Body Measurement Prediction Fairness

Alex J. Loosley [1], Amrollah Seifoddini[2], Alessandro Canopoli[2], and Meike Zehlike[1]

[1] *Zalando Research, Zalando SE, Valeska-Gert-Straße 5, 10243 Berlin, Germany*

[2] *Zalando Switzerland AG, Hardstrasse 201, 8005 Zürich, Switzerland*

Abstract

E-Commerce size and fit recommendations can be improved by using customer images to estimate body measurements. We are developing a body measurement fairness evaluation to determine if Zalando's AI/ML driven body measurement pipelines systematically underperform for customers of certain gender, body shape, or skin tone. The fairness evaluation produces four unfairness scores. The first is a gender unfairness score: the difference in average performance between genders. The second and third are body shape and skin tone unfairness scores: the difference in average performance between smallest and largest bodies, and lightest and deepest skin tones, respectively. The fourth is an intersectional unfairness score: the number of customers that are members of clusters associated with significant underperformance. We demonstrate the fairness pipeline on one body measurement pipeline candidate in the development stage showing that body shape receives the most significant unfairness score. This work allows us to catch unfair body measurement pipelines during experimentation and development stages to help our team avoid deploying unfair models into production.

Keywords

fairness, body segmentation, body measurements, skin tone labeling

1. Introduction

Zalando aims to provide high quality size and fit recommendations for all customers. Such recommendations are often made based on purchase history, but body measurement estimates based on images of customers in tight fitting clothes can further improve these recommendations. Zalando employs a mobile app-based body measurements pipeline for doing this in which customers are asked to take front and side pose photos of themselves, a segmentation model converts the two photos to binary body silhouettes, and the two body silhouettes are sent to the cloud for 3D body reconstruction and inference of body measurements for use in improving size and fit recommendations.

In conjunction with Zalando's values, the body measurements pipeline should be fair by not systematically underperforming for customers based on protected attributes such as gender, body shape, and so on. To that end, we are developing a body measurements fairness evaluation to track the unfairness of body measurements pipelines during the development stage so that deployment decisions can be made based on measures of fairness in addition to existing measures of performance and quality control.

2. Fairness Evaluation Dataset

2.1. Gender and Body Shape

Our work evaluates fairness with respect to three protected attributes, gender, body shape and skin tone, chosen by balancing the trade-offs between saliency and the feasibility of obtaining such data. We curated a fairness evaluation dataset by enriching a Zalando dataset composed of images of consenting customers posing in front and side positions and the gender they most identified with. We experimented with several techniques for quantifying body shape including using PCA based models, but opted to

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

EMAIL: aloosley@alumni.brown.edu (A. 1)

CEUR Workshop Proceedings (CEUR-WS.org)

use image body cross sectional area normalized by image height (referred to below as *normalized body x-section*) for fairness evaluation due to its simplicity and interpretability. Normalizing by image height removed variance caused by subjects standing at different distances from the camera. Assuming unbiased ground truth body silhouettes, using this simple, model-free, approach to encoding body shape reduced the likelihood of introducing new biases into the fairness evaluation dataset.

2.2. Skin Tone Labeling

Of the three protected attributes used for fairness evaluation, skin tone was the most challenging to obtain. Unfortunately, we could not ask the customers to self-report their skin tone. We abstained from automatically extracting skin tone using machine learning based models because such models are known to output biased results [1, 2]. We experimented with extracting skin tone by measuring the average pixel intensity of exposed skin based on skin silhouette annotations, henceforth skin albedo. Skin albedo as a quantity represents a combination of one's skin tone along with other factors such as lighting, skin reflectivity, skin redness, camera properties, image processing, and more. Therefore, evaluating for fairness using skin albedo does not enable one to disentangle which customers systematically receive inferior results based on skin tone.

To collect data for assessing skin tone fairness, we instead devised a labeling process aimed at mitigating some of the bias of skin tone labeling. The approach began with forming a small but diverse team consisting of four people: two men, two women, each having different racial backgrounds and expertise (one ethical data labeling expert, one computer vision expert working on size and fit problems, one responsible AI researcher, and one beauty product expert).

