Analyzing the Impact of Gender Misclassification on Face Recognition Accuracy

Afi Edem Edi Gbekevi[1], Paloma Vela Achu[1], Gabriella Pangelinan[1], Michael C. King[1], Kevin W. Bowyer[2]
[1]Florida Institute of Technology, [2]University of Notre Dame

Abstract

Automated face recognition technologies have been under scrutiny in recent years due to noted variations in accuracy relative to race and gender. Much of this concern was driven by media coverage of high error rates for women and persons of color reported in an evaluation of commercial gender classification ("gender from face") tools. Many decried the conflation of errors observed in the task of gender classification with the task of face recognition. This motivated the question of whether images that are misclassified by a gender classification algorithm have increased error rate with face recognition algorithms. In the first experiment, we analyze the False Match Rate (FMR) of face recognition for comparisons in which one or both of the images are gender-misclassified. In the second experiment, we examine match scores of gender-misclassified images when compared to images from their labeled versus classified gender. We find that, in general, gender misclassified images are not associated with an increased FMR. For females, non-mated comparisons involving one misclassified image actually shift the resultant impostor distribution to lower similarity scores, representing improved accuracy. To our knowledge, this is the first work to analyze (1) the FMR of one- and two-misclassification error pairs and (2) non-mated match scores for misclassified images against labeled- and classified-gender categories.

1. Introduction

Facial recognition algorithms are known to perform worse on biological females than males. In 2018, Buolamwini *et al*. [2] brought widespread attention to the gap in gender classification accuracy by gender and skin tone. In response, media outlets fueled public attention with provocative headlines like "Facial Recognition Is Accurate, if You're a White Guy" [12] and "How is Face Recognition Surveillance Technology Racist?" [4]. However, these stories generally failed to explain or even recognize a key point: gender classification and face matching algorithms operate differently.

Gender classification is a face analytics technique - a method to provide non-unique user attributes ("soft biometrics") that may be used for a variety of purposes [5, 9]. However, soft biometrics alone cannot adequately distinguish between two individuals in an identity match scenario [9]. The accuracy of face analytics tools, which can also estimate traits like hair color, height, weight, and other properties, has been shown to vary across demographic groups [15, 2, 14].

The primary aim of this paper is to provide experiment-backed clarity on the intersection of gender classification and face recognition. Previous research [15, 2, 14] has operated on the assumption that face images can be classified into the binary categories of "male" or "female", and a classifed label can be compared to the data's annotated label to determine classification accuracy. The following experiments rely on the manually annotated gender labels in [18], though we recognize that the binary scheme may fail to accurately convey the gender identity or intended presentation of each subject. The implication is interesting: if a person intentionally chooses to present as a different gender, a "misclassification" may actually be considered an accurate assignment by the person involved.

2. Literature Review

2.1. Disparities in Gender Classification

Buolamwini and Gebru [2] evaluated three commercial gender classifiers (Microsoft, Face++, and IBM) using the Pilot Parliaments Benchmark (PPB) dataset, a small, self-collected set of public images of African and European parliament members. All three classifiers were shown to be more accurate for males than females. Additionally, after grouping images based on manually-assigned skin tone ratings, they showed that the classification error rate was higher for darker-skin-tone subjects. The maximum error rate for darker-skin-tone females was 34.7%, while the maximum error rate for lighter-skin-tone males was only 0.8%.

In response to [2], the three companies released new versions of their classifiers. Raji and Buolamwini [17] re-evaluated the classifiers with updated performance metrics, and found that all three had reduced accuracy disparities with respect to gender and race. In particular, errors on the darker-skin-tone female group were reduced up to

30.4%. Performance of the improved classifiers was measured against that of two non-target classifiers (Amazon and Kairos) on the PPB dataset. On the darker-skin-tone female group, the non-targets minimally achieved a 22.50% error rate. The best target error rate was 1.52%. The worst was 16.97%.

Muthukumar et. al [13] followed up on the work of Buolamwini and Gebru [2] to test whether skin tone itself was the driving factor in observed accuracy differences in gender classification. They report results of color-theoretic experiments with face images that "...raise the possibility that broader differences in ethnicity, as opposed to the skin type alone, are what contribute to unequal gender classification accuracy in face images" [13].

