Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics

Ignacio Serna,[1] Aythami Morales,[1] Julian Fierrez,[1] Manuel Cebrian,[2] Nick Obradovich,[2] Iyad Rahwan[2]
[1]Universidad Autonoma de Madrid, Madrid, Spain
[2]Max Planck Institute for Human Development, Berlin, Germany
[1]{ignacio.serna, aythami.morales, julian.fierrez}@uam.es
[2]{cebrian, obradovich, sekrahwan}@mpib-berlin.mpg.de

Abstract

The most popular face recognition benchmarks assume a distribution of subjects without much attention to their demographic attributes. In this work, we perform a comprehensive discrimination-aware experimentation of deep learning-based face recognition. The main aim of this study is focused on a better understanding of the feature space generated by deep models, and the performance achieved over different demographic groups. We also propose a general formulation of algorithmic discrimination with application to face biometrics. The experiments are conducted over the new DiveFace database composed of 24K identities from six different demographic groups[1]. Two popular face recognition models are considered in the experimental framework: ResNet-50 and VGG-Face. We experimentally show that demographic groups highly represented in popular face databases have led to popular pre-trained deep face models presenting strong algorithmic discrimination. That discrimination can be observed both qualitatively at the feature space of the deep models and quantitatively in large performance differences when applying those models in different demographic groups, e.g. for face biometrics.

1 Introduction

Face recognition algorithms are good examples of recent advances in Artificial Intelligence (AI). The performance of automatic face recognition has been boosted during the last decade, achieving very competitive accuracies in the most challenging scenarios (Grother, Ngan, and Hanaoka 2018). These improvements have been possible due to improved machine learning approaches (e.g., deep learning), powerful computation (e.g., GPUs), and larger databases (e.g., at scale of millions of images). However, the recognition accuracy is not the only aspect to consider when designing biometric systems. Algorithms have an increasingly important role in the decision-making of several processes involving humans. These decisions have therefore increasing effects in our lives. Thus, there is currently a growing need for studying AI behavior to better understand its impact in our society (Rahwan et al. 2019).

Face recognition systems are especially sensitive due to the personal information present in face images (e.g., identity, gender, ethnicity, and age). Previous works suggested that face recognition accuracy is affected by demographic covariates. In (Klare et al. 2012; Cook et al. 2019), authors demonstrated that the performance of commercial face recognition systems varies according to demographic attributes. In (Acien et al. 2018; Lu et al. 2019), the authors evaluated how covariates affect the performance of face recognition systems based on deep neural network models. Among the different covariates, the skin color is repetitively remarked as a factor with high impact in the performance (Cook et al. 2019; Lu et al. 2019). However, ethnic face attributes are beyond skin color. The shape and size of facial features are partially defined by the ancestry origin. These differences can be used to accurate classify subjects according to their ancestry origin (Acien et al. 2018).

The number of published works pointing out the biases in the results of face detection (Buolamwini and Gebru 2018) and recognition algorithms is large (Klare et al. 2012; Acien et al. 2018; Alvi, Zisserman, and Nellåker 2018; Cook et al. 2019; Lu et al. 2019; Hupont and Fernandez 2019). Yet, only a limited number of works analyze how biases affect the learning process of these algorithms. The aim of this work is to analyze face recognition models using a discrimination-aware perspective. Previous studies have demonstrated that ethnicity and gender affect the performance of face recognition models (Gong, Liu, and Jain 2019). However, there is a lack of understanding regarding how this demographic information affects the model beyond the performance. The main contributions of this work are:

- A general formulation of algorithmic discrimination for machine learning tasks. In this work, we apply this formulation in the context of face recognition.
- Discrimination-aware performance analysis based on a new dataset (Morales, Fierrez, and Vera-Rodriguez 2019), with 24K identities equally distributed between six demographic groups.
- Study of the effects of gender and ethnicity in the feature representation of deep models.
- Analysis of the demographic diversity present in some of the most popular face databases.

