Cross-Gender and 1-to-N Face Recognition Error Analysis of Gender Misclassified Images

Paloma Vela Achu
Cross-Gender and 1-to-N Face Recognition Error Analysis of Gender Misclassified Images

by

Paloma Vela Achu

A thesis submitted to the Department of Computer Engineering and Sciences of Florida Institute of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

Melbourne, Florida
May, 2022
We the undersigned committee hereby approve the attached thesis,

“Cross-Gender and 1-to-N Face Recognition Error Analysis of Gender Misclassified Images”

by

Paloma Vela

________________________
Michael C. King, Ph.D.
Associate Professor
Computer Engineering and Sciences
Major Advisor

________________________
Kevin Bowyer, Ph.D.
Professor
Computer Science and Engineering
University of Notre Dame

________________________
Vanessa A. Edkins, Ph.D.
Professor
School of Psychology

________________________
Philip J. Bernhard, Ph.D.
Associate Professor and Department Head
Computer Engineering and Sciences
Abstract

Title: Cross-gender and 1-to-N Face Recognition Error Analysis of Gender Misclassified Images

Author: Paloma Vela Achu
Major Advisor: Michael King, Ph.D.

A number of recent research studies have shown that face recognition accuracy is meaningfully worse for females than males. Gender classification algorithms also perform worse: one commercial classifier gives a 7% error rate for African-American females vs. 0.5% for Caucasian males. In response to these observations, we consider one primary question: do errors in gender classification lead to errors in facial recognition? We approach this question by focusing on two main areas (1) do gender-misclassified images generate higher similarity scores with different individuals from the false-gender category versus their true-gender category? (2) What is the impact of gender misclassified images on the performance accuracy of the system? We find that (1) for all demographic groups, except for the African American Male, non-mated pairs of subjects with at least one gender-misclassified image have a higher False Match Rate (FMR) with their ground truth gender compared to their erroneously projected gender group. Similarly, on average and across demographics groups, gender-misclassified subjects still have higher similarity scores with subjects of their true gender than those of the falsely classified gender. (3) There was no significant impact on the 1-to-N accuracy when using the open-source algorithm, ArcFace, whereas for the commercial matcher, there seems to be a decline in performance accuracy for misclassified images.
To our knowledge, this is the first work to analyze and match scores for gender misclassified images against both the false-gender category and the true-gender category and extend the work from an identification standpoint.
Table of Contents

Abstract .................................................................................................................... iii

List of Figures ........................................................................................................ vii

List of Tables ......................................................................................................... ix

Chapter 1 Introduction ......................................................................................... 1

  Background ........................................................................................................... 4

Chapter 2 Literature Review .............................................................................. 10

  Accuracy Disparities in Gender Classification ................................................. 10

  Accuracy Disparities in Face Recognition ....................................................... 11

  Classification and Recognition ........................................................................ 12

Chapter 3 Experiment Design ............................................................................ 14

  MORPH Dataset ............................................................................................... 15

  Gender Classification Algorithms ................................................................ 17

  Face Recognition Matchers ........................................................................... 18

Chapter 4 Results: Cross-Gender Verification Rate Error Analysis ............ 19

  Analysis ............................................................................................................. 19

    Face Verification ............................................................................................... 19

    Analysis of African American Female Cohort ........................................... 21

    Analysis of African American Male Cohort ............................................... 27

    Analysis of Caucasian Female Cohort ......................................................... 32

    Analysis of Caucasian Male Cohort ............................................................ 36

    Discussion of Cross-Gender Verification Rate Error Analysis ............. 41
Chapter 5 Results: 1-to-N Identification Error Rate Analysis........ 46

Challenges and Limitations .............................................................. 60

Chapter 6 Conclusion................................................................. 62

References ................................................................................... 64
List of Figures

Figure 1: Score distribution and box plot of results for ArcFace Matcher with Open Source (a,b), Amazon (c,d) and Microsoft (e,f) ......................... 23
Figure 2: Score distribution and box plot of results for COTS-A Matcher with Open Source (a1,b1), Amazon (c1,d1) and Microsoft (e1,f1) .............. 24
Figure 3: Score distribution and box plot of results ArcFace Matcher with Open Source (a2,b2), Amazon (c2,d2) and Microsoft (e2,f2) .............. 28
Figure 4: Score distribution and box plot of results COTS-A Matcher with Open Source (a3,b3), Amazon (c3,d3) and Microsoft (e3,f3) .............. 29
Figure 5: Score distribution and box plot of results ArcFace Matcher with Open Source (a4,b4), Amazon (c4,d4) and Microsoft (e4,f4) .............. 33
Figure 6: Score distribution and box plot of results COTS-A Matcher with Open Source (a5,b5), Amazon (c5,d5) and Microsoft (e5,f5) .............. 34
Figure 7: Score distribution and box plot of results ArcFace Matcher with Open Source (a6,b6), Amazon (c6,d6) and Microsoft (e6,f6) .............. 37
Figure 8: Score distribution and box plot of results COTS-A Matcher with Open Source (a7,b7), Amazon (c7,d7) and Microsoft (e7,f7) .............. 38
Figure 9: AAF gender misclassified subjects who failed to rank 1 in 1-to-N matching ................................................................. 48
Figure 10: AAM gender misclassified subjects who failed to rank 1 in 1-to-N matching ................................................................. 49
Figure 11: Performance Accuracy of Gender Misclassified Images from the African demographic group using COTS-A and open-source gender classifier ................................................................. 50
Figure 12: Performance Accuracy of Gender Misclassified Images from the Caucasian demographic group using COTS-A and open-source gender classifier ................................................................. 51
Figure 13. Performance Accuracy of Gender Misclassified Images of all demographic groups using COTS-A and open-source gender classifier. .......................................................... 52

Figure 14. Performance Accuracy of Gender Misclassified Images from the African demographic group using COTS-A and Microsoft gender classifier. ................................................................................................................ 53

Figure 15. Performance Accuracy of Gender Misclassified Images from the Caucasian demographic group using COTS-A and Microsoft gender classifier. ................................................................................................. 54

Figure 16. Performance Accuracy of Gender Misclassified Images of all demographic groups using commercial matcher and Microsoft gender classifier. ........................................................................................................... 55

Figure 17. Performance Accuracy of Gender Misclassified Images from the African demographic group using commercial matcher and Amazon gender classifier. ......................................................................................................... 56

Figure 18. Performance Accuracy of Gender Misclassified Images from the Caucasian demographic group using COTS-A and Amazon gender classifier. ........................................................................................................... 57

Figure 19. Performance Accuracy of Gender Misclassified Images of all demographic groups using commercial matcher and Amazon gender classifier. ........................................................................................................... 58
List of Tables

Table 1: Number of errors individually and in common across classifiers. 17
Table 2: Gender classification accuracy for MORPH3 dataset for all three classifiers................................................................. 18
Table 3: ArcFace - AAF average impostor scores, FMR and total number of comparisons across classifiers ................................................................. 26
Table 4: COTS-A - AAF average impostor scores, FMR and total number of comparisons across classifiers ................................................................. 26
Table 5: ArcFace -AAM average impostor scores, FMR and total number of comparisons ......................................................................................... 30
Table 6: COTS-A - AAM average impostor scores, FMR and total number of comparisons......................................................................................... 31
Table 7: ArcFace - CF average impostor scores, FMR and total number of comparisons ............................................................................................. 35
Table 8: COTS-A - CF average impostor scores, FMR and total number of comparisons ............................................................................................. 36
Table 9: COTS-A - CM average impostor scores, FMR and total number of comparisons ............................................................................................. 40
Table 10: COTS-A - CM average impostor scores, FMR and total number of comparisons ............................................................................................. 40
Table 11: Each demographic’s GMD value when using ArcFace and gender classifiers ......................................................................................... 41
Table 12: Lowest False Match Rate impact group by demographic and gender classifier when using ArcFace ...................................................... 43
Table 13: Each demographic’s GMD value when using COTS-A and all gender classifiers...................................................................................... 43
Table 14. FMR lowest value between true gender vs cross-gender group between gender misclassified images ....................................................... 45
Table 15. Number of subjects with at least one gender misclassified image and number of subjects with all sample images resulting in gender error by gender classifier ........................................................................................................ 59
Table 16. Mean and Median Age difference between sample images of subjects .......................................................................................................................... 59
Chapter 1
Introduction

In a study in 2018, Buolamwini et al. [1] brought widespread attention to the gap in gender classification error between demographic groups. Her investigation was incited after a gender classifier from her university failed to associate her with her biological gender [2]. Even though the study revealed results regarding gender classification error, claims targeting the accuracy of face recognition technologies followed immediately. In ensuing years, media outlets fueled public attention with provocative headlines like “Facial Recognition Is Accurate, if You’re a White Guy” [3] and “How is Face Recognition Surveillance Technology Racist?” [4]. Even though the study revealed results regarding gender classification error, claims targeting the accuracy of face recognition technologies followed immediately. Such claims ignited increasing concern amongst civil liberty advocates regarding the variability of face recognition performance across demographic groups. This concern has fueled political debate and led to the use of the technology being banned in several municipalities across the US. This is due partly to unrest caused by what has been labeled as police misconduct in communities of color. For example, San Francisco was the first major municipality to ban the use of face recognition technologies to avoid harmful consequences to communities of color [5]. Other major cities, including Boston, Massachusetts, and Portland, Oregon soon followed. Moreover, those voting not to ban the technology outright have openly discussed measures to limit law enforcement's use of the technology.

