POST-COMPARISON MITIGATION OF DEMOGRAPHIC BIAS IN FACE RECOGNITION USING FAIR SCORE NORMALIZATION

A PREPRINT

Philipp Terhörst[12], Jan Niklas Kolf[2], Naser Damer[12], Florian Kirchbuchner[12], Arjan Kuijper[12]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]TU Darmstadt, Darmstadt, Germany
Email:{philipp.terhoerst, naser.damer, florian.kirchbuchner, arjan.kuijper}@igd.fraunhofer.de

July 12, 2019

ABSTRACT

Current face recognition systems achieved high progress on several benchmark tests. Despite this progress, recent works showed that these systems are strongly biased against demographic sub-groups. Consequently, an easily integrable solution is needed to reduce the discriminatory effect of these biased systems. Previous work introduced fairness-enhancing solutions that strongly degrades the overall system performance. In this work, we propose a novel fair score normalization approach that is specifically designed to reduce the effect of bias in face recognition and subsequently lead to a significant overall performance boost. Our hypothesis is built on the notation of individual fairness by designing a normalization approach that leads to treating "similar" individuals "similarly". Experiments were conducted on two publicly available datasets captured under controlled and in-the-wild circumstances. The results show that our fair normalization approach enhances the overall performance by up to 14.8% under intermediate false match rate settings and up to 30.7% under high security settings. Our proposed approach significantly reduces the errors of all demographic groups, and thus reduce bias. Especially under in-the-wild conditions, we demonstrated that our fair normalization method improves the recognition performance of the effected population sub-groups by 31.6%. Unlike previous work, our proposed fairness-enhancing solution does not require demographic information about the individuals, leads to an overall performance boost, and can be easily integrated in existing biometric systems.

***K*eywords** Bias · Biometrics · Face recognition

1 Introduction

Large-scale face recognition systems are spreading worldwide and are increasingly involved in critical decision-making processes, such as in forensics and law enforcement. Consequently, these systems also have a growing effect in everybody's daily life. However, current biometric solutions are mainly optimized for maximum recognition accuracy [15] and are heavily biased for certain demographic groups [23, 1, 9, 24, 4, 10]. This means that, for example, specific demographic groups can be falsely identified as a black-listed individual more frequently than other groups. Consequently, there is an increased need that guarantee fairness for biometric solutions [4, 11, 33] to prevent discriminatory decisions.

From political perspective, there are several regulations to guarantee fairness. Article 7 of the Universal Declaration of Human Rights and Article 14 of the European Convention of Human Rights ensure people the right to non-discrimination. Also the General Data Protection Regulation (GDPR) [31] aims at preventing discriminatory effects (article 71). In spite of these political efforts, several works [24, 4, 23, 1, 9, 10] showed that commercial [4] as well as open-source [23] face recognition systems are strongly biased towards different demographic groups. Consequently, there is an increased need in fair and unbiased biometric solutions [23, 10]. Recent works [22, 32] proposed solutions to reduce the bias in face recognition systems. However, these approaches increased the fairness at cost of the total system performance.

In this work, we propose a novel fair score normalization approach that is specifically designed to reduce bias in face recognition decisions. Unlike previous work, increasing the fairness also lead to an improved performance of the system in total. Our theoretical motivation is based on the notation of individual fairness [7], resulting in a solution that treats similar individuals similarly and thus, more fairly. Using clustering in the embedding space, similar identities are categorized without pre-defined demographic classes. Using optimal local cluster thresholds, a score normalization approach is developed to ensure a more individualized, unbiased, and thus, fair treatment.

We evaluate our approach on publicly available datasets captured under controlled and in-the-wild conditions. To justify the concept of our fair normalization approach, we provide a visual illustration that demonstrates the suitability of the notation of individual fairness for face recognition.

The experiments show that the overall recognition performance improves by 12.8% under controlled circumstances and up to 30.7% under a more challenging in-the-wild scenario. Furthermore, our proposed approach is able to significantly reduce the error rates of all demographic groups. Especially under real-world conditions, the experiments demonstrated that our proposed method enhances the recognition performance of the most effected sub-groups of the population by 31.6%.