[1] Available at GitHub: https://github.com/BiDAlab/DiveFace

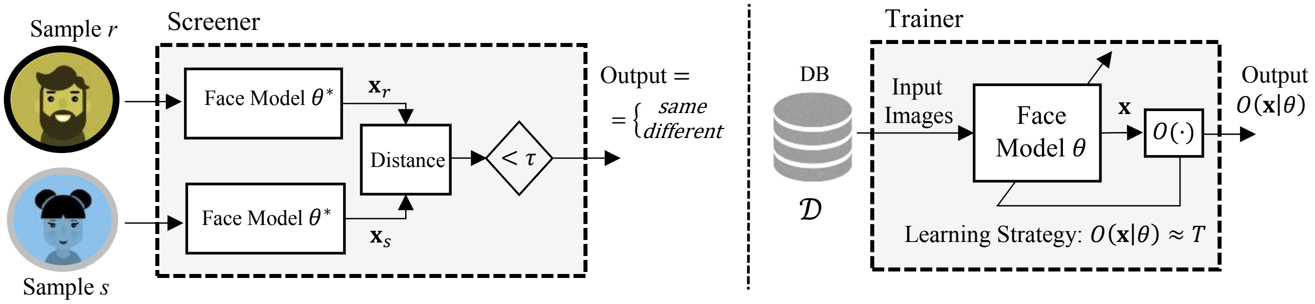

Figure 1: Face recognition block diagrams. The screener is an algorithm that given two face images decides if they belong to the same person. The trainer is an algorithm that generates the best data representation for the screener.

The rest of the paper is structured as follows: Section 2 presents our general formulation of algorithmic discrimination. Section 3 analyzes some of the most popular face recognition architectures and the experimental protocol followed in this work. Section 4 evaluates the causes and effects of biased learning in face recognition algorithms. Finally, Section 5 summarizes the main conclusions.

2 Formulation of Algorithmic Discrimination

Discrimination is defined by the Cambridge Dictionary as treating a person or particular group of people differently, especially in a worse way than the way in which you treat other people, because of their skin color, sex, sexuality, etc.

For the purpose of studying discrimination in artificial intelligence at large, we now formulate mathematically algorithmic discrimination based on the previous dictionary definition. Even though similar ideas to the ones embedded in our formulation can be found elsewhere (Calders and Verwer 2010; Raji and Buolamwini 2019), we didn't find this kind of formulation in related works. We hope that formalizing these concepts can be beneficial to foster further research and discussion in this hot topic.

Let's begin with notation and preliminary definitions. Assume $\mathbf{x}_s^i$ is a learned representation of individual i (out of I different individuals) corresponding to an input sample s (out of S samples) of that particular subject. That representation $\mathbf{x}$ is assumed to be useful for task T, e.g., face authentication or emotion recognition. That representation $\mathbf{x}$ is learned using an artificial intelligence approach with parameters θ. We also assume that there is a goodness criterion G on that task maximizing some performance real-valued function f in a given dataset $\mathcal{D}$ (collection of multiple samples) in the form:

$$G(\mathcal{D}) = \max_{\theta} f(\mathcal{D}, \theta) \tag{1}$$

The most popular form of the previous expression minimizes a loss function $\mathcal{L}$ over a set of training samples $\mathcal{D}$ in the form:

$$\theta^* = \arg\min_{\theta} \sum_{\mathbf{x}_s^i \in \mathcal{D}} \mathcal{L}(O(\mathbf{x}_s^i|\theta), T_s^i) \tag{2}$$

where O is the output of the learning algorithm that we seek to bring closer to the target function (or groundtruth) T defined by the task at hand. On the other hand, the I individuals can be classified according to D demographic criteria C_d, with $d = 1, ..., D$, which can be the source for discrimination, e.g., C_1 = *Gender* = {*Male, Female*} (demographic criterion 1 = *Gender* has two classes in this example). The particular class $k = 1, ..., K$ for a given demographic criterion d and a given sample is noted as $C_d(\mathbf{x}_s^i)$, e.g., $C_1(\mathbf{x}_s^i)$ = *Male*. We assume that all classes are well represented in dataset $\mathcal{D}$, i.e., the number of samples for each class in all criteria in $\mathcal{D}$ is significant. $\mathcal{D}_d^k \in \mathcal{D}$ represents all the samples corresponding to class k of demographic criterion d.

