Fairness of Exposure in Forensic Face Rankings

Discussion Paper

Andrea Atzori, Gianni Fenu and Mirko Marras*

Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, 09124 Cagliari, Italy

Abstract

In information forensics, (police) agents are usually presented with a ranking of suspects similar to a certain face probe whose identity should be determined. Used for estimating the relevance score of possible suspects, deep face models have been proven to lead to undesirable discriminatory outcomes for certain demographic groups. Despite other non-personalised person rankings being actively investigated, forensic face rankings still represent an underexplored, yet important and peculiar, domain. In this ongoing project, we propose a framework consisting of six state-of-the-art face models and a public data set to quantify (disparate) exposure of demographic groups in forensic face rankings. Our results show that biases in this domain are not negligible and urgently call for ad hoc fairness notions and mitigation.

Keywords

Identification, Bias, Biometrics, Fairness, Equity, Exposure, Forensics, Rankings.

1. Introduction

Rankings have become one of the dominant forms in which digital systems present results to users. The prevalence of rankings ranges from search engines [1] and online stores [2], to music [3] and news feeds [4]. One notable task based on rankings is the identification of suspects based on their face biometrics [5]. Under this task, (police) agents are presented with a ranking of suspects similar to the face probe. Deep face recognition models are supporting the generation of these rankings, thanks to their impressive performance in terms of accuracy [6, 7].

However, deep models adopted to extract a latent face representation for ranking purposes have been proven to be susceptible to biases [8, 9, 10]. For instance, adopting such latent representations for face authentication has led the system to fail more often for subjects with darker skin tones [11, 12]. As a consequence, considerable efforts have been made to analyse discriminatory results for groups created on the basis of protected attributes (e.g., gender and ethnicity) [13, 14]. Unfortunately, these analysis have focused on pure biometric authentication and identification [15], without considering undesired impacts from a ranking perspective.

Indeed, certain forensic face ranking techniques rely on hand-created sketches from possible eyewitnesses [16]. Such partial information is often used in combination with other attributes, such as textual descriptions [17]. To be functional, this type of approach can also be combined with text-to-image (to generate the query) and image-to-text (to generate search-useful galleries)

IIR2023: 13th Italian Information Retrieval Workshop, June 8th - 9th, 2023, Pisa, Italy

andrea.atzori@unica.it (A. Atzori); fenu@unica.it (G. Fenu); mirko.marras@acm.org (M. Marras)

0000-0002-6910-206X (A. Atzori); 0000-0003-4668-2476 (G. Fenu); 0000-0003-1989-6057 (M. Marras)

techniques [18]. Existing approaches only evaluate whether the offender appears in the first position of the ranking [19], often ignoring the rest of the ranking. However, analysing the ranking composition and the (disparate) exposure of demographic groups is fundamental, as being exposed to (police) agents as suspects might lead to undesired consequences (e.g., being wrongly investigated). Current research has assessed and mitigated unfairness in other non-personalised people rankings [20, 21, 22], but forensic face ranking still represents an underexplored domain, characterized by key peculiarities (e.g., normative, content, model).

In this ongoing project, we have the ambitious objective of bridging the face biometrics and information retrieval research communities by analyzing whether deep face recognition models lead to unfair exposure across demographic groups in forensic ranking systems. Our novel contribution is twofold. First, we propose an assessment framework with six state-of-the-art face recognition models and a public face data set labeled with two protected attributes (i.e., gender and ethnicity). Second, we conduct an exploratory study aimed at quantifying the (disparate) exposure of demographic groups in the resulting rankings, depending on the demographic group of the face probe (RQ1), and which demographic groups are most likely to incorrectly appear in the top positions of the ranking (RQ2). Our results show what biases forensic rankings are exposed to, emphasizing the importance of devising mitigation methods in this domain.

2. Method

Using a dataset annotated with two protected attributes, we first trained six deep face recognition models. Then, we evaluated the utility and fairness of the rankings generated by these models.

Data Preparation. Our experiments were carried on using the `DiveFace` [23] data set, consisting of 140,000 images belonging to 24,000 identities. It is properly annotated with sensitive information (gender and ethnicity) and balanced in terms of both attributes. Ethnicity labels include Asian, Black, and Caucasian. Gender labels include Women and Men. There are therefore six demographic groups represented in the data set: Asian Men, Asian Women, Black Men, Black Women, Caucasian Men and Caucasian Women. The original authors split the entire dataset into two sets: a training set and a test set, each of which contained 70% and 30% of the identities. To the best of our knowledge, this data set is one of the state-of-the-art sources for fairness analysis in the biometric field. In order to crop and resize the original images, the DeepFace toolkit [24] was used to detect the bounding box enclosing the face.

