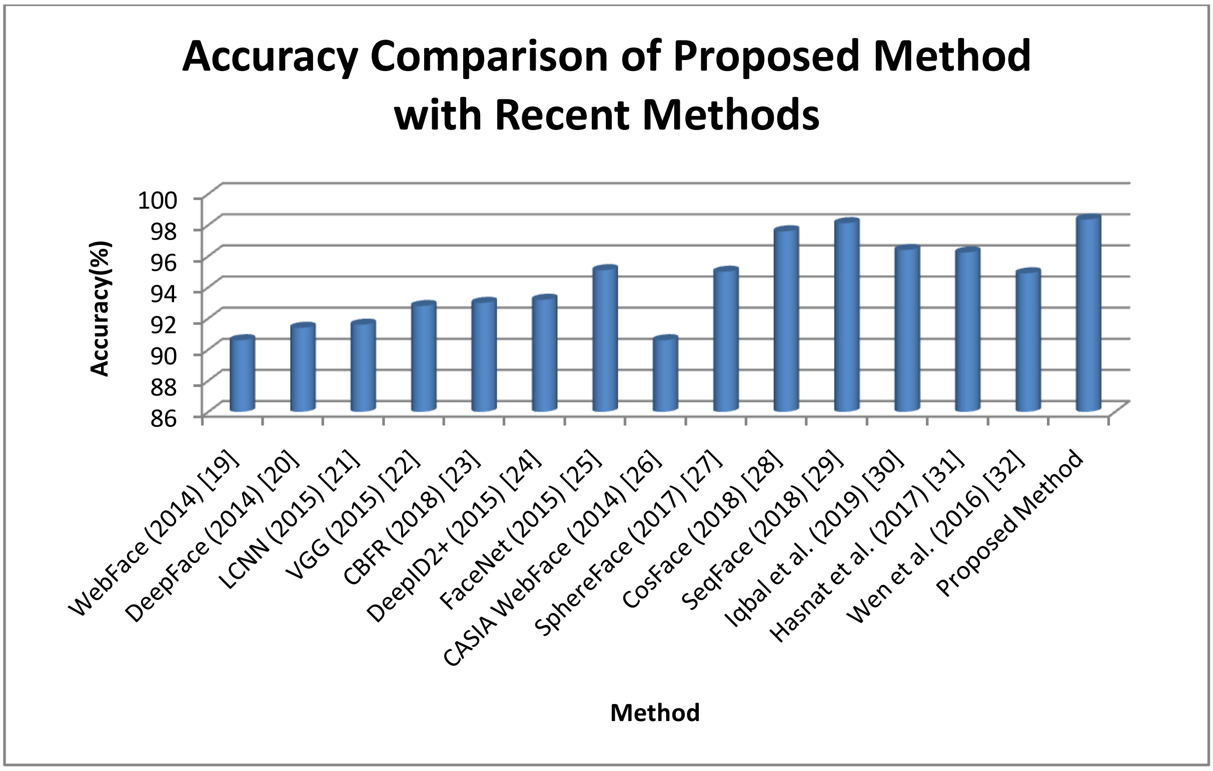


Fig. 3 Bar Chart Comparison of the proposed method with recent method

5. Conclusion

Face in Video Recognition is vital in all security departments. The poor lightning conditions and the poor quality frames makes the system worse in terms of computation time. In this paper, two phases are introduced to reduce the computational overhead. The first phase is the Keyframe Extraction phase which reduces the number of frames. Another one is Face Segmentation which removes the extra background for recognition. The proposed method is tested on YTF dataset and the results are compared with other recent methods. The proposed method achieves a accuracy of 98.36% which is higher than other methods.

References:

[1] Aulestia P.S., Talahua J.S., Andaluz V.H., Benalcázar M.E. (2017) Real-Time Face Detection Using Artificial Neural Networks. In: Lintas A., Rovetta S., Verschure P., Villa A. (eds) Artificial Neural Networks and Machine Learning, ICANN 2017, Lecture Notes in Computer Science, vol 10614. Springer, Cham

[2] WaelAbdAlmageeda,YueWua, Stephen Rawlsa,ShaiHarelc Tal Hassnera, IacopoMasibJongmooChoibJatuporn Toy LeksutbJungyeon Kim PremNatarajana, Ram Nevatiab Gerard Medionib," Face Recognition Using Deep Multi-Pose Representations", 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)

[3] Tsung-Han Chan, KuiJia, ShenghuaGao, Jiwen Lu, ZinanZeng, and Yi Ma, "PCANet: A Simple Deep Learning Baseline forImage Classification", IEEE Transactions on Image Processing, Volume: 24 , Issue: 12 , Dec. 2015, pp. 5017 – 5032

[4] Changxing Ding, Dacheng Tao, "Robust Face Recognition via Multimodal DeepFace Representation", IEEE TRANSACTIONS ON MULTIMEDIA, 2015

[5] ShaoqingRen, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processingsystems, pages 91–99, 2015.