Public disclosure over the risks associated with the use of face recognition technologies in applications involving public safety also led research scholars to place a greater emphasis on understanding what performance differentials
may exist. Subsequent studies of face recognition systems have shown that
global. Subsequent studies of face recognition systems have shown that
algorithms perform worse, in terms of the false match rate, on biological
females than males [6][7]. However, there has been limited research to reveal
the true relationship between gender classification and face recognition
technologies.

The aim and motivation behind this study is to provide experiment-backed
clarity on the intersection of gender classification and face recognition. The
remaining of the document provides the Background and the Literature
Review identifying knowledge gaps and potential contributions from
previous works. We then detail the Experiment set-up and provide the
Experiment Results with corresponding analysis. Finally, the study ends with
a Conclusion.

Research Questions

The compelling results showed in gender shades called for immediate action,
shaping a continuous debate around equity in facial recognition. The study
revealed three commercial gender classifiers showing poorer performance on
darker-skinned subjects and higher accuracy on lighter-skinned subjects.

In this study, we explore the following questions that have yet to be answered:
(1) Are misclassified images of men more likely to falsely match with women
and vice versa?
(2) What is the 1-to-N match accuracy of subjects with gender misclassified
images?

The structure of this report is shaped by these two main questions. We extend
the work of previous studies and go a step beyond to measure the impact of
gender error on the False Match Rate for gender misclassified images above
the 1-in-10,000 FMR threshold of the MORPH3 dataset with cross-gender analysis. Furthermore, we measure the impact of gender error on facial identification or 1-to-N matching.
Background

This section aims to provide an overview of the concepts underlying the topic of this thesis. We explain how biometric technologies have become so pervasive in our daily lives. We begin by discussing the importance of gender error and then draw our attention to describe current biometric modalities.

Perspectives on Gender and Automated Classification

One subtle factor that may contribute to sustained gender-based inaccuracy is the fact that, increasingly, facial features are not necessarily easy to classify as exclusively male or female. That is, there is a growing “genderqueer” population: individuals who do not ascribe to traditional gender norms and often do not “present” as (i.e. attempt to be perceived as) stereotypically male or female.

For transgender individuals, for example, gender identity differs from sex assigned at birth [8]. In 2017, Meerwijk et al. [8] estimated the US population of transgender adults at about 1 in every 250 adults - nearly 1 million. Non-binary individuals, who identify as neither male nor female exclusively, comprise an even larger group: a 2019 survey conducted by the UCLA School of Law estimated that 1.2 million LGBTQ Americans identify as nonbinary - 11% of LGBTQ adults [9]. Without even considering the remaining spectrum of non-gender-abiding presentations and practices, these two groups alone represent millions of individuals whose presented (and often desired) facial features are not represented or accounted for in the typical creation or training of facial recognition algorithms.

Individuals who identify as transgender or nonbinary may choose to alter their appearance to minimize the presentation of their birth gender. One popular method of birth-gender obfuscation is hormone therapy, which
suppresses feminizing features in males and masculinizing features in females. Some individuals seek gender-affirming surgical procedures like facial feminization surgery, which focuses on sculpting and contouring the facial skeleton [10]. Makeup application is perhaps the most accessible and least invasive face-focused approach [11]. These methods alter facial appearance and can therefore impact facial recognition accuracy.

The problem of gender-binary in facial recognition is not new. A 2018 analysis [12] of automated gender recognition (AGR) research found that 94.8% of papers treated gender as binary (having only male and female gender classes) and 72.4% of papers treated gender as immutable (once a gender class is assigned it cannot change). These usages of gender do not necessarily represent reality for many individuals, whose needs are not being considered given these findings. Albiero [13] showed that only 60.3% of models surveyed were physiology-based, even though many researchers referenced “the physiological features they were trying to use, or explicitly [identified] sexual dimorphism as their intended mechanism of distinction.”

Researchers have long acknowledged that face structure is a key component in gender recognition, both automated and human perception-based [14], [15], [16]. Using certain aspects of the face, we can assign “both unfamiliar and familiar faces to general semantic categories such as gender” [14]. The geometry of facial features differs greatly in male and female biology: Ghojogh provide a list ranging from side outline of eyebrows and palate to vertical distance between eyes and nose-tip or height of face [17]. As a research community, we must proactively seek to include a broader range of facial structures and non-gender-based classifications in the constructions of our models to evolve with the changing cultural understanding of gender.

In addition, we understand and acknowledge the controversy of assessing gender classification algorithms and the discussion of its error patterns.
Following previous work related to this area, we assume that a face image can be analyzed and result in being correct or incorrect when looking at gender classification models that consider gender as “female” or “male”. We do not intend to disrespect individuals who may disagree with gender binaryism.

**Biometric technologies**

The increased risk posed to information systems by cybercriminals has led to rapid advances and implementation of biometric technologies. Biometrics can be defined as the measurable characteristics of individuals based on physiological features or behavioral traits which can be used to authenticate or identify their identity [18]. Most people associate the term “biometrics” with fingerprint or face recognition; however, biometrics come in many forms to authenticate or identify an individual. We provide a list of human traits that participate in biometric systems. The more noteworthy or well-known human traits that participate in biometric systems are enumerated below.

1. Hand Geometry, fingerprints, palm prints are one of the oldest biometrics features that are used for identification. The government uses them for establishing unique identity cards which can be used to trace back to the individual
2. Facial regions have gained more interest in the following year due to 3D technology. Due to the complexity of the facial structure, this seems to make a rather difficult problem in pattern recognition. In addition to the face, other attributes used for biometrics purposes are the nose, mouth structure, teeth, ear and eyes [19].
3. The ocular region is one of the regions that contains the most reliable and stable signatures which are the retina, iris, sclera, and vasculature
These are biometric signatures that are almost impossible to forge.

4. Behavioral systems put emphasis on how humans behave in certain activities such as walking, keystroke dynamics, and vocal features. However, even though research is being performed on speech and other signatures, these human behavior features have been linked to emotional and external factors which could be used to impersonate.

5. Soft Biometrics have been gaining increased interest in the following years due to errors in hard traits such as the face. Features such as ethnicity, gender, height, scars, marks or tattoos can be used as part of the soft attributes. These features lack uniqueness and permanence since they are common attributes among humans. In spite of that, studies have combined hard and soft biometrics attributes which have achieved noteworthy improvement.

Face Processing

Among the species, humans are adept at recognizing faces, and we are a social species able to classify, identify and memorize faces for extended periods of time. This phenomenon, observed since childhood, has raised studies on these types of capabilities. Researchers have traced back the ability of facial processing to the first days of birth. For example, newborns show preferences to human faces. A baby will gradually have a visual preference that matches the gender and ethnicity of his caregiver.

The brain region responsible for human facial processing has been researched to investigate questions regarding perception and behavior, nature versus nurture. If these are learned through experience, are humans innately willing to focus on and interpret facial cues? In addition to these questions,
can we diagnose and treat neural diseases that are related to recognition and gain knowledge on why it is difficult to recognize people from other races? This study carries over into neuroscience, human developmental science and psychology.

Two types of information are used for recognition by facial processing. Featural attributes isolate internal features such as eyes, nodes and mouth with external ones such as hairstyle and jawline. The external features refer to spatial relations between certain elements such as distance between the mouth and nose. The combination of two will result in the representation of the face [25]. Research has confirmed that while processing begins at birth, it becomes more efficient with age and experience which includes environment and exposure on the preference and non-familiar facial recognition.

Overall, facial processing has been agreed upon by studies that it is a quite complex and difficult task that requires processing of several levels, whether beginning at pregnancy or experience. However, the information collected from the way the brain represents and processes faces assisted in creating computers and algorithms that emulate the ability of the brain for facial recognition. In addition, it assisted in automating the recognition process with landmark features, representation of the face and neural algorithms.

**Face Verification**

In the field of Biometrics, face verification is the task of comparing two distinct facial images to determine if they correspond to the same subject, it is also commonly referred to as one-to-one matching (1:1). For instance, facial verification is used when a person unlocks their phone using their face or when verifying their identity when boarding a plane.
One indicator for a weaker verification system is a higher False Match Rate (FMR). Furthermore, a higher FMR means a higher number of impostor subjects score over the threshold, fooling the system. For this study, since we are studying the effect of gender on facial recognition, we will use the 1:10,000 threshold value of the impostor distribution of the entire MORPH3 dataset [26]. By doing this, we are targeting a gender-neutral threshold value and using it as the benchmark for each demographic group.

