2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-7281-8470-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLA51294.2020.00018

Single Camera Masked Face Identification

Vivek Aswal, Omkar Tupe, Shifa Shaikh and Nadir N. Charniya

Department of Electronics and Telecommunication Engineering (EXTC)
Vivekananda Education Society's Institute of Technology (VESIT), Mumbai, India
2016.vivek.aswal@ves.ac.in, 2016.omkar.tupe@ves.ac.in, 2016shifa.shaikh@ves.ac.in, nadir.charniya@ves.ac.in

***Abstract*—In light of the novel Covid-19 pandemic, wearing masks has been declared mandatory in several institutions and public places for its widespread prevention and public health safety. Under given circumstances, person identification for security purposes including smart-phones face unlock has been a challenging task since the previous practices including both the human authentication by a person as well as by face recognition systems have heavily relied on complete facial features. However, the emergence of large datasets of masked images led to the rapid development of occluded face detection techniques. This paper focuses on single camera masked face detection and identification via the following two approaches: (i) single-step pre-trained YOLO-face/trained YOLOv3 model on the set of known individuals; and (ii) two-step process having pre-trained one stage feature pyramid detector network RetinaFace for localizing masked faces and VGGFace2 that generates facial feature vectors for efficient mask face verification. The dataset employed consists of real-world video examples comprising of 7 individuals with various orientations, illuminations, and occlusions. Experimental results show that RetinaFace and VGGFace2 achieve state-of-the-art results of 92.7% on overall performance, 98.1% face detection, and 94.5% face verification accuracy respectively in 1:1 face mask verification on our custom dataset.**

Keywords—Covid-19, face detection, face recognition, YOLOv3, RetinaFace, VGGFace2

I. INTRODUCTION

The spread of Novel Coronavirus disease 2019 (Covid-19) pandemic has unveiled a new reality that not only disrupted lives but also brought the world to a standstill. Covid-19, being a highly infectious respiratory disease has prompted governments and institutions that have urged people to wear face masks as an effective preventive measure against the virus. These circumstances necessitate the need for reforms in face recognition techniques that have been arguably the most important means of contactless identification.

Masked face facial feature-based identification refers to the process of identifying a masked person using his/her face. Whereas, person identification uses face verification for retrieving a target person via a uniform corresponding label across multiple non overlapping cameras. Since majority of face recognition techniques have been ineffective causing serious difficulty in accurate person identification that is essential for robust face authorization and attendance systems, wide-region tracking for security and surveillance, facilitating secure payments, contact tracing of Covid-19 suspects, etc.

This work focuses on developing an identification and verification system capable of retrieving a masked face person's identity based stored database. With standard face detectors obsolete on masked faces, two methodologies have been adopted. In the first approach, YOLOv3 [1] is trained on the custom dataset images for recognition of specific individuals. In the second method, a more generalized two-step approach has been considered wherein first, the faces are localized via the YOLO-face [2] or RetinaFace [3] that performs joint extra-supervised and self-supervised learning. In the next step, VGGFace2 [4], ResNet-50 [5] architecture computes the facial vectors in conjunction with a nearest-neighbor identity recognition for person verification.

II. RELATED WORK

Previous face detectors such as MTCNN [6] achieved reliable performance on unconstrained faces. However, for scaled, illuminated, and occluded faces, these detectors did not produce similar results. The availability of large-scale datasets like MS-Celeb-1M [7] and WIDER FACE [8] provided diverse facial features and annotations, motivated new algorithms to solve these challenges. Two-stage detectors had good performance but low speed. For faster approaches, one-stage detector located objects with a single detector by predefining the aspect ratios of anchor boxes.

