Survey on Face Recognition Methods

Shriya Kak[1] and Dr. Alain Pagani[2]

[1] shriyakak@gmail.com
[2] Alain.Pagani@dfki.de

Abstract. The last decade witnessed an increase in the acceptance of bio-metrics for authentication of virtual and physical platforms. The underlying concept for this wide scale acceptance is the unique biological attributes of each individual [8]. Among all the physiological and behavioral attributes that define bio-metrics in humans, Face recognition edges out [9]. Besides being independent on voluntary action for authentication, face recognition also enables easy acquisition of data with inexpensive apparatus such as fixed cameras. With the advent of Deep learning, the research landscape of face recognition is reshaped primarily because of the breakthroughs of Deep face method. Deep FR techniques which leverages hierarchical architecture has dramatically improved the performance of existing systems and fostered successful real-world applications [19]. In this paper, we provide a detailed review of recent development and challenges faced in face recognition methods which will cover the architecture, the loss function proposed and comparing them on commonly used databases for training with its complexity and accuracy. Also, some of the measures which are taken to detect original faces in Deep fake videos.

Keywords: Triplet loss, variable group convolution, embeddings, 3D Face modeling, anti-spoofing

1 Introduction

Face recognition (FR) has been a topic of interest for scientific research because of its numerous potential applications in many fields including social media, surveillance, security etc. For instance, many FR systems have been enabled to reduce human efforts and enhance security at border controls, in companies, ATMs etc. Previously, from the year 1990 - 2000, the FR system was based on holistic learning which includes the Eigenface approach, Fisherface, Bayes etc. However, these approaches failed to address the uncontrolled facial changes which gave rise to local feature-based FR in 2010. With some invariant properties of local filtering, they achieved robust performance. Unfortunately, handcrafted feature faced lack of distinctiveness and compactness [19]. In general, these traditional methods attempted to recognise face to great extent but were not capable of recognizing a face in the unconstrained environment due to lack of invariance to pose, illumination, ageing, expression, occlusion, low-resolution images. But all that changed with the emergence of deep learning methods such as convolution neural networks, which uses multiple layers for feature extraction and transformation. This level of hierarchy has proven a strong invariance toward different face poses, illumination or disguise. [19]. Deep learning has proven a boon to many computer vision tasks. There have been several scientific research on FR methods, here we are going to discuss the concepts which are based on deep learning methods and have achieved significant performance from last 4-5 years. We have tried to cover some of the basic yet important concepts of face recognition method which can be further enhanced by some modification. We have broken down this survey in many parts that follows: section 1) some basics concepts, section 2) the widely used benchmark dataset, section 3) Network Architecture, 4) some shortcomings of deep learning techniques and in last the experimental results.

2 Background concepts and Terminology:

The conventional pipeline for face recognition consists of three stages: 1) Face detection 2)Face alignment 3) Face recognition. Within the initial stage, the face detector is employed to localize faces in images or videos. Secondly, face landmarks are extracted and then the FR module is implemented with these aligned faces. Furthermore, the primary tasks of the FR system are Face verification and Face identification [14]. Face verification includes one to one matching of face images, for example, when given two face images it will ascertain whether it is same or not. In Face Identification there is one to many matching of a face image. E.g. Given an image of an unknown individual determining the person's identity by comparing the image with the database of known images. Face clustering is also one of the important tasks of face recognition system which groups the collection of face images when no external labels are associated with the image. It is efficient for large scale face retrieval.

3 Problems and Challenges

Early Face Recognition systems were modelled on controlled and small scale dataset which were not capable of handling real-world situations. The FR technology has achieved tremendous success in facial authentication domain. However, face recognition in many scenarios is challenging due to severe blur, richer pose, dramatic occlusion and illuminations [10]. Ohlyan et al. have broadly categorized the source cause of the variation in facial appearance in two groups: Intrinsic factors and Extrinsic factors [14].

A) Intrinsic Factors: are due to the physical nature of the face and it is independent of the observer. E.g facial expression, ageing etc.

