Casual Conversations: A dataset for measuring fairness in AI

Caner Hazirbas
hazirbas@fb.com

Joanna Bitton
jbitton@fb.com

Brian Dolhansky
bdol@fb.com

Jacqueline Pan
jackiepan@fb.com

Albert Gordo
agordo@fb.com

Cristian Canton Ferrer
ccanton@fb.com

Abstract

This paper introduces a novel "fairness" dataset to measure the robustness of AI models to a diverse set of age, genders, apparent skin tones and ambient lighting conditions. Our dataset is composed of 3,011 subjects and contains over 45,000 videos, with an average of 15 videos per person. The videos were recorded in multiple U.S. states with a diverse set of adults in various age, gender and apparent skin tone groups. A key feature is that each subject agreed to participate for their likenesses to be used. Additionally, our age and gender annotations are provided by the subjects themselves. A group of trained annotators labeled the subjects' apparent skin tone using the Fitzpatrick skin type scale [6]. Moreover, annotations for videos recorded in low ambient lighting are also provided. As an application to measure robustness of predictions across certain attributes, we evaluate the state-of-the-art apparent age and gender classification methods. Our experiments provides a through analysis on these models in terms of fair treatment of people from various backgrounds.

1. Introduction

Fairness in AI is an emerging topic in computer vision [3, 18] and has proven indispensable to develop unbiased AI models that are fair and inclusive to individuals from any background. Recent studies [5, 9, 19] suggest that top performing AI models trained on datasets that are created without considering fair distribution across sub-groups and thus quite unbalanced, do not necessarily reflect the outcome in real world. On the contrary, they may perform poorly and may be biased towards certain groups of people.

To address the aforementioned concerns, we propose a dataset composed of video recordings containing 3,011 individuals with a diverse set of age, genders and apparent skin types. Participants, who were paid actors gave their permission for their likeness to be used for improving AI, in the video recordings casually speak about various topics and sometimes depict a range of facial expressions. Thus, we call the dataset *Casual Conversations*. The dataset includes a unique identifier and age, gender, apparent skin type annotations for each subject. A distinguishing feature of our dataset is that age and gender annotations are provided by

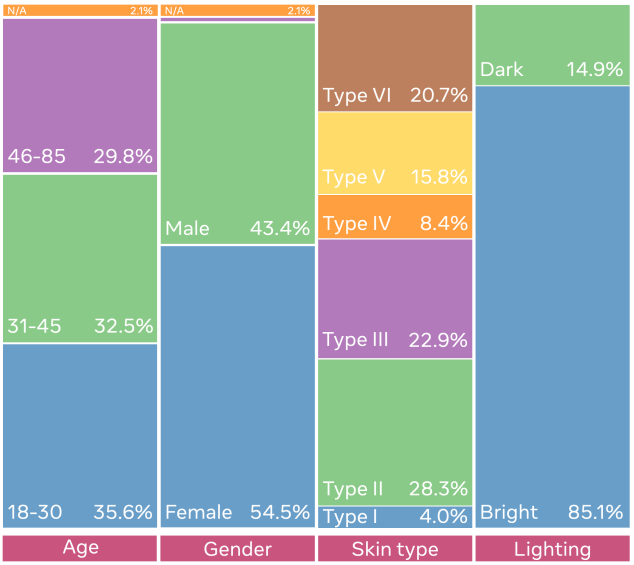

Figure 1: **Casual Conversations dataset** per-category distributions. Age and gender distributions are pretty balanced. Only 0.1% of participants identified themselves as *Other* gender (purple bar in the "Gender" column). Consecutive pairs of skin types can be grouped into three sub-categories for a uniform distribution. To balance lighting, we also provide sub-sampled dataset consists of two videos per actor, where one is *Dark* when possible.

This is the shorter version of our full paper at
https://arxiv.org/abs/2104.02821