Model Preparation. With the face images in the training set, we built and trained a range of deep face models by combining a collaborative-margin head network and six convolutional neural networks (CNNs), namely `MobileFaceNet`, `ResNet` [25], `AttentionNet` [26], `ResNeSt` [27], `RepVGG` [28], `HRNet` [29]). These neural architectures have been proven to yield state-of-the-art performance in recent face benchmarks [7]. For consistency, our experiments followed the same training procedures described in [30]. More specifically, each model was trained using 64-sized batches for a maximum of 80 epochs (early stopping, patience 5). The optimizer was SGD, with momentum 0.9, weight decay 1e-8, initial learning rate 0.1, and decays at 5, 25, and 68 epochs. The loss function was categorical cross-entropy.

Ranking Generation. With the face images in the test set (disjoint from the training set),

we created a ranking system that, given as a query the latent representation of an individual (probe), ranks all the identities in the gallery and provides the most similar $K = 10$ identities to the query. For this purpose, we considered only individuals with at least $N = 10$ face images in the test set and sampled only N images for each individual. Due to the uneven number of images per identity, and in order to equally represent each demographic group in our test set, we selected $P = 32$ identities from each group (since the less represented one, Black Women, had only P identities with at least K images), taking into consideration a total of $1,920$ images from 192 identities. Given the N images for an individual, 30% of them were assumed to be face probes (images to be used as a query), and the remaining 70% were included in the gallery. The latent representations of all the face images of an individual were averaged to obtain a single averaged representation of that individual. For each probe, we computed the cosine similarity (range: [-1, 1]) between its latent representation and the latent representations of the identities in the gallery. We ranked the identities in the gallery according to their decreasing similarity with the probe and considered only the K identities most similar to each query.

3. Experimental Results

The models' accuracy was between 98% and 99%. Experiments analyzed demographic group exposure per probe's group (RQ1) and overall exposure (RQ2) in rankings across groups.

Exposure for each probe's demographic group (RQ1). In a first analysis, for each probe's group, we computed the averaged exposure of each demographic group across rankings and models (Fig.1), adopting the definition of [31]. As expected, for each probe's group, the demographic group with the highest exposure corresponds to the probe's one (between 50% and 90%). Going beyond this case, probes of Asian men (top left) led to disparate exposure for Caucasian men (same gender) and Asian women (same ethnicity) compared to the other groups. Similarly, the rankings emerging from Asian women's probes (top center) disproportionately represent Caucasian women (same gender) and Asian males (same ethnicity). Probes from Black men (top right) make Caucasian men and Asian men more prominent; in both cases, the gender seemed to be the main causing factor. Black women's probes (bottom left), conversely, led to higher than more equal representation across certain groups beyond that of the probe. For instance, Asian women and Caucasian females and males are unfairly more at risk of incorrectly appear

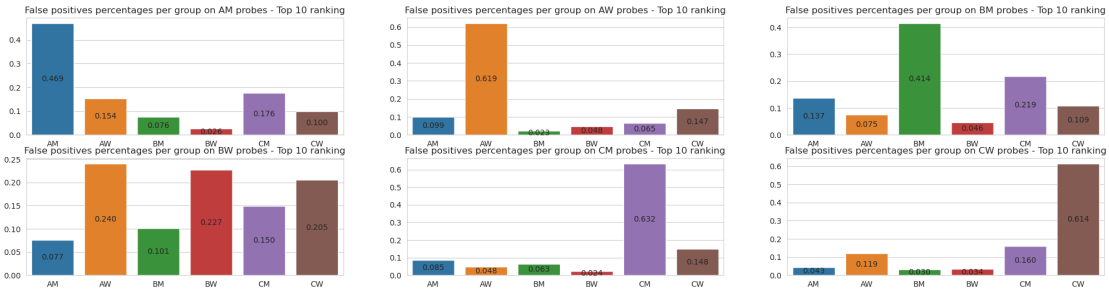

Figure 1: Demographic groups exposure in top-10 rankings for each probe's group. **Legend**: A:Asian; B:Black; C:Caucasian; M:Man; W:Woman.

