How (Not) to Measure Bias in Face Recognition Networks

Stefan Glüge[1 [0000−0002−7484−536X]], Mohammadreza Amirian[1,2 [0000−0003−0047−6802]], Dandolo Flumini[1 [0000−0003−3699−4871]], and Thilo Stadelmann[1,3 [0000−0002−3784−0420]]

[1] Zurich University of Applied Sciences ZHAW, Datalab, Wädenswil and Winterthur, Switzerland
{glue, amir, flum, stdm}@zhaw.ch
[2] Ulm University, Institute of Neural Information Processing, Ulm, Germany
[3] European Centre for Living Technology ECLT, Venice, Italy

Abstract. Within the last years Face Recognition (FR) systems have achieved human-like (or better) performance, leading to extensive deployment in large-scale practical settings. Yet, especially for sensible domains such as FR we expect algorithms to work equally well for everyone, regardless of somebody's age, gender, skin colour and/or origin. In this paper, we investigate a methodology to quantify the amount of bias in a trained Convolutional Neural Network (CNN) model for FR that is not only intuitively appealing, but also has already been used in the literature to argue for certain debiasing methods. It works by measuring the "blindness" of the model towards certain face characteristics in the embeddings of faces based on internal cluster validation measures. We conduct experiments on three openly available FR models to determine their bias regarding race, gender and age, and validate the computed scores by comparing their predictions against the actual drop in face recognition performance for minority cases. Interestingly, we could not link a crisp clustering in the embedding space to a strong bias in recognition rates—it is rather the opposite. We therefore offer arguments for the reasons behind this observation and argue for the need of a less naïve clustering approach to develop a working measure for bias in FR models.

Keywords: Deep learning · convolutional neural networks · fairness

1 Introduction

FR has improved considerably and constantly over the last decade [25, 40, 13, 17], giving rise to numerous applications ranging from services on mobile consumer devices, applications in sports, to the use by law enforcement agencies [43, 42, 35, 32]. The increased deployment has triggered an intense debate on the ethical downsides of pervasive use of biometrics [6, 34, 29, 39] up to the point where regulation [23] and bans on the technology are discussed[1] and partially enforced[2].

[1] https://www.banfacialrecognition.com/
[2] https://www.bbc.com/news/technology-48276660

This debate on ethical usage of FR technology is part of a larger trend in the machine learning field to account for ethical aspects of the methodology [7, 28], which includes the aspects of trustworthiness[3], transparency (interpretability) [16, 3] and fairness (bias) [4].

The issue of *bias* in machine learning is especially relevant in the area of FR, where we legitimately expect machine learning models to be unbiased because of their potentially large impact (e.g., for crime prediction [20]). The huge diversity due to race, gender and age in the appearance of human faces is however contrasted by a respective homogeneity of the data collections used to train such models. This leads to observations like the one that face recognition only works reliably for white grown-up males [8]. As working face recognition is increasingly relied on to grant individuals access to services and locations, and to predict people's behaviour, bias against certain people groups easily results in prohibitive discrimination.

The source of bias is usually the training material. Therefore, the community created datasets with known biases for race, skin color, gender and age, such as Racial Faces in-the-Wild (RFW) [45] and Diversity in Faces [33]. Given the bias in the data we are able to study the issue in the final models on two concrete levels: by (a) *quantifying* the amount of bias that exists in any trained FR model; and by (b) *reducing* identified bias in models by adequate countermeasures.

In this paper, we perform an in-depth exploration of a certain methodology for measuring the specific amount of bias that exists in any trained FR CNN. The underlying idea is appealing due to its intuitive approach and similar reasoning has already been used to argue for specific bias removal algorithms in the past [2]. The quantification itself relies on internal cluster validation measures for clusterings of embeddings based on labels for race, gender and age. It is agnostic towards the specific architecture and training procedure of the model and thus applicable to any FR system that exposes its embeddings; it is also non-invasive with respect to model training and does not effect the model's performance. Counterintuitively, our experiments speak against the validity of the idea and confirm the contrary: higher bias, as expressed in a drop in face recognition accuracy for minority cases, goes along with worse clustering, i.e. less "awareness" / more "blindness" of the model with respect to distinguishable features of the respective minority. We thus offer potential reasons for our observations, leading to preliminary results on how to better quantify bias in FR.