B) Extrinsic factors: are due to different conditions in which images are taken. Eg. Illumination, pose, resolution, scale, noise etc.

Another challenge which has attracted probing in FR domain is detecting real face among fake ones. In particular, four types of facial manipulation are reviewed in this study: i) entire face synthesis ii) face identity swap iii) facial attributes manipulation and iv) facial expression manipulation [18].

4 Benchmark Dataset

Since previously the FR system was based on the constrained environment there was a need for a dataset which could have unconstrained images. In 2007, LFW dataset was introduced which marked the beginning of FR under unconstrained conditions. **Labeled Faces in the Wild (LFW)**: is provided as an aid in studying the unconstrained, recognition problem. The database contains labelled face photographs spanning the range of conditions typically encountered in everyday life. The database exhibits “natural” variability in factors such as pose, lighting, race, accessories, occlusions, and background [7]. The data set contains more than 13,000 images of faces collected from the web. Each face has been labelled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set [7].

Social Face classification dataset (SFC): is well known Facebook's private dataset which was used to train Taigman's, et al. DeepFace model. The dataset includes 4.4 million labelled faces from 4,030 people each with 800 to 1200 faces. There is a large number of images per person which provides a unique way to learn invariance against the current challenges in Face Recognition [17].

YouTube Face Dataset (YTF): All the model which are trained for face recognition also being tested on various video dataset. It is the most widely used dataset for FR in videos, designed for studying the unconstrained condition in videos. It includes 3,425 videos of 1,595 different people.

An average of 2.15 videos is available for each subject. The shortest clip duration is 48 frames, the longest clip is 6,070 frames, and the average length of a video clip is 181.3 frames [20].

Face Anti-spoofing Database [13]: Due to manipulation of faces and this increase in face synthesis, variety of anti-spoof databases are created. Most widely used anti-spoof databases are CASIA-FASD [23], Replay-Attack [4], OULU-NPU [3], and SiW [12]. for print/replay attacks, and HKBU-MARs [11] for 3D mask attacks. Early databases such as CASIA-FASD and Replay-Attack have limited subject variety, pose/expression/lighting variations, and video resolutions. Recent databases improve those aspects, and also set up diverse evaluation protocols. However, up to now, all databases focus on either print/replay attacks or 3D mask.

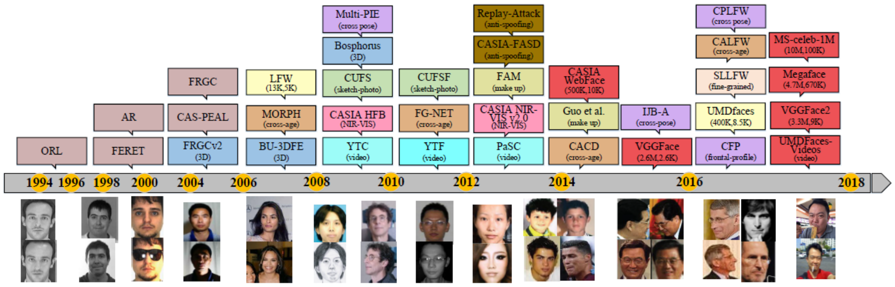


Fig. 1. The evolution of FR dataset from 2007 till 2019[19]

5 Network Architecture and training loss

Every human face has a unique feature but all of them share a similar shape and texture through which representation is learnt. The straight and easiest way is to have as many as ID included in the training set. Taigman, et al.[17] and Schroff, et al. [15] as mentioned in their paper and have trained their FR system by $10^6 - 10^7$ IDs [19] which have almost closed the majority of the remaining gap to human-level performance. Unfortunately, these datasets are not publicly available or accessible for academic use. So, Instead, the effort has been made in developing deeper architecture and adopt effective loss function which is more discriminative and compact. In this section, we have discussed the network architecture and method adopted by Google and Facebook researchers.