[6] Xudong Sun, Pengcheng Wu, Steven C.H. Hoi, "Face Detection using Deep Learning: An Improved Faster RCNN Approach"

[7] Ran Hea, TieniuTana, Larry Davisb, and ZhenanSuna," Learning Structured Ordinal Measures for Video based Face Recognition"

[8] Jun-Cheng Chen, Rajeev Ranjan, Swami Sankaranarayanan,AmitKumar,Ching-Hui Chen, Vishal M. Patel, Carlos D. Castillo, Rama Chellappa, "Unconstrained Still/Video-Based Face Veri_cation with DeepConvolutional Neural Networks"

[9] JingxiaoZheng, Rajeev Ranjan, Ching-Hui Chen, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, "An Automatic System for UnconstrainedVideo-Based Face Recognition"

[10] Thai Hoang Le, "Applying Artificial Neural Networks for Face Recognition", Hindawi Publishing Corporation Advances in Artificial Neural Systems Volume 2011.

[11] Sowmyayani, S, ArockiaJansi Rani, P., 'Adaptive GOP structure to H.264/AVC based on Scene change', ICTACT journal on image and video processing: special issue on video processing for multimedia systems, 2014, 5, (1), pp. 868-872,.

[12] Olivier Barnich, Marc Van Droogenbroeck, "ViBe: A Universal Background Subtraction Algorithm for Video Sequences",IEEE Transactions on Image Processing, Vol. 20, Iss. 6, June 2011, pp. 1709 – 1724.

[13] Y. L. Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Hand-written digit recognition with a back-propagation network. In *NIPS*, 1990.

[14] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[15] H. A. Rowley, *Neural Network Based Face Detection*, Neural network Based Face Detection, School of Computer Science, Computer Science Department, Carnegie Mellon University , Pittsburgh, Pa, USA, 1999.

[16] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 2006.

[17] Lior Wolf, Tal Hassner, and ItayMaoz, "Face recognition in unconstrained videos with matched background similarity," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011,pp. 529–534.

[18] T. Sawangsri, V. Patanavijit, and S. Jitapunkul, "Segmentation using novel skin-color map and morphological technique," in Proceedings of the World Academy of Science, Engineering and Technology, vol. 2, January 2005.

[19] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," arXiv preprint arXiv:1411.7923, 2014.

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

[21] Xiang Wu, Ran He, and Zhenan Sun, "A lightened cnn for deep face representation," arXiv preprint arXiv:1511.02683, 2015.

[22] M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in Proceedings of the British Machine Vision Conference (BMVC), 2015.

[23] Heng-Wei Hsu, Tung-Yu Wu, Wing Hung Wong and Chen-Yi Lee, "Correlation-Based Face Detection For Recognizing Faces In Videos", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018

[24] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2892– 2900.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch", 2014.

[27] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., And Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In CVPR, IEEE Computer Society, 6738–6746.

[28] Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., And Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition.

[29] Wei Hu, Yangyu Huang, Fan Zhang, Ruirui Li, Wei Li and Guodong Yuan, "SeqFace: Make full use of sequence information for face recognition", 2018

[30] MansoorIqbalM.,ShujahIslamSameem, NuzhatNaqvi, ShamsaKanwal, ZhongfuYe, "A deep learning approach for face recognition based on angularly discriminative features", Pattern Recognition Letters, Volume 128, 1 December 2019, Pages 414-419.

[31] AbulHasnat, JulienBohne, Jonathan Milgram, StephaneGentric and Liming Chen," DeepVisage: Making face recognition simple yet with powerful generalization skills", 2017.

[32] Wen, Y., Zhang, K., Li, Z., AndQiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In European Conference on Computer Vision, 499–515