Survey on the compatibility of Face Detection and Recognition Algorithms after a Pandemic

Rajkumar Yadav
Computer Science and Engineering Technology
MIT Academy of Engineering
Pune, India
rkyadav2710@gmail.com

Kunal Sakhare
Computer Science and Engineering Technology
MIT Academy of Engineering
Pune, India
kunalsakhare0313@gmail.com

Shantanu Choubey
Computer Science and Engineering Technology
MIT Academy of Engineering
Pune, India
santanuchoubey@gmail.com

Rutik Pol
Computer Science and Engineering Technology
MIT Academy of Engineering
Pune, India
rppol@mitaoe.ac.in

Gaurav Mahato
Computer Science and Engineering Technology
MIT Academy of Engineering
Pune, India
rautrocky25@gmail.com

Anil Kumar Gupta
Centre for Development of Advanced Computing
Senior Member IEEE
anilkgupta@ieee.org

***Abstract* – Face recognition has attracted many researchers knowing the vast amount of services it can provide. Many computer vision algorithms have been worked upon for extracting unique features from facial images to help differentiate between those unique 2D and 3D(among the most recent once) image representations. The research work ranges from hardcoded elements to modern machine learning methodology and deep learning-based feature engineering and extraction, due to the pandemic caused by covid-19 and safety reasons its almost obvious to see any concerned subject crossing by wearing a mask, which hugely affects the performance of such facial feature extracting models. This survey provides a review of different types of strategies used by various state-of-the-arts and the effects of the changing appearance of subjects in the pandemic and post-pandemic era on pre-pandemic algorithms. First, we summarize the algorithms based on the type of strategies used and describe their importance in performance change caused due to the changing appearance of faces in public places. Second, generalize the core problem in changing results for face detection and recognition algorithms which can be categorized into classes.**

Keywords—deep learning, feature engineering and extraction (keywords)

1. INTRODUCTION

Face recognition is becoming more and more critical in human-computer interaction. There have been several developments in the application and algorithms which are being used in the last few decades and with the dominance of deep learning in computer vision for studying facial features there have been many state-of-the-art models that have been built some of which can compete with humans capacity of identifying an individual. Amongst the most demanding face recognition applications are building robust and real-time security and surveillance systems that will help reduce human resources used by various agencies for such purposes. Face recognition is generally considered as a combination of face detection and recognition and face tracking in some cases.

Due to the pandemic caused by covid-19, it has become necessary for people moving in public places to cover their faces with masks and have less interaction with others. Wearing a mask while going in public places has become a routine which has added more to the challenges that are to be faced by facial recognition systems based on computer vision. Occlusion has become a significant concern for both detection and recognition purposes as the lower half of the Face, especially lips region has played an essential role in both types of feature extraction. The effects of occlusions on two kinds of models will be discussed in Section 2.2 and 3.3.

Face detection is mostly a more straightforward job than face recognition or face tracking throughout its appearance in continuous frame inputs. It has a set of problems to deal with before being qualified as a well-designed detection model some of which are occlusions, illumination, and scale which enforces most practitioners to use various classification models pre-trained on backbones like VGG16, ResNet50/152 [10, 11, 12], etc. that is used for transfer learning on training with facial datasets. Detection is very similar to the classification of objects into two classes, Face and background, generally with an additional regression part consisting of predicted bounding box parameters. With the problem of drastic change in pandemic requirements due to covid-19, the occlusion factor has become a significant concern among all. Face detection is mostly generalized in two types: a) single-stage detectors, that have a combined stage of classification and bounding box regression for achieving anchor-based and multi-scale detection simultaneously and b) two-stage sensors, with region proposal selection and regression. Most prominent of the features that affect the most after occlusion is the scale of the facial region. Various approaches have been proposed for scale-invariant output using an image pyramid and feature pyramid. Image pyramids require more inference time, and the feature pyramid has dominated the subject since its arrival, which uses sliding anchors on multi-scale feature maps. In recent years, pixel-wise face classification has made remarks, not much in computation cost but on accuracy and recall to a greater extent. Another most popular amongst all is the use of the landmark-based approach.