Finally, **our definition of algorithmic discrimination**: an algorithm discriminates the group of people represented with class k (e.g., *Female*) when performing the task T (e.g., face verification, or emotion recognition), if the goodness G in that task when considering the full set of data $\mathcal{D}$ (including multiple samples from multiple individuals), is significantly larger than the goodness $G(\mathcal{D}_d^k)$ in the subset of data corresponding to class k of the demographic criterion d.

The representation $\mathbf{x}$ and the model parameters θ will typically be real-valued vectors, but they can be any set of features combining real and discrete values. Note that the previous formulation can be easily extended to the case of varying number of samples S_i for different subjects, which is a usual case; or to classes K that are not disjoint. Note also that the previous formulation is based on average performances over groups of individuals. Different performance across specific individuals is usual in many artificial intelligence tasks due to diverse reasons, e.g., specific users who were not sensed properly (Alonso-Fernandez, Fierrez, and Ortega-Garcia 2011), even for algorithms that on average may perform similarly for the different classes that can be the source of discrimination.

3 Face Recognition Algorithms

A face recognition algorithm, as other machine learning systems, can be divided into two different algorithms: screener and trainer. Both algorithms are used for a different aim and therefore should be studied with a different perspective (Kleinberg et al. 2019).

The screener (see Fig. 1) is an algorithm that given two face images generates an output associated to the probability that they belong to the same person. This probability is obtained comparing the two learned representations obtained from a face model defined by the parameters θ. These parameters are trained previously based on a training dataset $\mathcal{D}$ and the goodness criterion G (see Fig. 1). If trained properly, the output of the trainer would be a model with parameters θ^* capable of representing the input data (e.g., face images) in a highly discriminant feature space $\mathbf{x}$.

The most popular architecture used to model face attributes is the Convolutional Neural Network (CNN). This type of network has drastically reduced the error rates of face recognition algorithms in the last decade (Ranjan et al. 2018) by learning highly discriminative features from large-scale databases. In our experiments we consider two popular face recognition pre-trained models: VGG-Face and ResNet-50. These models have been tested on competitive evaluations and public benchmarks (Parkhi et al. 2015; Cao et al. 2018).

VGG-Face is a model based on the VGG-Very-Deep-16 CNN architecture trained on the VGGFace dataset (Parkhi et al. 2015). ResNet-50 is a CNN model with 50 layers and 41M parameters initially proposed for general purpose image recognition tasks (He et al. 2016). The main difference between ResNet architecture and traditional convolutional neural networks is the inclusion of residual connections to allow information to skip layers and improve gradient flow.

Before applying the face models, we cropped the face images using the algorithm proposed in (Zhang et al. 2016). The pre-trained models are used as embedding extractor where $\mathbf{x}$ is a l_2-normalised learned representation of a face image. The similarity between two face descriptors $\mathbf{x}_r$ and $\mathbf{x}_s$ is calculated as the Euclidean distance $||\mathbf{x}_r - \mathbf{x}_s||$. Two faces are assigned to the same identity if their distance is smaller than a threshold τ. The recognition accuracy is obtained by comparing distances between positive matches (i.e., $\mathbf{x}_r$ and $\mathbf{x}_s$ belong to the same person) and negative matches (i.e., $\mathbf{x}_r$ and $\mathbf{x}_s$ belong to different persons).

The two face models considered in our experiments were trained with the VGGFace2 dataset according to the details provided in (Cao et al. 2018). As we will show in Section 4.3, databases used to train these two models are highly biased. Therefore, it is expected that the recognition models trained with this dataset present algorithmic discrimination.

3.1 Experimental protocol

Labeled Faces in the Wild (LFW) is a database for research on unconstrained face recognition (Learned-Miller et al. 2016). The database contains more than 13K images of faces collected from the web. In this study we consider the aligned images from the test set provided with view 1 and its associated evaluation protocol. This database is composed of images acquired in the wild, with large pose variations, and varying face expressions, image quality, illuminations, and background clutter among other variations. The performance achieved by the VGG-Face and ResNet-50 models for the LFW database is 4.1% and 1.7% Equal Error Rate respectively. These performances serve as a baseline for the models and the rest of experiments. We can observe the superior performance of the ResNet-50 model, with a performance ca. 3 times better than the VGG-Face model.