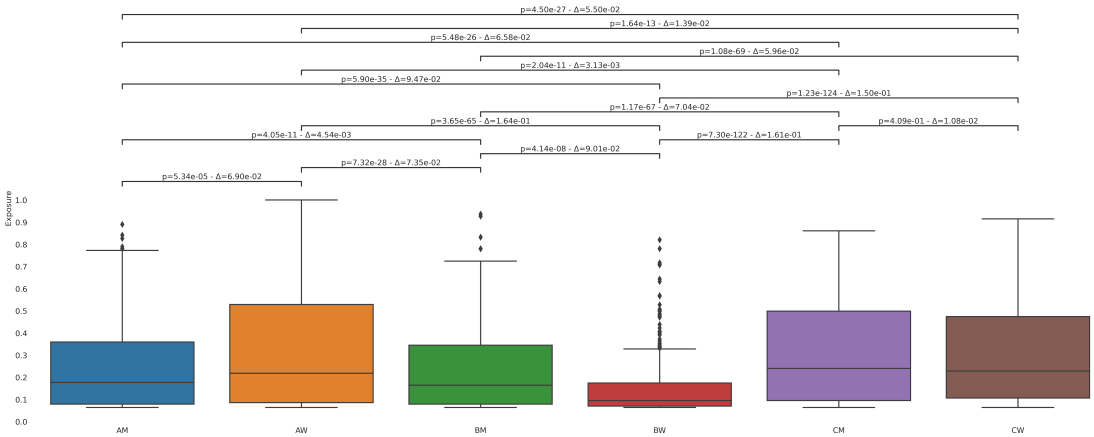

Figure 2: Group exposure distribution, averaged among models. Statistical significance assessed by means of the Kolmogorov-Smirnov test, which rejects the hypothesis that the exposure distributions are identical between two demographic groups in case the p-value is lower than 0.05.

in the top ranking. Caucasian probes make same ethnicity counterparts more prominent, along with Asians of the same gender. We can conclude that, based on the probe's group, certain demographic groups (beyond that of the probe) are overexposed in the rankings.

Overall disparate exposure of demographic groups (RQ2). In a second analysis, we investigated whether the observed disparate exposure across demographic groups is even more evident if we consider the position in which a certain possible suspect appears. Since our results were consistent across models, Fig. 2 shows the exposure distribution of groups across rankings, averaged among models. Comparing exposure across groups, it can be observed that certain groups tend to have a statistically significant disparate exposure compared to others. In particular, Asian women often tended to be unfairly represented in the top positions, even when the probe's group was different. Caucasian men and women were more often and incorrectly exposed at the top, followed by Asian women, Black men, and Black women. We can conclude that, in general, women tend to be more often over-exposed in the rankings.

4. Conclusions and Future Work

In this paper, we investigated the extent to which state-of-the-art face models adopted for forensic face ranking are subject to biases across demographic groups. Our results highlighted that dark-skinned individuals (especially women) have lower exposure under probes belonging to other demographic groups (especially female ones). Even more noticeably, dark-skinned female probes produce the least disparate results compared to all other groups. In addition to this, Asian and Caucasian are overexposed in the rankings, while women appear to be the most likely to hold prominent positions. In the next steps, we plan to investigate whether, and possibly to what extent, other factors (e.g., pose, lighting, expression) influence the considered forensic face rankings. We also plan to devise potential countermeasures regarding the disparities in treatment we uncovered through our study in this paper.

References

[1] R. Gao, C. Shah, Toward creating a fairer ranking in search engine results, Information Processing & Management 57 (2020) 102138.

[2] M. Wan, J. Ni, R. Misra, J. McAuley, Addressing marketing bias in product recommendations, in: Proceedings of the 13th international conference on web search and data mining, 2020, pp. 618–626.

[3] B. L. Pereira, A. Ueda, G. Penha, R. L. Santos, N. Ziviani, Online learning to rank for sequential music recommendation, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 237–245.

[4] L. Sanchez, J. He, J. Manotumruksa, D. Albakour, M. Martinez, A. Lipani, Easing legal news monitoring with learning to rank and bert, in: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42, Springer, 2020, pp. 336–343.

[5] M. Jacquet, C. Champod, Automated face recognition in forensic science: Review and perspectives, Forensic science international 307 (2020) 110124.

[6] M. Wang, W. Deng, Deep face recognition: A survey, Neurocomputing 429 (2021) 215–244.