Facial recognition-based systems have dominated the market of surveillance and biometric systems. The in-depth study of the subject breaks it into two parts. First, designing a proper architecture to convert images into embedded mathematical structures conveys meaning from input images and gives a platform to compare images later. The output must be optimal and informative enough to help differentiate different subjects in images with ease. Second, use an

appropriate loss function to compare embedded facial features represented in a mathematical structure and identify the extent of similarity among those. An ideal face recognition model needs to be scale-invariant, age invariant, illumination invariant, expression invariant and many more based on the type of application of the target system. Most algorithms in the decade of 2000-2010 aimed at one of those unconstrained facial changes and only had 1-3 layers for feature extraction; none had ways for integrated feature engineering. There was a drastic change in the technological approach used for image-based recognition after the success of AlexNet, which introduced deep learning for image classification. More research on the subject helped achieve human-level performance and in most recent times even more. Face detection is a part of pre-processing required before applying face recognition for generating feature embeddings from input images. The general trend of research in face recognition and related field uses neural networks categorized into two types, backbone and assembled where backbone models like AleNet, GoogleNet, VGGNet, etc., are introduced with the state-of-the-art results. The Assembled Network is designed by combining some new extensions to the existing backbone architectures. In addition to these, we need an appropriate loss function and Face matching mechanism to do justice with the amount of feature engineering done on facial images by the proposed models. Different sets of loss functions developed so far include Euclidean-distance based loss like triplet loss and contrastive loss, angular/cosine-margin based loss, softmax loss, and variations amongst the significant contributors. An appropriate loss function is a significant key for training any architecture because it directs the Network's parameter adjustments. The concept of the extent of error or loss is defined with respect to the output of loss function used. For Face matching similar to the loss function, compares the deep features of two inputs which uses either methods like cosine distance and L1 or L2 distance, etc.

2. FACE DETECTION

Existing approach producing state-of-the-art results:

Detection systems developed till date has seen a lot of growth and improvements in results since the work of Viola-Jones[9]. Since then the work in features engineering has moved from developing handcrafted features to designing network architectures for automating them to learn those by use of relevant datasets. Some of those dataset sets used today include WiderFace which has a wide range of facial variations. RetinaFace[11] make use of multi-task loss function(1) which consists of linear combination of classification loss, face box regression loss, facial landmark regression loss and dense regression loss where the landmark represents the five facial landmark points and dense regression loss is calculated by pixel by pixel loss which contributes to increased recall but to an extent increase the computation cost. For any training anchor *I*, we represent the multi-task loss function as:

$$L = L_{cls}(p_i, p_i^*) + X_1 p_i^* L_{box}(t_i, t_i^*) + X_2 p_i^* L_{pts}(l_i, l_i^*) + X_3 p_i^* L_{pixel} \quad (1)$$

In equation (1), $L_{cls}(p_i, p_i^*)$ is for classification loss with p_i as predicted probability of i^{th} sample being a face and p_i^* Is 1 for the sample being a face and 0 for not being a face.

Similarly, for face box regression loss $L_{box}(t_i, t_i^*)$ predicted set of bounding box values is t_i and ground-truth is t_i^*. And l_i represents a set of predicted coordinates of facial landmarks with l_i^* As its ground truth.

RetinaFace also make use of face mesh representation to extract texture information a differentiable 3D renderer and this section is trained with the help of dense regression loss denoted by L_{pixel} . It also uses feature pyramid constructed with pre-trained modules connected in U-shape.

One of the known approach mentioned by Peiyun Hu *et al.* [10], consists of creating scale-specific detectors, i.e. making discrete object class much like in region proposal network. The model's implementation is based on ResNet101 as its backbone designed with the fact that facial regions having smaller region can only be detected if the context is clear as facial area is tiny to extract accurate parts without increasing false positives. Another set of proposals that claims improvements and state-of-the-art performance on occluded face images make use of attention mechanism one such proposal makes use of Face Attention Network (FAN)[12] which is useful for locating faces with large occlusion along with multi-scale detection. The Network includes a combination of any of single-stage detector and a designed anchor level attention mechanism, in the feature pyramid after each feature layer an attention subnet which has their classification and regression loss segment corresponding to each defined scale of facial region extraction and trained using same multi-task loss function as in RetinaFace.