The experiments with DiveFace will be carried out following a cross-validation methodology using three images for each of the 4K identities from each of the six classes available in DiveFace (72K face images in total). This results in 72K genuine comparisons and near 3M impostor comparisons.

3.2 DiveFace database: an annotation dataset for face recognition trained on diversity

DiveFace was generated using the Megaface MF2 training dataset (Kemelmacher-Shlizerman et al. 2016). MF2 is part of the publicly available Megaface dataset with 4.7 million faces from 672K identities and it includes their respective bounding boxes. All images in the Megaface dataset were obtained from Flickr Yahoo's dataset (Thomee et al. 2015).

DiveFace contains annotations equally distributed among six classes related to gender and ethnicity (see Fig. 4 for example images). Gender and ethnicity have been annotated following a semi-automatic process. There are 24K identities (4K per class). The average number of images per identity is 5.5 with a minimum number of 3 for a total number of images greater than 120K. Users are grouped according to their gender (male or female) and three categories related to ethnic physical characteristics:

- **Group 1**: people with ancestral origins in Europe, North-America, and Latin-America (with European origin).
- **Group 2**: people with ancestral origins in Sub-Saharan Africa, India, Bangladesh, Bhutan, among others.
- **Group 3**: people with ancestral origin in Japan, China, Korea, and other countries in that region.

We are aware of the limitations of grouping all human ethnic origins into only three categories. According to studies, there are more than 5K ethnic groups in the world. We categorized according to only three groups in order to maximize differences among classes. Automatic classification algorithms based on these three categories show performances of up to 98% accuracy (Morales, Fierrez, and Vera-Rodriguez 2019).

4 Causes and Effects of Biased Learning in Face Recognition Algorithms

4.1 Performance of face recognition: role of demographic information

This section explores the effects of biased models in the performance of face recognition algorithms. Table 1 shows the performances obtained for each demographic group present in DiveFace. Traditional face recognition benchmarks usually do not explore this kind of demographic covariates. Results reported in Table 1 exhibit large gaps between performances obtained by different demographic groups, suggesting that both gender and ethnicity significantly affect the

Table 1: Performance (False Match Rate in % @ False Non-Match Rate = 0.1%) of Face Recognition Models on the DiveFace dataset. We show in brackets the relative error growth rates with respect to the best class (*Group 1 Male*).

Model	**Group 1**		**Group 2**		**Group 3**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
VGG-Face	7.99	9.38 (↑17%)	12.03 (↑50%)	13.95 (↑76%)	18.43 (↑131%)	23.66 (↑196%)
ResNet-50	1.60	1.96 (↑22%)	2.15 (↑34%)	3.61 (↑126%)	3.25 (↑103%)	5.07 (↑217%)

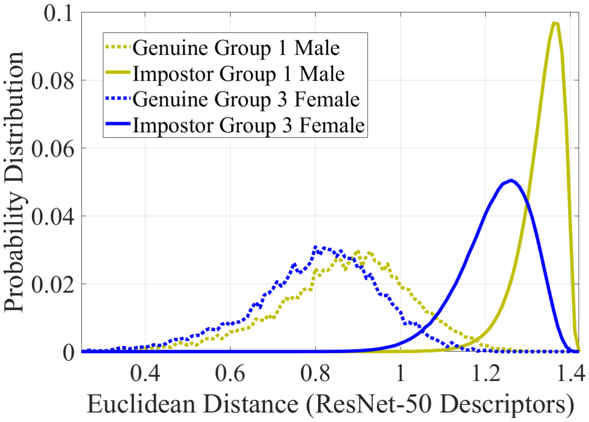

Figure 2: ResNet-50 face recognition score distributions for Group 3 females and Group 1 males.

performance of biased models. These effects are particularly high for ethnicity, with a very large degradation of the results for the class less represented in the training data (*Group* 3 *Female*). This degradation produces a relative increment of the Equal Error Rate (EER) of 196% and 217% for VGG-Face and ResNet-50, respectively, with regard to the best class (*Group* 1 *Male*). These differences are important as they mark the percentage of faces successfully matched and faces incorrectly matched. These results suggest that your ethnic origin can highly affect your possibilities to be incorrectly matched (false positives).