Taigman's, et al. DeepFace method has incorporated 3D face modelling which is based on the face alignment process. It follows the pipeline, detect-align-represent-classify. Fiducial point detector is used to direct the alignment process. Since 2D face alignment is not variant to pose and would not work when there is plane rotation mostly in unconstrained conditions therefore, the 3D face alignment was extracted. The generic 3D shape model and 3D affine camera are used to warp the 2D aligned crop to the image plane of the 3D shape. 3D face modelling is proven to be the correct solution in the unconstrained environment. The paper describes a method called frontalization which uses these fiducial points to warp a detected facial crop to a 3D frontal mode. In a nutshell, this method created a 3D face model for the incoming image and then uses a series of affine transformations of the fiducial points to "frontalize" the image as shown in fig.2. The frontalization is achieved by a piece-wised affine transformation from source (2D) to target (3D) directed by the Delaunay triangulation derived from the 67 fiducial points.

In Fig.3, There is front-end of single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colours illustrate feature maps produced at each layer[17].

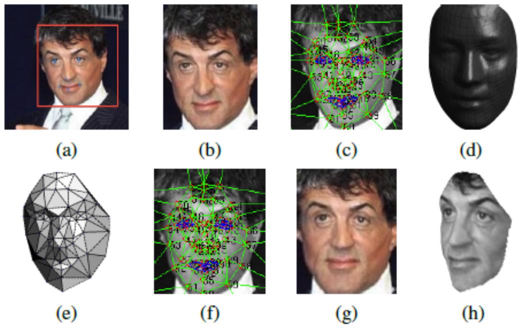


Fig. 2. The alignment pipeline from 2D to 3D shape model. (Fig description: (a) The detected face, with 6 initial fiducial points. (b) The induced 2D-aligned crop. (c) 67 fiducial points on the 2D-aligned crop with their corresponding Delaunay triangulation, we added triangles on the contour to avoid discontinuities. (d) The reference 3D shape transformed into the 2D-aligned crop image-plane. (e) Triangle visibility w.r.t. to the fitted 3D-2D camera; darker triangles are less visible. (f) The 67 fiducial points induced by the 3D model that is used to direct the piecewise affine warping. (g) The final frontalized crop. (h) A new view generated by the 3D model (not used in this paper[17]))

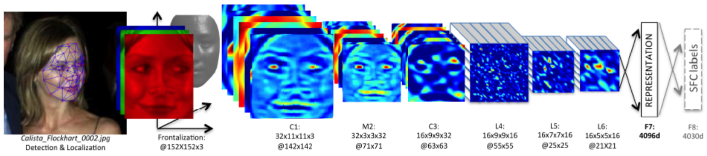


Fig. 3. The overall architecture of DeepFace[17]

The DNN is trained on the multi class face recognition task. A 3D-aligned 3-channels (RGB) face image of size 152 by 152 pixels is given to a convolutional layer (C1) with 32 filters of size 11x11x3. The resulting 32 feature maps are then fed to a max-pooling layer (M2) which takes the max over 3x3 spatial neighbourhoods with a stride of 2, separately for each channel [17]. This is followed by another convolution layer(C3) that has 16 filters of size 9*9*16. The overall network includes more than 120 million parameters, where more than 95% come from the local and fully connected layers. The goal of training is to maximize the probability of the correct face id. For that, they are using cross entropy based on softmax loss by which DeepFace has achieved an accuracy of 97.3%. It was also tested on LFW dataset, without 3D face alignment accuracy was decreased to 94.3% and without alignment, at all, it decreased to 87.9%. The model is generalized as it also worked on YTF dataset achieving 92.5% of accuracy[17].

Schroff's, et al. FaceNet introduced a new approach that directly learns mapping from face images to a compact Euclidean space where distance directly correspond to a measure of face similarity.

Basically, it provides embedding which is a feature vector for face recognition that are mapped to euclidean space. It is a technique to reduce the dimension of the input data and it is very useful in image data since it corresponds to huge dimension data (e.g 20 megapixel camera picture with 3 RGB layers means 60 millions of integers as the total info stored in the image). Once this embedding is done it becomes easy to calculate the distance between the image in euclidean space.FaceNet uses deep convolutional neural network [22] [16]. Data is passed into the batches to the deep CNN which is followed by L2 normalization and producing embedding nothing but feature maps.