2.2. Disparities in Face Recognition

Cook et al. [3] examined the performance of eleven commercial matchers on demographically-divided data. They demonstrated that face matching accuracy and efficiency (as measured in transaction times) are affected by a combination of often co-occurring biological and behavioral demographic factors. These include gender, age, height, and skin reflectance. The individual impact and effect of each factor varies between systems, and all become less impactful as a system's overall accuracy increases. The study found that average mated scores significantly decreased for three categories: subjects who were younger in age, self-identified as female, or had lower skin reflectance. The decrease indicates a lower False Non-Match Rate (FNMR) for these groups, increasing the likelihood that a genuine match will be rejected.

Krishnapriya et al. [10] analyzed FNMR and False Match Rate (FMR) to evaluate face recognition accuracy by race and gender. They discovered that, generally, both the mated and non-mated distributions for the African American cohort were shifted toward higher similarity scores. Thus, for a given decision threshold, the African American cohort had a higher FMR and a lower FNMR than the Caucasian male, whose values traditionally provide the baseline. Krishnapriya et al. showed that, despite the higher African American FMR value, the d-prime value given by some matchers showed that the ability to "cleanly" divide mated and non-mated scores is about equal across cohorts.

Albiero et al. [1] investigated commonly speculated causes of the "gender gap" in recognition accuracy. They reported that females typically exhibited a greater range of facial expressions. Males more often had neutral expressions, resulting in higher similarity scores with other neutral males. Females tended to have more facial occlusions, typically associated with hairstyle. Removing these occlusions improved the female d-prime values, indicating enhanced system ability to distinguish between image instances. Female mated and non-mated distributions remained less separable than male distributions - even with a matcher trained on an explicitly gender-balanced dataset.

2.3. Combined Analysis

There is very little work related to the intersection of gender classification and face recognition. In 2021, Qiu et al. [16] first examined the issue, reporting that the relationship between the two tasks varied across demographics. The study included three gender classifiers and two face-matching algorithms. To evaluate the combined effect, they recorded the mean scores and score distributions of pairs containing zero, one, or two gender-misclassified images. They reported that, on average, non-mated pairs with one misclassified image had lower similarity scores. Pairs with two misclassifications had higher similarity scores. From this result, they postulated that one-error pairs should have the lowest False Match Rate (FMR).

In this work, we analyze the same dataset and gender classifiers as [16] to provide the FMR values associated with each demographic and number of errors. In addition to the same open-source matcher, we include results from the newest version of a top-performing commercial matcher. Finally, we consider another angle, analyzing how the match scores of gender-misclassified images vary when compared to their labeled versus classified gender.

3. Experiment Design

3.1. The MORPH Dataset

The MORPH dataset [18] has been used extensively in demographic-focused face recognition research. [10, 11, 1] It contains mugshot images acquired with controlled lighting and an 18% gray background. They nominally feature a front pose and a neutral expression. Subjects are divided into four cohorts based on manually annotated demographic labels. This study uses a subset of MORPHv3 curated to remove duplicates, twins, and mislabeled images. The curated version contains 24,857 images of 5,929 African American females (AAF), 56,245 images of 8,839 African American males (AAM), 10,941 images of 2,798 Caucasian females (CF), and 35,276 images of 8,835 Caucasian males (CM).

3.2. Gender Classification

We produce gender classification results with three algorithms: two commercial (Amazon, "AM" and Microsoft, "MS") and one open-source ("OS") based on ArcFace [7].

Sample misclassified images are given in Figure 1. The top and bottom rows represent the African American and Caucasian cohorts, respectively. The images in the left column were labeled as female in the dataset, but classified as male. In the right column, the opposite case is shown.

