2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) | 978-1-6654-2354-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIBGRAPI54419.2021.00041

Bias and Fairness in Face Detection

Hanna F. Menezes, Arthur S. C. Ferreira
Academic Unity of Systems and Computing
Federal University of Campina Grande
Campina Grande, PB, Brazil
hanna@copin.ufcg.edu.br, arthur.ferreira@ccc.ufcg.edu.br

Eanes T. Pereira, and Herman M. Gomes
Academic Unity of Systems and Computing
Federal University of Campina Grande
Campina Grande, PB, Brazil
{eanes,hmg}@computacao.ufcg.edu.br

***Abstract*—Processing of face images is used in many areas, for example: commercial applications such as video-games; facial biometrics; facial expression recognition, etc. Face detection is a crucial step for any system that processes face images. Therefore, if there is bias or unfairness in this first step, all the processing steps that follow may be compromised. Errors in automatic face detection may be harmful to people as, for instance, in situations where a decision may limit or restrict their freedom to come and go. Therefore, it is crucial to investigate the existence of these errors caused due to bias or unfairness. In this paper, an analysis of five well-known top accuracy face detectors is performed to investigate the presence of bias and unfairness in their results. Some of the metrics used to identify the existence of bias and unfairness involved the verification of demographic parity, verification of existence of false positives and/or false negatives, rate of positive prediction, and verification of equalized odds. Data from about 365 different individuals were randomly selected from the Facebook Casual Conversations Dataset, resulting in approximately 5,500 videos, providing 550,000 frames used for face detection in the performed experiments. The obtained results show that all five face detectors presented a high risk of not detecting faces from the female gender and from people between 46 and 85 years old. Furthermore, the skin tone groups related with dark skin are the groups pointed out with highest risk of faces not being detected for four of the five evaluated face detectors. This paper points out the necessity of the research community to engage in breaking the perpetuation of injustice that may be present in datasets or machine learning models.**

I. Introduction

Face detection algorithms are a classical problem in computer vision. Some possible reasons for the high interest in face detection are: before the execution of any processing involving face biometrics, the face must be detected; faces own a regular pattern among all human beings; and facial expressions are the most important human aspect for expressing emotions.

The first researches aiming to solve the face detection problem were published in the decade of 1970 [1]. In that period, the techniques were based on heuristics and anthropomorphic measures. Clearly, those heuristic techniques were not capable of adequately dealing with all the variations a face image can be subjected to. Despite these relevant problems to be investigated, the research interest in face detection remained stagnant until the beginning of decade of 1990.

In the history of computer vision, there were many important contributions to solve the face detection problem, such as: PCA (Principal Component Analysis) applied as features to train face classifiers [2], artificial neural networks trained with raw pixel gray values [3], and integral features used to train weak-classifiers combined by Adaboost [4]. The approach proposed by Viola and Jones [4] used an Adaboost algorithm with Haar-like features extracted via integral image to train a cascade of face classifiers. The use of integral images allowed the face detection to be fifteen times faster, in average, when compared to other competing approaches at the time, and the false positive rate was bellow 10^{-6}.

After the success of the ImageNet Large Scale Visual Recognition Challenge [5], the community attention turned to the Deep Convolutional Neural Networks (DCNN). Face detection research has made significant progress in the last years due to the use of DCNN [6]. One important contribution of DCNN approaches is the possibility of training the algorithm to extract the specific features for the classification problem at hand. Therefore, it is possible to have a face detector trained end-to-end without the necessity of hand-crafted features. Contemporary face detectors, such as PyramidBox [7], benefit from DCNN characteristics and have better performance than the Viola and Jones' approach [4].

Machine learning algorithms are being employed as solution for many problems, from commonplace situations such as movie recommendation and product recommendation in shopping websites to situations where there is some type of risk to human life or to financial status of companies. For instance, the IBM Watson [1] promises to help health professionals to obtain fast answers for patient care. These intelligent systems operate learning from data and producing decisions as output, which can vary from relatively trivial to highly significant for patient health and for company economy [8]. As these systems are based on human data and they are becoming ubiquitous in different applications for human activity, safety and fairness concerns are arising. For, as human beings, algorithms are also prone to biases which can turn their decisions unfair.