Face Identification
Face identification is the process of verifying that a human face image can be found at a particular database. Unlike facial verification where there are only two human face images being compared (1:1), face identification is sometimes referred to as 1:N matching because one image is being compared to N human face images stored in a database. The performance of facial recognition systems can be evaluated by looking at the identification performance of each demographic group. In other words, how accurate is the system to identify a particular subject in a gallery or database that contains one image of that subject and many other images of different subjects?
Chapter 2
Literature Review

We provide literature review on demographic effects in gender classification and face recognition.

Accuracy Disparities in Gender Classification

Carcagni et al. [27] studied optimal algorithmic configurations for estimation of soft biometrics. They demonstrated that choice of a specific configuration can impact the accuracy of face analytic systems.

Buolamwini and Gebru [1] evaluated three commercial gender classifiers (MSFT, Face++, and IBM) using a self-collected dataset of public images of African and European Parliament members. All three classifiers were shown to be more accurate for males than females. Additionally, after dividing images into skin-tone groups using manually-assigned Fitzpatrick 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 a whopping 34.7%, as compared to the maximum error rate for the lighter-skin-tone male demographic: a mere 0.8%. (Since the use of manual skin tone ratings introduces observer bias, we elect to use an automated rating scheme to provide consistent rating values based solely on image pixel content.) [28]

The demographic discrepancies noted by Buolamwini et al. [1] sparked heated dialogue in the public sphere and research community alike. Media outlets released incendiary headlines decrying “bias in facial recognition” without explaining a key nuance of the issue: gender classification and face matching algorithms operate independently and uniquely. One aim of this
paper is to provide experiment-backed clarity on the intersection of gender classification and face recognition.

Kush [29] performed a study on gender classification error reporting African American male subjects being the most accurate demographic group when using gender classification algorithms. The study also revealed that under-sampling of dark-skinned people in training datasets yields systematic disadvantages against them [30]. The results of this study counter-argue the initial claim that darker-skinned subjects are more error prone to gender misclassification due to an imbalanced training dataset of classifiers with more than 75 percent being male and more than 80 percent white. While the discussion regarding skin-tone is relevant, it is beyond the scope of this thesis.

Accuracy Disparities in Face Recognition

Cook [31] examined the performance of eleven commercial matchers on demographically-divided data. They demonstrated that accuracy and efficiency (as measured in transaction times) of matching are affected by a combination of often co-occurring biological and behavioral demographic factors, including gender, age, eye-wear, 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.

Krishnapriya et al. [32] analyzed False Match Rate (FMR) and False Non-Match Rate (FNMR) to evaluate face recognition accuracy by race and gender, using the annotated MORPH dataset [7]. They discovered that, generally, both the impostor and genuine 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 showed that, despite the “disadvantage” of the African American FMR value, the d-prime value given by some matchers showed that the ability to “cleanly” divide impostor and genuine scores is about equal across cohorts.

Albiero [13] also investigated gender-based disparities in face recognition accuracy, focusing on five speculated explanations: facial expression, head pose, forehead occlusion, facial makeup, and a balanced training dataset. However, even when controlling for these factors, accurate raters were still lower for females than males. The authors suggested the need to dig further into the impact of face morphology between men and women.

Classification and Recognition

There is little work related to the question of possible correlation in errors between gender classification and face recognition. Qiu et al. [33] were the first to investigate this question, finding that the relationship between gender classification and face recognition errors varied across the different demographic groups. They conducted their experiment with three gender classifiers and two face-matching algorithms, looking at the FMR and FMNR of image-pairs when either one or both images is gender-misclassified. Additionally, they recorded that (1) images with gender-classification errors were present in fewer false-match pairs than false non-match pairs and (2) gender-classification-error pairs represent an insignificant proportion of the impostor distribution (~2%). Gbekevi [34] expanded Qiu’s [33] work by providing evidence of the non-effect of gender classification on the face matching comparisons between misclassified images and False Match Rate pairs. It is also the first investigation that studies skin tone as a driving factor in face processing results using an automated skin tone rating algorithm.
This study echoes the questions of Qiu et al. and Gbekevi’s work, introducing a new dimension with the inclusion of cross-gender match distributions. Furthermore, this study shifts from 1:1 matching analysis to analyzing the impact on face identification on 1:N matching. To my knowledge, this is the first study to analyze this type of error in 1-to-N matching.
Chapter 3
Experiment Design

Our experiments aim to answer the ongoing question of whether gender error carries over to facial recognition error. Are misclassified images of men more likely to falsely match with women and vice versa? How does gender error affect the performance accuracy of a system? In other words, what is the performance accuracy of gender misclassified subjects compared to subjects with no gender misclassified images? To our knowledge, there are no previous studies that provide sufficient research to answer these questions. Previous work shows that images participating in a matching pair with at least one of the images resulting in gender error can sometimes produce a shift in the impostor distribution. Nevertheless, a shift in the impostor distribution does not negatively impact facial recognition unless analyzing the shift past the 1 in 10,000 False Match Rate. In this study, we further the investigation by analyzing the average impostor scores beyond the 1 in 10,000 False Match Rate (FMR) value between gender misclassified images between their true gender group and cross-gender group. The goal is to show whether misclassified images of a certain demographic group provide higher average impostor scores with their true gender or with their cross-gender group. If the latter holds true, further investigation should investigate similarities between the images. The 1 in 10,000 False Match Rate Threshold value is evaluated for the entire MORPH3 dataset and will be used for all analyses.

We used three gender classifiers, one open-source algorithm, and two commercial off-the-shelf APIs, Microsoft, and Amazon. The second part of the experiment will assess the performance accuracy of the gender misclassified images. For facial recognition, we utilized one open-source algorithm (ArcFace)[35] and one commercial off-the-shelf matcher.
(COTS-A). To our knowledge, all previous work related to gender error is focused on facial verification. The contribution of this study is to extend the investigation to facial identification. The MORPH3 dataset is suitable for identification investigation, however, there were a few subjects that only contained one sample image in the dataset. Five of these subjects fall under the Caucasian Male group, four on African American male, one Caucasian female subject, and four were part of the African American Female group. These subjects were not part of this analysis because they do not have an additional sample image to include in the gallery, hence not suitable for our identification analysis.

MORPH Dataset

The MORPH dataset is a collection of face-images segmented into four demographic cohorts by race and gender. Originally, the MORPH dataset was collected and employed in the investigation of adult age progression or facial aging. We know of several studies that utilized this dataset for analysis of gender error. Nowadays, MOPRH3 is notoriously known for its contribution to several research areas. Unlike other available datasets (i.e. PPB dataset), MORPH is a suitable dataset to perform facial verification (1:1 matching) and facial identification (1:n matching) analysis.

The dataset was collected in a controlled environment getting mugshot-style images that are normally front-pose, neutral expression, controlled lighting, and an 18% gray background. For this thesis, a curated version of MORPH was used to remove duplicate images, twins, and mislabeled images. This updated version, MORPH3, contains 127,319 images, where 35,276 images fell into the Caucasian male group, 10941 images were labeled as Caucasian
female, and 56,245 images as African American male the remaining 24,857 images were African American females.
Gender Classification Algorithms

We used three gender classifiers throughout our study, two commercial algorithms (Amazon and Microsoft) and one open-source algorithm. The open-source algorithm is a Convolutional Neural Network (CNN) composed of fifty layers implemented in PyTorch. It is trained using ArcFace modified with ResNet-50 and features parallel acceleration on both feature and centers, releasing the intra-class compactness constraint and weighted binary cross-entry loss function. Each gender is weighted in proportion to the number of samples provided. The following collection of datasets were used for training and validation: AFAF, AgeDB, CACD, IMBD-WIKI, IMFDB, MegaAgeAsian, and UTKFace.

We provide a table that summarized the number of individual gender errors and in common across gender classifiers. All three gender classifiers obtain errors on some of the same images, however, there is also a gap.