2 Related work

In the data mining community, unbiased or fair decision making is known as discrimination-aware data mining (DADM) [2]. In [2, 3, 13], these solutions apply decision rules to prevent discriminatory decisions. Calders et al. presented three modified versions of a naive Bayes classifier [5] to compensate discriminatory effects. In [17], Kehrenberg et al. showed that decision bias can naturally be handled with Gaussian processes. They introduced a latent target output to modify the class probability distributions such that their solution is able to capture several notations of fairness. All these works [2, 3, 13, 5, 17] incorporate fairness in their solutions, however only on the decision level.

In general representation learning research, some works tried to incorporate fairness in the learning process. In [33], Zemel et al. formulated fairness as an optimization problem to learn suitable representations. These representations enabled fair classification for the notations of group fairness and individual fairness. Raff et al.[26] improved the fairness of representations by presenting a gradient-based neural network optimizer. It simultaneously predicts the target class and the protected attributed. Furthermore, it is able to reverse a weight-updated if the update leads to a discriminatory classification behaviour. In [16], Jia et al. trained an agnostic network. Their model tried to learn representations for image recognition while minimizing contextual information such as background. All these works [33, 26, 16] aim at learn fair representation, but limit their contribution to classification tasks.

In face biometrics, biased systems are usually induced by non-equally distributed classes in training data. Klare et al. [19] showed that the performance of face recognition algorithms is strongly influenced by demographic attributes. In [4, 23], the authors came to the same conclusions for commercial and open-sources face recognition algorithms. They demonstrated that the persons gender and ethnicity strongly determines their face recognition performance. Alvi et al. [1] tried to prevent the bias in soft-attribute classification. They proposed a domain and task-adaptation approach to learn new representation for classification tasks that are blind to the known bias. Similar in [6], Das et al. aimed at fair soft-attribute classification (gender, age, and ethnicity) from face images. They proposed multi-task convolutional neural network approach for this tasks. All these solutions [19, 23, 1, 6, 4] demonstrated that there is bias in face biometrics and the works [1, 6] further showed that it is possible to achieve a more fair performance, however, in attribute classification and not face verification.

Recently, some works focused on enhancing the privacy of face biometrics. The goal of these works was to suppress soft-biometric pattern in the data representations. Consequently, this improves the demographic fairness, since suppressed demographic attributes in the representations can not be used for a discriminatory behaviour [22]. Terhörst et al. focused on finding supervised [28] and unsupervised [29] templated-based solutions, while Mirjalili et al. [20, 21] proposed approaches on image level using semi-adversarial neural networks. In [22], Morales et al. incrementally removed attribute information while retaining most of the face recognition performance with an information-aware triplet loss. Wang et al. [32] proposed a deep information maximization adaptation network to reduce the effect of racial bias in face recognition. All these works [28, 29, 20, 21, 22, 32] proposed solutions for fair face recognition. However, developing these solutions needs the knowledge about the demographic attributes of each individual. More importantly, these solutions strongly degrades the total recognition performance. In this work, we propose a fair normalization solution that introduces fairness to an existing system while significantly improving the total system performance.

3 Methodology

The goal of this work is to enhance the fairness of existing face recognition systems. In this work, we follow the notation of individual fairness [7]. This notation emphasizes that similar individuals should be treated similarly. We transfer this idea to the embedding and score level to propose a novel fair group-based score normalization method, without the need for pre-defined demographic groups. The proposed approach is able to treat all identities more individually and therefore, increase the group-related, as well as the total, recognition performance.

3.1 Fair group score normalization

Our proposed solution is presented assuming a set of face embeddings $X = (X_{train} \cup X_{test})$ with the corresponding identity information $y = (y_{train} \cup y_{test})$, both partitioned into test and training set.