The process was broken down into two calibration labeling tasks (discussed below), one main labeling task where most of the images were labeled, and one post calibration labeling task to validate label quality and fix obvious labeling mistakes (Fig. 1). The goal of the calibration tasks was severalfold. First was to debug the labeling process including making sure our labelers had clear instructions and a clear view of label examples while labeling. Second was to establish labeler baselines so we could see if labeling statistics changed over time. The final goal was to have a reflection process for identifying and raising awareness of potential biases each labeler experienced so that each label could be mindful of such potential biases during the main labeling task. Some identified biases included: the tendency to choose a skin tone based on facial features associated with particular ethnicities, and the potential biasing effect of background objects on selecting skin tone. A total of 59 images were set aside for both calibration tasks, which took place several weeks apart. Thirty of the images from *Calibration 1* were relabeled during *Calibration 2* in order to assess labeler consistency.

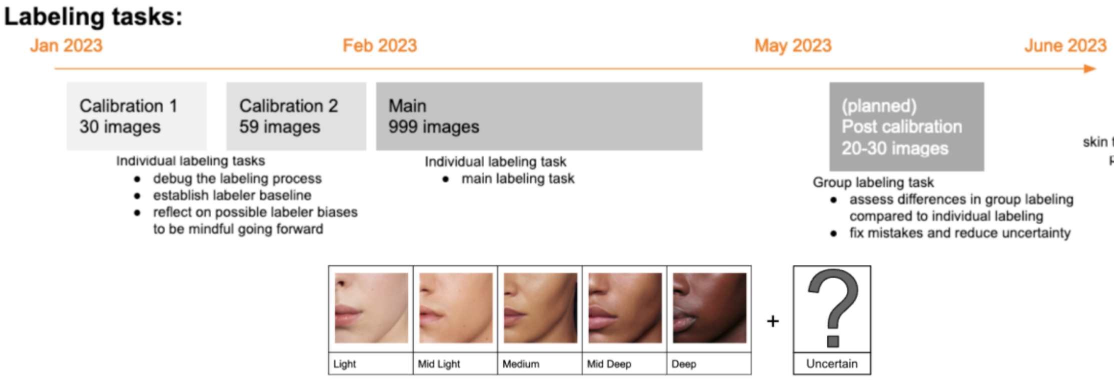
Labeling tasks:
Jan 2023
Feb 2023
May 2023
June 2023
Calibration 1
30 images
Calibration 2
59 images
Main
999 images
(planned)
Post calibration
20-30 images
skin tone labeling playbook
Individual labeling tasks
debug the labeling process
establish labeler baseline
reflect on possible labeler biases to be mindful going forward
Individual labeling task
main labeling task
Group labeling task
assess differences in group labeling compared to individual labeling
fix mistakes and reduce uncertainty
Light
Mid Light
Medium
Mid Deep
Deep
+
?
Uncertain

Figure 1. Skin tone labeling methodology. The labeling job was executed via multiple tasks to acquire labeling baselines and give labelers the chance to recognize sources of bias before labeling the bulk of images. Skin tone was labeled using a 5-point Zalando Beauty scale and labelers had the option to choose an uncertain label whenever they were not reasonably certain of their label choice (bottom panel).

Choosing a skin tone scale comprised of balancing the need for it to be representative, meaningful to others, and not too complicated for skin tone labelers. Until present, most skin tone data have been collected using the Fitzpatrick scale [3, 4]. Recently, the *Monk Skin Tone Scale* was proposed as a more representative scale than Fitzpatrick for fairness assessments [5]. However, the *Monk Skin Tone Scale* is complex, with ten different labels[2]. We found a middle ground by choosing the five-point *Zalando Beauty Skin Tone Scale* (Fig. 1, bottom) because five labels were simpler for labelers than ten labels, the labels were thoroughly tested with our customers, customers research found the label names inclusive, and labeling with this label set meant having labels that would be interpretable across Zalando. Labelers were allowed to label with two consecutive values if they thought the truth lay somewhere in the middle (i.e. both *mid-light* + *medium* was a valid label selection), and labelers were always allowed to choose *uncertain* if they did not have confidence in their ability to choose an accurate skin tone label. If there was uncertainty caused by more than one observable skin tone, labelers were asked to focus on the subject's cheeks.

