Bias Analysis of Deep Face Recognition with Masked Faces

Benjamin Džubur, Peter Peer, Žiga Emeršič

University of Ljubljana, Faculty of Computer and Information Science
E-mail: beni.dzu@gmail.com, {peter.peer, ziga.emersic}@fri-uni-lj.si

Abstract

Demographic bias in face recognition systems is an increasing concern as these systems become more and more prevalent in society. Additionally, due to the COVID-19 pandemic, face masks have shown to be troublesome for many such systems. In this paper we focus on how masked faces influence the demographic bias of such systems. We took a pretrained convolutional neural network (VGGFace2) and used it for feature encoding of faces from a demographically balanced dataset. We generated a new dataset of masked faces based on the original one. For each of the two datasets, we performed verification by constructing matches and non-matches and computing similarity scores between pairs of faces' feature vectors. We compared the results for groups based on gender (male, female) and ethnicity (white, black, indian, asian), and discovered that our system is biased towards some demographic groups. Due to the imbalanced data used to train our neural network, our system performs the best on images corresponding to male and white subjects. We found that masking the faces does not in any significant manner change those biases, but the verification rates, however, drop significantly.

1 Introduction

With increasing importance of artificial intelligence solutions that are being used on a daily basis, more and more ethical and legal concerns are being raised regarding the transparency, accountability, explainability and fairness of such systems [1]. When it comes to different demographic groups of individuals (e.g. differences in gender, ethnicity), the concern of the bias of such intelligent systems is often raised, especially in relation to biometric pipelines such as face recognition.

These and similar concerns might have especially been amplified recently, where face masks have been frequently worn due to the global COVID-19 pandemic. Many studies have reported that face recognition systems have trouble with recognition of individuals with masks, with their accuracy dropping noticeably when compared to non-masked individuals [2]. These issues can then be addressed by either using recognition system robust to occlusions or combining face recognition with other modalities, such as ears [3, 4], eyes [5], etc.

For face recognition specifically, more and more systems are adopting Convolutional Neural Networks (CNN) as the algorithmic backbone in decision-making, due to their increasing accuracy and decreasing computational cost. Such systems require massive amounts of training data and employ large numbers of local, non-linear computations.

This paper focuses on evaluating the differences in bias when recognizing masked and non-masked individuals of different demographic groups. Individuals are grouped based on two, arguably most controversial demographic properties – ethnicity and gender. We use established bias estimation approaches on a CNN-based system for the above mentioned groups in order to depict the situation in modern systems.

1.1 Nomenclature

We acknowledge the terms used to describe demographic groups and concepts, such as gender and ethnicity can be very diverse and bear different cultural, social or political implications. We do not wish to redefine the terms, but we have observed some patterns in the use of these terms in cited papers. Terms sex and gender are used interchangeably. Similarly, little distinction is made between race and ethnicity. In this paper, we therefore decided to use the terms gender and ethnicity. More importantly, however, the terms gender and sex are usually used in a binary manner and we regard them in that way in our paper as well, but merely due to the statistical prevalence of binary genders.

1.2 Background & related work

Over the last few years, many concerns have been raised regarding the fairness of various automated systems. Many studies of risk assessment tools have found the issue of systemic bias against some demographic groups (e.g. dark-skinned people). The consequences of the decisions of such systems can be detrimental, with subjects experiencing difficulties, such as being denied bail or welfare payments [6]. While arguably being mostly unintentional, bias is a common occurrence in automated computer algorithms and machine learning.

In this context, many causes of bias exist. The most prominent cause originates from the training data, which can be imbalanced, incomplete, outdated, or of varying

quality [7]. All of these factors are detrimental to the training our algorithm and propagate the present biases. Another cause of bias can be in the implementation of the chosen algorithm, which can be flawed due to poor design or data preprocessing steps [7].

Many studies have already been conducted in the field of demographic bias of face recognition systems. Garcia et. al. [8] have discovered that when evaluating the FaceNet deep learning based model, evaluated on the MultiPIE dataset, the model produces the most errors on images of Asian female subjects, performing best on images of Caucasian male subjects. On the other hand, Beveridge et. al. [9] have achieved better performance on Asian male subjects than on Caucasian males. When it comes to the age covariate, Michalski et. al. [10] achieved better performance on older subjects and with larger variation in performance for children.

