NISTIR 8052

Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms

Mei Ngan
Patrick Grother

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8052

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8052

Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms

Mei Ngan
Patrick Grother
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8052

April 2015

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director

Face Recognition Vendor Test (FRVT)

Performance of Automated Gender Classification Algorithms

NIST Interagency Report 8052

Mei Ngan and Patrick Grother

Information Access Division
National Institute of Standards and Technology

Executive Summary

Introduction

Facial gender classification is an area studied in the Face Recognition Vendor Test (FRVT) Still Facial Images Track. While peripheral to automated face recognition, it has become a growing area of research, with potential use in various applications. The motivation for gender classification systems has grown in recent years, with rise of the digital age and the increase in human-computer interaction. Gender-based indexing of face images, gender-targeted surveillance (e.g., monitoring gender-restricted areas), gender-adaptive targeted marketing (e.g., displaying gender-specific advertisements from digital signage), and passive gender demographic data collection are potential applications of automated gender classification.

NIST performed a large scale empirical evaluation of facial gender classification algorithms, with participation from five commercial providers and one university, using large operational datasets comprised of facial images from visas and law enforcement mugshots, leveraging a combined corpus of close to 1 million images. NIST employed a lights-out, black-box testing methodology designed to model operational reality where software is shipped and used "as-is" without subsequent algorithmic training. Core gender classification accuracy was baselined over a large dataset composed of images collected under well-controlled pose, illumination, and facial expression conditions, then assessed demographically by gender, age group, and ethnicity. Analysis on commonly benchmarked "in the wild" (i.e., unconstrained) datasets was conducted and compared with those from the constrained dataset. The impact of number of image samples per subject was captured and assessments of classification performance on sketches and gender verification accuracy were documented.

Key Results

Core Accuracy and Speed: Gender classification accuracy depends strongly on the provider of the core technology. Broadly, there is a threefold difference between the most accurate and the least accurate algorithm in terms of gender classification error, which is the percentage of images classified incorrectly. The most accurate algorithm (E32D from NEC) can correctly classify the gender of a person over a constrained database of approximately 1 million images 96.5% of the time. All algorithms can perform gender classification on a single image in less than 0.25 seconds with one server-class processor. The most accurate algorithm, on average, performs classification in 0.04 seconds. *Section 3.1*.

Impact of Demographic Data on Accuracy: For a dataset of 240 thousand visa images, it is empirically observed that, overall, gender classification is more accurate in males than females. All of the algorithms show a significant decrease in gender classification accuracy as age increases for adult females, with an empirically decreasing trend in accuracy seen in females past age 50. Gender classification is more accurate in adult males (ages 21-60) than young boys (ages 0-10) for all of the algorithms. For females, gender classification is the most accurate in young adults (ages 21-30). A majority of the algorithms demonstrate the lowest gender classification accuracy on subjects from Taiwan and Japan, with males being more often misclassified than females. *Section 3.2 and 3.3*.

These results state empirical observations for the particular dataset, but they do not determine cause. The impact of factors potentially driving the observed results between age and ethnicity, such as facial features innate to certain ethnic groups, facial changes with age, and hairstyles are not studied in this report. Further research would be required to objectively verify these conjectures.

Gender Verification Accuracy: In addition to a binary male or female decision, algorithms provide a real-valued score of maleness-femaleness. This score can be used to tradeoff the type 1 error versus type 2 error. For a system with an objective to verify that a person is female (e.g., in gender-restricted scenarios), to ensure that 99% of the time, the subject is indeed a female would result in 10% of the true female population being falsely classified as males, or inconveniently rejected as being female. The same Detection Error Tradeoff (DET) analysis can be done by gender and for different verification thresholds to support specific applications. *Section 3.7.1*.

Impact of Number of Image Samples on Accuracy: The FRVT Application Programming Interface (API) [12] supports multiple still image input to the algorithm software for gender classification, which enables the analysis of gender classification performance versus the number of image samples of the same person. For contemporaneous mugshot images of the

same subject collected within a one year period, the results show gender classification accuracy monotonically increasing as the number of image samples provided increased for most of the algorithms. Misclassification rates drop by as much as 50% between one and four input images, depending on the algorithm. Gender classification times increase linearly with respect to the number of image samples, which is the expected behavior. *Section 3.4*.

Comparison against Academic Methods: A performance evaluation was done with two commonly benchmarked in the wild datasets composed of face images that show a large range of variation in factors such as pose, illumination, facial expression, and background. The study attempts to compare FRVT gender classification participants with published methods from the academic literature. The top performing FRVT participants achieved comparable or higher overall gender classification accuracy when compared to the published methods that tested their algorithms with the same number of images for both datasets.

Straight comparison of performance results between the FRVT participants and published methods remains to be a challenge given some methods are tested on only a subset of the images and others use different testing protocols, which in all cases, allowed the implementation to train on a set of images through every iteration of testing. For the FRVT participant results documented, all images from the datasets were used for testing, and NIST employed a lights-out, black-box testing methodology that did not involve any type of algorithm training during evaluation. This is designed to model operational reality where software is shipped and used as-is without algorithmic training. *Section 3.5*.

Constrained versus In the Wild: Comparison of algorithmic performance on constrained versus in the wild images indicate that the qualities of in the wild images (e.g., variations in pose, illumination, facial expression, and background) does have a negative impact on algorithm certainty of gender classification. Greater average gender classification certainty differences are observed in males between constrained and in the wild images. *Section 3.6*.