Table 1: Number of errors individually and in common across classifiers

<table>
<thead>
<tr>
<th></th>
<th>AAF</th>
<th>AAM</th>
<th>CF</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>926</td>
<td>405</td>
<td>111</td>
<td>41</td>
</tr>
<tr>
<td>Amazon</td>
<td>1759</td>
<td>1053</td>
<td>231</td>
<td>171</td>
</tr>
<tr>
<td>Open-Source</td>
<td>4181</td>
<td>1155</td>
<td>919</td>
<td>283</td>
</tr>
<tr>
<td>%Microsoft</td>
<td>58%</td>
<td>51%</td>
<td>58%</td>
<td>36%</td>
</tr>
<tr>
<td>%Amazon</td>
<td>30%</td>
<td>19%</td>
<td>28%</td>
<td>8%</td>
</tr>
<tr>
<td>%Open Source</td>
<td>12%</td>
<td>18%</td>
<td>7%</td>
<td>5%</td>
</tr>
</tbody>
</table>

The table below portrays the gender classification accuracy on the MORPH3 dataset by classifier. Each classifier has a substantial fraction of its errors that are not errors for the other classifiers. Looking at the numbers across the demographic cohorts, all three classifiers demonstrate a higher accuracy on males compared to females and higher for Caucasians than African
Americans. The group with the lowest accuracy rate is African American females.

**Table 2: Gender classification accuracy for MORPH3 dataset for all three classifiers**

<table>
<thead>
<tr>
<th></th>
<th>OSA</th>
<th>MS</th>
<th>AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAF</td>
<td>83.2</td>
<td>92.9</td>
<td>96.3</td>
</tr>
<tr>
<td>AAM</td>
<td>97.9</td>
<td>98.1</td>
<td>99.2</td>
</tr>
<tr>
<td>CF</td>
<td>91.6</td>
<td>97.9</td>
<td>99.0</td>
</tr>
<tr>
<td>CM</td>
<td><strong>99.2</strong></td>
<td><strong>99.5</strong></td>
<td><strong>99.8</strong></td>
</tr>
</tbody>
</table>

In each column, the highest value is in bold and the lowest value is underlined. As expected, we see the consistent trend of highest accuracy for Caucasian Males and lowest for African American Female.

**Face Recognition Matchers**

This study performs facial recognition using one commercial off-the-shelf algorithm (COTS-A) and one open-source matcher (ArcFace). ArcFace is a state-of-the-art deep CNN, which uses publicly available weights trained on a more refined, curated version of MS1M, named MS1MV2. Due to confidentiality reasons, we are not allowed to provide details about COTS-A. The impostor distribution for all demographic groups of the MORPH3 dataset are sampled at the 1-in-10,000 False Match Rate (FMR) threshold setting ("high-similarity tail") to analyze the impact of gender error on FMR between ground truth gender and projected gender. The 1-in-10,000 FMR threshold for the entire dataset is used as the baseline for all demographic cohorts.
Chapter 4
Results: Cross-Gender Verification Rate Error Analysis

The following chapter aims to report the experimental results of this study. First, we report results related to our first research question, do images of a gender group have higher probability of falsely matching with their true gender group or with the erroneously projected gender? Then, we proceed to document results that answer the second question of this study which assesses the impact of gender error on identification accuracy.

Face Verification

The first part of this study attempts to answer the question whether gender misclassified images have on average higher impostor scores with their ground-truth gender or with their erroneously projected gender. Additionally, we measure the impact on the False Match Rate (FMR) on both true and cross-gender groups. A higher FMR translates to a higher degree of impact of gender misclassified images on the FMR of that group. The FMR was calculated using the 1-in-10,000 FMR threshold value of the impostor distribution of the entire dataset. For the open-source algorithm, ArcFace, the threshold value was 37.44, whereas for the commercial algorithm, COTS-A, the threshold value was 76.67.

The analysis is broken down in accordance with the four demographic groups identified in the Morph dataset: African-American Female (AAF), African American Male (AAM), Caucasian Female (CF), and Caucasian Male (CM). For each demographic group, face recognition matcher, and gender classifier, we provide results in three forms:
1. Impostor score distribution of gender misclassified images of a demographic group with their true gender group and cross-gender group. For instance, the impostor distribution of gender misclassified images of African American Female group with the African American Female group (ground truth gender) and the African American Male group (projected gender). For every graph, the green dotted line represents the impostor scores of Gender error (GE) images with the ground truth gender and blue represents impostor scores of GE images with cross-gender group. The differences in shape and height of the two distributions reflect the imbalanced nature in the number of images of the MORPH3 dataset.

2. Boxplot representing the same set of data of 1.

3. Table illustrating values of average impostor scores and FMRs. The False Match Rate is the rate at which two biometric signals from different people are identified as coming from the same person. The FMR was determined by taking the number of scores above the 1-in 10,000 threshold value and dividing it by the total number of impostor comparisons to then multiply it by 100. For the first column in the table, green values indicate higher average impostor score of misclassified images with their true gender vs cross-gender group and for the second column, green values highlight demonstrates a higher FMR score of misclassified images with their true gender group.
Analysis of African American Female Cohort

We begin our analysis with the AAF demographic group. The AAF group consists of 24,857 images of 5,929 subjects. As noted in Table 2, the gender classification error rate is 16.8% using the Open-source algorithm, 3.7% for Amazon, and 7.1% for Microsoft. This translates into 4181 images of 1799 subjects of the AAF group being misclassified as male by the open-source algorithm, 1759 images of 824 subjects by Amazon, and 926 images of 364 subjects by Microsoft. We should note that most (90%) of the subjects with an image incorrectly classified as male, had at least one image correctly classified as female. Given the difference in error rates for the gender classifiers, we should note that images that have been incorrectly classified as male by one algorithm may be correctly classified as female by another.

Next, we examine whether an AAF image misclassified as male will produce higher similarity scores when compared to other images/subjects of its ground-truth labeled group or with its counterpart. In Figure 1, the score distribution represents impostor scores of genders misclassified African American female images with ground truth gender (African American Female) and erroneously projected gender (African American Male). The displayed results utilize ArcFace as the face recognition matcher, and the Open-source, Amazon, and Microsoft gender classifier algorithms respectively. Along the horizontal axis, the match score values are presented.
Along the vertical axis, the probability density of the scores is presented. The impostor distribution with ground truth gender is represented with green color, whereas the impostor distribution with the cross-gender group is shown in blue. The red dashed line represents the FMR value evaluated at the 1-in-10,000 threshold for the entire MORPH3 dataset. The differences in shape and height of the distributions reflect the imbalanced nature of the number of images in the MORPH3 dataset. This is detailed in Table 3 and Table 4, under total number of impostor comparisons. The boxplots shown at the right of every score distribution provide another representation of the impostor scores between gender misclassified images with true and cross-gender group. The orange line inside the box represents the median of the scores, while the red dashed line, just like the score distributions, mark the 1-in-10,000 FMR threshold value. Ultimately, we are interested on studying the degree of impact of scores that lie above the red dashed line on the boxplot and to the right of the score distribution. The degree of impact is determined by the impact on the False Match Rate with ground truth gender and projected gender.

Similarly, the results for a commercial face recognition algorithm (COTS-A) are displayed in Figure 2.
Figure 1: Score distribution and box plot of results for ArcFace Matcher with Open Source (a,b), Amazon (c,d) and Microsoft (e,f)
Figure 2: Score distribution and box plot of results for COTS-A Matcher with Open Source (a1,b1), Amazon (c1,d1) and Microsoft (e1,f1)
As mentioned earlier, we are interested in the analysis of scores that cross the 1-in 10,000 FMR threshold value set by the MORPH3 dataset. Table 3 and Table 4 summarize the analysis including average impostor scores above the 1-in-10,000 FMR threshold, the False Match Rate and the total number of comparisons between gender misclassified and their true and cross-gender groups. Results that do not imply a negative impact on facial recognition error are highlighted in green, whereas scores that contribute to facial recognition error are highlighted in red. That is, higher average impostor scores of misclassified images with their labeled gender are seen as green. Similarly, higher False Match Rate values with the true gender group in contrast to the cross-gender group are also green. Looking at Table 3 and Table 4, the False Match Rate values for the African American Female misclassified images are consistently higher with the African American Female group than with the African American Male for all gender classifiers. African American Female subjects that display more male features are more likely to participate in a false match with African American Female subjects than with African American Male subjects. Gender error is not echoing facial recognition error since the false match rate for misclassified images is still higher with the true gender for both matchers and gender classifiers.
Table 3: ArcFace - AAF average impostor scores, FMR and total number of comparisons across classifiers

<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>AAF (GE) vs AAF</td>
<td>40.451</td>
<td>0.13143</td>
<td>103883506</td>
</tr>
<tr>
<td></td>
<td>AAF (GE) vs AAM</td>
<td>39.750</td>
<td>0.0062</td>
<td>235164526</td>
</tr>
<tr>
<td>Amazon</td>
<td>AAF (GE) vs AAF</td>
<td>40.504</td>
<td>0.13226</td>
<td>43703323</td>
</tr>
<tr>
<td></td>
<td>AAF (GE) VS AAM</td>
<td>39.775</td>
<td>0.00899</td>
<td>98936714</td>
</tr>
<tr>
<td>Microsoft</td>
<td>AAF (GE) vs AAF</td>
<td>40.336</td>
<td>0.07732</td>
<td>23004848</td>
</tr>
<tr>
<td></td>
<td>AAF (GE) vs AAM</td>
<td>39.735</td>
<td>0.010594</td>
<td>52083796</td>
</tr>
</tbody>
</table>