The unique part about FaceNet method is its learning objective which is based on euclidean distance loss called Triplet Loss. It is a combination of three parameters called Anchor, Positive,

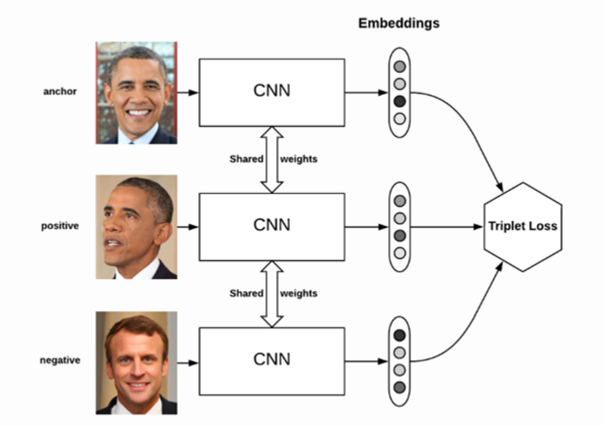


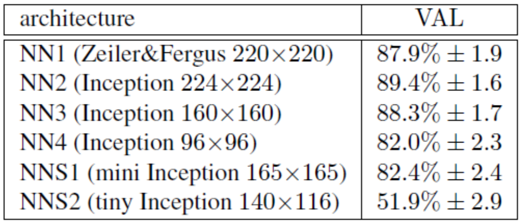
Fig. 4. Representation of triplet loss on two positive faces (Obama) and one negative face (Macron)[1]

Negative. As the terminology of triplet loss, if there is an Anchor image then the distance with the Anchor and positive (i.e the same person) should be smaller and the distance between the anchor and the negative(i.e the different person) should be farther apart. That is Why it is called a Triplet loss because you will always be looking at the three images at a time [2]. Triplet loss can be visualised as below in fig. 5. On widely used LFW dataset, FaceNet has achieved accuracy of 99.63% and 95.12% on YTF dataset. With this architecture, FaceNet has 140M parameters and requires 1.6 billion FLOPS. Later, different architecture shown in fig.6 was tried to reduce the computation resources, performance drop was noticed and it came to this conclusion that there is trade-off between computation and accuracy depicted in fig.7 which means if the complexity of architecture increases it will lead to accuracy drop. This idea can be further used to reduce the computation resource without any performance drop so that it can easily run on mobile phones. As the need of lightweight face recognition network is increasing with limited computational cost system such as mobile phones. Many novel concepts were proposed to reduce the consumption of computational resources and were also successful in achieving good performance on various applications, however, optimization problems on embedded system still remained. To handle this conflict Yan et, al.[21] proposes a variable group convolution which can efficiently solve the unbalance computational intensity inside the block. This model has achieved stat-of-the-art performance on LFR challenge [6] which requires network whose FLOPs is under 1G and memory footprints under 20M. The principle of VarGFaceNet is based on 2 things: 1) Small intermediate feature map between the blocks 2) Balanced computational intensity inside a block The architecture is designed to improve the discriminative ability and generalization ability of the lightweight models using two blocks named Head setting and Embedding setting. In order to reserve the discriminative ability in lightweight networks, 3X3 conv with stride 1 at the start of the network and output feature size will remain the same as input. To maintain the generalization ability Angular distillation loss was employed. This loss was proposed to make learned feature separable with a larger angular/cosine distance.

$$\mathcal{L}(A,P,N) = \max\left(\|\mathrm{f}(A) - \mathrm{f}(P)\|^2 - \|\mathrm{f}(A) - \mathrm{f}(N)\|^2 + \alpha, 0\right)$$

Fig. 5. Triplet Loss [15]

With the emerging deep learning techniques, the FR techniques has achieved tremendous results. But some of the techniques, GAN in particular, with free access to large scale public databases have



architecture	VAL
NN1 (Zeiler&Fergus 220×220)	87.9% ± 1.9
NN2 (Inception 224×224)	89.4% ± 1.6
NN3 (Inception 160×160)	88.3% ± 1.7
NN4 (Inception 96×96)	82.0% ± 2.3
NNS1 (mini Inception 165×165)	82.4% ± 2.4
NNS2 (tiny Inception 140×116)	51.9% ± 2.9