In order to gain notable detection speed, the single-stage detectors sacrifice a small amount of performance. YOLO [9] algorithm achieves this fast detection speed by splitting the image into cells and then locates objects in each cell. Since YOLO depends on fixed receptive field feature maps, the algorithm does not work well for small objects. This has led to Single Stage Headless (SSH) [10] face detector method which is scale invariant as it uses multi-scale detection and relies on image pyramid consisting of several feature maps. Face Attention Network (FAN) [11] uses an anchor level attention technique for detecting faces to improve the recall of occluded faces at a low false positive rate. By introducing a novel focal loss and combining SSD with Feature Pyramid Network (FPN) [12], RetinaNet [13] was built to solve the class imbalance complications.

Face verification is a complex multidimensional task and developed by relying on Eigenvectors. Early methodology [14] consisted of feature extraction using PCA and recognition using Feed Forward back propagation neural networks. The intrinsically 2-D approach [15] represented significant face characteristics known as Eigenfaces and could identify faces by unsupervised learning. Similarity metrics [16] were used for verification and recognition for large categories and small training examples. Modern techniques such as DeepFace [17], employs a four-stage pipeline of detecting, aligning, representing, and classifying. This process combined 3-D face modeling with piecewise affine transformation to obtain a face representation vector. FaceNet [18] mapped the extracted face images to Euclidean space that generated FaceNet embeddings as a measure of similarity. It introduced a novel online mining triplet technique to increase representational efficiency and achieve state-of-the-art results.

DOI 10.1109/ICMLA51294.2020.00018

III. METHODOLOGY

The person identification system can be achieved with two different procedures. In the single-step approach, the YOLOv3 model is trained on the previous database images of the person to be recognized. Whereas, the sequential two-step process is based on YOLO-face and RetinaFace that employ face detection which first localizes the face in an image and then performs face verification via VGGFace2.

A. YOLOv3

YOLOv3 [1] is a more advanced version of YOLO [9] which incorporates: (i) Logistic regression for predicting the objectness score of each bounding box. (ii) Multi-label class predictions by using cross-entropy loss and independent logistic classifier instead of softmax for improving performance. (iii) Scaling of prediction boxes at three different levels. To obtain useful semantic information, concatenation features maps are produced by merging earlier and unsampled features. (iv) Darknet-53 with shortcut connections for better GPU utilization, efficient evaluation, and faster performance.

B. YOLO-face

YOLO-face [2] is a face detector that is based on the one-stage YOLOv3 object detector. The objective of YOLO-face is to increase face detection performance at a fast detection speed. The YOLO-face achieves this performance by concentrating on a better selection method of suitable anchor boxes and a more effective loss function while using 106 layered Deeper DarkNet framework.

The anchor boxes used for face detection are modified and two different types of anchor boxes are selected. The first ones are taken from YOLOv3 but are transformed into narrower boxes by increasing the height with respect to the width. The second type is obtained by k-means clustering on the WIDER FACE training dataset for obtaining the box dimensions. Generalization of IoU (GIoU) metric is used to optimize losses of non-overlapping bounding boxes.

C. RetinaFace

RetinaFace architecture [3] uses ResNet-50 backbone for masked face detection for obtaining the feature maps by extracting the information present in the images. FPN along with an addition operation combines the extracted high-level semantic information to the information in the feature maps of the previous layers. Multi-scale detection technique for predicting FPN feature maps is used to detect different object sizes by receptive fields. Then, every generated feature map is passed to a detection head, inside which a context attention module adjusts the respective field size to concentrate on specific regions. The final detection head output is obtained through a CNN to reduce the network parameters and provide five landmark facial features.

The RetinaFace detector focuses on minimizing the Multi-task loss [3] for training anchor i:

$$L = L_{\text{cls}}(pi,pi^*) + \lambda_1 pi^* L_{box}(ti,ti^*) + \lambda_2 p_i^* L_{pts}(li,li^*) + \lambda_3 pi^* L_{pixel} \quad (1)$$

(i) Face classification loss L_{cls} is the binary class softmax loss for the anchor i between the expected probability (pi) and ground-truth (pi^*) of a face (1) or not (0). (ii) Face box regression loss L_{box} associates the predicted (ti) and ground-truth (ti^*) coordinates of the bounding box. (iii) Facial landmark regression loss L_{pts} relates the five predicted facial landmarks (li) with the respective ground-truths (li^*) of positive anchors. (iv) Dense regression loss L_{pixel} increases the weightage to more suitable boxes as well as landmark locations by suitable loss-balancing parameters λ_1-λ_3.