According to Nelson [9], bias is a reflex of: the data chosen by programmers, approaches of combination and data wrangling, practices of model creation, methods of application and interpretation of results. According to Suresh and Guttag [10], the following types of bias are among the most important described in literature:

(i) **Historical bias** occurs when real-world pre-existing bias and socio-technical questions are infiltrated in the process of data generation;
(ii) **Representation bias** occurs when the data selection for model training misses important real-world elements. Due

[1] Available at https://www.ibm.com/cloud/ai, last access: June 25, 2021

978-1-6654-2354-0/21/$31.00 ©2021 IEEE
DOI 10.1109/SIBGRAPI54419.2021.00041

to representation bias, the development sample underrepresents and, consequently, fails in the generalization to production samples;

(iii) **Measurement bias** occurs during choice, collection, labelling and feature extraction for prediction problems. The set of features and labels may leave out important factors or insert noise which depends of the input and it affects the performance;

(iv) **Aggregation bias** occurs during model construction, when distinct populations are inadequately combined;

(v) **Evaluation bias** occurs during iteration and model evaluation, when test population or population of external reference do not equally represent the many parts of production population. Evaluation bias also occurs when the performance metrics are not appropriate to the context in which the model will be applied;

(vi) **Learning bias** occurs when modeling choices amplify performance disparities in data with underrepresented attributes;

(vii) **Deployment bias** occurs after the deployment of the model when the system is inadequately used or interpreted.

In the context of decision making, fairness may be defined as the absence of prejudice or partiality in relation to a subject or group based on their inherent or acquired characteristics [11]. Thus, an algorithm may be considered unfair when the decisions are directed to a specific group of people. To better understand how fairness-related failures are incorporated to algorithms, Mehribi et al. [11] listed the main types of discrimination as:

(i) **Direct discrimination** occurs when protected attributes of subjects explicitly imply in unfavorable results to them;

(ii) **Indirect discrimination** occurs when people appear to be treated based on apparently neutral and non-protected attributes, but groups or protected subjects may still be treated unfairly as a result of the implicit effects of their protected attributes;

(iii) **Systemic discrimination** occurs when politics, customs or behaviors that are part of the culture or structure of an organization perpetuate the prejudice against certain population subgroups;

(iv) **Statistical discrimination** occurs when the decision makers use obvious (e.g. average statistics of a group) and recognizable features of a subject as a proxy to hidden features or to features more difficult to determine, which could be more relevant to the aimed result;

(v) **Explainable discrimination** occurs when the differences in the treatment and results among different groups may be justified and explained by means of some attributes;

(vi) **Unexplainable discrimination** occurs when the prejudice against a group is unjustified and, therefore, considered illegal.

Although face detectors based on CNN have been extensively studied, major visual variations of faces, such as occlusions, pose and extreme illumination, are still challenging to real world applications [12]. Furthermore, some researchers as well as research agencies report that face detection systems are prone to work differently for distinct demographic groups [13].

As CNN's typically depend on large scale datasets, this may be a bad stimulus to the yielding of biased and unfair systems, because to adequately annotate large amounts of data is time consuming and expensive. Unbalanced datasets, in which there are historically underrepresented demographic groups, are often used in the training stage of models, which perpetuate the unfairness by inducing lower precision classification to the underrepresented demographic groups. This systemic discrimination also occurs in labeling and annotation of datasets, where the categories of race, ethnicity and gender are dynamic and reflect cultural norms and subjective categorizations that may lead to forms of scientific racism and prejudice [13].

Given that face detection applications have a direct effect on people lives and may be highly harmful if not correctly designed, it is important to evaluate and consider the fairness of systems which use face detection [11].

In this paper, it was performed an analysis of the bias and unfairness that may be present in the data sample selected for training and/or in the evaluated face detectors. Some of the metrics used to identify the existence of bias and unfairness involved the verification of demographic parity, verification of existence of false positive and/or false negative, rate of positive prediction, and verification of equalized odds. Those metrics are detailed in Section II.

II. Metrics to Measure Bias and Unfairness

Recently, the interest on face detection applications increased. Wójcik et al. [14] say that among the most important reasons for that interest is the concern about public security using applications such as: digital identity verification, facial analysis, modeling techniques for multimedia data and digital entertainment.

In the context of classification systems, in which the face detection is inserted, the analysis of discriminatory results may be performed by using metrics that evaluate the existence of prejudice in such systems. The distributional group metrics and error-based group metrics are the most applied for the analysis of discriminatory results [15]. Table I describes the most used metrics for bias and unfairness analysis. In addition to the metrics presented in Table I, Mehrabi et al. [11] and Saravanakumar [16] address other three, that are used for fairness analysis: (i) demographic parity, (ii) equal opportunity, and (iii) equalized odds.