Fig. 6. FaceNet model was tested on different architecture to increase the accuracy but with the increase in complexity of architecture, performance drop was experienced.[15]

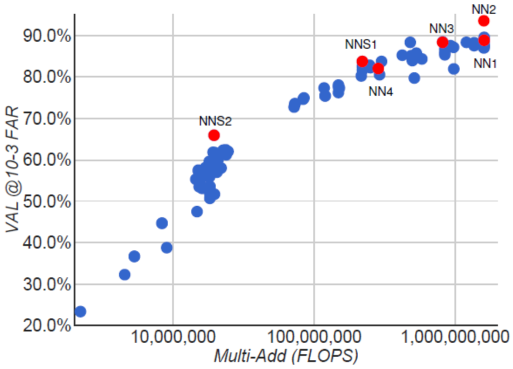


Fig. 7. FLOPS vs Accuracy trade-off [15]

led to the generation of some fake content. [18]. To overcome this issue, many methods were adopted but in this paper only state-of-the-art image spoofing detection method is summarized. Liu et al. defined the detection of unknown spoof attacks as Zero Short Face Anit-spoofing(ZSFA). They introduced a novel method called Deep Tree Network (DTN) to partition the spoof samples into semantic sub-groups in an unsupervised fashion. Previous methods were able to detect one or two types of spoof attacks but in the state-of-the-method they investigate wide range of 13 types of spoof attacks including print, replay, 3D mask and so on as shown in Fig.8.

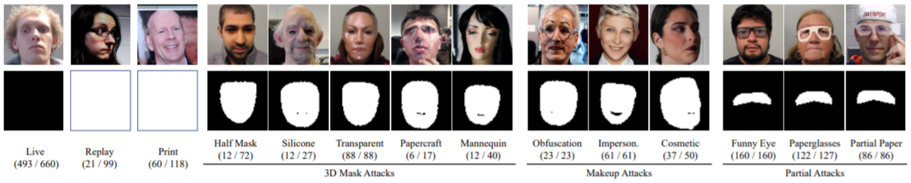


Fig. 8. The examples of the live faces and 13 types of spoof attacks. The second row shows the ground truth masks for the pixel-wise supervision Dk. For (m, n) in the third row, m/n denotes the number of subjects/videos for each type of data [13]

The basic idea is when a data sample arrives, being known or unknown attacks, DTN routes it to the most similar spoof cluster e.g in Fig.9 and makes the binary decision. The tree structured network has proven to be helpful in tackling language-related tasks such as parsing and translation[13]. It has a property for learning features hierarchically. The network learns the semantic embedding for known spoof attacks unsupervised. The partition of data naturally associates semantic attributes with sub-groups. During the testing, the unknown attacks are projected to the embedding to find the closest attribute for spoof detection.

6 Experiments

The experimental results shown in Fig 10 are based on testing of FR method on LFW dataset and it is quite clear that employing angular losses outperform the state-of-the-art method. This table aggregates results as reported in the various papers. One of the main problems of triplet loss is that

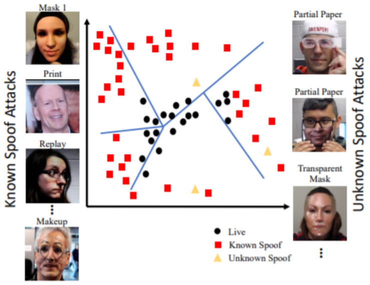
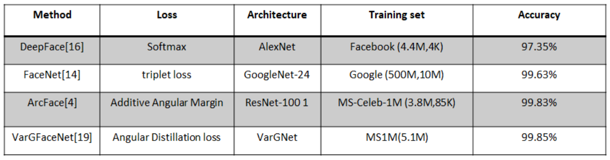