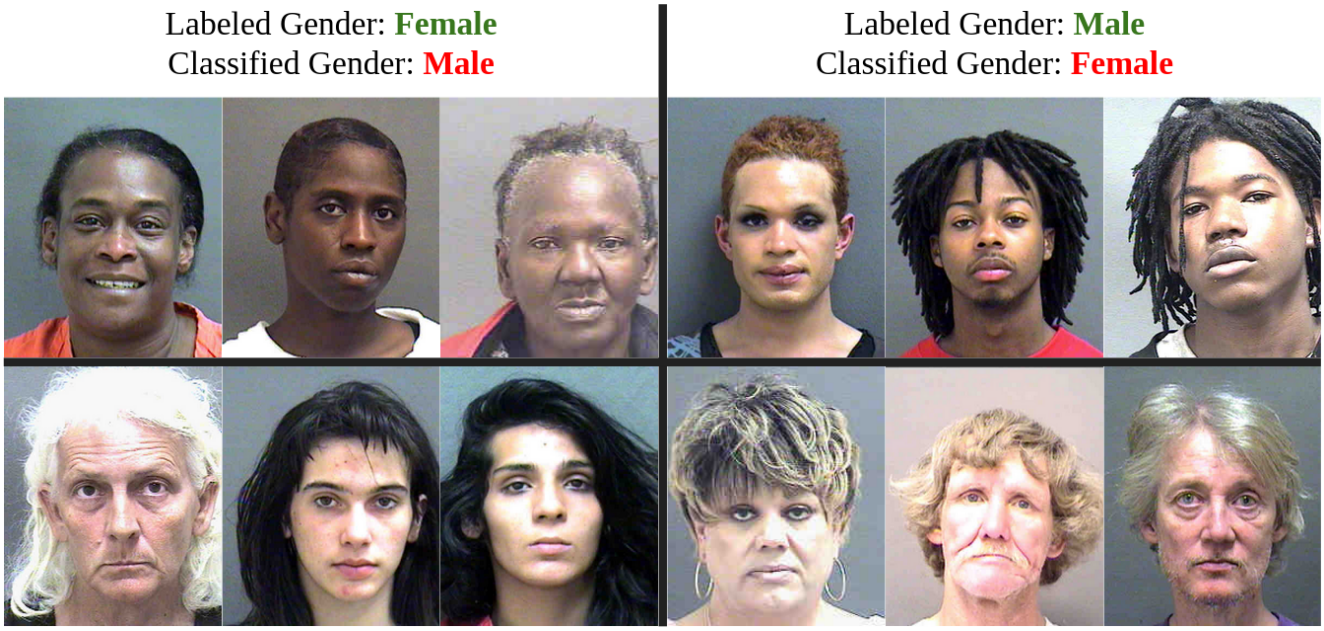


Figure 1: Sample gender-misclassified images for each demographic.

	OS		AM		MS	
	Cor	*Inc*	*Cor*	*Inc*	*Cor*	*Inc*
AAF	20676	4181	23098	1759	23927	926
AAM	55090	1155	55192	1053	55805	405
CF	10022	919	10710	231	10829	111
CM	34993	283	35105	171	35203	41

Table 1: Number of correct/incorrect classifications across demographics.

Table 1 gives the number of gender misclassification errors for each classifier, where "Cor" indicates a correct gender classification and "Inc" an incorrect classification. In general, Microsoft is found to be the best-performing classifier, and the open-source classifier the worst. Corresponding accuracy rates are provided in Table 2. For all three classifiers, accuracy is highest for CM and lowest for AAF.

	OS	**AM**	**MS**
	GC Accuracy (%)		
AAF	**83.2**	**92.9**	**96.3**
AAM	97.9	98.1	99.2
CF	91.6	97.9	99.0
CM	**99.2**	**99.5**	**99.8**

Table 2: Accuracy and error rates by gender classifier.

3.3. Face Recognition

For the face recognition task, we use two matchers: one open-source ("OS") and one commercial-off-the-shelf ("COTS"). The OS matcher is based on ArcFace and publicly-available weights trained on the MS1MV2 dataset, an accessible and curated version of MS1M, a large-scale recognition dataset [8]. The COTS matcher is one of the highest-performing commercially available, and we use the most recently updated version.

In order to perform face recognition, the target database must contain two or more images per subject to complete the matching process. Since MORPH is a longitudinal database containing multiple images of each subject, it is ideal for use in recognition experiments.

Gender classification analyzes one image to make a binary classification of that image, but face recognition analyzes a pair of images to compute a similarity (or dissimilarity) score. A mated (or genuine) pair of images is a comparison between two images of the same individual. A non-mated (or impostor) pair is a comparison between images of different individuals. Similarity scores resulting from comparisons of a pair of images are evaluated against predefined decision thresholds for each matcher. Non-mated pairs that exceed the threshold are said to generate a "false match" for the given matcher and threshold. The two images would be (falsely) identified as a match (i.e. containing the same individual).