Methods	Easy	Medium	Hard
RetinaNet	92.6	91.2	63.4
FAN Baseline	89.0	87.7	79.8

New proposals to tackle problems after pandemic:

Most of the architectures proposed by researchers have been computationally very expensive with hundreds of thousands of samples in datasets, especially for recognition purposes for GPU devices for training and real-time deployment for its efficient and practical use in the real world. With the increasing demand for such applications for deployment in edge devices, it has become necessary to come up with light-weight architectures that are just complex enough to solve the problems at hand. Keeping these in mind some light-weight networks are proposed like Light and Fast Face Detector[13]which introduces an anchor free model and yet gives similar results as some of the feature pyramid based models which has anchor-based algorithms. The paper's main focus persists with the working of receptive fields and the phenomenon of the effective receptive field (ERF). According to the paper, tiny faces involved in a frame of an image need some more of context information to help reduce the false positives with added improvement in recall hence leading to increasing the confidence of coming across a facial region. The proposed model has a CNN backbone with four segments working on a tiny part, a small part, medium part and large part, and 8 loss branches distributed along with the depth of the backbone architecture. Each loss branch has two sections Face classification and bounding box regression. For face classification softmax with cross-entropy loss is used over two classes and for regression on bounding box L2 loss has been adopted directly.

Datasets suitable for detection model after pandemic:

Today the primary requirement for face detection models is to increase recall on detecting partial facial regions which can be feasible only if the dataset used for training is descriptive enough for increasing detection confidence on partial visibility of target subjects. Till date, there have been many types of face detection datasets available for evaluating models. Datasets created in recent times has more descriptive annotations available while many in the early times, like FDDB has only face location mentioned in annotations. MALF and WIDERFACE provide many attributes including positions, gender, poses and most important, landmarks.

Amongst the most challenging dataset available, after the covid-19 pandemic is MAFA dataset which includes annotations based on the degree of occlusions, mask type being an addition on existing annotation generally available. The dataset also has proper proportions of all varieties depending occlusion type, face size, orientation and mask type as mentioned in Fig. X

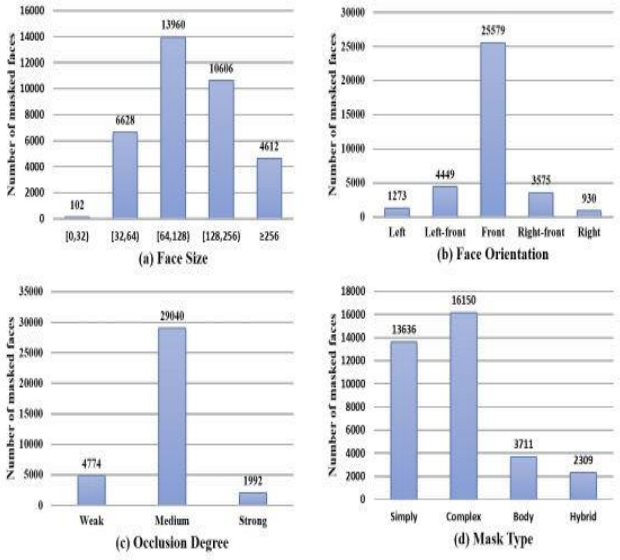


Fig. 1: statistical representation of MAFA dataset based on its distribution in various parameters as mentioned in [20]

3. FACE RECOGNITION

Using existing face recognition state-of-the-arts:

Amongst the most demanding application of face recognition is its use in building robust and real-time security and surveillance systems that will help reduce human resources used by the various agencies for such purposes. The deployment of such systems requires IoT to connect end devices to a network. Some of the traditional approaches aimed at sending real-time image frames from end camera devices to the centralized cloud, which then processes those images to serve the required purpose. Many deployment mechanisms and procedures have been proposed in the recent past to tackle cloud computing issues and favour the use of edge computing to improve the feasibility of using such applications. One of the most popular models using deep learning is FaceNet [21]. The core concept of FaceNet lies in mapping facial images from the database into a compact Euclidean space used to compare faces by finding the distance that measures the extent of face similarity. The model can be used for verification, recognition and clustering based on a feature vector(called Face embedding) extracted from input image frames. FaceNet produces 128D embedding using the triplet-loss function. The triplet consists of a person's anchor image, a positive image, i.e., a different image of the person in anchor image and negative image, i.e., image of a person with an identity separate from that of the anchor. The triplet loss function gives the most primary advantage of generating a decisive bounding region that predicts all vector lying inside the area as belonging to one identity. In most cases, interclass variation is some cases the interclass variation is less than intra-class variation.

Evolution in loss function used for training recognition models:

There have been various core neural network architecture for face recognition (or Face to feature embedding mapping) but for the model to be discriminative enough it must be trained with an appropriate loss function that directs the model to correct itself by defining the significance of deviation in feature embedding of different samples, and compactness in the embedding for samples belonging to the same identity. Many loss functions have been defined till date like softmax loss and its variant, angular margin loss, centre loss, contrastive loss, Euclidean-based losses and many more.

In the early years of deep learning, the strategies based on loss function were inherited from object classification state-of-the-arts proposed then. Later it was realized that softmax and any of its variants where not sufficient enough to tackle with the discriminative power and generalization ability that was required which led to enforcing exploration of a novel loss function some of which are Euclidean distance-based approach which tries to reduce gaps between intra-class and increase the gap between inter-class samples. Softmax loss has the disadvantage of linearly increasing linear transformation matrix size with the increase in the number of identities n. The features are separable in a closed set classification problem that is not compatible with face recognition. Similarly, with triplet loss, the question arises with an increasing possible combination of triplets and deciding an appropriate collection of good quality triplets in proper proportions with respect to the extent of similarity which is difficult for practical model training. The triplet loss function is represented as in equation (2) where A is an anchor input, P is a positive image input which belongs to the same class as an anchor, N belongs to a different class sample, a is margin defined between any pair of positive and negative pairs described using Euclidean distance. The function $f(X)$ is any embedding function that converts an input image to fixed size embedding vector.

$$L(A,P,N) = \max\left(\begin{matrix} \|f(A) - f(P)\|_2^2 \\ -\|f(A) - f(N)\| + a, 0 \end{matrix} \right) \quad (2)$$

Additive angular margin loss (ArcFace) [22] proposes loss function that improves the discriminative power of the model, which means having significant and sufficient decision margins between any two classes. Various other loss functions like SphereFace, cosFace, and Softmax have varying gaps between decision boundaries or changing decision margins with changing testing samples. ArcFace uses additive angular margin, cosFace uses additive cosine margin and SphereFace uses multiplicative angular margin. Light CNN framework introduces Max-Feature-Map (MFM) a variant of max out activation function used in each convolutional layer. MobiFace consists of fast downsampling and residual block with a bottleneck giving 99.7% on LFW. Other than the explored strategies, some baseline light-weight architectures like SqueezeNet, Xception, MobileNet, etc., that are yet to be discovered for its use in face verification and identification.

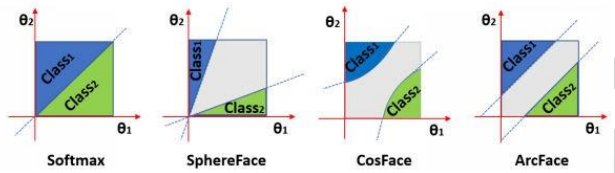
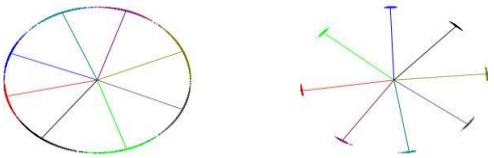


Fig. 2: decision boundary generation for different loss function presented by J. Deng *et al*[21]

In recent times since the incoming of AutoML, there has been various attempts to automate the loss function search called AM-LFS (AutoML for Loss Function Search)[22] which makes use of reinforcement learning to search loss function while in the training process but has lot of complexity to tackle hence reducing the benefits of generating optimal loss function. Xiaobo Wang *et al* [16] provides a proof of procedure to reduce the softmax probability being a key for feature discrimination, and they also propose a formulation for existing margin-based softmax loss.