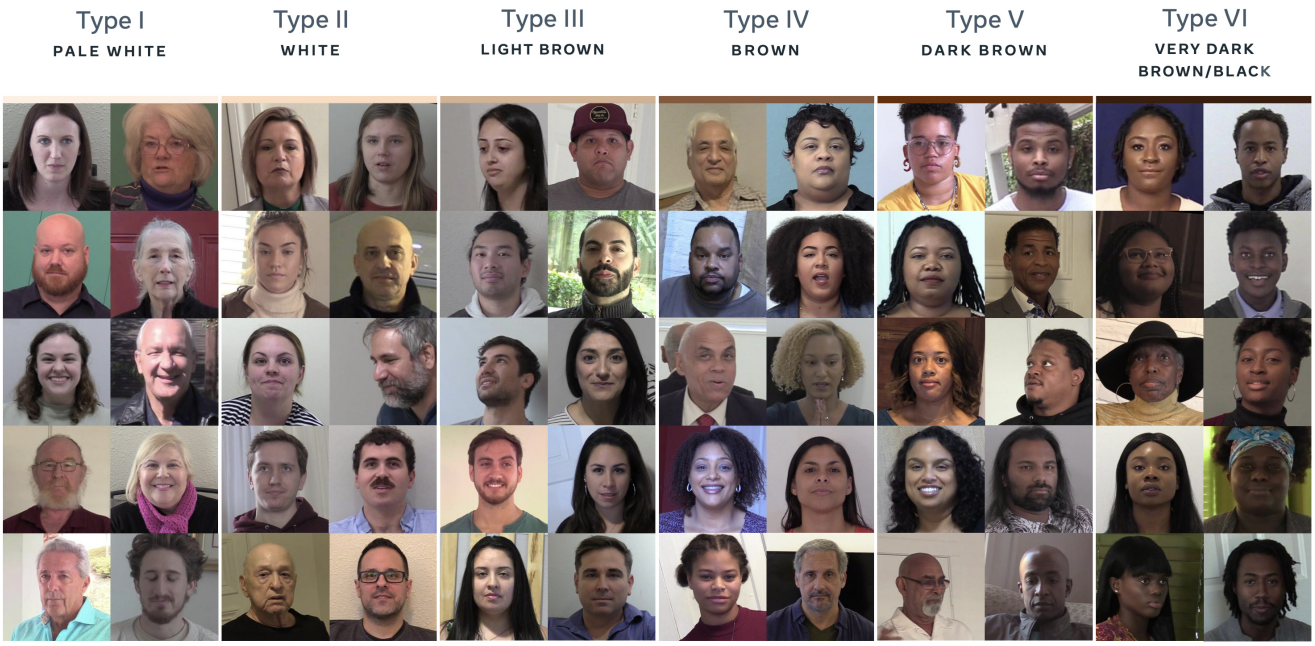

Figure 2: **Example face crops** from the *Casual Conversations* dataset, categorized by their apparent Fitzpatrick skin types.

the subjects themselves. We prefer this human-centered approach and believe it allows our data to have a relatively unbiased view of age and gender. As a third dimension in our dataset, we annotated the apparent skin tone of each subject using the Fitzpatrick[6] scale; we also label videos recorded in low ambient lighting. This set of attributes allow us to measure model robustness on four dimensions: age, gender, apparent skin tone and ambient lighting.

Although *Casual Conversations* is intended to evaluate robustness of AI models across several facial attributes, we believe that its value is greater and indispensable for many other open challenges. Image inpainting, developing temporally consistent models, audio understanding, responsible AI on facial attribute classification and handling low-light scenarios in the aforementioned problems are potential application areas of this dataset.

We organize the paper as follows; Section 2 provides a comprehensive background on fairness in AI, up-to-date facial attribute datasets, deepfake detection and current challenges in personal attribute classification. Section 3 describes the data acquisition process and the annotation pipeline for our dataset. Section 4 analyzes the biases of the state-of-the-art apparent age and gender classification models, using our dataset. Consequently, we finalize our findings and provide an overview of the results in Section 5.

2. Related Work

Fairness in AI challenges the field of artificial intelligence to be more inclusive, fair and responsible. Research has clearly shown that deep networks that achieve a high performance on certain datasets are likely to favor only sub-groups of people due to the imbalanced distribution of the categories in the data [10]. Buolamwini and Gebru [2] pointed out that the IJB-A [8] and Adience [5] datasets are composed of mostly lighter skin toned subjects. Raji *et al*. [13, 14] analyze the commercial impact Gender Shades [2] and discuss ethical concerns auditing facial processing technologies. Du *et al*. [4] provide a comprehensive review on recent developments of fairness in deep learning and discuss potential fairness mitigation approaches in deep learning. Meanwhile, Barocas *et al*. [1] are in the process of compiling a book that intends to give a fresh perspective on machine learning in regards to fairness as a central concern and discusses possible mitigation strategies on the ethical challenges in AI.

Facial attribute datasets [5, 9, 19] are created to train and validate face recognition, age, and gender classification models. However, provided facial attributes in these datasets are hand-labelled and annotated by third-parties. Although it has been claimed that the annotations are uniformly distributed over different attributes, *e.g*. age and gender, there is no guarantee on the accuracy of these annotations. An individual's visual appearance may differ significantly from their own self-identification which will thus

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
Levi & Hassner [12]	38.05	37.44	39.48	66.67	39.56	38.72	40.84	36.47	36.47	34.89	38.49	37.04
LMTCNN [11]	42.26	42.28	44.53	100.00	42.33	41.78	42.30	42.79	42.44	37.99	42.94	41.12
LightFace [16]	54.32	54.21	56.18	83.33	46.51	55.52	54.59	55.78	53.78	52.57	54.17	55.20

Table 1: **Precision** comparison of the apparent **age classification** methods, with a breakdown by fairness categories.