We use the 1-in-10,000 (1-in-10k) impostor threshold, calculated with respect to the Caucasian male impostor distribution, to make identity decisions. Similar demographic-

focused works [6, 10, 11] have used this threshold. It corresponds to the non-mated similarity score at which one error occurs for 10,000 non-mated pair comparisons. For the OS matcher, which scales scores from -1 to 1, the 1-in-10k CM threshold is 0.3483. The COTS matcher gives scores from 0 to 1, and yields a threshold of 0.7550. The impostor scores at and around the 1-in-10k threshold are in the "high-likelihood false match region", i.e. they are likely to generate a false match. [10]

Relative frequency histograms of each demographic's non-mated (genuine) and mated (impostor) distributions are given in Figures 2. The d-prime value measures the distance between the distributions: a higher d-prime value indicates greater separation of genuine and impostor distributions.

4. Experimental Results

4.1. Error Pair Analysis

For each demographic, we examine the FMR associated with (1) all non-mated pairs, (2) non-mated pairs containing one gender-misclassified image ("one-error pairs"), and (3) non-mated pairs containing two gender-misclassified images ("two-error pairs"). Table 3 provides the percentage of each demographic's full score distribution comprised by each error-pair type, based on the given gender classifier.

GC	Dem	1-Error	2-Error
OS	AAF	27.98%	2.83%
	AAM	4.02%	0.04%
	CF	15.39%	0.70%
	CM	1.59%	0.01%
AM	AAF	13.15%	0.50%
	AAM	3.67%	0.03%
	CF	4.13%	0.04%
	CM	0.96%	0.002%
MS	AAF	7.17%	0.14%
	AAM	1.43%	0.01%
	CF	2.01%	0.01%
	CM	0.23%	0.00%

Table 3: Proportion of one- and two-error pairs versus all image comparisons.

Figure 3 gives the FMR (%) for each pair type: all non-mated pairs (column 2) and non-mated pairs containing either one or two gender-misclassified images. For each demographic and matcher, the FMR for *all* non-mated pairs is taken as the baseline. The "OS Error" and "COTS Error" rows indicate whether FMR increased or decreased from the baseline for the one- and two-error pairs. Insufficient data is reflected as "N/A" in the respective sub-tables.

4.1.1 One-Error Pairs

In Figure 3, the OS- and COTS-matcher FMR results are consistent for each demographic and all three classifiers. For the two female groups, FMR decreases from the baseline when only one misclassified image is involved in a comparison. For the male groups, FMR either increases or does not change versus the baseline.

4.1.2 Two-Error Pairs

In Figure 3, two-error pairs generally yield an increased FMR, with the exception of the open-source classifier and matcher combination for African American females. The two-error FMR increase is very slight for the female groups, while African American males show a more significant increase (from a baseline of about 0.04% to 0.54% maximally). The Caucasian groups generally do not have enough data to report on a meaningful FMR change (other than the female open-source classifier case, which shows an on-trend increase).

4.2. Cross-Gender Comparison

We proceed with the most accurate gender classifier (Microsoft) and face matcher (COTS) from the first experiment. The Microsoft classifier gives 41 errors for CM, 111 for CF, 405 for AAM, and 926 for AAF. For each misclassified image, we consider its labeled gender (provided by the dataset) and its classified gender (output by the Microsoft classifier).

For each misclassified image, the highest non-mated match scores associated with (1) a labeled-gender image and (2) a classified-gender image are recorded. Misclassified male images are compared to all other male images and all female images, and vice versa. The resulting match scores are visualized as boxplots and relative frequency histograms in Figures 4 (Caucasian) and 5 (African American). In each plot, the COTS decision threshold of 0.755 is indicated with the dashed black line.

The CM histogram shows the greatest separation between the labeled-gender distribution (green) and its classified-gender counterpart (red). In fact, all of the classified-gender scores fall below the threshold, and would correctly be classified as impostors. For CF, however, there is significant overlap between the two score sets - with the *classified*-gender scores shifted towards higher values. The classified-gender scores also feature a slightly higher FMR than the labeled-gender.

For AAM, labeled-gender and classified-gender scores show significant overlap. AAF scores overlap slightly less. For both males and females, the majority of scores in both gender groups cross the decision threshold.