Transfer learning is used due to the restricted size of the dataset as feature learning is difficult for small datasets. Hence, the trained model weights based on the larger scale WIDER FACE [8] dataset have been considered.

D. Face Recognition via VGGFace

Face recognition is a classification problem in which the face of the person to be recognized is matched against the existing categories of faces in the database and video frames. For this purpose, a ResNet-50 CNN [5] was employed as the backbone architecture which was pre-trained using the softmax loss function on MS-Celeb-1M [7] and VGGFace2 [4] dataset that had 3.31 million images categorized into 9131 individuals, including variations in ethnicity, age, and pose. VGGFace2 model outputs a face embedding of 2048 vector dimensional descriptor. These are then L2 normalized and the similarity between the faces is measured by cosine distance.

IV. PERSON IDENTIFICATION SYSTEM

A. Dataset

The dataset consists videos of individuals wearing face masks in close proximity. A total of 17 videos having 7 different individuals were taken and split into video frames at 1 fps. The dataset contains variations in orientation, scale, occlusions, illumination, and appearance as shown in Fig. 1.

Fig. 1. Different individuals in the video dataset

B. Single-Step Person Identification

A custom dataset of 20 masked faces per subject was used for training YOLOv3 with multi-scale predictions, data augmentation, batch normalization and Categorical cross-entropy loss for label predictions. The advantage of using transfer learning on YOLOv3 pre-trained on COCO [19] dataset is that the training process requires fewer images per class. The training process predicts 3 boxes at each scale, resulting in a $N \times N \times [3*(4+1+7)]$ tensor. The tensor represents 4 bounding box offsets, 1 predicted object, and probabilities of the 7 different classes. K-means clustering determines bounding box priors by randomly choosing 9 clusters and 3 scales for uniformly partitioning the clusters.

C. Two-Step Recognition: Person Recognition from Database

First image of the person to be recognized is shown to the network as indicated by the flowchart in Fig. 2. The threshold confidence for the detection of faces is 0.5. Next, the YOLO-face/RetinaFace extracts the face from the images and resizes it to 224×224 dimensions. This extracted face is compared with the available known database faces by using VGGFace2 model which returns the Eigen-vectors for each face. Fig. 3 shows database containing faces of 10 different individuals without masks. Then the cosine distance is calculated between the Eigenvectors with threshold 0.55 for the person identified from the known person database. If the cosine distance is less than the threshold then the person is identified from the known database having least distance else the person is unknown.