2 Related Work

The problem of bias in machine learning is well documented in the literature [30]. Several reasons trigger this bias: bias in the final decision as imposed by algorithm design, training process and loss design is addressed by the term *algorithmic bias*, though the term can be problematic[4]. *Selection bias* is introduced

[3] https://liu.se/en/research/tailor/
[4] https://stdm.github.io/Algorithmic-bias/

when human biases lead to selecting the wrong algorithm or data for FR, leading to biased decisions. *Data bias* finally is introduced by a lack of diversity or present imbalance in a dataset used for training or evaluating a model.

The presence of bias in model predictions for FR – leading to discrimination against certain factors such as race, age and gender of individuals – motivates two strands of recent research: (a) to automatically quantify the amount of bias in a model, and (b) to reduce it by a range of methods. Regarding bias measurement (a), Hannak et al. give criteria to accurately measure bias with respect to price discrimination [19]. Garcia et al. identify demographic bias by investigating the drop in confidence of face matching models for certain ethnicities [15]. Cavazos et al. use three different identification thresholds, namely the thresholds at equal false accept rates (FARs) and the recognition accuracy, to quantify racial bias for face matching [11]. Serna et al. show that FR bias is measurable using normalized overall activation of the models for different races [41]. In this paper, we explore a novel method to measure (quantify) bias that differs threefold from these approaches: (i) it is applicable to *any* model that exposes its embeddings, (ii) it is independed of model training and (iii) it is not based on model performance, but rather on the way faces are represented in the network.

Regarding bias reduction (b), most of the research in FR aims at tackling racial bias. However, Li et al. propose optimizing a distance metric for removing age bias. The remainder of the literature focuses on racial bias by improving both algorithmic and data biases [26]. Steed and Caliskan attempt to predict appearance bias of human annotators using transfer learning to estimate the bias in datasets [44], and Kortylewski el al. introduce synthetic data to reduce the negative effects of imbalance in datasets [22]. Yu et al. propose an adaptive triplet selection for correcting the distribution shift and model bias [47]. Robinson et al. show that the performance gaps in FR for various races can be reduced by adapting the decision thresholds for each race [36]. Domain transfer and adversarial learning are the other methods to reduce racial bias by adapting the algorithms. Wang et al. use a deep Information Maximization Adaptation Network (IMAN) for unsupervised knowledge transfer from the source domain (Caucasian) to target domains (other races) [45]. To remove the statistical dependency of the learned features to the source of bias (racial group), Adeli et al. propose an adversarial loss that minimizes the correlation between model representations and races [1]. Finally, Wang et al. [46] propose an update to the "Hard Debias" algorithm that post-processes word embeddings for unbiased text analysis and state that the idea might be transferable to other domains.

3 An Intuitively Appealing Method to Measure Bias

Human FR is not unbiased at all: we recognize faces that are most familiar much better than others. This "other-race effect" is one of the most robust empirical findings in human FR and accompanied by the popular belief that other-race faces all look alike [31]. The source of this drop in recognition performance for faces of unfamiliar origin seems to be that we know a rich feature set to distin-

guish between akin faces, but only know very coarse features for very differently looking faces. This results in the effect, that unfamiliar races appear to be different in general, which overlays how they differ amongst each other.