The average, highest, and lowest non-mated scores for each comparison type are given in Table 4. In the average column, a red-highlighted value indicates that the average

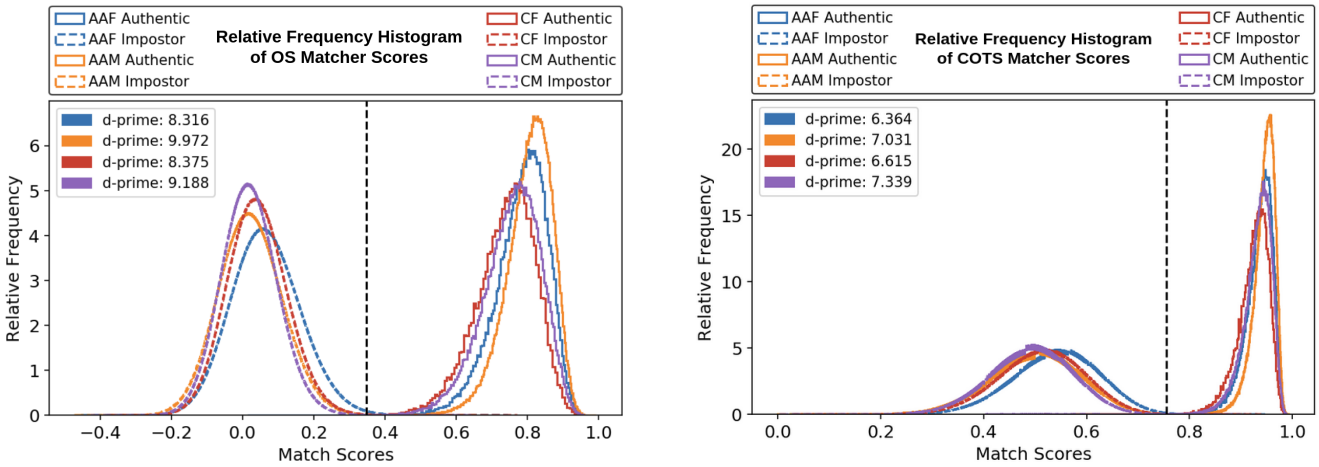


Figure 2: Distribution of genuine and impostor scores generated by the open-source (left) and commercial (right) matchers.

	Classifier		**Microsoft**		**Open-Source**		**Amazon**	
		All Pairs	1-Error	2-Error	1-Error	2-Error	1-Error	2-Error
AAF	OS FMR (%)	0.30	0.15	0.34	0.236	0.240	0.237	0.429
	OS Error	*Baseline*	Dec	Inc	Dec	Dec	Dec	Inc
	COTS FMR (%)	0.175	0.105	0.229	0.141	0.229	0.148	0.290
	COTS Error	*Baseline*	Dec	Inc	Dec	Inc	Dec	Inc
AAM	OS FMR (%)	0.039	0.046	0.54	0.049	0.25	0.050	0.33
	OS Error	Baseline	Inc	Inc	Inc	Inc	Inc	Inc
	COTS FMR (%)	0.0376	0.050	0.399	0.051	0.264	0.052	0.303
	COTS Error	Baseline	Inc	Inc	Inc	Inc	Inc	Inc
CF	OS FMR (%)	0.040	0.024	0.05	0.039	0.07	0.029	0.06
	OS Error	Baseline	Dec	N/A*	Dec	Inc	Dec	N/A*
	COTS FMR (%)	0.039	0.023	0.05	0.037	0.076	0.032	0.061
	COTS Error	Baseline	Dec	N/A*	Dec	Inc	Dec	N/A*
CM	OS FMR (%)	0.009	0.007	0	0.01	0.035	0.01	0.013
	OS Error	Baseline	N/A*	N/A*	Inc	N/A*	Inc	N/A*
	COTS FMR (%)	0.010	0.012	0	0.010	0.063	0.010	0.028
	COTS Error	Baseline	Inc	N/A*	No Change	N/A*	No Change	N/A*

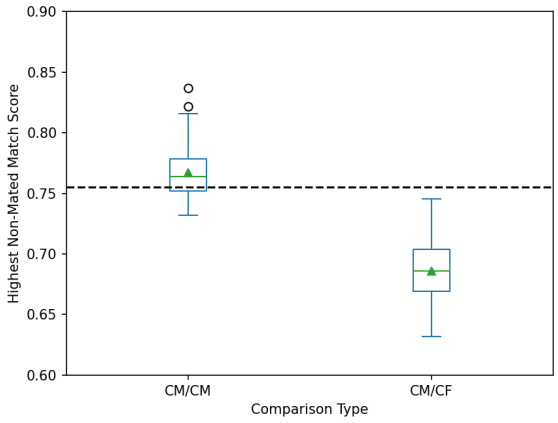
Figure 3: False match rate (FMR) for each demographic and error-pair type, divided by gender classifier and recognition system. Each system's baseline error is given by the FMR associated with all image pairs (match scores).

classified gender comparison score is higher than that of the labeled gender for the given demographic. A green value indicates the opposite.