4.2 Understanding biased performances

The relatively low performance in *Group* 3 seems to be originated by a limited ability to capture the best discriminant features for the groups underrepresented in the training databases. The results suggest that features capable of reaching high accuracy for a specific demographic group may be less competitive in others. Let's analyze the causes behind these degradations. Fig. 2 represents the probability distributions of genuine and impostor scores for *Group* 1 *Male* (the best group) and *Group 3 Female* (the worst group). The comparison between genuine and impostor distributions reveals large differences for the impostor's ones. The genuine distribution (intra-class variability) between *Group* 3 and *Group* 1 is similar, but the impostor distribution (inter-class variability) is significantly different. The model has difficulties to differentiate face attributes from different subjects.

Algorithmic discrimination implications: define the performance function f as the accuracy of the face recognition model, and $G(\mathcal{D}_d^k) = f(\mathcal{D}_d^k, \theta^*)$ the goodness considering all the samples corresponding to class k of the demographic criterion d, for an algorithm θ^* trained on the full set of data $\mathcal{D}$ (as described in Eq. 1). Results suggest large differences between the goodness $G(\mathcal{D}_d^k)$ for different classes, especially for classes $k = \mathit{Group}\,1, \mathit{Group}\,2, \mathit{Group}\,3$.

4.3 Bias in face databases

Bias and discrimination concepts are related to each other, but they are not necessarily the same thing. Bias is traditionally associated with unequal representation of classes in a dataset. The history of automatic face recognition has been linked to the history of the databases used for algorithm training during the last two decades. The number of publicly available databases is high, and they allow training models using millions of face images. Fig. 3 summarizes the demographic statistics of some of the most cited face databases. Each of these databases is characterized by its own biases (e.g. image quality, pose, backgrounds, and aging). In this work, we highlight the unequal representation of demographic information in very popular face recognition databases. As it can be seen, the differences between ethnic groups are severe. Even though the people in *Group* 3 are more than 35% of the world's population, they represent only 9% of the users in those popular face recognition databases.

Biased databases imply a double penalty for underrepresented classes. On the one hand, models are trained according to non-representative diversity. On the other hand, benchmark accuracies are reported over privileged classes and overestimate the real performance over a diverse society.

Recently, diverse and discrimination-aware databases have been proposed in (Buolamwini and Gebru 2018; Merler et al. 2019; Wang and Deng 2019). These databases are valuable resources to explore how diversity can be used to improve face biometrics. However, some of these databases do not include identities (Buolamwini and Gebru 2018; Merler et al. 2019), and face images cannot be matched to other images. Therefore, these databases do not allow to properly train or test face recognition algorithms.

Algorithmic discrimination implications: classes k are unequally represented in the most popular face databases $\mathcal{D}$.

4.4 Biased embedding space of deep models

We now analyze the effects of ethnicity and gender attributes in the embedding space generated by VGG-Face and

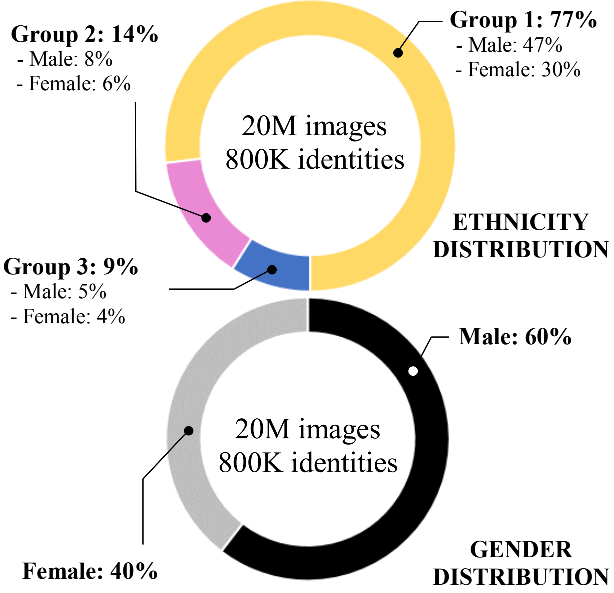

Figure 3: Demographic statistics of the 12 most cited face databases available in the literature. BioSecure (Ortega-Garcia et al. 2009), YouTubeFaces (Wolf, Hassner, and Maoz 2011), PubFig (Kumar et al. 2011), CasiaFace (Yi et al. 2014), VGGFace (Parkhi et al. 2015), CelebA (Yang et al. 2015), MS-Celeb-1M (Guo et al. 2016), Megaface (Kemelmacher-Shlizerman et al. 2016), LFW (Learned-Miller et al. 2016), UTKface (Zhang, Song, and Qi 2017), VGGFace2 (Cao et al. 2018), IJB-C (Maze et al. 2018), DiveFace (Morales, Fierrez, and Vera-Rodriguez 2019).