Comparing mean skin tone label across labelers to skin albedo provided some interesting insights about the differences between the two measures (Fig. 2A). Overall, there was an expected negative correlation with a wide variance of skin tone labels given a particular value of skin albedo (and vice versa) indicating the importance of differentiating between the two measurements. Skin tone data, unlike skin albedo data, demonstrates the need to collect more data on subjects with deeper skin tone.

For the majority of images (62%), there was *near consensus* amongst labelers, meaning labels were within one ordinal value of each other (Fig. 2B). Skin tone labeling is highly subjective and labeling customer curated images taken under a variety of lighting conditions makes the process even more challenging. The consensus rates were, however, comparable between calibration and main tasks giving us confidence in the data, despite the rates being lower than those from a recent skin tone labeling experiment by *Krishnapriya et al.* [4] where color corrected, well lit, close-up images of faces were labeled with a *near consensus* rate of 96% (using the Fitzpatrick scale and three labelers).

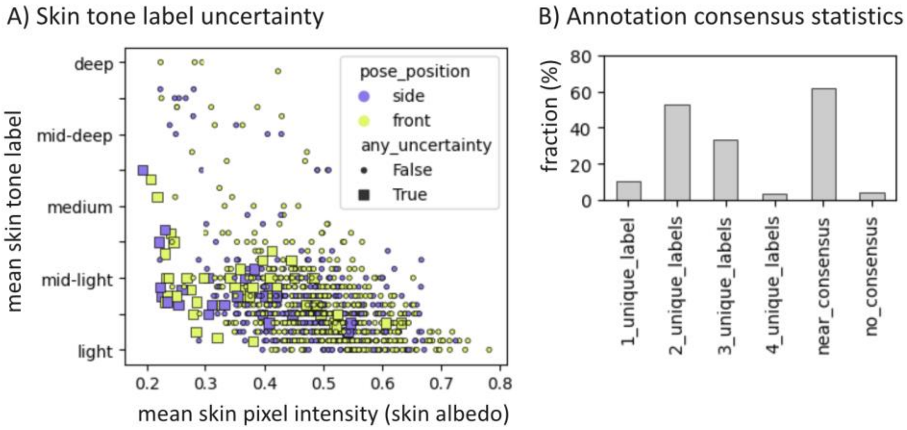

Figure 2. **Skin tone label uncertainty and inter labeler consensus.** (A) Mean skin tone label (across four labelers) vs. mean skin pixel intensity (0 is black, 1 is white). (B) Measures of agreement between labelers. Skin tone labels were all within one ordinal value of each other for 62% of images (*near consensus*).

3. Fairness Evaluation

Given the fairness evaluation dataset, we are implementing a fairness evaluation for the end-to-end body measurements pipeline as well as the individual processing steps (i.e. silhouette extraction and body reconstruction). From a systems engineering perspective, the former acts like an end-to-end test indicative of how fair the customer facing result might be when the system is deployed. The latter acts like a set of unit tests that help pinpoint causes of overall unfairness and provide applied scientists rapid model specific feedback during their experiments. For brevity, this article focuses on silhouette extraction fairness. To prevent potential application misuse that could occur by showing weaknesses in a live running Zalando application, our analysis and corresponding conclusions from this point on are based on a non-deployed demonstration model only.

[2] We should note that, after the experiment was designed and carried out, the authors behind the *Monk Skin Tone Scale* open sourced a skin tone examples dataset that should simplify the usage of this skin tone scale [6].

3.1. Performance Metrics

To measure silhouette extraction fairness, our evaluation calculates performance on each image in terms of an error length scale representing the average deviation between predicted and ground truth silhouette boundaries. Such an error length scale allows an evaluator to compare against an error threshold above which downstream size and fit recommendations might deviate by one size unit. We are currently trying to understand what a reasonable error threshold signifying underperformance impacting the end result should be (it is also garment specific). However, a conservatively low error threshold of *2 mm* is used for interpreting the results below.