Training phase During training phase, a k-means cluster algorithm [14] is trained on X_{train} to split the embedding space into k clusters ($k = 100$ in our experiment). For each cluster $c \in \{1, \ldots, k\}$, an optimal equal error rate (EER) threshold is computed using the genuine and imposter scores of cluster c

$$gen_c = \{s_{ij} \,|\, \mathit{ID}(i) = \mathit{ID}(j),\, i \neq j,\, \forall\, i \in C_c, \} \tag{1}$$

$$imp_c = \{s_{ij} \,|\, \mathit{ID}(i) \neq \mathit{ID}(j),\, \forall\, i \in C_c, \} \,. \tag{2}$$

The genuine score set gen_c of cluster c includes the all comparison scores of samples i and j that come from the same identity ($\mathit{ID}(i) = \mathit{ID}(j)$), where at least one sample lies within cluster c ($i \in C_c$). Conversely, the imposter score set imp_c of cluster c is defined as all comparison scores s_{ij} from different identity pairs ($\mathit{ID}(i) \neq \mathit{ID}(j)$) where at least one sample lies within cluster c ($i \in C_c$). The EER-threshold for each cluster c is denoted as $thr(c)$. Furthermore, the EER-threshold for the whole training set X_{train} is calculated and denoted as the global threshold thr_G.

Operation phase During operation phase, the normalized comparison score $\hat{s}_{ij}$ should be computed to determine if sample i and j belong to the same identity. Firstly, the corresponding clusters for both samples are computed. The cluster thresholds for sample i and j are denoted as thr_i and thr_j. Secondly, these cluster thresholds, as well as the global threshold thr_G, are used to calculate the normalized score

$$\hat{s}_{ij} = s_{ij} - \frac{1}{2} \left(\Delta thr_i + \Delta thr_j \right), \tag{3}$$

where

$$\Delta thr_i = thr_i - thr_G, \tag{4}$$

describes the local-global threshold difference for sample i.

Discussion The goal of this score normalization is to introduce individual fairness in a biometric system and thus, reduce the discriminatory behavior of face recognition systems. The notation of individual fairness emphasizes that similar individuals should be treated similarly. We incorporate this statement in our normalization method using clustering and local thresholds. Clustering in the embedding space identifies similar individuals and local thresholds enable an approximate individual treatment.

Assuming that an identity $\mathcal{I}$, with samples i and j, belongs to a biased group. Consequently, their comparison score s_{ij} has a high probability to be low and thus, there is a high probability that the samples are verified incorrectly. Since the normalization method is aware of the bias, due to the local thresholds, the local-global thresholds difference Δthr is negative. Therefore, the normalized score $\hat{s}_{ij}$ will be higher than s_{ij} and thus, there is a higher chance that the samples get correct verification decisions.

For unbiased samples, the distinction performance is very precise and thus, the comparison score for a genuine is high, while the comparison scores for an imposter is very low. Since their local threshold will be higher than the global threshold, in this case, the normalized score will decrease. However, the final verification decisions will be the same, since the comparison scores of genuine pairs are relatively high.

Matching samples from groups with different bias-levels will lead to a normalized score close to the original scores and thus, there is a high chance that the comparison will be correctly labeled as an imposter.

4 Experimental setup

Database - In order to evaluate the face recognition performance of our approach under controlled and unconstrained conditions, we conducted experiments on the public available ColorFERET [25] and Adience [8] databases. ColorFERET [25] consists of 14,126 images from 1,199 different individuals with different poses under controlled conditions. Furthermore, a variety of face poses, facial expressions, and lighting conditions are included in the dataset. The Adience dataset [8] consists of over 26.5k images from over 2.2k different subjects under unconstrained imaging conditions. While Adience contains additional information about gender and age, ColorFERET also provides labels regarding the