The demographic parity or statistic parity proposes that the proportion of each protected class segment receives positive result in equal rates, avoiding the tendency of the model to uneven prediction for a given label for any sensible group [11] [16]. Mathematically, demographic parity requires a predictor $\widehat{Y}$ satisfying demographic parity independently of the protected class, A, as in Equation 1.

$$P(\widehat{Y}|A=0) = P(\widehat{Y}|A=1) \tag{1}$$

The equal opportunity proposes that each group should obtain positive results in equal rates. The equal opportunity

TABLE I
METRICS FOR DETECTION OF BIAS AND UNFAIRNESS. ADAPTED FROM SALEIRO ET AL. [15]

Metrics	Formula	Description
False Discovery Rate	$FDR_g = FP_g/PP_g = Pr(Y = 0\|\hat{Y} = 1, A = a_i)$	fraction of false positives in a group in the prediction of positives of the group.
False Omission Rate	$FOR_g = FN_g/PP_g = Pr(Y = 1 - \hat{Y} = 0, A = a_i)$	fraction of false negatives of a group in the negatives predicted from the group.
False Positive Rate	$FPR_g = FP_g/LN_g = Pr(\hat{Y} = 1 - Y = 0, A = a_i)$	fraction of the false positives of a group in the labeled negatives of the group.
False Negative Rate	$FNR_g = FN_g/LP_g = Pr(\hat{Y} = 0 - Y = 1, A = a_i)$	fraction of the false negatives of a group in the labeled positives of the group.
Predicted Positive	PPg	number of entities in a group in which the decision is positive, $\hat{Y}$=1.
Total Predictive Positive	$K = \sum_{A=a_1}^{A=a_n} PP_{g(a_i)}$	total number of entities predicted as positive among the groups defined by A.
Predicted Negative	PNg	number of entities within a group for which the decision is negative, $\hat{Y}$ = 0.
Predicted Prevalence	$PP_{rev_g} = PP_g/\|g\| = Pr(\hat{Y} = 1 - A = a_i)$	fraction of entities within a group that were predicted to be positive.
Predicted Positive Rate	$PPR_g = PP_g/K = Pr(a = a_i\|\hat{Y} = 1)$	fraction of entities predicted as positive that belongs to a certain group.

requires the positive result to be independent of the protected class, A, conditioned to real positive Y, as in Equation 2.

$$P(\widehat{Y} = 1|A = 0, Y = 1) = P(\widehat{Y} = 1|A = 1, Y = 1) \quad (2)$$

The equalized odds proposes that the model should correctly identify the positive result in equal rates among the groups (it is similar with equity of opportunity metric), but also it should classify in equal proportion the false positives among the groups. The equalized odds require the positive result to be independent of the protected class, A, conditioned to a real number Y, as in Equation 3.

$$P(\widehat{Y} = 1|A = 0, Y = y) = P(\widehat{Y} = 1|A = 1, Y = y), \quad y \in \{0, 1\} \quad (3)$$

After performing experiments, this research concluded that the false negative, positive prediction and demographic parity were the metrics with the most significant results for a detailed analysis that is presented in Section IV.

III. MATERIALS AND METHODS

This section presents the software libraries used in this research (III-A), the datasets and its handling (III-B), together with the proposed methodology for analysing bias and fairness in the context of face detection (III-C). Figure 1 contains the flowchart of the proposed methodology, which comprises: data extraction and preparation, face detection using the selected approaches, estimation of face detection statistics, and analysis of bias and equity based on the detector's results.

A. Software libraries for the analysis of bias and fairness

Initially, three libraries were selected, namely: Aequitas [2], AI Fairness 360 [3] and Audit AI [4]. These are all open source tools designed to analyze labelled training data and machine learning model predictions.

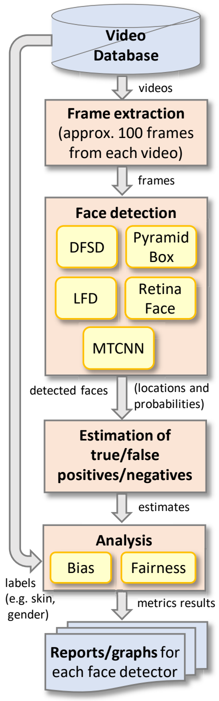

Video Database
videos
Frame extraction (approx. 100 frames from each video)
frames
Face detection
DFSD
Pyramid Box
LFD
Retina Face
MTCNN
detected faces
(locations and probabilities)
Estimation of true/false positives/negatives
estimates
Analysis
Bias
Fairness
labels (e.g. skin, gender)
metrics results
Reports/graphs for each face detector

Fig. 1. Flowchart of the proposed methodology.

Fairness 360 is a toolkit, developed by IBM Research, to examine, report and mitigate discrimination and bias in machine learning models across the entire AI application lifecycle [17].

Audit-AI is a Python library built on top of pandas[5] and

[2] http://aequitas.dssg.io/, last access: Juny 24, 2021
[3] https://aif360.mybluemix.net/, last access: Juny 24, 2021
[4] https://github.com/pymetrics/audit-ai, last access: Juny 24, 2021

[5] https://pandas.pydata.org/

sklearn[6] tools, which implements impartiality-aware machine learning algorithms. Developed by the team of data scientists of the Pymetrics[7] company, the tool is designed to measure and mitigate the effects of discriminatory patterns on training data and predictions made by machine learning algorithms trained for the purposes of socially sensitive decision processes.