3.2. Non-intersectional (Un)fairness Evaluation

Non-intersectional unfairness considers discrepancies in performance versus a single protected attribute. Gender unfairness was determined using a Kolmogorov–Smirnov two-sample test (p=0.05) to determine if the errors between male and female customers appeared to have been sampled from different distributions. If so, unfairness was deemed statistically significant and was scored as the difference of means between the two distributions. The severity of unfairness was determined by comparison to the *2 mm* error threshold. For scoring unfairness with respect to body shape and skin tone, linear regression was used. If the slope was significantly different from zero as determined by t-test, unfairness was deemed statistically significant and scored as the difference in performance between 5th and 95th percentile protected attribute values along the line of fit.

Given the demonstration model, no significant gender unfairness was observed for either side or front pose images (Fig. 3A). Body shape unfairness was significantly different from the null hypothesis for both side and front pose images (Fig. 3B). However, the corresponding unfairness scores were both well below the error threshold of *2 mm*, likely indicating no discernible difference in outcomes between small and large customers. Skin tone unfairness for side pose images was significantly different from the null hypothesis, but also well below the *2 mm* error threshold likely indicating no discernible difference in outcomes between customers with light and deep skin tone (Fig. 3C). The reliability of the latter result would be improved with more data on subjects with deeper skin tone.

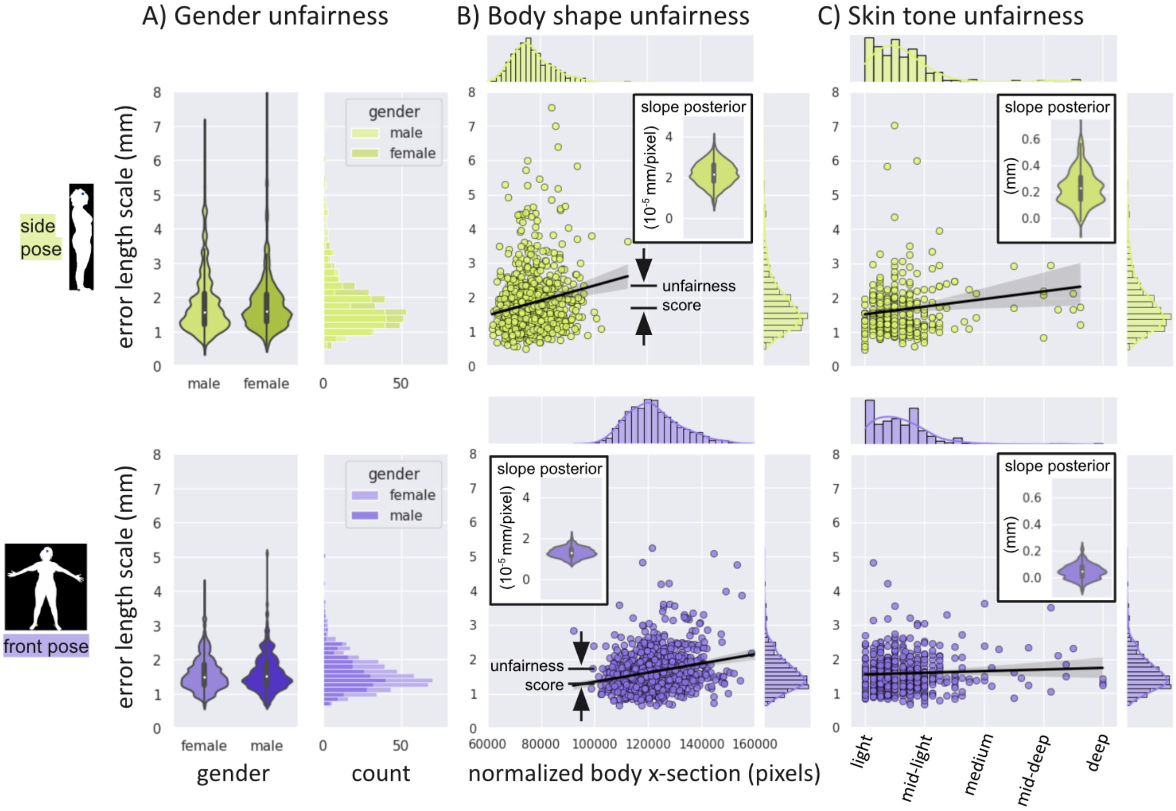

Figure 3. **Non-intersectional Fairness Evaluation.** Silhouette extraction fairness evaluation with respect to gender, body shape, and skin tone.