General approach to tackle occlusion in face recognition:

In General, there have been three types of approaches matching discarding occlusion and restoration, for solving the occlusion issue while building and training architectures. For matching, different identities are compared using the set of different patches on each identity, consisting of key points each and prediction based on the number of patches compared and the extent of the similarity. Now considering discarding of occluded regions whenever the detector model passes a partial probe face, the occluded section is identified and cropped from all gallery images temporarily. Since the incoming of Deep learning into use to solve computer vision problems, it has been the most flexible approach comparing samples of faces with different occlusion type and degree using some sort of attention mechanism which reduced the extra work of pre-processing discarding occluded regions. The restoration approach has been most challenging among the other methods used for higher-dimensional image samples. An example of this type would be the use of a depth map to detect occlusions and Principal Component Analysis (PCA)[24] or Iterative closest point(ICP)[25]. But we are mainly interested in 2D input solutions, so restoration is out of scope here.



a) Softmax loss b) ArcFace loss

Fig. 3: classification ability of softmax and ArcFace loss function as demonstrated in ArcFace [22].

3.3 *Datasets compatible with training recognition models for reduced subject size and partial appearance:*

Every face database has been structured and annotated for different and objective, and those matching our objective of occlusion based training are Aleix-Robert (AR), Bosphorus, Labeled Faces in the Wild (LFW), University of Milano Bicocca Database (UMB) face databases and Real-World-Masked-Face-Dataset(RWMFD). Among these, Bosphorus and UMB are 3D databases, and RWMFD is recently developed especially for masked face recognition. AR Face database consists of 4000 colour images of 126 people (70 men and 56 women). This database has frontal face appearance in other facial occlusions, expressions and illumination. Occlusions in the AR database include sun glass and scarf coverings only, which would not be suitable for masked facial feature identification. LFW database consists of 13,233 facial images of 5749 distinct subjects which have only 29% of the residents containing more than one face image. Identities with multiple faces have almost the same proportion of occlusions, illuminations, pose, and accessories, hence reducing the significance of occluded faces. Real-World-Masked-Face-Dataset has three types of face images: Masked Face Detection Dataset(MFDD), Real-World Masked Face Recognition Dataset(RMFRD) and Simulated Masked Face Recognition Dataset(SMFRD) where SMFRD contains 500000 face images of 10000 people and RMFRD contains images of 525 subjects with 5000 of those images wearing and 90000 images without masks.

4. CONCLUSION

Due to the pandemic situation, the value of a more generalized face recognition system has dropped. There has been a greater demand for strategies that can tackle the issue of increased average occlusions. Most of the algorithms that have been used till date have shown average result on masked face appearance. We presented some of the work on the 2D database that would be useful as a baseline for further improvement for both detection and recognition systems. The database with samples to train models for occlusion has a lower percentage of those samples; hence there is a requirement of new databases, one of which arrived recently called Real-World-Masked-Face-Dataset and many more to come.

REFERENCES

[1] M. Z. Khan, S. Harous, S. U. Hassan, M. U. Ghani Khan, R. Iqbal and S. Mumtaz, "Deep Unified Model For Face Recognition Based on Convolution Neural Network and Edge Computing," in IEEE Access, vol. 7, pp. 72622-72633, 2019, doi: 10.1109/ACCESS.2019.2918275.

[2] -C. D. Gürkaynak and N. Arica, "A case study on transfer learning in convolutional neural networks," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404642.

[3] Sermanet, Pierre & Eigen, David & Zhang, Xiang & Mathieu, Michael & Fergus, Rob & Lecun, Yann. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. International Conference on Learning Representations (ICLR) (Banff).

[4] Das, Sukhendu. (2016). Domain Adaptation with Soft-margin multiple feature-kernel learning beats Deep Learning for surveillance face recognition.

[5] Bashbaghi, Saman & Granger, Eric & Sabourin, Robert & Parchami, Armin. (2018). Deep Learning Architectures for Face Recognition in Video Surveillance.