Aequitas is an AI systems auditing tool, which investigates biased actions or results based on false or distorted inferences about demographic groups [15]. Aequitas operates on a command line interface, which makes use of a specific Python library, being able to load system data to be audited and configure metrics for protected attribute groups. The tool then generates bias reports according to what was configured by the user [15].

Thus, it is possible to assess the performance of machine learning models using various bias and fairness metrics, which can assess the (i) risks of biased actions or interventions that are not representatively allocated of the population and/or (ii) biased outputs from actions or interventions that result from the fact that the system is wrong about certain groups of people [15].

To conduct the experiments in this study, it was observed that the Fairness 360 and Audit AI tools are focused on mitigating prejudices and injustices, based on the implementation of a model for that purpose, which is not the focus of this study. The Aequitas library, on the other hand, has resources aimed at auditing and performing a detailed data analysis. Therefore, the Aequitas tool was selected because it meets the objectives of the proposed study.

In Aequitas, disparities are calculated as a proportion of a metric for an interest group, compared to a reference group (which in this study is the majority group) [15]. Thus, for example, the disparity in the false negative rate (FNR) for females (interest group) in relation to males (majority group) given by the quotient between FNR_{female} and FNR_{male}.

Differences in calculated metrics are explained based on the calculated disparity. The results are statistically significant at the 5% level (default level applied by the tool). It should be noted that the reference group always has a disparity equal to 1 (one). In this way, the other groups are displayed with at least 0.1 and no more than 10 times the size of the reference group.

B. Dataset

To conduct this study, the Casual Conversations dataset [18] from Facebook AI, was used. This dataset was originally designed to help researchers evaluate computer vision and audio applications. It contains videos captured from a diverse set of individuals, with varying age, gender, apparent skin tones and ambient lighting conditions . The dataset is comprised of over 45,000 videos (involving 3,011 participants), which feature paid individuals who have agreed to participate in the project and have explicitly provided age and gender labels. The apparent skin tone and lighting attributes were labeled by a group of human evaluators. Skin tone was assessed in the Fitzpatrick scale [19], which is based on the skin's reaction to ultraviolet light. The scale ranges from Type I (light skin that never tans, but always burns), to Type VI (very dark skin that always tans, but never burns). In this research, we adopted the same age ranges present in the dataset, which roughly leads a balanced distribution: age (46-85: 29.8%, 31-45: 32.5%, 18-30: 35.6%, N/A: 2.1%); gender (Female: 54.5%, Male: 43.4%, N/A: 2.1%).

Data from about 365 different individuals were randomly selected, resulting in approximately 5,500 videos (12%) to compose the data sample used in our evaluations. From each video, an average of 100 frames was extracted. The frame extraction was systematically performed: one frame was extracted for each 20 frames in the sequence until 100 frames were extracted for each video. However, some videos had an insufficient duration, which prevented the extraction of the full set of 100 frames. For those videos, the number of extracted frames was limited to the multiples of 20 frames found in the sequence. The final set of extracted frames was composed of 550,000 frames.

C. Face Detectors

Among the various facial detectors proposed in the literature and that have their implementation publicly available, five were used in this study. The main selection criterion was the time spent for inference. The training stage of the detectors was not considered in our experiments, thus pre-trained models were used.

One of the detectors used in this paper was the DSFD facial detector proposed by Jian Li et al. [12]. It is based on the Single Shot MultiBox Detector (SSD), by Wei Liu et al [6]. The face detector architecture contains a Feature Enhance Module (FEM) to strengthen the original feature maps and thus extend the single-shot detector to the double-shot detector. In addition, the concept of Progressive Anchor Loss (PAL) is introduced to improve the learning process by taking into account two different sets of anchors. Finally, an Improved Anchor Matching (IAM), integrating a new anchor assignment strategy with data augmentation, is introduced to provide a better initialization for the regressor. Extensive experiments on popular benchmarks, such as WIDER FACE [20] and FDDB [21], demonstrate the superiority of DSFD over existing state-of-the-art face detectors.

The RetinaFace face detector [22] is presented as a robust single-stage facial detector that performs pixel-wise face localisation on various scales by means of a set of feature pyramids with independent context modules. The detector takes advantage of joint extra-supervised and self-supervised multi-task learning. In the WIDER FACE hardware test suite, RetinaFace outperforms the best Average Accuracy (AP) results by 1.1% (reaching AP equal to 91.4%).

The RetinaFace [22] model was originally implemented with the ResNet152 architecture framework and was based on MXNet architecture. However, the experiments conducted in this work used a reimplementation of RetinaFace with the TensorFlow framework, built on top of the ResNet50 architecture. Although the original and modified models have different

[6] https://scikit-learn.org/

[7] https://www.pymetrics.ai/

structures, the implementation with TensorFlow presents a similar performance when compared with the one based on MXNet, achieving only 1% less precision than the original implementation in the validation stage with the WIDERFACE database [20] in the *easy* and *medium* subsets, and 2% in the *difficult*. This implementation was chosen because it had simplest configuration and usage for the inference task and presented faster execution time.