Similarly, due to the COVID-19 pandemic, many studies have already been conducted in the field of assessment and mitigation of the impact of wearing face masks on facial recognition systems. Damer et. al. [11] have noted a large drop in the verification performance and the significant effect on the genuine and imposter score distributions when wearing masks. Anwar & Raychowdhury [12] propose retraining the models on masked sets, generated from original training data, seeing a significant improvement in recognition performance.

However, to the best of our knowledge no studies have yet explored how, if at all, masked faces impact the inherent demographic biases of facial recognition systems. We will explore this issue in the scope of this paper.

CNNs are inherently biased due to the data they are trained on [13]. Aside from the quality and diversity of the data, the balance of the dataset in regards to important distinguishing properties proves to be a big factor, especially with smaller datasets. As a result, some systems which for example originate from Asia and are trained predominantly on Asian subjects, will have an easier time distinguishing between and consequently recognizing Asian subjects than Caucasians. The opposite may hold for a system trained on Caucasians [14], etc. This fact and the fact that CNN-based approaches are the most used approaches in face recognition by a large margin, makes them the ideal candidate for our analysis.

2 Method

For our study, we use a model based on the popular ResNet architecture. The model is trained on the VGGFace2 dataset, which consists of over 3 million faces and promises large variations in pose, age, illumination, ethnicity and profession [15]. The actual distribution of ethnicity seems to however heavily favor the Caucasian Latin group at around 79%, with the other three groups (African American, East Asian and Asian Indian) distributed relatively equally over the remainder. When it comes to genders, the dataset consists of approximately 62% male and 38% female subjects [16].

The ResNet-50 architecture is designed to output a 128-dimensional feature vector, encoding the relevant face attributes. The result of encoding two images, which belong to the same subject is therefore hopefully two similarly sized and oriented vectors and vice versa.

We choose the ResNet architecture due to its residual connections, which mitigate the vanishing gradient problem and consequently improve performance in deep networks. The 50-layer pretrained version presents a good trade-off between speed and performance [17] and is more than deep enough for our purposes.

3 Experimental Setup

In the case of evaluating bias in face recognition, the most useful approach proves to be verification, which begins with comparing a chosen image (probe) against a specific reference image [14]. We therefore base our bias analysis on verification in our experiments. The similarity score between the two images is computed, and based on a set threshold, we either evaluate the pair of images to be a match (belong to the same subjects) or non-match. The purpose of the CNN in this case is essential, as it takes care of encoding the high dimensional images to a low dimensional vector of discriminating features. Such vectors can then easily and more reliably be compared to one another.

3.1 Metrics

When we choose a specific threshold for the similarity score, errors will inevitably be made in regards to our match decision. We may recognize two different subjects as one and the same (false positive, FP), or not recognize the same subject in both images (false negative, FN), as seen in Figure 1. Depending on where we set the threshold, the sizes of these errors will vary. The metrics which quantify these errors are the false positive rate or FPR (also FAR, FMR):

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR \tag{1}$$

and the false negative rate or FNR (also FRR, FNMR):

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR, \tag{2}$$

where TPR and TNR represent the true positive rate and true negative rate [18]. We may also report the equal error rate or EER, which is achieved by setting the threshold in such a way, that both of the above mentioned error rates are equal.

3.2 Similarity score

In order to measure the similarity of the feature vectors u and v, we resort to the standard cosine similarity, which proves to be a good similarity measure in high-dimensional settings and especially for face verification [19]. It is defined as:

$$similarity = \cos\phi = \frac{u \cdot v}{\|u\| \, \|v\|} \tag{3}$$

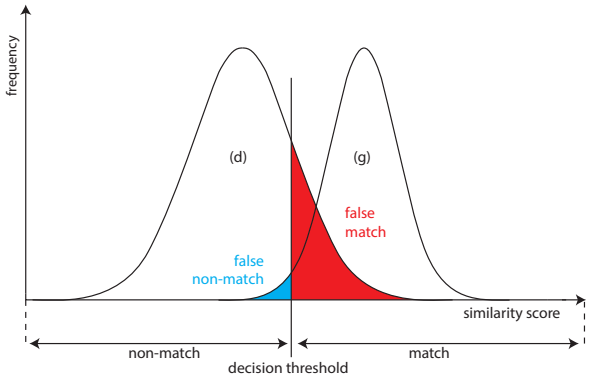

Figure 1: Similarity score distribution in verification. The red area represents FP, the blue FN. FPR can be calculated as the ratio between the red area and the area under the curve marked by (d). The FNR is the ratio between the blue area and the area under the curve marked by (g).