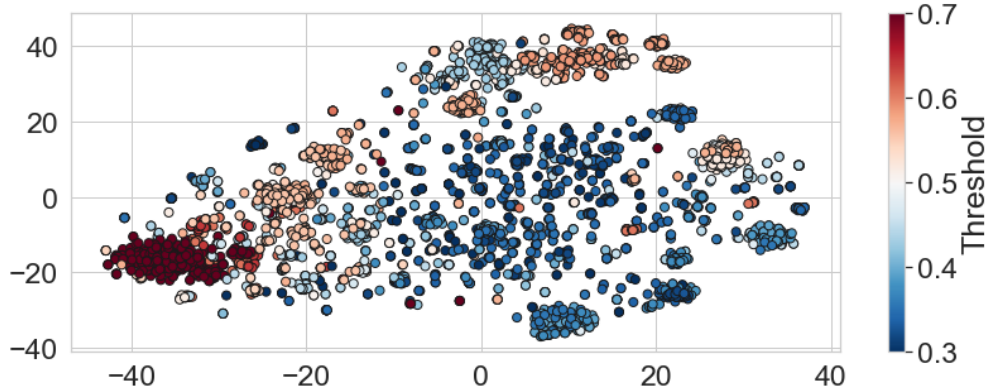


Figure 1: Visualizations of the Adience embeddings using t-distributed stochastic neighbor embedding (t-SNE) [30]. Each individual is represented as a point and each point is colored based on its optimal local threshold thr_k.

subjects ethnicities. In the experiments, this information is used to investigate the face recognition performance for several demographic groups.

Evaluation metrics - In this work, we will report our recognition performances in terms of false non-match rate (FNMR) at fixed false match rates (FMR). We also report the EER, which equals the FMR at the threshold where FMR = $1-$FNMR. The EER acts as a single-value indicator of the systems performance.

Workflow details For the comparison of two samples, firstly, both face images get aligned, scaled, and cropped as implemented in the Dlib-ml toolkit [18]. Secondly, the preprocessed images are passed into the FaceNet model[1] [27], which was pretrained on MS-Celeb-1M [12]. This creates 128-dimensional embeddings. Finally, the comparison of two embeddings was done using cosine-similarity. For all experiment scenarios, subject-disjoint 5-fold cross-validation is utilized.

Investigations - The goal of the proposed normalization approach is to reduce the demographic bias in face recognition decisions. To proof the individual steps of our solution, our experiments investigates the following issues:

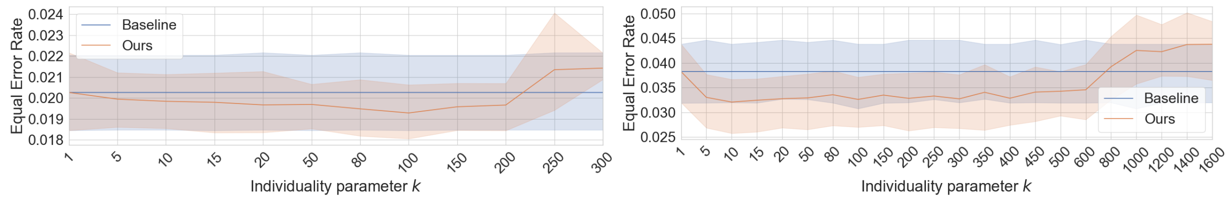
- Since our normalization method is based on the concept of individual fairness, a visual illustration is presented to demonstrate that this concept is suitable for face recognition.
- In order to justify our choice of the individual parameter k, a performance analysis is given over a wide parameter range.
- To point out the demographic bias issue and prove the effectiveness of our approach, the recognition performance is analyzed with and without our fairness-introducing normalization, under different security settings, and on different demographic groups.
- To evaluate the generalization of our approach, the experiments are conducted under controlled and uncontrolled scenarios (databases).

5 Results

Since our approach is based on the idea of individual fairness, we first want to visually demonstrate why this notation is suitable for face recognition. Figure 1 shows an t-SNE visualization of the embedding space for the dataset Adience. The t-SNE algorithm maps the high-dimensional embedding space into a two dimensional space such that similar samples in the high-dimension space lie closely together in two dimensions. Furthermore, each sample is colored based on the local thresholds computed by the proposed approach. Two observations can be made from this figure: first, it shows that there are several clusters with similar local thresholds in the embedding space. Consequently, our proposed approach is able to identify similar identities and to treat them similarly (through similar local thresholds). Second, it shows that the optimal thresholds for each cluster varies significantly from 0.3 to 0.7. This wide spread of optimal local thresholds demonstrates the need for a more individual treatment.