Xu Tang et al. [7] proposed a new context-assisted single-shot face detector called PyramidBox. Their work improves the use of contextual information in the following three aspects. First, a new context anchor is designed to supervise the learning of high-level contextual features by a semi-supervised method called Pyramid Anchors. Next, the Low-level Feature Pyramid Network is proposed to properly combine high-level context semantic elements and low-level facial elements, thus allowing the PyramidBox to predict faces at different scales in a single scene. Furthermore, a context-sensitive structure is presented to increase the capacity of the prediction network and, thus, to improve the accuracy of the final predictions. Finally, the data anchor sampling method is used to extend the training samples at different scales, which increases the diversity of training data for smaller faces. Among the various publicly available implementations, the latest stable version of the original implementation maintained with the PaddlePaddle framework[8] was used.

Another detector used in our study was the Light and Fast Face Detector (LFD) based on the work proposed by He et al. [23], being presented only as an evolution of the Light and Fast Face Detector for Edge Devices (LFFD) detector. LFD is implemented with the PyTorch framework and has code-level modifications that improve inference time and latency. LFD is free of anchors and belongs to the single stage category of face detectors. During development, the importance of the *receptive field (RF)* and *effective receptive field*(ERF) in the face detection task was reviewed, since the RF's of neurons in a given layer are regularly distributed in the input image and these RFs are naturally implicit anchors. Combining these RF anchors and appropriate RF *strides*, the proposed method can detect a wide range of continuous facial scales, with 100% coverage, in theory. Insightful understanding of the relationships between ERF and face scales motivated an efficient framework for single-stage detection. The architecture is structured into eight common detection branches and layers, which improves algorithm efficiency.

We also investigated the MTCNN face detector proposed by Kaipeng Zhang et al. [24], which has a deep cascading multitasking structure and exploits the inherent correlation between face detection and alignment in an unrestricted environment that is challenging due to various poses, illuminations and occlusions. The model adopts a cascade structure with three stages of carefully designed deep convolutional neural networks that predict the location of the face and reference point in an approximate way. In addition, in the learning process, a new online sample mining strategy that can automatically improve performance without manual sample selection was also proposed. All detectors adopted in this study were validated by their respective authors with the validation set of the WIDER FACE dataset [20], widely used and considered as a reference for face detection applications.

We used the following information obtained from the output of the aforementioned detectors: bounding box coordinates of faces found in an image as well as the confidence score. Knowing that all extracted images had at least one face, in order to facilitate data handling in the analysis of bias and injustice with the Aequitas library, a post-processing step took place, as explained next. Whenever the bounding boxes of multiple face detections within a single frame presented an intersection of 50% or less, we assumed as a true positive only the bounding box with highest detection confidence score. The remaining bounding boxes (presenting smaller confidence scores) were counted as false positives. In the case of no face being detected in a frame, the true positive and false positive counts were set as zero for that image. Finally, in the case of a single face being detected in a frame (or multiple detections with an intersection higher than 50%), the true positive counter for that image was set to one and the false positive counter was set to zero.

All detectors adopted in this study were validated by their respective authors with the validation set of the WIDER FACE dataset [20], widely used and considered as a reference for face detection applications.

IV. Analysis and Discussion of Results

This section presents the main results obtained from the experiments performed, which involved the use of bias and fairness analysis metrics. Among the metrics presented in the section II, group distribution metrics and error-based group metrics were addressed, in addition to demographic parity (statistics). As mentioned in the previous section, five face detectors were evaluated: Retina Face, DSFD, LFD, Pyramidbox e MTCNN. The objective of evaluating those detectors is to verify if they present some type of bias or unfairness. Three categories of attributes were analysed in the dataset: gender, age and skin tone.

First, face detection rate was evaluated for the categories of mentioned attributes. For age, data were classified in three groups: Group 1 (subjects between 18 and 30 years old), Group 2 (subjects between 31 and 45 years old), and Group 3 (subjects between 46 and 85 years old). For skin tone, as it was described in Subsection III-B, the Fitzpatrick scale was applied.

From the results presented by the face detectors, it is possible to verify that the five face detectors presented a highest risk of not detecting faces from the female gender and from the age category number 3 (which corresponds to people between 46 and 85 years old).

For skin tone, apparently, the LFD, DSFD and RetinaFace detectors had the skin tone group 4 as that with highest risk of not have detected faces (score = 0). The Pyramidbox face detector had the skin tone group 2 as that with highest risk, although the group 4 also had a high risk rate. The MTCNN

[8] https://github.com/PaddlePaddle/Paddle

face detector had the group 6 with the highest risk of not detecting faces, followed by group 2.