ResNet-50 models. CNNs are composed of a large number of stacked filters. These filters are trained to extract the richest information for a pre-defined task (e.g. face recognition). As face recognition models are trained to identify individuals, it is reasonable to think that the response of the models can slightly vary from one person to another. In order to visualize the response of the model to different faces, we consider the specific Class Activation MAP (CAM) proposed in (Selvaraju et al. 2017), named Grad-CAM. This visualization technique uses the gradients of any target concept, flowing into the selected convolutional layer to produce a coarse localization map. The resulting heat map highlights the activated regions in the image for the mentioned target (e.g. an individual identity in our case). Fig. 4 represents the heat maps obtained by the ResNet-50 model for faces from different demographic groups. Additionally, we include the heat map obtained after averaging results from 120 different individuals from the six demographic groups included in DiveFace. The activation maps show clear differences between ethnic groups with the highest activation for *Group* 1 and the lowest for *Group* 3. These differences suggest that features extracted by the model are, at least, partially affected by the ethnic attributes.

On a different front, we applied a popular data visualization algorithm to better understand the importance of ethnic features in the embedding space generated by deep models. t-SNE is an algorithm to visualize high-dimensional data. This algorithm minimizes the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. Fig. 5 shows the projection of each face into a 2D space generated from ResNet-50 embeddings and the t-SNE algorithm. Additionally, we have colored each point according to its ethnic attribute. As we can see, the resulting face representation results in three clusters highly correlated with the ethnicity attributes. Note that ResNet-50 has been trained for face recognition, not ethnicity detection. However, the ethnicity information is highly embedded in the feature space and a simple t-SNE algorithm reveals the presence of this information.

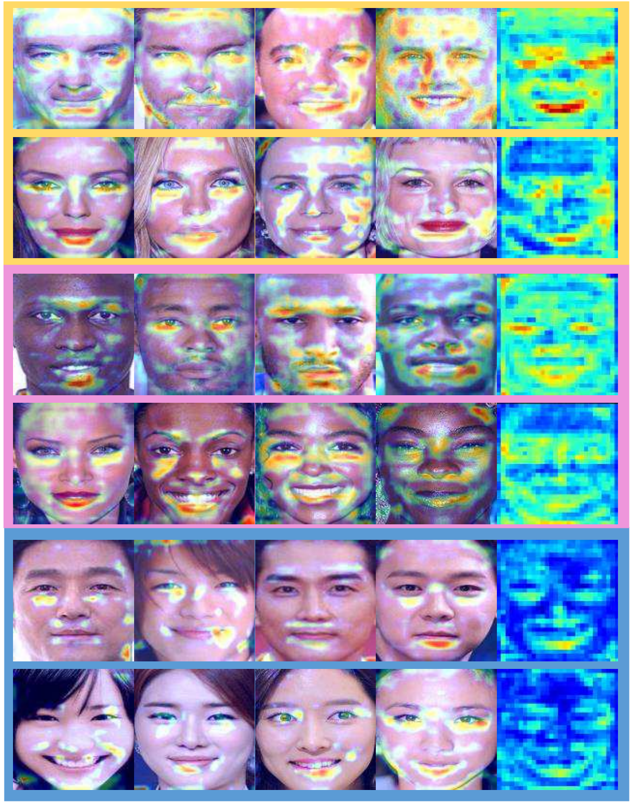

Figure 4: Examples of the six classes available in the DiveFace database (columns 1 to 4). Column 5 shows the averaged Class Activation MAP (first filter of the third convolutional block of ResNet-50) obtained from 20 random face images from each of the classes. Columns 1-4 show Class Activation MAPs for each of the face images. Maximum and minimum activations are represented by red and blue colors respectively. Average pixel value of the activation maps generated for the six classes (*Groups* 1 to 3, and *Male/Female*): G1M=0.23, G1F=0.19, G2M=0.21, G2F=0.18, G3M=0.12, G3F=0.13. (This is a colored image, see the digital version for a better quality.)