Figure 2 presents the effect of the individuality parameter k on the verification performance on both databases. Figure 2 shows the face recognition EER over a wide range of individuality parameters k. k here theoretically represents

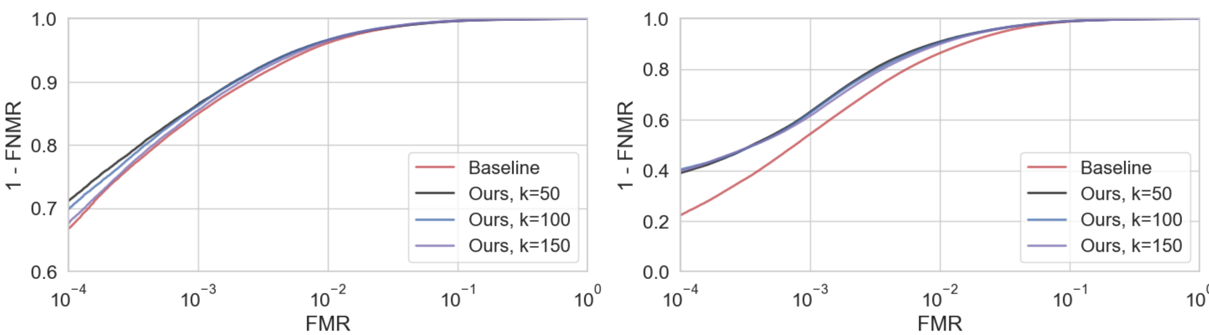
[1] https://github.com/davidsandberg/facenet



(a) ColorFERET (b) Adience

Figure 2: Recognition performance (EER) achieved when choosing several individuality parameter k. The proposed normalization approach (blue) is compared against the baseline (orange) without normalization on two datasets. The shaded areas represent the standard deviation over the 5 cross-validation folds.

different clusterable cases that appears in face images, such cases can be different demographic groups (as targeted by our paper), but also can represent different variations related to capture environment, pose, compression, among others. For $k = 1$, the normalization does not changes the scores and thus, the same performance is observed. For $k \geq 1$, the EER decreases, since our fair score normalization approach leads to a more individual treatment of each sample. If k is much bigger than the number of possible clusterable cases (detectable variations), more clusters would contain less samples and would represent less meaningful cases. Consequently, the local thresholds thr_k of these clusters become less reliable and thus, the goal of our proposed approach would not be optimally achieved.From Figure 2, it is noticeable that higher values of k keep achieving good performance in the Adience database in comparison to the ColorFERET database. This can be due to the larger size and the in-the-wild nature of the Adience database, resulting in a larger possibility of a higher number of clusterable cases.



(a) ColorFERET (b) Adience

Figure 3: ROC curves comparing our fair normalization approach for several individuality parameter k with the unnormalized baseline.

Figure 3 shows receiver operating characteristics (ROC) curves of the proposed normalization method for several individuality parameter k compared with the unnormalized baseline. These curves shows the FMR and 1-FNMR values at a wide range of operating points. It can be seen that our approach improves the recognition performance for all false match rates. Comparing both datasets, the achieved improvement by our normalization method is higher on the Adience dataset. This shows the effectiveness of our normalization method in challenging noise, pose, and lightning scenarios.

Different applications require varying security settings. In order to analyze the effectiveness of the proposed normalization method under diverse circumstances, Table 1 shows the FNMR at FMR ranging from 10^{-4} to 10^{-2}. Our approach improves the recognition performance by 8.3%-12.8% under controlled (ColorFERET) scenario and by 18.4%-30.7% under uncontrolled (Adience) scenario. These results demonstrates the effectiveness of our normalization approach, and thus individual fairness, for high security applications.