Humans associate the presence of bias with an observed *drop in recognition performance*; and the models seem to be the more biased the more *aware of the differences* between certain facial characteristics that are associated with potential discrimination. The method for measuring bias that we are concerned with in this paper builds upon both observations by exploiting them in the following way: first (a), bias in a specific model and for a specific characteristics (e.g., age, gender or race) is measured by quantifying how well the embeddings of a set of faces build clusters with respect to this characteristic. A good clustering into, for instance, age groups suggests that the model is very *aware* of the differences in age, which enables it to potentially discriminate age groups (in the two-fold meaning). Then (b), the resulting "score" is verified by experimentally checking for a *drop in FR performance* for faces with minority expressions for this characteristic. Alvi et al. argue along these lines in order to demonstrate the effect of an algorithm to remove bias: 'After unlearning gender, the feature representation is no longer separable by gender, demonstrating that this bias has been removed.'[2].

3.1 Quantifying Bias through Internal Cluster Validation Measures

A straight-forward way to perform the respective bias quantification in (a) is to use existing cluster validity measures on the embeddings of FR models. The embeddings, usually taken as the activations of the last fully connected layer of a trained CNN during evaluation on a respective image, form a mapping of a face image into a compact Euclidean space where distances directly correspond to a measure of face similarity [40]. As the internal representation of the model contains the facial discriminant information, its embedding forms the basis for all recognition, similarity computation etc. A model with embeddings which do not cluster well with respect to a certain facial characteristics can be said to be "blind" towards the features that distinguish between its different expressions, which seems to be a good starting point for unbiasedness.

To eliminate the effect of many hyperparameters on the evaluation of the methodology, we rely on ground truth labels (either human provided or predicted by reference models) rather than a clustering algorithm for the membership of embeddings to clusters of specific discriminative characteristic. Hence, cluster membership is dependent only on the characteristics of the dataset itself and not on the FR model under evaluation and the only model-dependent part entering the bias measurement are the embeddings themselves. How well they cluster can then be quantified by so-called internal cluster validation measures [27] that are well established to measure the "goodness" of a clustering compared to other ones. Internal cluster validation measures are the correct ones to use because regardless of the source of our cluster memberships, we want to compare different clusterings with each other and not a clustering to ground truth. Generally, the indices are suitable for measuring crisp clustering, where no

overlap between partitions is allowed [24]. For our evaluation, we compute the Mean Silhouette Coefficient [38], Calinski-Harabasz Index [9], Davies-Bouldin Index [12] and Dunn Index [14]. However, for the Dunn Index we observe very small values and large variance in all experiments, resulting in no meaningful distinctions between different FR models. Therefore, those results are omitted.

The *Mean Silhouette Coefficient* is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It is bounded between -1 and $+1$, whereas *scores around zero indicate overlapping clusters.* Negative values indicate that there may be too many or too few clusters, and *positive values towards* 1 *indicate well separable clusters.* The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficients per sample. The *Calinski-Harabasz Index*, also known as the Variance Ratio Criterion, is defined as the ratio of the between-cluster variance and the within-cluster variance. Well-defined clusters have large between-cluster and a small within-cluster variance, i.e. a *higher score relates to a model with better defined clusters.* Finally, the *Davies-Bouldin Index* computes for each cluster the other cluster that it is most similar to. Afterwards, it summarizes the maximum cluster similarities to create a single index. A low index indicates that the clusters are not very similar, i.e. *a low value relates to a model with better separation between the clusters.*

3.2 Models, Dataset and Experimental Setup

In the following, we describe our experimental setup to (a) measure bias in the embedding space based on internal cluster validation measures, and (b) validate the resulting score based on the drop in face recognition performance on a benchmark dataset. We choose three different FR models from the popular and well-established Visual Geometry Group (VGG) family, openly available from the VGG, perform measurements and validate on the RFW dataset to study the bias for race, gender and age.

We use trained models that are available directly from the authors[5]. They were pretrained on the MS-Celeb-1M [18] dataset and then fine-tuned on VGGFace2, which contains 3.31 million images of 9 131 identities [10]. All models follow the SE-ResNet-50 architectural configuration in [21], but differ in the dimensionality of embedding layer (128D/256D) which is stacked on top of the original final feature layer (2048D) adjacent to the classifier. All models were trained with standard softmax loss.