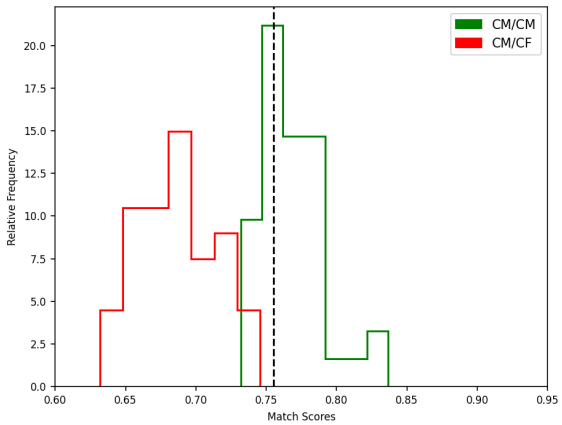
CF is the only group for which comparisons against the classified gender give higher average match scores than against the labeled gender. For the other three demographic groups, the labeled-gender averages are higher. In all cases, the discrepancy between averages is minimal.

5. Conclusions and Discussion

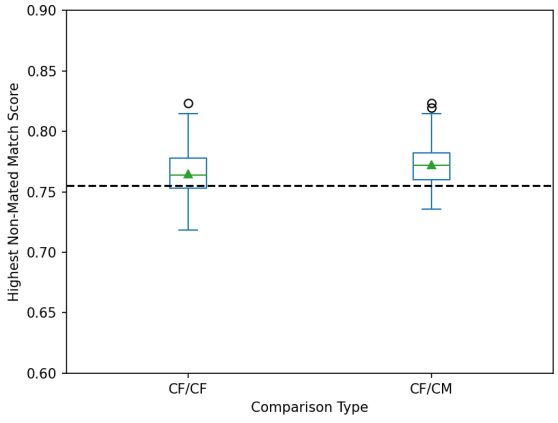
The primary goal of this work is to provide clarity on whether errors in the gender classification task lead to errors in the separate task of face recognition. First, we seek to understand the relationship between gender-misclassified face images and face recognition accuracy, as measured by False Match Rate (FMR). Our primary findings regarding one- and two-error pairs are summarized as follows:



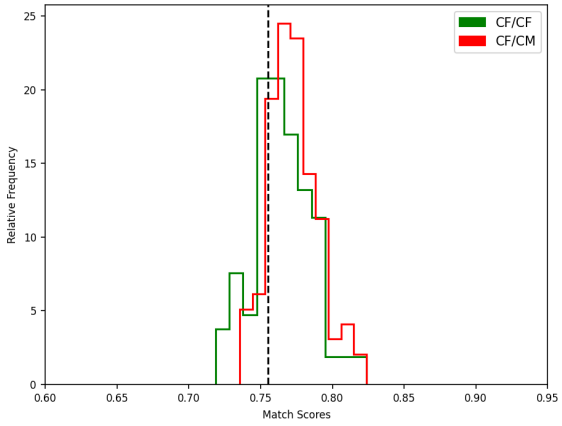
(a) Scores of misclassified CM images vs. images from the labeled (CM, left) and classified gender (CF, right).



(b) Score distribution for misclassified CM images vs. images from the labeled (CM, green) and classified gender (CF, red).



(c) Scores of misclassified CF images vs. images from the labeled (CF, left) and classified gender (CM, right).



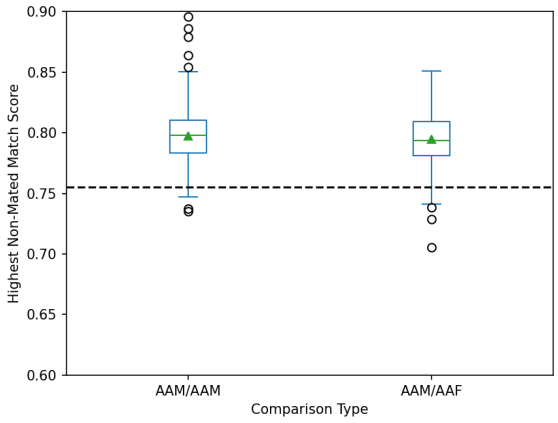
(d) Score distribution for misclassified CF images vs. images from the labeled (CF, green) and classified gender (CM, red).