Table 4: COTS-A - AAF average impostor scores, FMR and total number of comparisons across classifiers

<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>AAF (GE) vs AAF</td>
<td>78.829</td>
<td>0.169862</td>
<td>103883506</td>
</tr>
<tr>
<td></td>
<td>AAF (GE) vs AAM</td>
<td>78.522</td>
<td>0.017658</td>
<td>235164526</td>
</tr>
<tr>
<td>Amazon</td>
<td>AAF (GE) vs AAF</td>
<td>78.9121</td>
<td>0.193055</td>
<td>43703323</td>
</tr>
<tr>
<td></td>
<td>AAF (GE) VS AAM</td>
<td>78.5858</td>
<td>0.025887</td>
<td>98936714</td>
</tr>
<tr>
<td>Microsoft</td>
<td>AAF (GE) vs AAF</td>
<td>78.768</td>
<td>0.11594</td>
<td>23004848</td>
</tr>
</tbody>
</table>
Analysis of African American Male Cohort

We continue our analysis with the AAM demographic group. The AAM group consists of 56,245 images of 8,839 subjects. As noted in Table 2, the gender classification error rate is 2.1% using the Open source algorithm, 0.8% for Amazon, and 1.9% for Microsoft. This translates into 1155 images of 632 of the AAM group being misclassified as female by the open-source algorithm, 1053 images of 470 subjects by Amazon, and 405 images of 170 subjects by Microsoft. We should note, that most (92%) of the subjects with an image incorrectly classified as female, had at least one image correctly classified as male. Given the difference in error rates for the gender classifiers, we should note that images that have been incorrectly classified as male by one algorithm may be correctly classified as female by another.

The analysis conducted for the African American Male group follows the same procedure as details presented in the previous demographic group. Figure 3 and Figure 4 show the score distributions and box plot results for the open-source face algorithm (ArcFace) and commercial face recognition algorithm (COTS-A), respectively.
Figure 3: Score distribution and box plot of results ArcFace Matcher with Open Source (a2,b2), Amazon (c2,d2) and Microsoft (e2,f2)
Figure 4: Score distribution and box plot of results COTS-A Matcher with Open Source (a3,b3), Amazon (c3,d3) and Microsoft (e3,f3)
As mentioned earlier, we are interested in the analysis of scores that cross the 1-in 10,000 FMR threshold value set by the MORPH3 dataset. Table 5 and Table 6 summarize analysis including average impostor scores above the 1-in-10,000 FMR threshold, the False Match Rate and the total number of comparisons between gender misclassified and their true and cross-gender for ArcFace and COTS-A. The False Match Rate values for the African American Male misclassified images are consistently higher with the African American Female group than with the African American Male for all gender classifiers. African American Male subjects that display more female features are more likely to participate in a false match with African American Female subjects than with African American Male subjects. Gender error is echoing facial recognition error as the false match rate for misclassified images is higher with projected gender for both matchers and gender classifiers.

Table 5: ArcFace -AAM average impostor scores, FMR and total number of comparisons

<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open Source</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAM (GE) vs AAM</td>
<td>40.217</td>
<td>0.02565</td>
<td></td>
<td>64952743</td>
</tr>
<tr>
<td>AAM (GE) vs AAF</td>
<td>40.181</td>
<td>0.02762</td>
<td></td>
<td>28709835</td>
</tr>
<tr>
<td><strong>Amazon</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAM (GE) vs AAM</td>
<td>40.334</td>
<td>0.02688</td>
<td></td>
<td>59216394</td>
</tr>
<tr>
<td>AAM (GE) vs AAF</td>
<td>40.159</td>
<td>0.034812</td>
<td></td>
<td>26174421</td>
</tr>
<tr>
<td><strong>Microsoft</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAM (GE) vs AAM</td>
<td>40.455</td>
<td>0.02435</td>
<td></td>
<td>22775077</td>
</tr>
<tr>
<td>AAM (GE) vs AAF</td>
<td>40.290</td>
<td>0.06582</td>
<td></td>
<td>10067085</td>
</tr>
</tbody>
</table>
Table 6: COTS-A - AAM average impostor scores, FMR and total number of comparisons

<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>AAM (GE) vs AAM</td>
<td>78.736</td>
<td>0.043084</td>
<td>64952743</td>
</tr>
<tr>
<td></td>
<td>AAM (GE) vs AAF</td>
<td>78.806</td>
<td>0.06981</td>
<td>28709835</td>
</tr>
<tr>
<td>Amazon</td>
<td>AAM (GE) vs AAM</td>
<td>78.743</td>
<td>0.04057</td>
<td>59216394</td>
</tr>
<tr>
<td></td>
<td>AAM (GE) vs AAF</td>
<td>78.779</td>
<td>0.08148</td>
<td>26174421</td>
</tr>
<tr>
<td>Microsoft</td>
<td>AAM (GE) vs AAM</td>
<td>78.844</td>
<td>0.045145</td>
<td>22775077</td>
</tr>
<tr>
<td></td>
<td>AAM (GE) vs AAF</td>
<td>78.942</td>
<td>0.14086</td>
<td>10067085</td>
</tr>
</tbody>
</table>
Analysis of Caucasian Female Cohort

We further our analysis with the CF demographic group. The CF group consists of 10,941 images of 2,798 subjects. As noted in Table 2, the gender classification error rate is 8.4% using the Open source algorithm, 1% for Amazon, and 2.1% for Microsoft. This translates into 919 images of 466 subjects of the CF group being misclassified as male by the open-source algorithm, 231 images of 128 subjects by Amazon, and 111 images of 70 subjects by Microsoft. We should note, that most (95%) of the subjects with an image incorrectly classified as male, had at least one image correctly classified as female. Given the difference in error rates for the gender classifiers, we should note that images that have been incorrectly classified as male by one algorithm may be correctly classified as female by another.
Figure 5: Score distribution and box plot of results ArcFace Matcher with Open Source (a4,b4), Amazon (c4,d4) and Microsoft (e4,f4)
Figure 6: Score distribution and box plot of results COTS-A Matcher with Open Source (a5,b5), Amazon (c5,d5) and Microsoft (e5,f5)
We are interested in the analysis of scores that cross the 1-in 10,000 FMR threshold value set by the MORPH3 dataset. Table 7 and Table 8 summarizes analysis including average impostor scores above the 1-in-10,000 FMR threshold, the False Match Rate and the total number of comparisons between gender misclassified and their true and cross-gender. The False Match Rate values for the Caucasian Female misclassified images are consistently higher with the Caucasian Female group than with the Caucasian Male for almost all gender classifiers. Caucasian Female subjects that display more masculine features are more likely to participate in a false match with Caucasian Female subjects than with Caucasian Male subjects for all classifiers when using ArcFace. This observation does not hold true for all classifiers when using the COTS-A matcher. With the subset of Caucasian female misclassified subjects, the FMR is higher with the erroneously projected gender over the ground truth gender with the Microsoft classifiers. Images of Caucasian Female that resulted in gender error are more likely to falsely match with the cross-gender. Gender error is not echoing facial recognition error because the false match rate for misclassified images is still higher with the true gender for both matchers and most gender classifiers.

Table 7: ArcFace - CF average impostor scores, FMR and total number of comparisons

<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>CF (GE) vs CF</td>
<td>40.241</td>
<td>0.01964</td>
<td>10046225</td>
</tr>
<tr>
<td></td>
<td>CF (GE) vs CM</td>
<td>39.415</td>
<td>0.000391</td>
<td>32418644</td>
</tr>
<tr>
<td>Amazon</td>
<td>CF (GE) vs CF</td>
<td>40.062</td>
<td>0.01441</td>
<td>2525081</td>
</tr>
</tbody>
</table>
Analysis of Caucasian Male Cohort