		Age			Skin type						Lighting	
	Overall	18-30	31-45	46-85	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
Levi & Hassner [12]	39.42	39.29	40.65	54.00	47.51	56.81	55.97	53.97	35.89	35.30	40.21	38.12
LMTCNN [11]	41.56	41.57	43.14	56.29	50.38	58.64	58.22	55.85	38.62	39.68	41.44	41.75
LightFace [16]	44.12	44.33	45.40	59.85	55.73	62.39	61.12	61.81	41.90	41.66	44.29	43.86

Table 2: **Precision** comparison of the apparent **gender classification** methods, with a breakdown by fairness categories.

result as bias in the dataset. In contrast, we provide age and gender annotations that are *self-identified* by the subjects. Aside from age and gender, public benchmarks tend to also provide annotated ethnicity labels. However, we find that labeling the ethnicity of subjects could lead to inaccuracies. Raters may have unconscious biases towards certain ethnic groups that may reduce the labelling accuracy in the provided annotations. In the FairFace [9] dataset paper, the authors claim that skin tone is a one dimensional concept in comparison to ethnicity because lighting is a big factor when deciding on the skin tone as a subject's skin tone may vary over time. Although these claims sound reasonable, the ethnicity attribute is still ill-defined and can conceptually cause confusions in many aspects; for example there may no difference in facial appearance of African-American and African people, although, they may be referred to with two distinct racial categories. We, therefore, have opted to annotate the apparent skin tone of each subject. Our dataset is composed of multiple recordings (*i.e.* on avg. 15) per actor, so annotators voted based on the sampled frames of these videos. Since these videos were captured in varying ambient lighting conditions, we alleviate the aforementioned concerns stated in [9].

Apparent age and gender classification has been a rapidly growing research field over a decade but recently took more attention after tremendous increase in social media usage. Therefore, apparent age and gender prediction is still an active research field investigated in automated human biometrics and facial attribute classification methods. Levi & Hassner [12] proposed an end-to-end trained Convolutional Neural Network (CNN) on Adience benchmark [5] to predict apparent age and gender from faces. Lee *et al*. [11] further developed a system for mobile and proposed an efficient, lightweight multi-task CNN for simultaneous apparent age and gender classification. Serengil *et al*. [16] recently presented a hybrid face recognition framework composed of the state-of-the-art face recognition approaches. Nevertheless, none of these models were evaluated against apparent skin type variations and ambient lighting conditions. Therefore, we present a close look into the results of these methods on our dataset to measure robustness of the recent face technologies.

3. Casual Conversations Dataset

The *Casual Conversations* dataset is composed of approximately fifteen one minute video recordings for each of our 3,011 subjects. Videos are captured in the United States in the cities of Atlanta, Houston, Miami, New Orleans, and Richmond. The subjects that participated in this study are from diverse age (18+), gender and ethnicity groups. In most recordings, only one subject is present; however, there are videos in which two subjects are present simultaneously as well. Nonetheless, we only provide *one set of labels* and it is for the current subject of interest.

In this dataset, we provide annotations for age, gender, apparent skin tone and whether or not the video was recorded in low ambient lighting. Age and gender attributes of subjects are provided by ***subjects themselves***. All other publicly available datasets provide hand or machine labelled annotations and therefore introduce a drastic bias towards appearance of a person other than the actual age and gender. Gender in our dataset is categorized as Male, Female, Other and N/A (preferred not to say or removed during data cleaning). We are aware that this categorization is over simplistic and does not sufficiently capture the diversity of genders that exist, and that we hope in the future there is more progress on enabling data analysis that captures this additional diversity while continuing to respect people's privacy and any data ethics concerns.

In addition to ***self-identified*** age and gender labels, we also provide skin tone annotations using the Fitzpatrick

scale [6]. Although the debate on ethnicity versus skin tone is still disputed [9], we believe it is ill-defined considering that the apparent ethnicity of a person may differ from their actual ethnicity, thereby causing algorithms to classify incorrectly. On the other hand, skin tone is an expressive and generic way to group people, which is necessary to measure the bias of the-state-of-the-art methods. The Fitzpatrick scale [6] is commonly used in classification of apparent skin tones. The Fitzpatrick scale constitutes six skin types based on the skin's reaction to Ultraviolet light. The scale ranges from Type I to VI, where Type I skin is pale to fair, never tans but always burns whereas Type VI skin is very dark, always tans but never burns (see example face crops in Figure 2). Additionally, the Fitzpatrick scale has limitations in capturing diversity outside of the Western theories of race related skin tone and does not perform as well for people with darker skin tones [15, 17]. Three of out the six skin types cover white skin, two cover brown skin, and there is only skin type for black skin, which clearly does not encompass the diversity within brown and black skin tones. A common procedure to alleviate this bias is to group the Fitzpatrick skin types into three buckets of light [types I, II], medium [types III & IV], and dark skins [type V & VI]. Our annotations provide the full, non-bucketed skin types such that others can decide how they'd to group the skin types.