The RFW dataset was designed to study racial bias in FR systems [45]. It is constructed with four testing subsets, namely Caucasian, Asian, Indian and African. Each subset contains about 10k images of 3k individuals for face verification. We further added a gender and age label to each test image of the RFW dataset using a Wide Residual Network trained on the UTKFace [49] and IMDB-WIKI [37] datasets[6]. The age prediction is in the range of $0-100$. For the cluster evaluation, we split the age predictions into the three non-overlapping

[5] https://github.com/ox-vgg/vgg_face2
[6] https://github.com/yu4u/age-gender-estimation

groups < 30, 30–45 and 45+. The boundaries are chosen such that we have at least $3,000$ samples in each class. Gender prediction follows the same procedure as age prediction. The model yields a continuous gender score s_{gender} between 0 and 1, whereas lower values indicate male and higher values indicate female. In order to use all samples from the dataset, we split it at $s_{\text{gender}} < 0.5$ for male and $s_{\text{gender}} > 0.5$ for female. Tab. 1 gives an overview of the resulting number of samples per cluster with respect to the characteristic race, age and gender. As one can see, the race clusters are nicely balanced, whereas for gender we have a strong imbalance towards "male", and for age the 30–45-group is dominant.

Table 1: Number of samples per cluster regarding different facial characteristics in the RFW dataset that are associated with bias.

Face characteristic	**Clusters**	**#samples**
Race (human annotation)	Caucasian; Indian; Asian; African	$10,099$; $10,221$ $9,602$; $10,397$
Age [years] (predicted)	< 30; 30–45; 45+	$4,815$; $32,530$; $3,046$
Gender (predicted)	male; female	$28,928$; $11,463$

For our evaluation we extract the embeddings of the approximately 40k face images from the RFW testset for each of the VGG2 models. Face detection and alignment is done using the MTCNN approach[7] proposed by Zhang et al. [48]. Based on the embeddings, we report the FR rates and the cluster validation measures as per the dimensions race, gender and age. For face recognition, we report a match if the sample that is the nearest neighbor to the test face comes from the same person.

4 Results

So far, we have discussed two proxies for quantifying bias. (a) *goodness of clustering* of the embeddings w.r.t to the different expressions of a facial characteristic like age, gender or race—the higher, the more bias. It can be measured for any model that exhibits its embeddings, given that a dataset with labels for these expressions exists. It is thus a candidate for a *measurement methodology* to quantify bias in general. (b) *face recognition rate* for cases that belong to the minority expression of said characteristics—the lower, the more bias. This approach needs labels and multiple samples of persons, but serves as a measure of the real-world impact of bias/discrimination (as people from minority groups are less well handled); it is thus our candidate to *validate* the bias as measured by proxy (a).

Tab. 2 shows face recognition rates per model and expressions of facial characteristic (b), alongside the introduced cluster validation indices (a). We highlight the best recognition rates and the lowest percentual difference compared to the mean over the different expressions of a characteristic (i.e., lowest actual bias).

[7] https://github.com/YYuanAnyVision/mxnet_mtcnn_face_detection

Further, we highlight the worst clustering according to each index (i.e., lowest measured bias as predicted by the method under consideration here).

Table 2: Bias measurement results (bad Clustering Score) vs. bias validation results (good Recognition Rate and low Relative Difference) per model and characteristic/expression. Clustering scores are the Mean Silhouette Coefficient (MS), Calinski-Harabasz Index (CH) and Davies-Bouldin Index (DB); ↑ and ↓ depict if a high or low value indicate a good clustering, respectively. Lowest bias according to each type of score is highlighted.