Figure 4: Highest non-mated match scores of gender-misclassified Caucasian images versus other images in their labeled and classified gender categories. The 1-in-10k CM threshold (0.755) is shown as a dashed black line.

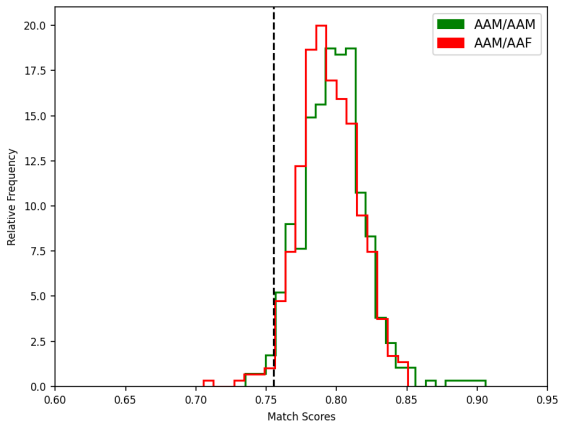
Comparison	Average	Highest	Lowest
CF vs. CF	0.765	0.824	0.719
CF vs. CM	**0.772**	0.824	0.735
CM vs. CM	**0.767**	0.837	0.732
CM vs. CF	0.686	0.746	0.632
AAM vs. AAM	**0.798**	0.906	0.735
AAM vs. AAF	0.795	0.805	0.705
AAF vs. AAF	**0.799**	0.907	0.741
AAF vs. AAM	0.782	0.857	0.718

Table 4: Average, highest, and lowest non-mated match scores for each demographic's labeled-gender and classified-gender comparisons.

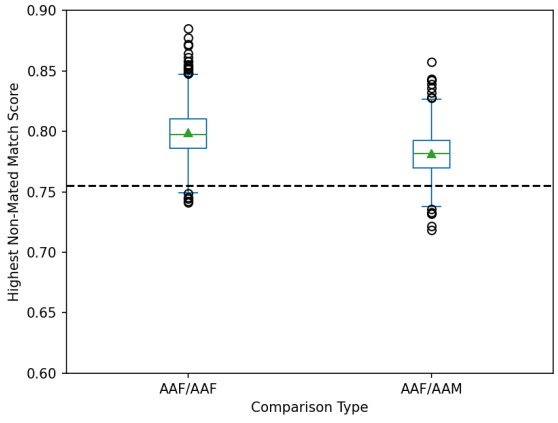
For females, non-mated comparisons involving only one misclassified image have a slightly lower FMR. Figure 3 shows that one-error pairs consistently give a lower FMR than the all-pair baseline for Caucasian and African American female groups. This finding aligns with that of [16]: that an image generating a gender-classification error is slightly less likely to participate in a false match-inducing image pair. Interestingly, this does not seem to be the case for males. The African American male FMR increases for one-error pairs across all three classifiers and both matchers. The Caucasian male FMR follows the same trend, though there was insufficient data to characterize the effect well.



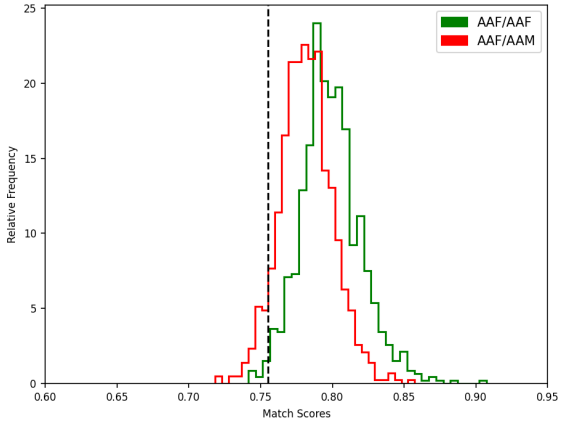
(a) Scores of misclassified AAM images vs. images from the labeled (AAM, left) and classified gender (AAF, right).



(b) Score distribution for misclassified AAM images vs. images from the labeled (AAM, green) and classified gender (AAF, red).