[6] Yang, Shuo & Luo, Ping & Loy, Chen Change & Tang, Xiaoou. (2017). Faceness-Net: Face Detection through Deep Facial Part Responses. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 10.1109/TPAMI.2017.2738644.

[7] Saez-Trigueros, Daniel et al. "Face Recognition: From Traditional to Deep Learning Methods." ArXiv abs/1811.00116 (2018): n. pag.

[8] L. He, H. Li, Q. Zhang and Z. Sun, "Dynamic Feature Learning for Partial Face Recognition," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7054-7063, doi: 10.1109/CVPR.2018.00737.

[9] P. Viola and M. J. Jones. Robust real-time face detection. International Journal of Computer Vision (IJCV), 57(2):137-154, 2004.

[10] P. Hu and D. Ramanan, "Finding Tiny Faces," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1522-1530, doi: 10.1109/CVPR.2017.166.

[11] J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5202-5211, doi: 10.1109/CVPR42600.2020.00525.

[12] Wang, Jianfeng & Yuan, Ye & Yu, Gang. (2017). Face Attention Network: An effective Face Detector for the Occluded Faces.

[13] He, Yonghao et al. "LFFD: A Light and Fast Face Detector for Edge Devices." *ArXiv* abs/1904.10633 (2019): n. pag.

[14] I. Masi, Y. Wu, T. Hassner and P. Natarajan, "Deep Face Recognition: A Survey," 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, 2018, pp. 471-478, doi: 10.1109/SIBGRAPI.2018.00067.

[15] Wu, Xiang & He, Ran & Sun, Zhenan & Tan, Tieniu. (2018). A Light CNN for Deep Face Representation with Noisy Labels. IEEE Transactions on Information Forensics and Security. 13. 1-1. 10.1109/TIFS.2018.2833032.

[16] Wang, Xiaobo & Wang, Shuo & Chi, Cheng & Zhang, Shifeng & Mei, Tao. (2020). Loss Function Search for Face Recognition.

[17] Shepley, Andrew. (2019). Deep Learning For Face Recognition: A Critical Analysis.

[18] Zhao, Jian & Cheng, Yu & Cheng, Yi & yang, yang & Lan, Haochong & Zhao, Fang & Xiong, Lin & Xu, Yan & Li, Jianshu & Pranata, Sugiri & Shen, Shengmei & Xing, Junliang & Liu, Hengzhu & Yan, Shuicheng & Feng, Jiashi. (2018). Look Across Elapse: Disentangled Representation Learning and Photorealistic Cross-Age Face Synthesis for Age-Invariant Face Recognition. 10.13140/RG.2.2.24847.23204.

[19] Z. Lu, X. Jiang and A. Kot, "Deep Coupled ResNet for Low-Resolution Face Recognition," in IEEE Signal Processing Letters, vol. 25, no. 4, pp. 526-530, April 2018.

[20] S. Ge, J. Li, Q. Ye and Z. Luo, "Detecting Masked Faces in the Wild with LLE-CNNs," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 426-434, doi: 10.1109/CVPR.2017.53.

[21] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering", In Proc. CVPR, 2015.

[22] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 4685-4694. doi: 10.1109/CVPR.2019.00482

[23] Li, C., Yuan, X., Lin, C., Guo, M., Wu, W., Yan, J., and Ouyang, W. Am-lfs: Automl for loss function search. In Proceedings of the IEEE International Conference on Computer Vision, pp. 8410–8419, 2019.

[24] Parama Bagchi, Debotosh Bhattacharjee, and Mita Nasipuri. Robust 3d face recognition in presence of pose and partial occlusions or missing parts. arXiv preprint arXiv:1408.3709, 2014.

[25] Ashwini S Gawali and Ratnadeep R Deshmukh. 3d face recognition using geodesic facial curves to handle expression, occlusion and pose variations. International Journal of Computer Science and Information Technologies, 5(3):4284–4287, 2014.

[26] Walid, Hariri. (2020). Efficient Masked Face Recognition Method during the COVID-19 Pandemic. 10.21203/rs.3.rs-39289/v1.