Architecture (#features)	**Metric**	**Race**				**Expr. Avg.**	**Clustering Score**		
		Caucasian	Indian	Asian	African		MS ↑	CH ↑	DB ↓
VGG2 (128)	Rec. Rate	**0.8906**	**0.8531**	**0.8310**	**0.7998**	**0.8436**	0.062	1,812	3.85
	Rel. Diff. (%)	5.5648	1.1271	-1.5011	-5.1907	-			
VGG2 (256)	Rec. Rate	0.8787	0.8265	0.7981	0.7597	0.8158	**0.029**	**766**	**6.20**
	Rel. Diff. (%)	7.7157	1.3208	-2.1693	-6.867	-			
VGG2 (2048)	Rec. Rate	0.8799	0.8472	0.8305	0.7959	0.8384	0.050	1,473	4.49
	Rel. Diff. (%)	**4.9542**	**1.0523**	**-0.9427**	**-5.0638**	-			
		Gender							
		Male	Female						
VGG2 (128)	Rec. Rate	**0.8381**	**0.8576**			**0.8479**	0.0048	135.3	15.56
	Rel. Diff. (%)	**-1.1507**	**1.1507**			-			
VGG2 (256)	Rec. Rate	0.8088	0.833			0.8211	**0.0017**	**64.44**	**22.55**
	Rel. Diff. (%)	-1.5039	1.5039			-			
VGG2 (2048)	Rec. Rate	0.8327	0.8525			0.8426	0.0063	143.7	15.11
	Rel. Diff. (%)	-1.1725	1.1725			-			
		Age							
		< 30	30–45	45+					
VGG2 (128)	Rec. Rate	**0.8671**	**0.8358**	**0.8907**		**0.8645**	0.0002	21.92	34.34
	Rel. Diff. (%)	0.2971	-3.3234	3.0263		-			
VGG2 (256)	Rec. Rate	0.8424	0.8063	0.8749		0.8412	**-0.0003**	**8.51**	**52.90**
	Rel. Diff. (%)	0.139	-4.1481	4.009		-			
VGG2 (2048)	Rec. Rate	0.8638	0.8305	0.8821		0.8588	0.0006	25.40	32.26
	Rel. Diff. (%)	**0.5789**	**-3.2981**	**2.7192**		-			

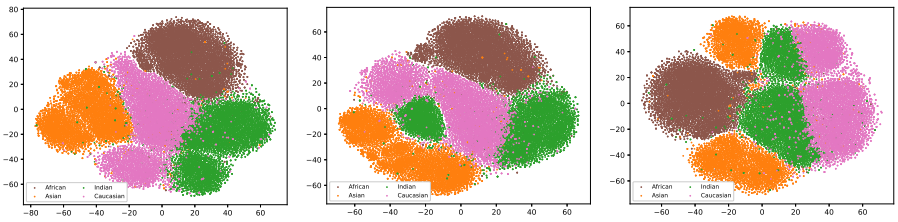
For race and age, the VGG2 (128) model performs best regarding pure recognition rates, whereas VGG2 (2048) shows the lowest performance drop, i.e. is the least biased model regarding race and age. For gender we observe only a marginal difference in recognition rates and performance drops between those models. The VGG2 (256) model is the worst option with respect to recognition rates as well as performance drops (actual bias). Looking at the clustering scores (measured bias), much to our surprise, this same VGG2 (256) model produces the worst clustering with respect to all validation indices and therefore can be considered to be the model with the least distinctive face representations regarding race, gender and age (lowest measure bias). However, this is not reflected in the class-based performance drop for the recognition rates (that should be small). Regarding the age groups, the Mean Silhouette Coefficient takes very small or negative values which indicates overlapping and/or an unsuitable number of clusters. Given the continuous nature of age, this is to be expected.

5 Discussion and Conclusions

In general, we could not link a crisp clustering in the embedding space to a strong bias in the recognition rates. In our experiments we found quite the opposite. This spawns discussion on three levels:

First, on the level of the *underlying reasoning*: it needs to be checked how much the two proxies used here to quantify bias (namely, face recognition performance on minority examples as a sign for the practical effect of bias; and well-defined clusters in the embedding space with respect to typically biased characteristics as a way to measure / predict this real-world influence) are actually correlated with what is meant by "bias". This is especially relevant in the light of the fact that parts of this reasoning have already been adopted in the literature as a way to show the effectiveness of debiasing algorithms.