In order to demonstrate the fairness aspect of the proposed normalization method, Table 2 and 3 present the normalized and baseline (unnormalized) face recognition performance for the different attribute classes and thus, demographic groups. The baseline clearly shows decision bias between demographic groups. Furthermore, the table presents the improvement ratio introduced by our solution and the ratio of each demographic class in the database (sample distribution). Table 2 shows the face recognition performance on the ColorFERET dataset. It is observed that without exception for all gender, age, and ethnicity classes our normalization method leads to a clear improvement up to 10.4%.

Table 1: FNMR at several FMR for varying security conditions on both ColorFERET (CF) and Adience (AD) databases.

		FNMR@FMR					
		FMR	Improvement	FMR	Improvement	FMR	Improvement
		10^{-4}		10^{-3}		10^{-2}	
CF	Baseline	33.3 %	-	15.0 %	-	3.8 %	-
	Ours	30.2 %	**9.1 %**	13.8 %	**8.3 %**	3.3 %	**12.8 %**
AD	Baseline	77.7 %	-	45.7 %	-	13.6 %	-
	Ours	59.7 %	**29.6 %**	37.3 %	**18.4 %**	9.4 %	**30.7 %**

Table 3 presents the face verification performance for several gender and age classes on the Adience benchmark. The images of the Adience dataset were captured under uncontrolled conditions. Therefore, the error rates for this database are higher than the error rates of ColorFERET. Concerning the gender classes, it can be observed that applying the proposed fair normalization methodology leads to a significant face recognition improvement of 17.5% and 12.4% for male and female respectively. For the majority of the age classes (81.4% of the dataset), the recognition performance increases up to 31%. Especially for the subgroup class (age 48-53), with only 5% of the database samples belonging to this group, a performance increase of 31.6% is observed. In total, applying the proposed fair normalization approach leads to enhanced verification performance across different demographic groups.

Table 2: ColorFERET: EER values with and without our proposed fairness approach for different demographic groups.

		EER			
Attribute	Class	Baseline	Ours	Improvement	Sample distribution
Gender	Male	1.67 %	1.54 %	**7.78 %**	64.63 %
	Female	3.00 %	2.83 %	**5.67 %**	35.37 %
Age	10-20	3.07 %	2.75 %	**10.42 %**	34.55 %
	21-30	2.16 %	2.00 %	**7.41 %**	23.84 %
	31-40	1.59 %	1.53 %	**3.77 %**	5.64 %
	41+	1.14 %	1.09 %	**4.39 %**	35.97 %
Ethnicity	Asian	2.02 %	1.94 %	**3.96 %**	23.30 %
	Black	2.81 %	2.72 %	**3.20 %**	7.82 %
	Other	2.58 %	2.42 %	**6.20 %**	6.40 %
	White	1.83 %	1.81 %	**1.09 %**	62.48 %
	All (EER)	2.03 %	1.93 %	**4.93 %**	100.00 %

6 Conclusion

Despite the progress achieved by current face recognition systems, which are optimized towards the total recognition accuracy only. Recent works showed that biometric systems impose a strong bias against subgroups of the population. Consequently, there is an increased need for solutions that increase the fairness of such systems. Previous work proposed fairness-enhancing solutions at the cost of degraded overall recognition performances. In this work, we propose a novel fair score normalization approach that is explicitly designed to mitigate decision bias against variations in the processed data (including demographic groups). Integrating the idea of individual fairness, our solution aims at treating similar individuals similarly. Our experiments were conducted on two public available datasets captured under controlled and in-the-wild conditions. The results showed an enhancement of the overall system performance by up to 14.77% under intermediate security settings and an improvement of 30.7% under high security settings. Since this approach aims at reducing the bias in face recognition, the results showed a significant reduction in the recognition error rates of all demographic groups. Under in-the-wild conditions, the recognition performance of the effected subgroups were increased by up to 31.62%. Unlike previous work, our solutions does not need demographic information about individuals and is not only limited to face biometrics.