These two simple experiments illustrate the presence and importance of ethnic attributes in the feature space generated by face deep models.

Algorithmic discrimination implications: popular deep models trained for task T on biased databases (i.e., unequally represented classes k for a given demographic criterion d such as gender) result in feature spaces (corresponding to the solution θ^* of the Eq. 1) that introduce strong differentiation between classes k. This differentiation affects the representation $\mathbf{x}$ and enables classifying between classes k using $\mathbf{x}$, even though $\mathbf{x}$ was trained for solving a different task T.

5 Conclusions

This work has presented a comprehensive analysis of face recognition models according to a new discrimination-aware perspective. This work presents a new general formulation of algorithmic discrimination with application to face recognition. We have shown the high bias introduced when training the deep models with the most popular databases employed in the literature, and testing with the DiveFace dataset with well balanced data across demographic groups[2]. We have evaluated two popular models according to the proposed formulation. Biased models based on competitive deep learning algorithms have been shown to be very sensitive to gender and ethnicity attributes. This sensitivity results in different feature representations and a large gap between performances depending on the ethnic origin. This gap between performances reached up to 200% of relative error degradation between the best class (*Group* 1 *Male*) and the worst (*Group* 3 *Female*). These results suggest that false positives are 200% more likely in *Group* 3 *Female* than in *Group* 1 *Male* for the models evaluated in this work. These results encourage training more diverse models and developing methods capable to deal with the differences inherent to demographic groups. Future work will go in line with this approach, as authors do in (Wang and Deng 2019).

Acknowledgments

This work has been supported by projects: BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), Bio-Guard (Ayudas Fundacion BBVA a Equipos de Investigacion Cientifica 2017) and Accenture.

[2] Available at GitHub: https://github.com/BiDAlab/DiveFace

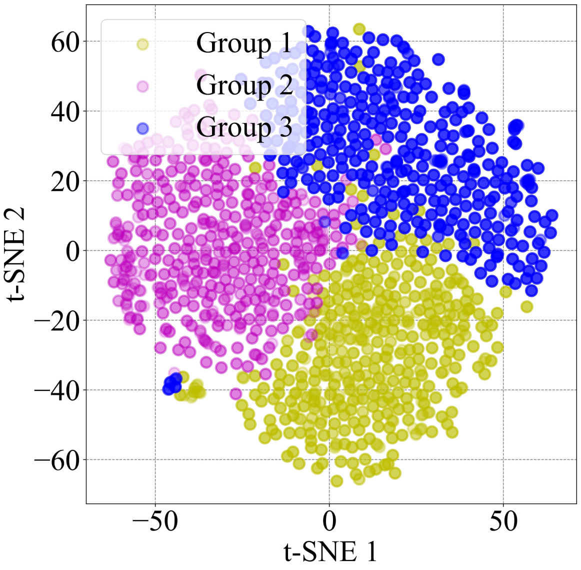

Figure 5: Projections of the ResNet-50 embeddings into the 2D space generated with t-SNE.

References

Acien, A.; Morales, A.; Vera-Rodriguez, R.; Bartolome, I.; and Fierrez, J. 2018. Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition. In *Iberoamerican Congress on Pattern Recognition (IAPR)*, 584–593. Madrid, Spain: Springer.

Alonso-Fernandez, F.; Fierrez, J.; and Ortega-Garcia, J. 2011. Quality Measures in Biometric Systems. *IEEE Security & Privacy* 10(6):52–62.

Alvi, M.; Zisserman, A.; and Nellåker, C. 2018. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Buolamwini, J., and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A., and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91.

Calders, T., and Verwer, S. 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A Dataset for Recognising Faces Across Pose and Age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 67–74.

Cook, C. M.; Howard, J. J.; Sirotin, Y. B.; Tipton, J. L.; and Vemury, A. R. 2019. Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1(1):32–41.

Gong, S.; Liu, X.; and Jain, A. K. 2019. DebFace: Debiasing Face Recognition. *arXiv preprint arXiv:1911.08080*.