Fig. 9. The network is trained on 12 out of 13 types of spoof attack and then testing it on remaining one which is never seen by network and hence considered as unknown spoof type. [13]



Method	Loss	Architecture	Training set	Accuracy
DeepFace[16]	Softmax	AlexNet	Facebook (4.4M,4K)	97.35%
FaceNet[14]	triplet loss	GoogleNet-24	Google (500M,10M)	99.63%
ArcFace[4]	Additive Angular Margin	ResNet-100 1	MS-Celeb-1M (3.8M,85K)	99.83%
VarGFaceNet[19]	Angular Distillation loss	VarGNet	MS1M(5.1M)	99.85%

Fig. 10. The Accuracy of different methods on LFW Dataset

it's very expensive. There is a combinational explosion with large dataset as large number of images leads to exponential number of pairings. Additionally, Triplet loss requires semi-hard sampling in order to learn effectively. Furthermore, softmax was not sufficient by itself to learn feature with a large margin and to further increase the accuracy of the FR system. In the paper [19], the Euclidean-distance-based loss, angular/cosine[5] loss explicitly adds discriminative constraints but Wang et al. showed that angular/cosine-margin based loss, is vulnerable to noise and becomes worse in the high-noise region [19]. The effectiveness of state-of-the-art [21] approach is well evaluated with the limitation of 1G FLOPS in champion of Deepglint light track of LFR (2019)[6].

7 Conclusion

In this paper, we have analysed different face recognition techniques that use deep learning approach and their shortcomings. According to experimental results mentioned in section 6, it is quite clear that Yan's method [21] outperforms the state-of-the-art approach with 99.85% accuracy with the limitation of 1G FLOPS and parameters less than 20M. Moreover, the effectiveness of this method is also published in Light Face Recognition challenge in 2019 which requires networks whose FLOPs under 1G and memory footprints under 20M which makes this method computationally efficient. Therefore, with the reduced complexity of the network, installation of FR system on remote devices becomes not only light and easy but also makes it ubiquitous across applications. As explained above,the antispoof detection method, although is not directly comparable because of different evaluation metrics (dataset) but it has achieved 95.9% accuracy when tested on Replay, CASIA and MSU-MFSD dataset[13]. Deep FR has dramatically improved the stat-of-the-art performance by taking advantage of big annotated data, deep learning methods and GPUs. In conclusion, probing the practical and commercial applications of the FR systems has resulted in falsifying assumptions and reaching to the core of real world issues. The most critical issue that has impacted the security of the FR systems is the Adversarial attack[19]. The root cause of adversarial attack is still unclear and this area can be used for further investigation.

References

1. https://omoindrot.github.io/triplet-lossa-better-implementation-with-online-triplet-mining. N.D.
2. https://towardsdatascience.com/image-similarity-using-triplet-loss-3744c0f67973. N.D.
3. Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 612–618. IEEE, 2017.
4. Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.
5. Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
6. Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
7. Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.
8. Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *Jips*, 5(2):41–68, 2009.
9. Anil K Jain. Biometric recognition: how do i know who you are? In *International Conference on Image Analysis and Processing*, pages 19–26. Springer, 2005.
10. Meng Zhang1 Rujie Liu, Hajime Nada, Hidetsugu Uchida, Tomoaki Matsunami, and Narishige Abe. A pairwise learning strategy for video-based face recognition.
11. Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 100–106, 2016.
12. Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018.
13. Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019.
14. Sonia Ohlyan, Sunita Sangwan, and T Ahuja. A survey on various problems & challenges in face recognition. In *International Journal of Engineering Research & Technology (IJERT)*, volume 2, 2013.
15. Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
16. C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, and A Rabinovich. Going deeper with convolutions, corr abs/1409.4842. *URL http://arxiv. org/abs/1409.4842*, 2014.
17. Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.
18. Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.
19. Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.
20. Lior Wolf, Tal Hassner, and Itay Maoz. *Face recognition in unconstrained videos with matched background similarity.* IEEE, 2011.
21. Mengjia Yan, Mengao Zhao, Zining Xu, Qian Zhang, Guoli Wang, and Zhizhong Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
22. Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
23. Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012.