Table 3: Adience: EER values with and without our proposed fairness approach for different demographic groups.

		EER			
Attribute	Class	Baseline	Ours	Improvement	Sample distribution
Gender	Male	4.29 %	3.54 %	**17.48 %**	52.9 %
	Female	3.38 %	2.96 %	**12.43 %**	47.1 %
Age	0-2	4.39 %	4.53 %	-3.19 %	8.6 %
	4-6	5.30 %	5.23 %	**1.32 %**	13.0 %
	8-12	5.03 %	3.95 %	**21.47 %**	12.9 %
	15-20	3.86 %	4.25 %	-10.10 %	10.0 %
	25-32	2.67 %	2.21 %	**17.23 %**	30.8 %
	38-43	2.35 %	2.25 %	**4.26 %**	14.3 %
	48-53	1.36 %	0.93 %	**31.62 %**	5.0 %
	60-100	1.11 %	0.94 %	**15.32 %**	5.3 %
	All	3.86 %	3.29 %	**14.77 %**	100.0 %

Acknowledgments

This research work has been funded by the German Federal Ministery of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the Counterdrug Technology Development Program Office.

References

[1] M. S. Alvi, A. Zisserman, and C. Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *CoRR*, abs/1809.02169, 2018.

[2] B. Berendt and S. Preibusch. Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 344–351, Dec 2012.

[3] B. Berendt and S. Preibusch. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law*, 22(2):175–209, Jun 2014.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[5] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, Sep 2010.

[6] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In L. Leal-Taixé and S. Roth, editors, *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11129 of *Lecture Notes in Computer Science*, pages 573–585. Springer, 2018.

[7] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

[8] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Trans. Info. For. Sec.*, 9(12):2170–2179, Dec. 2014.

[9] N. Furl, P. J. Phillips, and A. J. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26:797–815, 2002.

[10] C. Garvie, G. U. C. on Privacy, Technology, and G. U. L. C. C. on Privacy & Technology. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

[11] B. Goodman and S. Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation", 2016. cite arxiv:1606.08813Comment: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

[12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *CoRR*, abs/1607.08221, 2016.

[13] S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté. Discrimination prevention in data mining for intrusion and crime detection. In *2011 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, pages 47–54, April 2011.

[14] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[15] A. K. Jain, K. Nandakumar, and A. Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016.

[16] S. Jia, T. Lansdall-Welfare, and N. Cristianini. Right for the right reason: Training agnostic networks. *CoRR*, abs/1806.06296, 2018.

[17] T. Kehrenberg, Z. Chen, and N. Quadrianto. Interpretable fairness via target labels in gaussian process models. *CoRR*, abs/1810.05598, 2018.

[18] D. E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, Dec. 2009.

[19] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, Dec 2012.

[20] V. Mirjalili, S. Raschka, A. M. Namboodiri, and A. Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. *CoRR*, abs/1712.00321, 2017.

[21] V. Mirjalili, S. Raschka, and A. Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. *CoRR*, abs/1807.11936, 2018.

[22] A. Morales, J. Fiérrez, and R. Vera-Rodríguez. Sensitivenets: Learning agnostic representations with application to face recognition. *CoRR*, abs/1902.00334, 2019.

[23] M. Orcutt. Are face recognition systems accurate? depends on your race. *MIT Technology Review 2016*, 2016.

[24] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.*, 8(2):14:1–14:11, Feb. 2011.

[25] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, Oct 2000.

[26] E. Raff and J. Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In F. Bonchi, F. J. Provost, T. Eliassi-Rad, W. Wang, C. Cattuto, and R. Ghani, editors, *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 189–198. IEEE, 2018.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[28] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper. Suppressing gender and age in face templates using incremental variable elimination. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. IEEE, 2019.

[29] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper. Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. *Applied Intelligence*, Feb 2019.

[30] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[31] P. Voigt and A. v. d. Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[32] M. Wang, W. Deng, J. Hu, J. Peng, X. Tao, and Y. Huang. Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. *CoRR*, abs/1812.00194, 2018.

[33] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.