We end our analysis on face verification with the CM demographic group. The CM group consists of 35,276 images of 8,835 subjects. As noted in Table 2, the gender classification error rate is 0.8% using the Open source algorithm, 0.2% for Amazon, and 0.5% for Microsoft. This translates into 283 images of 217 subjects of the CM group being misclassified as female by the open-source algorithm, 171 images of 137 subjects by Amazon, and 41 images of 27 subjects by Microsoft. We should note, that most (94%) of the subjects had an image incorrectly classified as female, had at least one image correctly classified as male. Given the difference in error rates for the gender classifiers, we should note that images that have been incorrectly classified as male by one algorithm may be correctly classified as female by another.
Figure 7: Score distribution and box plot of results ArcFace Matcher with Open Source (a6,b6), Amazon (c6,d6) and Microsoft (e6,f6)
Figure 8: Score distribution and box plot of results COTS-A Matcher with Open Source (a7,b7), Amazon (c7,d7) and Microsoft (e7,f7)
We are interested in the analysis of scores that cross the 1-in 10,000 FMR threshold value set by the MORPH3 dataset. Table 9 and Table 10 summarizes analysis including average impostor scores above the 1-in-10,000 FMR threshold, the False Match Rate and the total number of comparisons between gender misclassified and their true and cross-gender. The False Match Rate values for the Caucasian Male misclassified images are consistently higher with the Caucasian Female group than with the Caucasian Male for almost all gender classifiers. Caucasian Female subjects that display more masculine features are more likely to participate in a false match with Caucasian Female subjects than with Caucasian Male subjects for open source and amazon classifiers when using ArcFace. This observation does not hold true for all classifiers when using the COTS-A matcher. With the subset of Caucasian female misclassified subjects, the FMR is higher with the erroneously projected gender over the ground truth gender with the Microsoft classifiers. Images of Caucasian Female that resulted in gender error are more likely to falsely match with the cross-gender. Gender error is not echoing facial recognition error as the false match rate for misclassified images is still higher with the true gender for both matchers and most gender classifiers.
<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>CM (GE) vs CM</td>
<td>40.377</td>
<td>0.00537</td>
<td>9980155</td>
</tr>
<tr>
<td></td>
<td>CM (GE) vs CF</td>
<td>39.398</td>
<td>0.00132</td>
<td>3096303</td>
</tr>
<tr>
<td>Amazon</td>
<td>CM (GE) vs CM</td>
<td>40.131</td>
<td>0.004808</td>
<td>6031014</td>
</tr>
<tr>
<td></td>
<td>CM (GE) vs CF</td>
<td>39.491</td>
<td>0.002779</td>
<td>1870911</td>
</tr>
<tr>
<td>Microsoft</td>
<td>CM (GE) vs CM</td>
<td>39.624</td>
<td>0.003042</td>
<td>1445965</td>
</tr>
<tr>
<td></td>
<td>CM (GE) vs CF</td>
<td>39.244</td>
<td>0.00557</td>
<td>448581</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender classifier</th>
<th>Number of Gender misclassified Images = 4</th>
<th>Average Impostor Scores above 1 in 10,000 Threshold</th>
<th>False Match Rate (FMR)</th>
<th>Total number of comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>CM (GE) vs CM</td>
<td>78.038</td>
<td>0.007535</td>
<td>9980155</td>
</tr>
<tr>
<td></td>
<td>CM (GE) vs CF</td>
<td>78.309</td>
<td>0.005523</td>
<td>3096303</td>
</tr>
<tr>
<td>Amazon</td>
<td>CM (GE) vs CM</td>
<td>78.558</td>
<td>0.027645</td>
<td>6031014</td>
</tr>
<tr>
<td></td>
<td>CM (GE) vs CF</td>
<td>78.284</td>
<td>0.003743</td>
<td>1870911</td>
</tr>
<tr>
<td>Microsoft</td>
<td>CM (GE) vs CM</td>
<td>79.833</td>
<td>0.002766</td>
<td>1445965</td>
</tr>
<tr>
<td></td>
<td>CM (GE) vs CF</td>
<td>78.213</td>
<td>0.01092</td>
<td>448581</td>
</tr>
</tbody>
</table>
Discussion of Cross-Gender Verification Rate Error Analysis

We proceed to capitulate the average impostor scores in the form of a table. We then introduce the term of Gender Mean Difference (GMD), which is defined as the difference between the mean match scores of true-versus false gender comparisons above the 1 in 10,000 False Match Rate value. A lower GMD-value indicates a higher degree of similarity between the true gender and the false gender group.

**Table 11. Each demographic’s GMD value when using ArcFace and gender classifiers**

<table>
<thead>
<tr>
<th></th>
<th>AAF Value</th>
<th>AAM Value</th>
<th>CF Value</th>
<th>CM Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Microsoft</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAF Mean</td>
<td>40.336</td>
<td>40.455</td>
<td>40.174</td>
<td>39.624</td>
</tr>
<tr>
<td>AAM Mean</td>
<td>39.735</td>
<td>40.290</td>
<td>39.776</td>
<td>39.244</td>
</tr>
<tr>
<td>GMD</td>
<td>0.600</td>
<td>0.1653</td>
<td>0.3987</td>
<td>0.3797</td>
</tr>
<tr>
<td><strong>Amazon</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAF Mean</td>
<td>40.504</td>
<td>40.334</td>
<td>40.062</td>
<td>40.131</td>
</tr>
<tr>
<td>AAM Mean</td>
<td>39.775</td>
<td>40.159</td>
<td>39.613</td>
<td>39.491</td>
</tr>
<tr>
<td>GMD</td>
<td>0.7291</td>
<td>0.1753</td>
<td>0.4490</td>
<td>0.639</td>
</tr>
<tr>
<td><strong>OSA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAF Mean</td>
<td>40.451</td>
<td>40.217</td>
<td>40.241</td>
<td>40.377</td>
</tr>
<tr>
<td>AAM Mean</td>
<td>39.750</td>
<td>40.181</td>
<td>39.415</td>
<td>39.398</td>
</tr>
<tr>
<td>GMD</td>
<td>0.700</td>
<td>0.0361</td>
<td>0.8259</td>
<td>0.9785</td>
</tr>
</tbody>
</table>

Table 11 summarizes the average impostor scores of gender misclassified images when compared with their true and cross-gender, as well as the computation of the gender mean difference between the two average scores recorded. At first sight, we can observe that the Gender Mean Difference is positive for all demographic groups and for all gender classifiers.
Furthermore, the GMD value is positive when the average impostor score of the gender misclassified images with their true gender is higher than the average impostor score of the cohort. It is interesting to analyze the variations of the GMD values across demographic groups and across gender classifiers. Starting at the top of the table and with Microsoft classifier, the African American Female group has the highest GMD value, followed very closely by the Caucasian cohort groups. The African American Male group lies at the bottom of the ranking with the lowest GMD value. African American Male misclassified produce similar impostor scores when compared to their true and cross-gender. Moving on to the second classifier, Amazon’s GMDs value follow the same pattern as the previous classifier. The African American Female misclassified images have the highest GMD value, followed by the Caucasian Female Group, Caucasian Male having a slightly lower GMD value, and African American Male having the lowest GMD value. It is interesting to note that both the highest and lowest GMD values are contained by the cohorts of a demographic group. Finally, at the very bottom of the table we analyze the GMD values of the gender misclassified images produced by the open-source algorithm within demographic group. In this case, the top-ranked GMD values are seen in the Caucasian group with the Caucasian Male misclassified images laying at the top, followed very closely by Caucasian Female images. Again, we see a similar trend to the commercial off-the-shelf gender classifiers, where the African American Male group not only has the lowest value within demographic groups when using the open source algorithm but also across all gender classifiers.
Table 12: Lowest False Match Rate impact group by demographic and gender classifier when using ArcFace

<table>
<thead>
<tr>
<th>Gender Misclassified Images</th>
<th>OSA</th>
<th>Amazon</th>
<th>Microsoft</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAF</td>
<td>PG</td>
<td>PG</td>
<td>PG</td>
</tr>
<tr>
<td>AAM</td>
<td>GT</td>
<td>GT</td>
<td>GT</td>
</tr>
<tr>
<td>CF</td>
<td>PG</td>
<td>PG</td>
<td>PG</td>
</tr>
<tr>
<td>CM</td>
<td>PG</td>
<td>PG</td>
<td>GT</td>
</tr>
</tbody>
</table>

Table 12 reports the group that shows the lowest FMR value when compared with gender misclassified images. PG, stands for Projected Gender whereas GT stands for Ground Truth. The African American Male group show consistency across all classifiers with a lower FMR with the Ground Truth group. We see another instance with the Caucasian Male where the previous statement is also true, in other words, the Caucasian Male misclassified images have a lower impact on the FMR of the ground truth group than with the erroneously projected group. For all remaining combinations of demographic and classifiers, misclassified images have a lower degree of impact on the erroneously projected group.