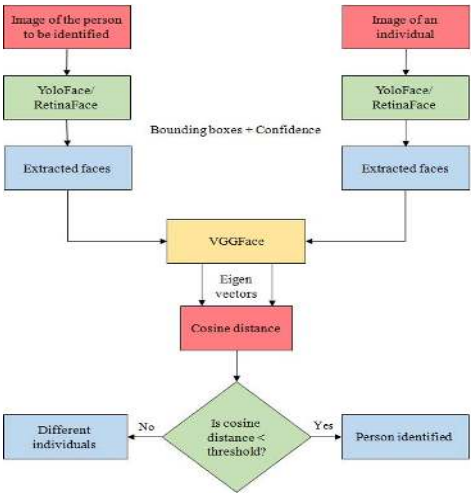

Fig. 2. Flowchart of the proposed identification system

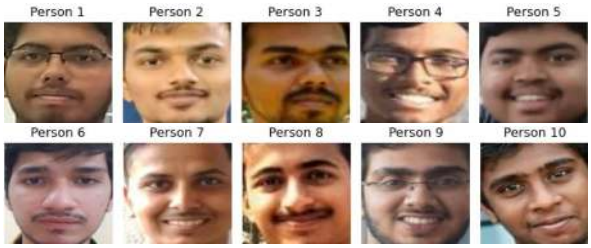

Fig. 3. Extracted faces of different individuals in the known database

D. Two-Step Recognition: Person Verification from Video

Initially, YOLO-face/RetinaFace model is applied for face extraction with a threshold confidence of 0.5. The extracted faces from the frames are compared against the person to be recognized and the Eigen-vectors are compared. The cosine distance of these vectors is then calculated having a threshold of 0.53. Sometimes, although the YOLO-face/RetinaFace model has detected the face with greater than threshold confidence, due to face mask, blurring, and large variations in the video frames, the VGGFace2 model fails to match that particular frame's face to the person to be identified. Hence, in this case, the following reprocessing procedure is employed:

1) The face of the current frame is matched with the previous correctly matched video frame face instead of matching with the person to be identified.

2) If the two video frames match, then the person in both the frames is same since the previous video frame has been identified correctly.

However, if the first frame does not match, as there cannot be any previous correctly identified frame, the reprocessing procedure is further modified:

3) In this case, VGGFace2 matches the next frame and if not matched, continues matching until any of the frame matches with the image of the person to be identified.

4) When the person is identified in any frame, VGGFace2 again reprocesses the first frame with the matched frame. Similarly, all the successive frames until the recognized frame are reprocessed as mentioned in step 1.

V. Results and Discussion

Fig. 4 shows the verification results of person recognition from the database. This matching can be either done on full masked faces or upper half faces excluding the masked portion for two different backbone architectures, the ResNet-50 and SENet-50 [38]. In case of full-face identification, the average SENet-50 distance scores of the identified person are lower compared to the ResNet-50. However, the ResNet-50 has higher distance separation between different individuals. This implies that although SENet-50 has a better recognizing capability, ResNet results in superior person differentiation. For half face recognition, the average cosine distances obtained are less than full faces but this leads to identity misclassification as the distance between different people is reduced. Hence, ResNet-50 for full faces achieves the best distance difference followed by SENet-50 for full faces, ResNet-50, and SENet-50 for half faces, respectively.

The detection results comparison in Fig. 5 clearly indicate that the YOLO-face algorithm does not capture faces that are cropped, blurred, and faces in low lighting conditions. On the other hand, YOLOv3 works fine in these cases but fails in appearance changes since the algorithm has been trained on masked faces in normal conditions which the generalized YOLO-face and RetinaFace algorithms detect successfully. The bounding boxes generated by YOLOv3 are of arbitrary shape and cover regions other than the face. RetinaFace when compared to YOLO-face generates more compact and precise bounding boxes with much higher confidence.

The performance of all the algorithms is compared on the masked face dataset in Table I. The results show that the detection accuracy of RetinaNet is superior for each subject except person 2 that is marginally outperformed by YOLOv3. RetinaFace has the lowest accuracy on Person 6 since wearing a mask with protective shield makes detections difficult for generalized YOLO-face and RetinaFace. YOLOv3 on the other hand has poor detection accuracy for Person 1 and Person 5 due to high degree of appearance changes. Based on the identification results, YOLOv3 again performs poorly on Person 1, whereas YOLO-face with VGGFace2 outperforms RetinaFace with VGGFace2 since the RetinaFace network is able to capture all kinds of faces from a variety of conditions and thus making it difficult of the VGGFace2 to verify the person to be identified in the video frames.