(c) Scores of misclassified AAF images vs. images from the labeled (AAF, left) and classified gender (AAM, right).



(d) Score distribution for misclassified AAF images vs. images from the labeled (AAF, green) and classified gender (AAM, red).

Figure 5: Highest non-mated match scores of gender-misclassified African American images versus other images in their labeled and classified gender categories. The 1-in-10k CM threshold (0.755) is shown as a dashed black line.

For African American males, non-mated comparisons containing at least one gender-misclassified image have a slightly higher FMR. Each combination of pair-type, matcher, and classifier in Figure 3 yields an increased FMR versus the baseline for African American males. While one-error pairs give only a slight increase in FMR (from about 0.04% to 0.05%), two-error pairs increase it up to 0.54%.

Our second question involves the relationship between a gender-misclassified image and its labeled versus classified gender. Using the best classifier and matcher from the previous experiment, we report the following observations:

Misclassified Caucasian female images produce slightly but insignificantly higher similarity scores with Caucasian male images. In Table 4, the "CF vs. CM" comparison gives the average highest non-mated match score for misclassified Caucasian female images versus all male images. This average is higher than that of the labeled-gender "CF vs. CF" comparison, though the score disparity is insignificant.

Misclassified males and African American females produce slightly but insignificantly higher similarity scores with their labeled-gender images. For each of these groups, the average score of the labeled gender comparison is higher than that of the classified gender

(Table 4). Again, however, the score disparity is minimal enough to be insignificant.

The second experiment only considered the highest non-mated score associated with each misclassified image "probe". In future work, we will augment the analysis to include each demographic's full impostor distribution with its labeled versus classified gender.

The lack of Caucasian male data complicates the task of accurately evaluating the relationship between gender classification and recognition accuracy. The small count of misclassified images does not seem related to representation in the data. With 56k images, CM is the second-largest cohort, and essentially ties with African American males for most unique subjects (8,835). For the algorithms and data used in this experiment, Caucasian males are simply found to be the most gender-classifiable. Even with the worst-performing classifier, only 1.59% of non-mated pairs contain a misclassified image - *especially* low considering the African American female group's corresponding 27.98%. As a result of this discrepancy, there were not enough Caucasian male classification errors to draw any substantive conclusions.

Though media reports on recent research have often conflated the two, our results suggest that errors in gender classification do *not* cause errors in recognition. In fact, pairs involving one misclassified image actually improve the FMR for two demographics. For the other two groups, FMR increase is minimal. In the two-error case, FMR increase is relatively consistent but also minimal. Given the very small portion of scores two-error comparisons represent, it is unlikely that this case would cause significant impact in a real-world recognition scenario.

References

[1] Vitor Albiero, K. S. Krishnapriya, Kushal Vangara, Kai Zhang, Michael C. King, and Kevin W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 81–89, 2020.

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[3] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 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, 2019.

[4] Kade Crockford. How is face recognition surveillance technology racist? *ACLU News and Commentary*, Jun 2020. https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist/.

[5] Antitza Dantcheva, Petros Elia, and Arun Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2016.

[6] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. 2019.

[7] Jia Guo and Jiankang Deng. Insightface. `https://github.com/deepinsight/insightface`, 2021.

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, 2016.

[9] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Can soft biometric traits assist user recognition? In *Biometric technology for human identification*, volume 5404, pages 561–572. Spie, 2004.

[10] K. S. Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[11] K. S. Krishnapriya, Gabriella Pangelinan, Michael C. King, and Kevin W. Bowyer. Analysis of manual and automated skin tone assignments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 429–438, January 2022.

[12] Steve Lohr. Facial recognition is accurate, if you're a white guy. *The New York Times*, Feb 2018. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

[13] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilović, and Kush R. Varshney. Color-theoretic experiments to understand unequal gender classification accuracy from face images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2286–2295, 2019.

[14] Vidya Muthukumar, Tejaswini Pedapati, Nalini K. Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. Understanding unequal gender classification accuracy from face images. *CoRR*, abs/1812.00099, 2018.

[15] Mei Ngan and Patrick Grother. Face recognition vendor test (frvt) - performance of automated gender classification algorithms. 2015.

[16] Ying Qiu, Vítor Albiero, Michael C King, and Kevin W Bowyer. Does face recognition error echo gender classification error? *arXiv preprint arXiv:2104.13803*, 2021.

[17] 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*, pages 429–435, 2019.

[18] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006.