Table 13. Each demographic’s GMD value when using COTS-A and all gender classifiers

<table>
<thead>
<tr>
<th></th>
<th>AAF Value</th>
<th>AAM Value</th>
<th>CF Value</th>
<th>CM Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAF Mean</td>
<td>78.76</td>
<td>78.84</td>
<td>78.51</td>
<td>79.8333275</td>
</tr>
<tr>
<td>AAM Mean</td>
<td>78.60</td>
<td>78.94</td>
<td>78.38</td>
<td>78.21034285</td>
</tr>
<tr>
<td>GM D</td>
<td>0.1669</td>
<td>-0.097</td>
<td>0.1325</td>
<td>1.622</td>
</tr>
<tr>
<td>Amazon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAF Mean</td>
<td>78.91</td>
<td>78.74</td>
<td>78.55</td>
<td>78.94533</td>
</tr>
</tbody>
</table>

43
We conducted the same procedure to analyze the results obtained when using the commercial off-the-shelf facial matcher, COTS-A. Table 14 represents the average impostor scores above the 1 in 10,000 False Match Rate between the gender misclassified images of a demographic group against their true and cross-gender group. We documented our results in the form of a table where the gender classifiers are listed on the left. The Gender Mean Difference (GMD) was lower than the values observed when using ArcFace for every demographic group and classifier. Lower GMD values indicate a higher similarity between the true and false gender group. The red values highlight that the average impostor score above the 1 in 10,000 False Match Rate Threshold between gender misclassified images of a demographic group and their cross-gender group are higher than when compared to their true gender. In other words, among the demographics, African-American males who are gender misclassified have the most in common, as measured by match score, with their false-gender counterpart (AAF). This statement holds true for all gender classifiers when looking at African American male misclassified images. It is interesting to highlight that for both facial matcher, ArcFace and COTS-A, the African-American Male group displayed the lowest GMD values. A negative GMD also implies a shift in the impostor distribution and in consequence a higher False Match Rate.

<table>
<thead>
<tr>
<th>OSA</th>
<th>AA M Mean</th>
<th>78.58</th>
<th>AAF Mean</th>
<th>78.77</th>
<th>CM Mean</th>
<th>78.28</th>
<th>CF Mean</th>
<th>78.4864</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM D</td>
<td>0.326</td>
<td>GMD</td>
<td>-0.036</td>
<td>GM D</td>
<td>0.274</td>
<td>GM D</td>
<td>0.4583</td>
<td></td>
</tr>
<tr>
<td>AAF Mean</td>
<td>78.82</td>
<td>9</td>
<td>AAM Mean</td>
<td>78.73</td>
<td>CF Mean</td>
<td>78.62</td>
<td>CM Mean</td>
<td>79.037</td>
</tr>
<tr>
<td>AA M Mean</td>
<td>78.52</td>
<td>2</td>
<td>AAF Mean</td>
<td>78.80</td>
<td>CM Mean</td>
<td>78.22</td>
<td>CF Mean</td>
<td>78.308</td>
</tr>
<tr>
<td>GM D</td>
<td>0.306</td>
<td>8</td>
<td>GMD</td>
<td>-0.069</td>
<td>GM D</td>
<td>0.399</td>
<td>GM D</td>
<td>0.7292</td>
</tr>
</tbody>
</table>
Table 14 reports the group that shows the lowest FMR value when compared with gender misclassified images. PG, stands for Projected Gender whereas GT stands for Ground Truth, which. The African American Male group show consistency across all classifiers with a lower FMR with the Ground Truth group. We see another instance with the Caucasian Male for Amazon and Microsoft classifiers, where the previous statement is also true, in other words, the Caucasian Male misclassified images have a lower impact on the FMR of the ground truth group than with the erroneously projected group. For all remaining combinations of demographic and classifiers, misclassified images have a lower degree of impact on the erroneously projected group. Future work should focus on analyzing those subjects that challenge the high FMR value with cross-gender group.
Chapter 5
Results: 1-to-N Identification Error Rate Analysis

We turn our focus to analyzing the impact of gender error on the performance accuracy of ArcFace and COTS-A algorithms. We compare the identification performance by matcher of subjects with no gender misclassified images to the subjects with at least one misclassified image for all demographic groups across the three gender classifiers. During facial identification, a probe image is compared against a gallery of images for which the image with the highest similarity score is associated with the probe image.

For every subject, the probe image corresponded to the most recent image present in the dataset for that subject. In the case of subjects containing at least one misclassified image, the probe image was the most recent misclassified image. The gallery consisted of all remaining images of the dataset, however neglecting the sample images of the probe subject except for the least recent image of that subject. For instance, for a particular subject with 10 sample images, the probe image corresponds to the latest picture taken of that person (image number 10) and the only mated image present in the gallery is the first-ever picture present of that subject in the database. In real-life identification systems, it is uncommon to encounter more than one image of the same subject in the database. Hence, we try to attain a realistic system by only considering one image in the database and we chose the most recent image of each subject as the probe. For those subjects whose least recent image was their only gender misclassified image, the gender misclassified image was used as the probe and the most recent image of the subject in the gallery. Few subjects were not considered for the identification evaluation as there was only one image present of these subjects in the
dataset. None of these subjects presented gender error across any of the gender classifiers.

We proceed to illustrate the performance of the open-source algorithm as well as the commercial off-the-shelf algorithm. ArcFace displayed high genuine scores for all subjects. In fact, this facial recognition system reached nearly absolute precision for subjects with at least one gender misclassified image in the MORPH3 dataset, with an accuracy higher than 99.5 percent for gender misclassified subjects. In addition, only 6 subjects out of the 3114 subjects (when using the open-source classifier) did not obtain a genuine score as their highest score when comparing it to images in the gallery. For the African American Female group, out of a total of 5,929 subjects, 1799 subjects had at least one gender misclassified image, which only 2 subjects failed to rank 1 in the 1-to-N matching. Moreover, for the African American Male demographic group, out of a total of 8,839 subjects, 632 subjects had at least one gender misclassified image, which only 4 subjects failed to obtain rank 1 in the 1-to-N matching. In the Caucasian demographic group, for both the female and male demographic group, subjects with at least one gender misclassified image were successfully identified with rank 1 accuracy.

We now take a closer look at the probe and gallery images of the subjects that failed to rank 1 in the 1-to-N matching analysis. Figure 9 shows the African American Female subjects, where subject a) obtained rank 2 and subject b) rank 4. At first sight, there is a variation in the pose and facial expression between probe and gallery image. For instance, when looking at Subject a) forehead wrinkles are visible in the probe image, whereas in the gallery image they are covered by the change in hairstyle. Subject b’)s probe image appears to be smiling, while the gallery image holds a neutral expression. Such variations may play a role in the deficit accuracy of 1-to-N matching.
Moving on to the African American Male group, Figure 10 shows gender misclassified subjects c), d), e) and f) who obtained rank 7, 4, 4 and 6, respectively. It is interesting to note the evident difference in appearance of the subjects between the probe and gallery images. Visually, probe images show more male characteristic features, whereas the images of the subjects used in the gallery show more female characteristics, including longer hair and make-up. The remarkable changes between probe and gallery images may explain one more time the deficit in accuracy of the subjects in question. Even with this, the 1-to-N matching accuracy of gender misclassified subjects when using ArcFace was better than rank 10.
Figure 10: AAM gender misclassified subjects who failed to rank 1 in 1-to-N matching

Moving onto the analysis of commercial off-the-shelf algorithms. The Open-Source algorithm (OSA) was the gender classifier that produced the highest gender inaccuracy on the demographic groups. We first analyze each demographic group independently plotting the performance of subjects containing at least one sample image resulting in gender error vs subject with no gender misclassified images of the same demographic group.
Figure 11: Performance Accuracy of Gender Misclassified Images from the African demographic group using COTS-A and open-source gender classifier.

The graph on the left of Figure 11 portrays the performance accuracy of African American Female subjects containing at least one gender misclassified image (Gender Error >=1) and subjects with no gender misclassified images (Gender Error = 0). Similarly, the graph on the right conveys results for the African American Male group. The identification performance for both African American demographic groups result higher for subjects with no gender misclassified images. The number of subjects in the African American Female and Male group that contained at least one gender misclassified image was 1799 and 632, whereas subjects with no gender misclassified images were 4127 and 8205, respectively.
Figure 12: Performance Accuracy of Gender Misclassified Images from the Caucasian demographic group using COTS-A and open-source gender classifier.

The graph on the left of Figure 12 portrays the performance accuracy of Caucasian Male subjects containing at least one gender misclassified image (Gender Error >=1) and subjects with no gender misclassified images (Gender Error = 0). Similarly, the graph on the right conveys results for the Caucasian Female group. The identification performance for both Caucasian demographic groups resulted in higher for subjects with no gender misclassified images. The number of subjects in the Caucasian Male and Female group that contained at least one gender misclassified image was 217 and 466, whereas subjects with no gender misclassified images were 8613 and 2330, respectively.
Figure 13. Performance Accuracy of Gender Misclassified Images of all demographic groups using COTS-A and open-source gender classifier.

Figure 13 summarizes the performance accuracy of gender misclassified subjects of all four demographic groups using ArcFace as the facial matcher and the open-source algorithm as the gender classifier. At the bottom of the performance accuracy ranking, we found the African American Female group with just over 90 percent performance accuracy. At the top of the ranking, we found the Caucasian Female group with the best performance with over 96 percent of accuracy.
Figure 14. Performance Accuracy of Gender Misclassified Images from the African demographic group using COTS-A and Microsoft gender classifier.