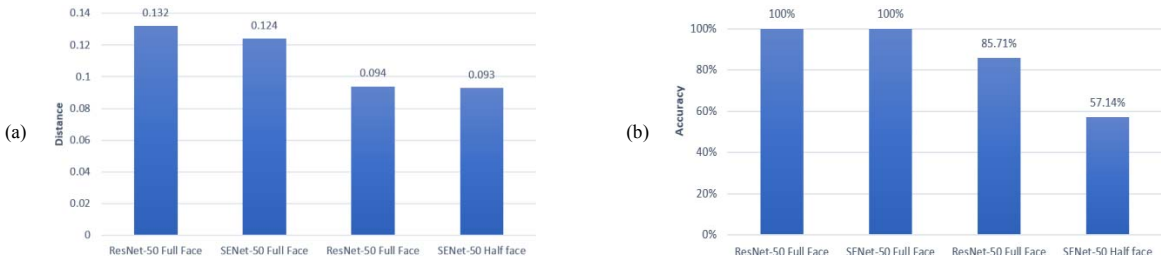

Fig. 4. Database person recognition comparison by VGGFace2 backbones. (a) The average identification cosine distance. (b) The identification accuracy.

Fig. 5. Detection results of different algorithms on the dataset video frames. (a) YOLO-face. (b) YOLOv3. (c) RetinaFace.

TABLE I. ACCURACY OF DEEP LEARNING MODELS ON DATASET (RED INDICATES THE BEST TOTAL PERFORMANCE)

Person	Frames	Detection Accuracy			Verification (Identification*) Accuracy			Overall Performance (Detection×Verification)		
		YOLOv3	YOLO-face	RetinaFace	YOLOv3*	YOLO-face+ VGGFace2	RetinaFace+ VGGFace2	YOLOv3	YOLO-face+ VGGFace2	RetinaFace+ VGGFace2
Person 1	152	0.276	0.842	0.912	0.523	0.898	0.908	0.146	0.757	0.888
Person 2	125	0.950	0.912	0.944	1.000	1.000	0.992	0.946	0.912	0.944
Person 3	131	0.929	0.947	0.977	1.000	0.992	0.976	0.929	0.939	0.977
Person 4	178	0.653	0.388	0.910	0.925	0.930	0.903	0.600	0.361	0.885
Person 5	190	0.373	0.558	1.000	0.972	1.000	1.000	0.363	0.558	1.000
Person 6	98	0.810	0.418	0.841	0.971	0.976	0.836	0.786	0.408	0.755
Person 7	134	0.602	0.970	0.985	0.800	0.977	1.000	0.481	0.978	0.985
Total	1013	0.650	0.749	0.981	0.930	0.968	0.945	0.604	0.725	0.927

VI. CONCLUSION AND FUTURE WORK

RetinaFace outperforms other algorithms in overall performance, confidence, and preciseness of the bounding box. It also detects the landmark facial features in cases of severe occlusions, appearances, and illumination. Using VGGFace2 based ResNet-50 backbone, full masked face recognition is achieved with high accuracy, making the identification system suitable for practical applications.

Future work can focus on YOLOv4 [20] algorithm that offers optimal speed as well as accuracy. Also, in the case of two-step identification, the verification results can be improvised by replacing the VGGFace2 model with the ARCFace [21] algorithm that obtains highly discriminative features by assigning an additive angular margin loss function that increases the distance between similar classes.

REFERENCES

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv: 1804.02767*.

[2] W. Chen, H. Huang, S. Peng, C. Zhou and C. Zhang, "YOLO-face: A real-time face detector," *The Visual Computer*, 2020. Available: https://doi.org/10.1007/s00371-020-01831-7.

[3] J. Deng, J. Guo, Y. Zhou, I. Kotsia and S. Zafeiriou, "RetinaFace: Singe-stage dense face localisation in the wild," 2019, *arXiv: 1905.00641*.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv: 1512.03385*.

[6] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499-1503, 2016.

[7] Y. Guo, L. Zhang, Y. Hu, X. He and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," 2016, *arXiv: 1607.08221*.

[8] S. Yang, P. Luo, C. C. Loy and X. Tang, "WIDER FACE: A face detection benchmark," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, "SSH: Single stage headless face detector," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] J. Wang, Y. Yuan and G. Yu, "Face attention network: An effective face detector for the occluded faces," 2017, *arXiv: 1711.07246*.

[12] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1991.

[15] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[16] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

[17] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *European Conference on Computer Vision (ECCV)*, 2014.

[20] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv: 2004.10934*.

[21] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.