3.3 Evaluation dataset and matching

In order to gather reliable results, our dataset for evaluation would have to be balanced in terms of the chosen demographic properties. For this reason we chose the Balanced Faces in the Wild (BFW) dataset, which consists of 8 subgroups, composed of individuals with different combinations of gender (male vs. female) and ethnicity (white vs. asian vs. indian vs. black). For each subgroup, there are 100 distinct subjects, described by 25 (non-masked) images each [13]. This totals to 2500 images per subgroup and 20000 total images. From this we are able to construct $\frac{25\times24}{2} \times 800 = 240000$ (genuine) matches. We then produced the same amount of non-matches (impostor matches) for each subject by sampling the probe image uniformly at random from the rest of the population. When we group results by the ethnicity or gender of the reference images, we end up with the same amount of matches as well as non-matches which simplifies the analysis.

The images in the dataset vary noticeably in resolution, illumination and aspect.

3.4 Masked dataset

To perform unbiased comparisons of performance between masked and non-masked individuals, we need to have a dataset of masked faces, equivalent to BFW. We do this by generating a masked version of the dataset using a deep learning tool [12]. The generated faces wore a generic and popular blue surgical mask. This tool is not perfect, as it sometimes misaligns the mask, and in some extreme cases where the input face is captured from an off angle (e.g. profile) the tool is unable to generate a masked face.

We managed to produce a masked dataset which consists of 16911 out of the original 20000 faces. Generated images can be seen in Figure 2. We repeated the match and non-match generation for this dataset as well, which this time resulted in a little over 174000 matches and the same number of non-matches.

Figure 2: Examples of masked (generated) and original faces from BFW.

4 Results and Discussion

We first compared distributions of similarity scores between masked and non-masked datasets. The distributions across all demographic groups can be seen in Figure 3. It is evident that the distributions of true and false matches for non-masked individuals have a smaller intersection than those of masked individuals, which implies lower error rates for the non-masked population. This result is expected, as masking individuals obstructs an important part of their faces and prevents extraction of discriminating features. However, we might have expected a noticeable shift to the right of both genuine and impostor distributions when it comes to masked faces, due to the features extracted from the masked part of faces being similar across all matches. This is however not the case, implying that our neural network could capable of detecting and ignoring similarities due to face masks.

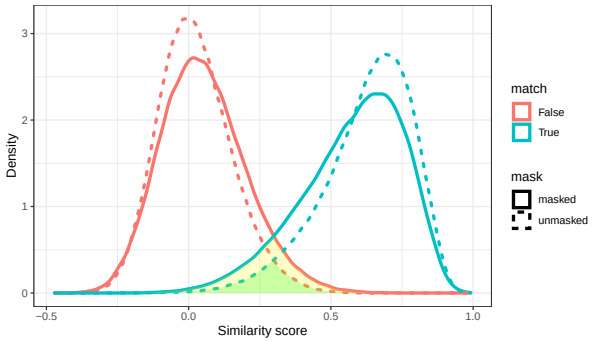

Figure 3: Distribution of masked and non-masked matches vs. non-matches. From the areas under the intersection of the match and non-match distributions (indicated by the colored areas), it is apparent that the equal error rate is noticeably lower for non-masked individuals.

Next, we take a look at true positive rates (verification rates) at different values of FPR across different demographic groups. This gives us an insight of how well

different demographic groups perform at the face verification task. In Figure 4 the two groups based on gender are compared, whereas in Figure 5, four groups are compared based on ethnicity.

We immediately notice that the drop in TPR when using masked faces as opposed to non-masked ones is quite large across all demographic groups, especially at lower FPR (10^{-4} to 10^{-3}). This would render such a basic system using no error mitigation strategies useless in practice for verification of masked faces.

When it comes to gender, males clearly achieve better verification rate than females regardless of masking. At FPR 0.1%, the TPR for males and females is 0.82 and 0.79 respectively (non-masked), or 0.62 and 0.49 respectively (masked). This disparity between genders can be attributed to the imbalance of training data based on gender. We do not observe any meaningful shifts in bias based on gender when comparing masked and non-masked performance.