Grother, P. J.; Ngan, M. L.; and Hanaoka, K. K. 2018. *Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification*. NIST Internal Report. U.S. Department of Commerce, National Institute of Standards and Technology.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Msceleb-1m: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*, 87–102. Amsterdam, The Netherlands: Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hupont, I., and Fernandez, C. 2019. DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition. In *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The Megaface Benchmark: 1 Million Faces for Recognition at Scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4873–4882.

Klare, B. F.; Burge, M. J.; Klontz, J. C.; Bruegge, R. W. V.; and Jain, A. K. 2012. Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security* 7(6):1789–1801.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Sunstein, C. R. 2019. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10:113–174.

Kumar, N.; Berg, A.; Belhumeur, P. N.; and Nayar, S. 2011. Describable Visual Attributes for Face Verification and Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(10):1962–1977.

Learned-Miller, E.; Huang, G. B.; RoyChowdhury, A.; Li, H.; and Hua, G. 2016. Labeled Faces in the Wild: A Survey. In Kawulok, M.; Celebi, M. E.; and Smolka, B., eds., *Advances in Face Detection and Facial Image Analysis*. Springer. 189–248.

Lu, B.; Chen, J.-C.; Castillo, C. D.; and Chellappa, R. 2019. An experimental Evaluation of Covariates Effects on Unconstrained Face Verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1(1):42–55.

Maze, B.; Adams, J.; Duncan, J. A.; Kalka, N.; Miller, T.; Otto, C.; Jain, A. K.; Niggel, W. T.; Anderson, J.; Cheney, J.; et al. 2018. IARPA Janus Benchmark-C: Face Dataset and Protocol. In *International Conference on Biometrics (ICB)*, 158–165. Gold Coast, Australia: IEEE.

Merler, M.; Ratha, N.; Feris, R. S.; and Smith, J. R. 2019. Diversity in Faces. *arXiv preprint arXiv:1901.10436*.

Morales, A.; Fierrez, J.; and Vera-Rodriguez, R. 2019. SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. *arXiv preprint arXiv:1902.00334*.

Ortega-Garcia, J.; Fierrez, J.; Alonso-Fernandez, F.; Galbally, J.; Freire, M. R.; Gonzalez-Rodriguez, J.; Garcia-Mateo, C.; Alba-Castro, J.-L.; Gonzalez-Agulla, E.; Otero-Muras, E.; et al. 2009. The Multiscenario Multienvironment Biosecure Multimodal Database (BMDB). *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(6):1097–1111.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; et al. 2015. Deep Face Recognition. In *British Machine Vision Conference (BMVC)*.

Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.-F.; Breazeal, C.; Crandall, J. W.; Christakis, N. A.; Couzin, I. D.; Jackson, M. O.; Jennings, N. R.; Kamar, E.; Kloumann, I. M.; Larochelle, H.; Lazer, D.; McElreath, R.; Mislove, A.; Parkes, D. C.; Pentland, A. S.; Roberts, M. E.; Shariff, A.; Tenenbaum, J. B.; and Wellman, M. 2019. Machine Behaviour. *Nature* 568(7753):477–486.

Raji, I. D., and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *AAAI/ACM Conf. on AI Ethics and Society (AIES)*.

Ranjan, R.; Sankaranarayanan, S.; Bansal, A.; Bodla, N.; Chen, J.-C.; Patel, V. M.; Castillo, C. D.; and Chellappa, R. 2018. Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. *IEEE Signal Processing Magazine* 35(1):66–83.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks Via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2015. The New Data and New Challenges in Multimedia Research. *arXiv preprint arXiv:1503.01817*.

Wang, M., and Deng, W. 2019. Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning. *arXiv preprint arXiv:1911.10692*.

Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face Recognition in Unconstrained Videos with Matched Background Similarity. In *Computer Vision and Pattern Recognition (CVPR)*, 529–534.

Yang, S.; Luo, P.; Loy, C.-C.; and Tang, X. 2015. From Facial Parts Responses to Face Detection: A Deep Learning Approach. In *Proceedings of the IEEE International Conference on Computer Vision*, 3676–3684.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning Face Representation from Scratch. *arXiv preprint arXiv:1411.7923*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23(10):1499–1503.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5810–5818.