Second, on the level of *implementation*: Even if the notion of bias is reflected in the embedding space, the adopted naïve clustering approach using only broad expression types for races, gender and age groups can be reconsidered. We hypothesize that a cluster like "male" or "African" is too general and rather formed of multiple sub-clusters in the embedding space. Thus, a cluster validation index on male/female cannot reflect the actual awareness of the model of these ultimately relevant sub-clusters. This intuition is supported by a visualisation of the embeddings as show in Fig. 1. For all three models, one can see that an expression of race such as "Caucasian" is comprised of at least two sub-clusters. However, one has to keep in mind that the t-SNE representation is just a projection into 2D space from the original 128/256/2048D space and generates slightly different results each time on the same data set. Furthermore, for the age characteristic, the distribution of embeddings into three clusters was somewhat arbitrarily based on a balance argument and could be chosen differently. Additional "hyperparameters" of the methodology and hence candidates for further experiments are the tested model architectures and their training details (especially the used loss functions have a large effect on the embeddings and the space spanned by them).



(a) (b) (c)

Fig. 1: T-SNE visualisations of embeddings from VGG2 (128) (a), VGG2 (256) (b) and VGG2 (2048) (c). The samples are colored according to their race.

Third, on the level of *insight / explanation*: focusing now solely on the example of racial bias for illustrative reasons, the idea of how to measure bias in

this paper relies on the the assumption that the main source of bias in FR is the separation of races in the embedding space (the better separated, the more awareness, hence the more biased); this could be measured by clustering quality with respect to different expressions of race. The failure to observe any such correlation could be due to, we conjecture, the between-cluster separation being less important to explain bias than the *within-cluster distribution* (i.e., it doesn't mean too much how e.g. "Africans" are separated from "Asians" in the embedding space – it is much more important how the "African" embeddings are distributed amongst each other). To underpin this hypothesis, we present the distribution of pairwise distances between test embeddings from various races in Fig. 2.

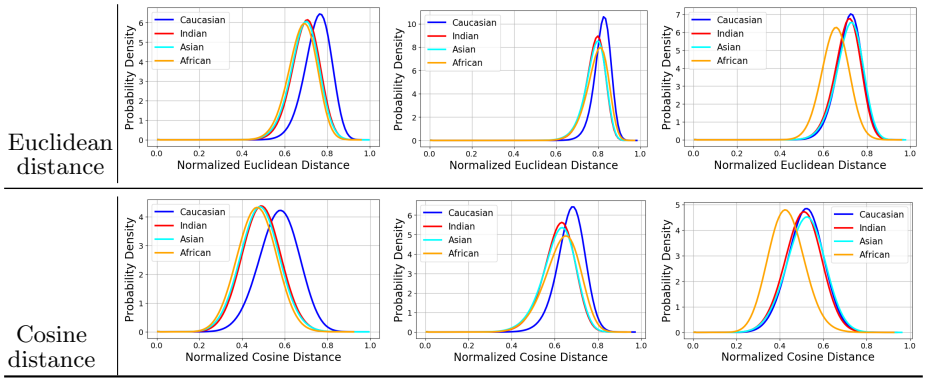


Fig. 2: Probability density distribution of pairwise (Euclidean and Cosine) distances between test image embeddings of different races.

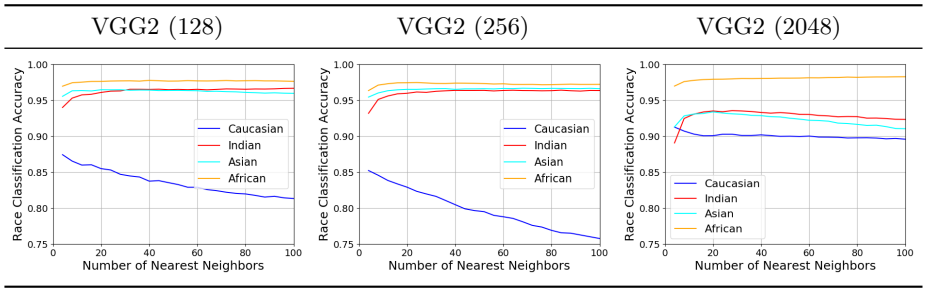


Fig. 3: Accuracy of k-nearest neighbors classifiers for race classification in the embedding space.