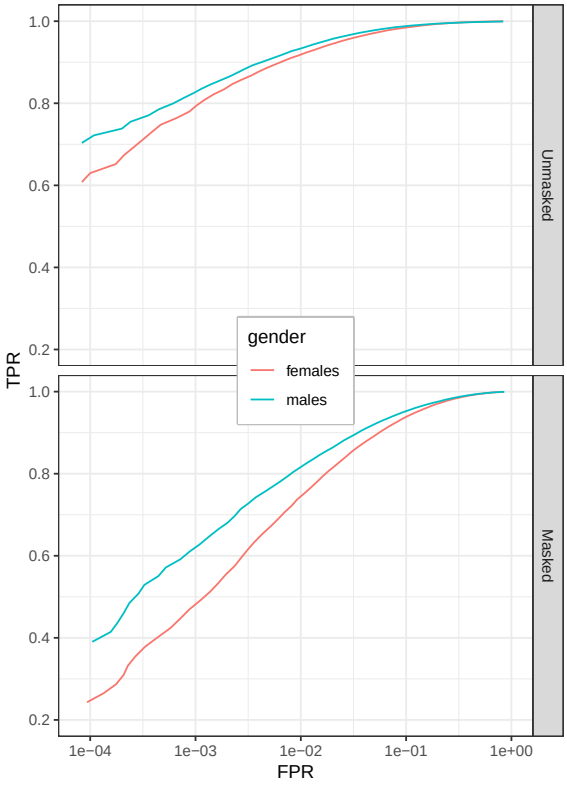

Figure 4: True positive rate at different false positive rates based on gender, showing a better verification rate of males than females.

When it comes to ethnicity, we observe a similar pattern, with whites achieving the best result across both datasets. When comparing indians and blacks, indians seem to perform noticeably better than blacks when masked, and about the same as blacks when not wearing masks at higher FPR. The impostor and genuine distributions for non-masked faces are similar for blacks and indians, however the genuine distribution of indians on masked faces is significantly more skewed to the right with smaller variance than the one of blacks. The genuine distribution of black faces might have changed significantly when dropping images to create the masked dataset. However, we cannot disregard the possibility that the facial features covered by the mask are more discriminative when comparing faces of black subjects as opposed to indian subjects. However, the drops in TPR when comparing non-masked and masked faces at different FPR are more or less proportional to the biases observed for each ethnicity group (see Table 1). Therefore, we cannot with confidence conclude that the discriminating features of the masked area might be more important for verification of one ethnicity group than another.

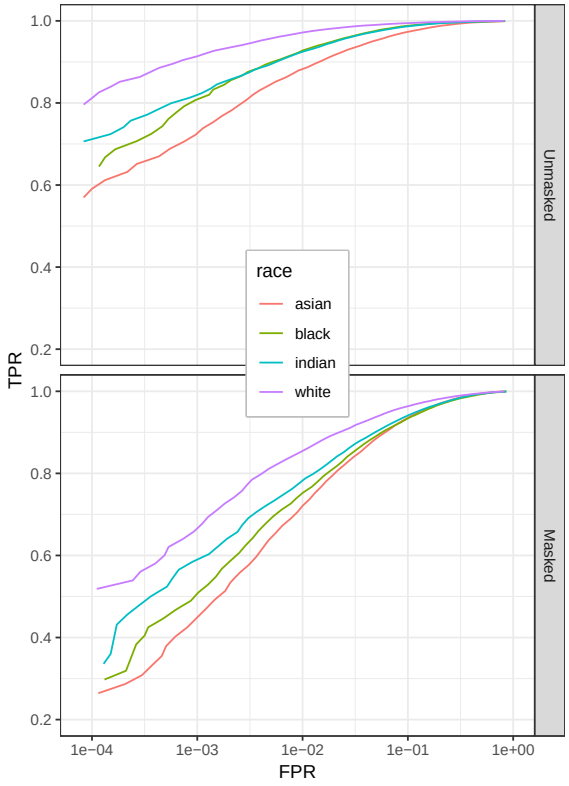

Figure 5: True positive rate at different false positive rates based on ethnicity, showing the best verification rate of whites, followed by indians, blacks and finally asians.