In order to annotate for apparent skin types, eight individuals (raters) were appointed to annotate all subjects and to also flag the subjects that they are not confident about. As the final skin type annotations, we accumulated the weighted histograms over eight votes (uncertain votes are counted as half) and pick the most voted skin type as the ground-truth annotation.

Figure 1 shows the per-category distributions over our 3,011 subjects. As shown in the figure, we have decently balanced distributions over gender and age groups. For the skin type annotations, each paired group of types I & II, III & IV and V & VI would be almost equal to one-third of the dataset. Uniform distributions of the annotations allow us to reduce the impact of bias in our measurements and hence let us better evaluate model robustness.

In Figure 1 the percentage of bright versus dark videos over all 45,186 videos is also depicted. To have a balanced lighting distribution, we sub-sample our dataset to include only one pair of videos per subject, a total of 6,022 videos. When possible, we chose one *dark* and one *bright* video. Note that sub-sampling only affects the lighting distribution because there is only one set of labels per subject in the dataset. After re-sampling, we end up with 37.3% *dark* videos in the smaller dataset. In all experiments, we use the mini Casual Conversations dataset.

Our dataset will be publicly available[1] for general use and we encourage users to extend annotations of our dataset for various computer vision applications, in line with our data use agreement.

[1] https://ai.facebook.com/datasets/casual-conversations-dataset

4. Experiments

We compared the apparent age and gender prediction results of the three state-of-the-art models, evaluated on our dataset. In the following experiments, we used the reduced (mini) dataset. We first detect faces in each frame with DLIB [7] and evaluate the models on the sampled 100 face crops per video. Final predictions are calculated by aggregating results over these samples (most voted gender and median age). Levi & Hassner [12] and LMTCNN [11] predicts age in predefined brackets and therefore we map their age prediction to our predefined age groups in 1.

Tables 1 and 2 show the precision of the models on apparent age and gender, respectively. Levi & Hassner [12] is one of the early works that used deep neural networks. It is comparatively less accurate method among all, however, almost as good in apparent gender classification as LMTCNN [11]. LightFace [16], on the other hand, is more successful on predicting accurate apparent age and gender. Nevertheless, state-of-the-art methods' apparent gender precision on darker skin types (Type V & VI) is drastically lower by more than 20% on average.

5. Conclusions

We presented the *Casual Conversations Dataset*, a dataset designed to measure robustness of AI models across four main dimensions, *age*, *gender*, *apparent skin type* and *lighting*. As previously stated, a unique factor of our dataset is that the *age* and *gender* labels are provided by the participants themselves. The dataset has uniform distributions over all categories and could be used to measure various AI methods, such as face detection, apparent age and gender classification, or to assess robustness to various ambient lighting conditions.

As an application of our dataset, we presented an analysis of the recent apparent age and gender prediction models on our dataset. In both of the applications, we noticed an obvious algorithmic bias towards lighter skinned subjects. Apparent gender classification methods are most successful on older people (+45 years old) and generally as good on darker videos as on brighter ones.

Beyond aforementioned research topics, our dataset enables researchers to develop and also thoroughly evaluate models for more inclusive and responsible AI.

Acknowledgements. We would like to thank Ida Cheng and Tashrima Hossain for their help in regards to annotating the dataset for the Fitzpatrick skin type.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

[3] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[4] M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 2020.

[5] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 2014.

[6] Thomas B. Fitzpatrick. "Soleil et peau" [Sun and skin]. *Journal de Médecine Esthétique (in French)*, 2:33–34, 1975.

[7] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, Dec. 2009.

[8] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[9] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[10] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 2020.

[11] Jia-Hong Lee, Yi-Ming Chan, Ting-Yen Chen, and Chu-Song Chen. Joint estimation of age and gender from unconstrained face images using lightweight multi-task cnn for mobile applications. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018.

[12] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, June 2015.

[13] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[14] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

[15] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, and Vinodkumar Prabhakaran. Non-portability of algorithmic fairness in india, 2020.

[16] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020.

[17] Olivia R. Ware, Jessica E Dawson, M. Shinohara, and S. Taylor. Racial limitations of fitzpatrick skin type. *Cutis*, 2020.

[18] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020.

[19] Zhang Zhifei, Song Yang, and Qi Hairong. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.