3.3. Intersectional (Un)fairness Evaluation

Intersectional unfairness was evaluated by seeking out clusters of subjects (with similar combinations of body shapes and skin tones) associated with significant underperformance. To do this, values of body shape and skin tone values were standardized by removing the mean and scaling to unit variance. Given standardized data points, the density-based clustering algorithm DB Scan [7] was used to search for clusters associated with underperformance. An ensemble of DB Scan models was trained varying the density hyperparameter $\varepsilon=10^{-1.5}, 10^{-1.49}, \ldots, 10^{-0.01}, 10^{0}$ (the lower the ε-value, the closer data points must be to be considered part of the same cluster). Any clusters with a minimum of 10 data points where at least 70% of data points had an error length scale $\geq 1.7\ mm$ were considered underperforming clusters. Given these underperformance criteria, DB Scans with density hyperparameters $10^{-0.1} \leq \varepsilon \leq 10^{-0.01}$ consistently identified one underperforming cluster: side posing subjects with deeper skin tones (Fig. 4). Although the exact underperformance criteria used here were somewhat arbitrary, applying stricter underperformance criteria led to no findings. Thus, this result represents a worst-case scenario for unfair performance.

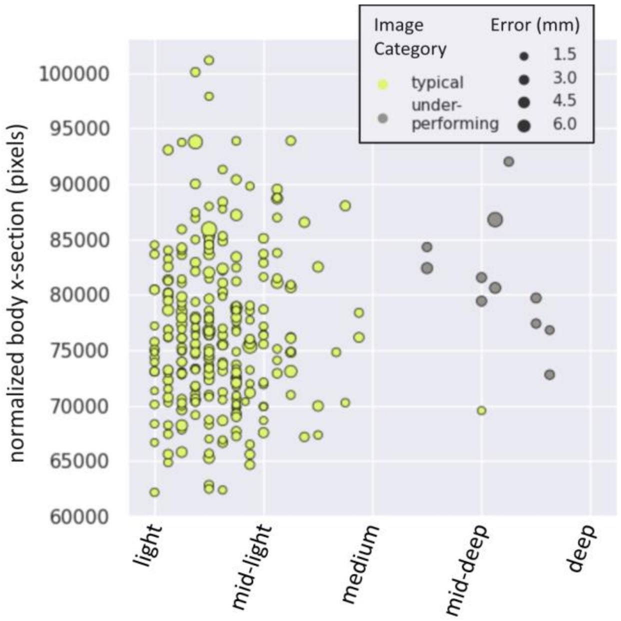

Figure 4. Intersectional Fairness Evaluation. DB Scan was used to identify neighborhoods of similar subjects encoded by body shape (normalized body x-section) and skin tone (see Fig. 2 for details) associated with significant underperformance. Point size represents error length scale. Only one underperforming cluster (gray points) was consistently identified: side posing subjects with deeper skin tones.

4. Conclusion

Ultimately, we believe this work will contribute to a higher customer satisfaction with size and fit recommendations as we are now able to identify and deploy body measurement pipelines that not only perform well, but also perform fairly. We hope this work can more generally provide a fairness evaluation playbook for others developing human-centric computer vision systems.

References

[1] H. Feng, T. Bolkart, J. Tesch, M. Black, V. Abrevaya, Towards Racially Unbiased Skin Tone Estimation via Scene Disambiguation (2022) doi: 10.48550/ARXIV.2205.03962

[2] J. J. Howard, Y. B. Sirotin, J. L. Tipton, A. R. Vemury, Reliability and Validity of Image-Based and Self-Reported Skin Phenotype Metrics, IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 4, pp. 550-560, Oct. 2021, doi: 10.1109/TBIOM.2021.3123550

[3] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. Archives of Dermatology, 124(6):869–871, 1988.

[4] K. S. Krishnapriya, M. C. King, K. W. Bowyer, Analysis of Manual and Automated Skin Tone Assignments for Face Recognition Applications, (2022), doi: 10.48550/ARXIV.2104.14685

[5] E. Monk, Monk Skin Tone Scale, URL: https://skintone.google, Accessed 2023/06/04

[6] E. Monk, MST-E dataset, URL: https://skintone.google/mste-dataset, Accessed 2023/06/04

[7] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996.