The graph on the left of Figure 14 portrays the performance accuracy of African American Female subjects containing at least one gender misclassified image (Gender Error $\geq 1$) and subjects with no gender misclassified images (Gender Error = 0). Similarly, the graph on the right conveys results for the African American Male group. For the first time in our results, we see a shift in identification accuracy in the African American Female group where the subset of subjects with at least one gender misclassified image performs better than subjects with no gender classified image. Moreover, in the African American Male group we see again the repeated pattern of subjects with at least one misclassified image performing worse than subjects with no gender misclassified images. The number of subjects in the African American Female and Male group that contained at least one gender misclassified image was 364 and 170, whereas subjects with no gender misclassified images were 5562 and 8667, respectively.
Figure 15. Performance Accuracy of Gender Misclassified Images from the Caucasian demographic group using COTS-A and Microsoft gender classifier.

The graph on the left of Figure 15 portrays the performance accuracy of Caucasian Male subjects containing at least one gender misclassified image (Gender Error \( \geq 1 \)) and subjects with no gender misclassified images (Gender Error = 0). Similarly, the graph on the right conveys results for the Caucasian Female group. The identification performance for both Caucasian demographic groups result higher for subjects with no gender misclassified images. The number of subjects in the Caucasian Male and Female group that contained at least one gender misclassified image was 27 and 70, whereas subjects with no gender misclassified images were 8803 and 2726, respectively.
Figure 16. Performance Accuracy of Gender Misclassified Images of all demographic groups using commercial matcher and Microsoft gender classifier.

Figure 16 summarizes the performance accuracy of gender misclassified subjects of all four demographic groups using ArcFace as the facial matcher and Microsoft as the gender classifier. At the bottom of the performance accuracy ranking, we found the African American Male group with a performance accuracy lower than 90 percent. The top performer when looking at Rank One identification is the Caucasian Female group, however followed very closely by the Caucasian Male group.
Figure 17. Performance Accuracy of Gender Misclassified Images from the African demographic group using commercial matcher and Amazon gender classifier.

In Figure 17, the graph on the left portrays the performance accuracy of African American Female subjects containing at least one gender misclassified image (Gender Error $\geq 1$) and subjects with no gender misclassified images (Gender Error $= 0$). Similarly, the graph on the right conveys results for the African American Male group on the same regard. Looking at the African American Female group it is impossible to discern if subjects with or without gender misclassified images have a better performance. On the other hand, by looking at the graph of African American Male demographic group, it is evident that subjects with no gender misclassified images perform better than subjects with at least one misclassified image. The number of subjects in the African American Female and Male group that contained at least one gender misclassified image were 824 and 470, whereas subjects with no gender misclassified images were 5102 and 8367, respectively.

56
Figure 18. Performance Accuracy of Gender Misclassified Images from the Caucasian demographic group using COTS -A and Amazon gender classifier.

The graph on the left of Figure 18 portrays the performance accuracy of Caucasian Male subjects containing at least one gender misclassified image (Gender Error $\geq 1$) and subjects with no gender misclassified images (Gender Error = 0). Similarly, the graph on the right conveys results for the Caucasian Female group. The identification performance for both Caucasian demographic groups resulted in higher for subjects with no gender misclassified images. The number of subjects in the Caucasian Male and Female group that contained at least one gender misclassified image was 137 and 128, whereas subjects with no gender misclassified images were 8693 and 2668, respectively.
Figure 19. Performance Accuracy of Gender Misclassified Images of all demographic groups using commercial matcher and Amazon gender classifier.

Figure 19 summarizes the performance of gender misclassified subjects of all four demographic groups by the Amazon gender classifier. At first sight, it is unclear to discern between the rank performance of each demographic group. It may be interesting to note that past rank 10, African American Female misclassified subjects tend to increase at a slightly slower rate than the other demographic groups.

Table 15 summarizes the number of subjects with at least one gender misclassified image and the number of subjects with all sample images resulting in gender error.
Table 15. Number of subjects with at least one gender misclassified image and number of subjects with all sample images resulting in gender error by gender classifier.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Total Number of Subjects</th>
<th>Number of subjects with Gender Error &gt;=1</th>
<th>Number of subjects with all sample images with Gender Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OSA</td>
<td>MS</td>
</tr>
<tr>
<td>AAM</td>
<td>8,839</td>
<td>632</td>
<td>170</td>
</tr>
<tr>
<td>AAF</td>
<td>5,929</td>
<td>1799</td>
<td>364</td>
</tr>
<tr>
<td>CM</td>
<td>8,835</td>
<td>217</td>
<td>27</td>
</tr>
<tr>
<td>CF</td>
<td>2,798</td>
<td>466</td>
<td>70</td>
</tr>
</tbody>
</table>

We investigated potential factors that may explain the disparities in the performance accuracy of gender misclassified subjects within demographic groups. A factor that may explain the performance variation between misclassified subjects of demographic groups is the age difference between the sample images of the subjects. In our study, the probe image is the most recent image of the subject or the most recent misclassified image of the subject. MORPH3 contains subjects with a variety of sample images with some subjects having more than 90 sample images. Table 16 summarizes the mean and median differences between sample images of subjects within demographic groups. The African American Male group has the biggest age difference with an average mean of 3.6 or 3 as the median. The number in parenthesis defines the average age of the demographic group.

Table 16. Mean and Median Age difference between sample images of subjects

<table>
<thead>
<tr>
<th>MORPH3</th>
<th>AAF(30)</th>
<th>AAM(32)</th>
<th>CF(32)</th>
<th>CM(32)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean</td>
<td>2.36</td>
<td>3.6</td>
<td>1.86</td>
<td>1.99</td>
</tr>
</tbody>
</table>
Challenges and Limitations

The experiment procedure for this study was overall smooth and successful. As with any procedure, we faced several challenges and limitations that affected our timeline, and which are interesting to highlight. To start with, due to the nature of the length of the MORPH3 dataset, the impostor distribution consisted of over 5 billion comparison scores.

There was a shortlist of subjects with only one image present in the dataset. While these subjects were suitable to be used in the verification analysis of the study, they were not eligible for use when assessing the 1-to-N accuracy of the system, or in other words, during the identification analysis. During facial identification, the image of the individual in question should appear one or more times in the known gallery.

The MORPH3 dataset is an imbalanced dataset with an unequal number of images per demographic group, in other words, unequal distribution of classes. The data set is biased towards the African American Male group, with more than 56 thousand images (8,839 subjects), accounting for almost half of the size of the entire dataset (44%). On the other end lies the Caucasian Female group with barely 10 thousand images (2,798 subjects), which makes it only 8.5% of the total number of images. The imbalanced nature of the MORPH3 dataset can question the identification discrepancies analysis produced in this study. Although all probe subjects are compared to an equal gallery size, probe subjects from a demographic with higher order of images have, on average, higher impostor scores present in their gallery. This can lead to question the results of our analysis. In addition, future work should be focused on performing the analysis on a balanced version of the MORPH3 dataset with an equal number of subjects and number of images. We can compare how much a balanced dataset improved the performance of our
initial experimental set up. This degree of accuracy is only possible in ideal
conditions where there is consistency in lighting and positioning, and where
the facial features of the subject are clear and unobscured.
Chapter 6
Conclusion

The goal of this study is twofold: (1) determine whether gender misclassified images of men are more likely to falsely match with women and vice versa; and (2) contribute to past studies on the impact of gender error on facial recognition by extending the research to 1-to-N matching for gender misclassified images.

We used a curated version of MORPH dataset (MORPH3 dataset) that contains a collection of 127,319 images including four demographic groups: African American Female, African American Male, Caucasian Female and Caucasian Male.

In our study, we used one commercial facial matcher, COTS-A, and one open-source algorithm, ArcFace. The FMR values for both matchers showed that misclassified images of men have lower FMR values with their ground truth gender than with their projected gender, meaning, they have a higher likelihood of falsely matching with women. On the contrary, for female demographic groups, the FMR values of misclassified images with their projected gender are lower than with their ground truth, meaning there is a lower probability that women falsely match with men. For the second part of the experiment, we concluded that when looking at ArcFace as a facial matcher, only 6 images from subjects containing at least one misclassified image ranked less than one. To our surprise, those subjects portrayed a big difference in appearance between the probe image and the gallery image of that subject. Therefore, we conclude that ArcFace neglects gender error. When using COTS-A as the facial matcher we see that for the most part, gender misclassified subjects perform slightly worse than subjects with no gender misclassified images.
Finally, with this study and others that follow, we hope to call awareness to a long-standing “given” in the field of biometrics that is due for re-evaluation: the treatment of gender as binary. The binary classification of “male” and “female” fail to capture our world’s changing perspectives on attitudes toward gender norms and identities. Questions raised by our study and similar research are imperative in ensuring that biometric technology contributes to, rather than detracts from, race and gender equity. It is important that we continue to critically examine the ways our research and its outputs impact the world at large.
References