Figure 2 allows the conclusion that *the average distance of the embeddings for "Caucasians" is higher* than the one for other races. This means in turn

that the embeddings of "Caucasians" are distributed with a lower density in the embedding space. At the same time, these embeddings are the ones with the best recognition accuracies. The same observation is supported by the experiment behind Fig. 3 that uses the k-nearest neighbor classifier with varying k for *race* classification. "Africans" show the highest race recognition rate, suggesting that the embeddings are concentrated with high density in a specific region (similar to the lower average pairwise distance according to Fig. 2). The race recognition accuracy of "Caucasian" embeddings appropriately is the smallest and drops with the number of nearest neighbors, suggesting a low-density distribution of the embeddings for this race.

In summary, we presented an intuitively motivated idea on how to measure bias in any existing FR model that exposes its embeddings, and how to validate it based on FR accuracy. A similar reasoning has been used in the past in the literature to argue for the benefits of certain debiasing methods. This is why the presented results, though "negative" (they did not confirm the validity of the method, but testified to the opposite effect), are still very important: they show that similarly to what is known as the "curse of dimensionality" [5], intuition fails in this complex scenario, and assumptions need to be more thoroughly checked. Nevertheless, the given explanatory approaches show a way to turn the underlying reasoning into usable measures of bias in the future.

Future work will thus first focus on finding answers to the questions raised in the discussion. Then, a next step is to *calibrate* any resulting measure: to have the differently scaled clustering indices combined into a single bounded measure between, say, -1 and 1 which allows interpretations similar to the meaning the correlation coefficient can provide (i.e, certain ranges of values mean "biased" or "bias-free").

Acknowledgements

The authors are grateful for the support through CTI grant 25256.1 PFES-ES "LIBRA" and the collaboration with Gilbert F. Duivesteijn.

References

1. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M.: Bias-resilient neural network. ArXiv **abs/1910.03676** (2019)
2. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Workshop on Bias Estimation in Face Analytics, ECCV. pp. 556–572 (2018)
3. Amirian, M., Schwenker, F., Stadelmann, T.: Trace and detect adversarial attacks on cnns using feature response maps. In: ANNPR. pp. 346–358 (2018)
4. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 1–15 (2019)
5. Bellman, R.: Dynamic Programming. Princeton University Press (1957)

6. Bernal, P.: Data gathering, surveillance and human rights: recasting the debate. Journal of Cyber Policy **1**(2), 243–264 (2016)
7. Brundage, M., Avin, S., Clark, J., Toner, H., et al.: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. ArXiv **abs/1802.07228** (2019)
8. Buolamwini, J.A.: Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Master's thesis, MIT (2017)
9. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974)
10. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: Int. Conf. on Automatic Face Gesture Recognition. pp. 67–74 (2018)
11. Cavazos, J.G., Phillips, P.J., Castillo, C.D., O'Toole, A.J.: Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? ArXiv **abs/1912.07398** (2019)
12. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979)
13. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699 (2019)
14. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics **4**(1), 95–104 (1974)
15. Garcia, R.V., Wandzik, L., Grabner, L., Krueger, J.: The harms of demographic bias in deep face recognition research. In: ICB. pp. 1–6 (2019)
16. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: Xai—explainable artificial intelligence. Science Robotics **4**(37) (2019)
17. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Computer Vision and Image Understanding **189**, 102805 (2019)
18. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV. pp. 87–102 (2016)
19. Hannak, A., Soeller, G., Lazer, D., Mislove, A., Wilson, C.: Measuring price discrimination and steering on e-commerce web sites. In: Conf. on Internet Measurement Conference. pp. 305–318 (2014)
20. Hashemi, M., Hall, M.: Criminal tendency detection from facial images and the gender bias effect. Journal of Big Data **7**(1), 1–16 (2020)
21. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
22. Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T.: Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: CVPR (2019)
23. Learned-Miller, E., Ordóñez, V., Morgenster, J., Buolamwini, J.: Facial recognition technologies in the wild: A call for a federal office. Tech. rep., Algorithmic Justice League (05 2020)
24. Legány, C., Juhász, S., Babos, A.: Cluster validity measurement techniques. In: Int. Conf. on Artif. Intell., Knowl. Engin. and Data Bases. pp. 388–393 (2006)
25. Li, S., Jain, A.: Handbook of Face Recognition. Springer London (2011)
26. Li, Y., Wang, G., Nie, L., Wang, Q., Tan, W.: Distance metric optimization driven convolutional neural network for age invariant face recognition. Pattern Recognition **75**, 51–62 (2018)
27. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: Int. Conf. on Data Mining. pp. 911–916 (2010)
28. Loi, M., Heitz, C., Christen, M.: A comparative assessment and synthesis of twenty ethics codes on AI and big data. In: Swiss Conference on Data Science (2020)