Among the analysed attributes, one may observe that the apparent skin tone presented the highest divergence of face detection results. It must be emphasized that, although there is divergence, the skin tone groups number 4 and 6, related with dark skin in the Fitzpatrick scale, are the groups pointed out with highest risk of faces not being detected for 4 of the 5 evaluated face detectors.

A. Levels of Disparity

The software library Aequitas allows to evaluate trends in all subgroups of the datasets by means of a confusion matrix for each subgroup. The confusion matrix provides important metrics to evaluate the performance of a classification algorithm, such as: false positive rate, group prevalence, and false omission rate.

The graphics in Figure 2 show the group absolute metrics for the False Negative Rate (FNR) computed with each attribute, for each face detector. The color is based on the magnitude of the absolute metric (computed using the number of samples in the attribute group). Darker color indicates higher rate. After a preliminary analysis, only the metrics that presented significant differences (those which presented disparity levels higher or lower than the reference group) were selected for detailed discussion in this paper.

It may also be observed in Figure 2 that for attribute ”age cat” the RetinaFace, Pyramidbox and DSFD detectors had the age group 1 (subjects between 18 and 30 years old) as the more prone to incorrect detection (when the detector is not capable of detecting faces). In this case, the false negative rate (FNR) was 0.21, 0.20 and 0.15 respectively. The LFD and MTCNN face detectors had the age group 3 as the most prone to not detect faces, with false negative rates of 0.22 and 0.28, respectively.

For the gender attribute, Pyramidbox, DSFD and LFD detectors did not present differences in false negative rates. For those classifiers, both genders male and female have the same probability of being incorrectly classified (face not detected). On the other hand, Retina Face and MTCNN presented results that indicates that the male gender has highest risk of incorrect detection, with false negative rates of 0.19 and 0.31, respectively.

The highest divergence among the detector results was for skin tone. For Retina Face and DSFD, the skin tone type 5 was the pronest to be incorrectly classified (face not detected), with false negative rates of 0.21 and 0.15, respectively. The Pyramidbox had the skin tone type 2 as the pronest to be incorrectly classified. The MTCNN had the skin tones type 6 and type 2 as the pronest, with results of false negative rate of 0.34 and 0.33, respectively. The LFD had the skin tone type 1 as the pronest to misclassification, with false negative rate of 0.24. Thus, one may observe that 3 out of 5 detectors presented results with high probability of incorrect detection for dark skin tones (types 5 and 6).

For the Positive Predictive Rate, for all detectors, the attributes age, skin tone and gender obtained similar results. In all cases, the age group 3 (subjects between 46 and 85 years old) are positively predicted more frequently than the other two age groups (PPR: 0.41), the skin tone 4 (PPR: 0.22) and the female gender (PPR: 0.52) had the highest positive predictive rates. The single observed difference was the MTCNN detector that obtained a PPR higher than the other detectors for the female gender (PPR: 0.56).

B. Levels of Equity

The level of equity was computed to measure the fairness of the face detector results for each model. Fairness was computed for a reference group (majority group), which will have a disparity of 1.0. As well as in the bias analysis, a previous analysis of the fairness (parity) results was performed and two metrics were selected for a more detailed analysis: False Negative Rate and Positive Predictive Rate. The graphics in Figures 3, 4 and 5 show the absolute group metric of the Positive Predictive Rate disparity. The colors are based on the fairness determination for each attribute group (green = ’True’, red = ’False’).