[7] J. Wang, Y. Liu, Y. Hu, H. Shi, T. Mei, Facex-zoo: A pytorch toolbox for face recognition, in: Proc. of ACM/MM 2021, 2021, pp. 3779–3782.

[8] M. Gwilliam, S. Hegde, L. Tinubu, A. Hanson, Rethinking common assumptions to mitigate racial bias in face recognition datasets, in: Proc. of CVPR 2021, 2021, pp. 4123–4132.

[9] V. Albiero, K. KS, K. Vangara, K. Zhang, M. C. King, K. W. Bowyer, Analysis of gender inequality in face recognition accuracy, in: Proc. of the IEEE/CVF Winter Conf. on App. of Computer Vision Workshops, 2020, pp. 81–89.

[10] V. Albiero, K. W. Bowyer, Is face recognition sexist? no, gendered hairstyles and biology are, in: Proc. of BMVC 2020, 2020.

[11] N. Srinivas, M. Hivner, K. Gay, H. Atwal, M. King, K. Ricanek, Exploring automatic face recognition on match performance and gender bias for children, in: Proc. of WACVW 2019, 2019, pp. 107–115.

[12] J. J. Howard, Y. B. Sirotin, A. R. Vemury, The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance, in: Proc. of BTAS 2019, IEEE, 2019, pp. 1–8.

[13] A. Atzori, G. Fenu, M. Marras, The more secure, the less equally usable: Gender and ethnicity (un) fairness of deep face recognition along security thresholds, Procedia Computer Science 210 (2022) 212–217.

[14] A. Atzori, G. Fenu, M. Marras, Explaining bias in deep face recognition via image characteristics, in: Proc. of IJCB 2022, IEEE, 2022.

[15] J. M. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, in: Proc. of ITCS 2012, volume 67, 2017, pp. 43:1–43:23.

[16] K. Ounachad, M. Oualla, A. Souhar, A. Sadiq, Face sketch recognition-an overview, in: Proceedings of the 3rd International Conference on Networking, Information Systems & Security, NISS2020, Association for Computing Machinery, New York, NY, USA, 2020.

[17] Y. Zhao, Y. Song, Q. Jin, Progressive learning for image retrieval with hybrid-modality queries, in: Proceedings of the 45th International ACM SIGIR Conference on Research and

Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1012–1021.

[18] L. Zhang, M. Yang, C. Li, R. Xu, Image-text retrieval via contrastive learning with auxiliary generative features and support-set regularization, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1938–1943.

[19] D. K. Sharma, A. S. Jalal, B. Sikander, Suspect face retrieval via multicriteria decision process, in: 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), 2022, pp. 849–853.

[20] A. J. Biega, K. P. Gummadi, G. Weikum, Equity of attention: Amortizing individual fairness in rankings, in: The 41st international acm sigir conference on research & development in information retrieval, 2018, pp. 405–414.

[21] A. Singh, T. Joachims, Fairness of exposure in rankings, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2219–2228.

[22] P. Lahoti, K. P. Gummadi, G. Weikum, Operationalizing individual fairness with pairwise fair representations, arXiv preprint arXiv:1907.01439 (2019).

[23] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, Sensitivenets: Learning agnostic representations with application to face images, IEEE Trans. on Pattern Analysis and Machine Intel. 43 (2020) 2158–2164.

[24] S. I. Serengil, A. Ozpinar, Lightface: A hybrid deep face recognition framework, in: Proc. of the Innovations in Intelligent Systems and Applications Conference (ASYU 2020), IEEE, 2020, pp. 1–5.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of CVPR 2016, 2016, pp. 770–778.

[26] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proc. of CVPR 2017, 2017, pp. 3156–3164.

[27] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A. J. Smola, Resnest: Split-attention networks, in: Proc. of CVPR 2022, 2022, pp. 2735–2745.

[28] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: Making vgg-style convnets great again, in: Proc. of CVPR 2021, 2021, pp. 13733–13742.

[29] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. on Pattern Analysis and Machine Intel. 43 (2020) 3349–3364.

[30] A. Atzori, G. Fenu, M. Marras, Demographic bias in low-resolution deep face recognition in the wild, IEEE Journal of Selected Topics in Signal Processing (2023) 1–13.

[31] E. Gómez, C. Shui Zhang, L. Boratto, M. Salamó, M. Marras, The winner takes it all: geographic imbalance and provider (un) fairness in educational recommender systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1808–1812.