29. Mann, M., Smith, M.: Automated facial recognition technology: Recent developments and approaches to oversight. University of New South Wales Law Journal **40**, 121–145 (2017)
30. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ArXiv **abs/1908.09635** (2019)
31. Meissner, C.A., Brigham, J.C.: Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review. Psychology, Public Policy, & Law pp. 3–35 (2001)
32. Merler, M., Mac, K.N.C., Joshi, D., et al.: Automatic curation of sports highlights using multimodal excitement features. IEEE Trans. on Multimedia **21**(5), 1147–1160 (2019)
33. Merler, M., Ratha, N.K., Feris, R.S., Smith, J.R.: Diversity in faces. ArXiv **abs/1901.10436** (2019)
34. Norval, A., Prasopoulou, E.: Public faces? a critical exploration of the diffusion of face recognition technologies in online social networks. New media & society **19**(4), 637–654 (2017)
35. Robertson, D.J., Noyes, E., Dowsett, A.J., Jenkins, R., Burton, A.M.: Face recognition by metropolitan police super-recognisers. PlOS ONE **11**, 1–8 (2 2016)
36. Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: Face recognition: too bias, or not too bias? ArXiv **abs/2002.06483** (2020)
37. Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. of Computer Vision **126**(2-4), 144–157 (2018)
38. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comp. and Applied Mathematics **20**, 53–65 (1987)
39. Royakkers, L., Timmer, J., Kool, L., van Est, R.: Societal and ethical issues of digitization. Ethics and Information Technology **20**(2), 127–142 (2018)
40. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
41. Serna, I., Peña, A., Morales, A., Fierrez, J.: InsideBias: Measuring bias in deep networks and application to face gender biometrics. ArXiv **abs/2004.06592** (2020)
42. Smith, D.F., Wiliem, A., Lovell, B.C.: Face recognition on consumer devices: Reflections on replay attacks. IEEE Trans. on Information Forensics and Security **10**(4), 736–745 (2015)
43. Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., et al.: Deep learning in the wild. In: IAPR ANNPR. pp. 17–38. Springer (2018)
44. Steed, R., Caliskan, A.: Machines learn appearance bias in face recognition. ArXiv **abs/2002.05636** (2020)
45. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: ICCV. pp. 692–702 (2019)
46. Wang, T., Lin, X.V., Rajani, N.F., McCann, B., et al.: Double-hard debias: Tailoring word embeddings for gender bias mitigation. ArXiv **abs/2005.00965** (2020)
47. Yu, B., Liu, T., Gong, M., Ding, C., Tao, D.: Correcting the triplet selection bias for triplet loss. In: Computer Vision – ECCV 2018. pp. 71–87 (2018)
48. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
49. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR. pp. 4352–4360 (2017)