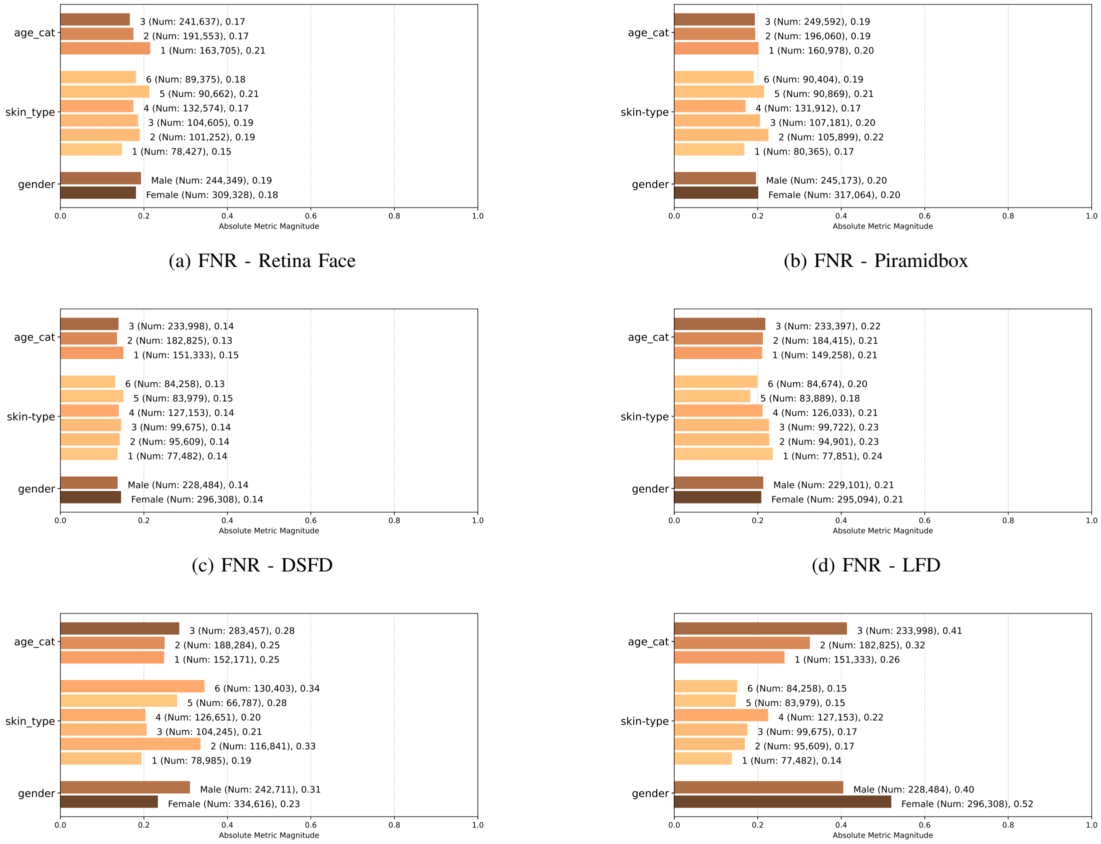
In Figures 3, 4 and 5, one can see that, for all face detectors, the female gender, age group 3 and skin tone 4 received fair detections. However, those are the reference groups, and this shows that the models are not fair in terms of statistical parity with anyone of the other groups. It should be noted that the MTCNN detector results differ only in skin tone results, in this case, the MTCNN was considered not fair for skin tones 5 and 1.

The results for absolute parity of the FNR metric are presented in Figure 6. The green color indicates that the DSFD and LFD detectors were considered fair for all attribute groups analyzed. The RetinaFace detector was not considered fair in relation to statistical parity only for age group 1. For skin tone results, the groups 2 and 5 were considered unfair for Pyramidbox detector and the groups 1, 3 and 4 were considered unfair for MTCNN detector. The results for the male gender obtained by the MTCNN detector were considered unfair. Considering the analyzed metrics, one can infer that the detectors presented some type of unfairness in their results related with at least one group of attributes. The most common attributes that presented unfair results were skin tone and age.

V. Conclusion

Considering the fact that face detection is the first step for any facial image processing, if this step is *contaminated* with bias or unfairness all the following steps will be affected. The analyses performed on face detection results presented in this paper sheds some light on the issues of bias and unfairness in facial biometrics. This paper contributes a methodology to objectively assess those issues in a context that can be easily extended to other pattern recognition problems.

Five face detectors were analyzed considering three types of sensitive attributes: age, skin tone and gender. In all cases, for all detectors, at least one category of attribute received



(a) FNR - Retina Face

(b) FNR - Piramidbox

(c) FNR - DSFD

(d) FNR - LFD

(e) FNR - MTCNN

(f) PPR - All detectors

Fig. 2. False Negative Rate (FNR) and Positive Predictive Rate (PPR)

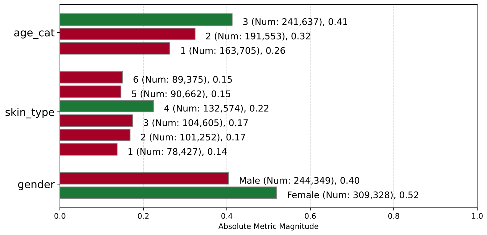


Fig. 3. Predicte Positive Rate (PPR) - Retina Face, Pyramidbox and DSFD.

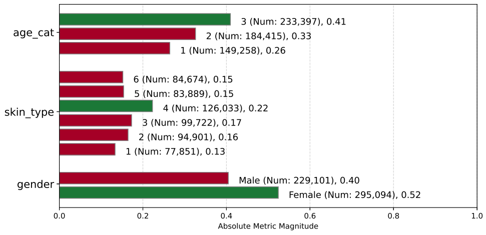


Fig. 4. Positive Predictive Rate (PPR) - LFD.

a treatment, in terms of face detection, considered unfair or biased. This should raise concerns in the facial biometrics community at least to evaluate their models to verify the existence of bias and unfairness before freely distribution or commercial usage. The effort to eliminate bias and unfairness from machine learning models should be accepted by all researchers as a way of cutting the perpetuating process of discrimination and injustice.