The disparity in performance for different ethnicity groups is again most likely due to the mentioned imbalance of the training data used for our neural network. As we initially believed, masking the faces does not in any significant manner change these biases. We were unable to confirm the hypothesis that the influence of wearing face masks on face recognition models differs significantly for different demographic groups. This would require more extensive experiments to be conducted. Aside from that theory, we found no real reason for biases to change significantly when the lower face region is masked. In practice, this means such models generally do not need to be additionally corrected for demographic bias due to masking the mouth and nose area, however the biases between demographic groups might become more apparent due to the significant drops in verification performance. Generally, we would in practice consider one of the pos-

sible error mitigation strategies due to face concealment if we wished to use such a system in a production environment.

FPR	Ethnicity	TPR (Original)	TPR (Masked)	TPR drop	TPR drop (%)
0.01%	white	0.81	0.52	0.29	35.80
	black	0.65	0.28	0.37	56.92
	indian	0.71	0.31	0.40	56.34
	asian	0.59	0.26	0.33	55.93
0.1%	white	0.91	0.68	0.23	25.27
	black	0.81	0.51	0.30	37.04
	indian	0.82	0.59	0.23	28.05
	asian	0.73	0.45	0.28	38.36
1%	white	0.97	0.85	0.12	12.37
	black	0.93	0.75	0.18	19.35
	indian	0.92	0.77	0.15	16.30
	asian	0.88	0.72	0.16	18.18

Table 1: True positive rate drop at different false positive rates based on ethnicity.

5 Conclusion

To summarize, we have analyzed and compared the results of face verification between masked and non-masked faces, for demographic groups based on two different covariates (gender, ethnicity). We used a single CNN, pretrained on millions of faces but somewhat imbalanced in the demographic groups we are interested in. We have observed differences in verification rates for different ethnicities and genders, but proportionally to the imbalance of the training data. We found that masking the faces does not in any significant manner change those biases.

In order to solidify our findings, additional testing should be performed on other face recognition models, trained on differently balanced datasets. The procedure of generating masked images, generating matches and non-matches for evaluation could be further looked into and improved to make the results more reliable and their interpretations more conclusive.

References

[1] J. G. Cavazos, P. Phillips, C. D. Castillo, and A. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *ArXiv*, vol. abs/1912.07398, 2019.

[2] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," 2020.

[3] Ž. Emeršič *et al.*, "The unconstrained ear recognition challenge 2019," in *2019 International Conference on Biometrics (ICB)*, 2019, pp. 1–15.

[4] Ž. Emeršič, J. Križaj, V. Štruc, and P. Peer, *Deep Ear Recognition Pipeline*. Cham: Springer International Publishing, 2019, pp. 333–362.

[5] P. Rot, M. Vitek, K. Grm, Ž. Emeršič, P. Peer, and V. Štruc, "Deep sclera segmentation and recognition," in *Handbook of vascular biometrics*. Springer, Cham, 2020, pp. 395–432.

[6] J. Beveridge, G. Givens, P. J. Phillips, and B. Draper, "Factors that influence algorithm performance in the face recognition grand challenge," *Computer Vision and Image Understanding*, vol. 113, pp. 750–762, 2009.

[7] C. Castelluccia, D. Métayer, E. P. D.-G. for Parliamentary Research Services, and E. P. E. P. R. S. S. F. Unit, *Understanding Algorithmic Decision-making: Opportunities and Challenges*. Publications Office of the European Union, 2019. [Online]. Available: https://books.google.si/books?id=L4WTwwEACAAJ

[8] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger, "The harms of demographic bias in deep face recognition research," in *2019 International Conference on Biometrics (ICB)*, 2019, pp. 1–6.

[9] J. Beveridge, H. Zhang, B. Draper, P. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, V. Štruc, J. Križaj, C. Ding, and P. J. Phillips, "Report on the fg 2015 video person recognition evaluation," 2015.

[10] D. Michalski, S. Y. Yiu, and C. Malec, "The impact of age and threshold variation on facial recognition algorithm performance using images of children," in *2018 International Conference on Biometrics (ICB)*, 2018, pp. 217–224.

[11] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, and A. Kuijper, "The effect of wearing a mask on face recognition performance: an exploratory study," 2020.

[12] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," 2020.

[13] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?" in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1–10.

[14] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[15] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," 2018, pp. 67–74.

[16] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset," *Machine Vision and Applications*, vol. 31, p. 67, 2020.

[17] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," 2017.

[18] A. Tharwat, "Classification assessment methods: a detailed tutorial," 2018.

[19] H. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," vol. 6493, 2010, pp. 709–720.