A direction for future work would be the development of methods for bias and unfairness mitigation in face detection models. One direct method should be the curating of datasets. However, the model retraining may be expensive, and the researcher should create methods for mitigating bias and unfairness as a post-processing, without the necessity of model retraining. One possibility for that may be inspired by the application of Causal Inference for bias selection [25].

REFERENCES

[1] R. Chellappa, C. L. Wilson, and S. Sirohey, “Human and machine recognition of faces: a survey,” *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, 1995.

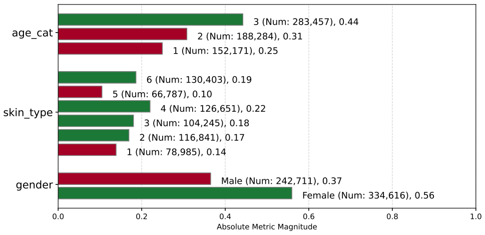
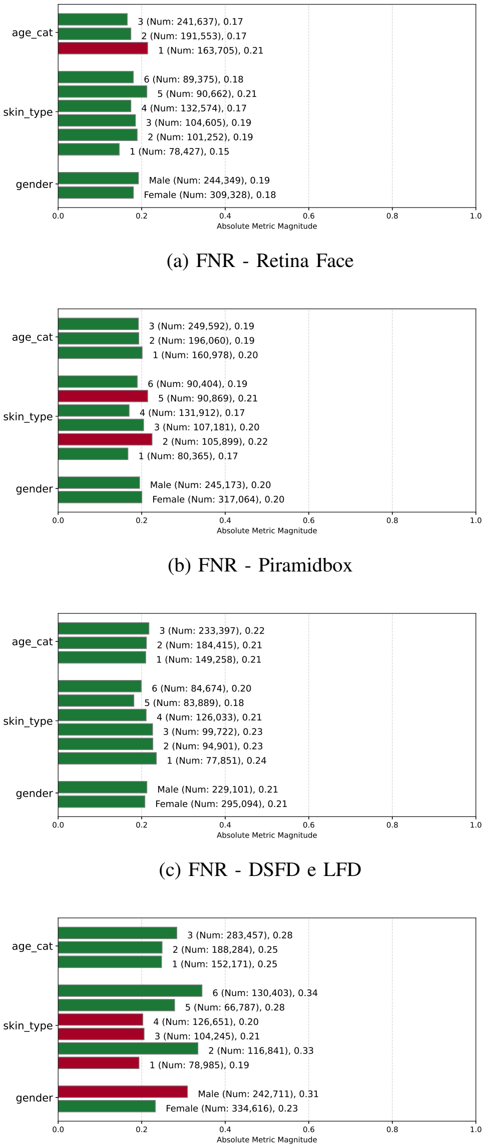


Fig. 5. Positive Predictive Rate (PPR) - MTCNN.



(a) FNR - Retina Face

(b) FNR - Piramidbox

(c) FNR - DSFD e LFD

(d) FNR - MTCNN

Fig. 6. False Negative Rates (FNR) for the evaluated face detectors

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, 1998.

[4] P. Viola and M. J. Jones, "Robust real-time face detection," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 747–747.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, p. 21–37.

[7] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer International Publishing, 2018, pp. 812–828.

[8] V. C. Müller, "Ethics of artificial intelligence and robotics," in *Stanford Encyclopedia of Philosophy*, E. Zalta, Ed. Palo Alto, Cal.: CSLI, Stanford University, 2020, pp. 1–70.

[9] G. S. Nelson, "Bias in artificial intelligence," *North Carolina medical journal*, vol. 80, pp. 220–222, 2019.

[10] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, 2019.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[12] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: Dual shot face detector," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5055–5064.

[13] D. Leslie, "Understanding bias in facial recognition technologies," *arXiv preprint arXiv:2010.07023*, 2020.

[14] W. Wójcik, K. Gromaszek, and M. Junisbekov, "Face recognition: Issues, methods and alternative applications," *Face Recognition-Semisupervised Classification, Subspace Projection and Evaluation Methods*, pp. 7–28, 2016.

[15] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[16] K. K. Saravanakumar, "The impossibility theorem of machine fairness–a causal perspective," *arXiv preprint arXiv:2007.06024*, 2020.

[17] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[18] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in ai: the casual conversations dataset," *arXiv preprint arXiv:2104.02821*, 2021.

[19] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[20] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525–5533.

[21] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[22] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019.

[23] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "Lffd: A light and fast face detector for edge devices," *arXiv preprint arXiv:1904.10633*, 2019.

[24] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[25] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Science (PNAS)*, vol. 113, 2016.