Sketches: By running the algorithms through a set of 856 non-operational sketch images that were derived from a set of photographs, the most accurate algorithm (B31D from Cognitec) achieves an overall classification accuracy of 93.8%. An important caveat is that gender classification on sketches was never declared to be part of this FRVT study and better algorithms may be available from the providers. That said, automated face recognition algorithms are being used to recognize sketches operationally, and gender classification on sketches may help reduce the search space for face recognition in those scenarios. *Section 3.7.2*.

Caveats

Nature of the data: The main dataset used for overall accuracy assessment is comprised of close to 1 million images collected under well-controlled pose, illumination, and facial expression conditions. Although image collection was subject to the guidelines published by the Department of State (DoS) [3] and the Federal Bureau of Investigation (FBI), the images are compressed JPEG files which exhibit artifacts of JPEG compression causing reduction in image detail. With more detail available in less compressed images, gender classification accuracy may improve, but errors will still likely exist due to variation driven by intrinsic and extrinsic factors.

Failure to compute: The results presented in this document are for cases where gender classification computation *did not* fail. That is, metrics did not include a penalty for cases where algorithms failed to generate a gender classification decision and score. As such, the algorithms were evaluated based on the goodness of the results that they were able to generate, which might make an algorithm with a high failure-to-compute or no-attempt-to-compute rate seem to be more effective than a robust algorithm that has a zero failure-to-compute rate.

Acknowledgements

The authors would like to thank the sponsors of this activity. These are the Criminal Justice Information Systems (CJIS) division and the Biometric Center of Excellence (BCOE) of the Federal Bureau of Investigation (FBI) and the Science and Technology (S&T) Directorate in the Department of Homeland Security (DHS).

Disclaimer

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Release Notes

- **Appendices**: This report is accompanied by a number of appendices which present exhaustive results on a per-algorithm basis. These are machine-generated and are included because the authors believe that visualization of such data is broadly informative and vital to understanding the context of the report.
- **Typesetting**: Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable LaTeX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.
- **Graphics**: Many of the figures in this report were produced using Hadley Wickham's ggplot2 [27] package running under  , the capabilities of which extend beyond those evident in this document.
- **Contact**: Correspondence regarding this report should be directed to FRVT2012 at NIST dot GOV.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose

Gender classification of a face in one or more images is an area investigated in the Face Recognition Vendor Test (FRVT) with Still Images Track. Similar to age, gender is viewed as a soft biometric trait [17], and its automated characterization has become a growing area of study that has applications in surveillance, human-computer interaction, and image retrieval systems. Some of the earliest attempts to perform gender classification using computer vision techniques occurred over two decades ago, and a number of methods have been published in the literature over the years addressing the problem of gender classification using facial images. In recent years, gender classification on "in the wild" (i.e., unconstrained) images has been investigated given an increased attention to face recognition in the wild and commercial applications of gender classification.

The main goals of this evaluation are to:

- Provide an objective, independent, open, and free assessment of current automated gender classification technology.
- Leverage massive operational corpora. The availability of images from large populations (around one million) supports statistical significance and repeatability of the studies. The use of operational images brings greater operational relevance to the test results.
- Investigate gender classification accuracy across various factors, including age, ethnicity, and constrained versus in the wild facial images.

1.2 Application Scenarios

The motivation for gender classification has grown in the last few decades given the rise of the digital age and the increase in human-computer interaction. The process of gender determination has potential application in at least the areas described below:

Gender as criterion for indexing into large-scale biometric databases for faster retrieval has been discussed [26] and can also apply to automatic sorting and image retrieval from digital photo albums and the internet. For example, gender can be used for one-to-many search partitioning where given thresholds for male and female scores that represent a certain level of gender classification certainty, a schema can be employed where if a score falls within the certainty threshold, it performs a search only on the partition specific to that gender, thereby reducing the search space. For gender scores that don't fall within a particular gender certainty threshold (representing less certainty in gender classification), the entire database can be searched.

Gender-targeted surveillance can assist with monitoring gender-restricted areas and elevated threat levels that might be associated with a specific gender. [18]

Gender-adaptive human-computer interaction is on the rise given the popularity of digital signage and the opportunity for targeted digital marketing. Targeted advertisements can be displayed based on the gender of the audience walking past a digital sign.

Passive gender demographic data collection [16] can be achieved and used to support decisions relating to, for example, the ratio of female versus male product offerings in a store.

2 Methodology

2.1 Test Environment

The evaluation was conducted offline at a NIST facility. Offline evaluations are attractive because they allow uniform, fair, repeatable, and large-scale statistically robust testing. However, they do not capture all aspects of an operational

system. While this evaluation is designed to mimic operational reality as much as possible, it does not include a live image acquisition component or any interaction with real users. Testing was performed on high-end server-class blades running the Linux operating system. Most of the blades were 12-core machines with dual processors running at 3.47 GHz with 192 GB of main memory. The test harness used concurrent processing to distribute workload across dozens of blades.

2.2 Algorithms

The FRVT program was open to participation worldwide. The participation window opened on July 25, 2012, and submission to the final phase for gender classification algorithms closed on October 4, 2013. There was no charge to participate. The process and format of algorithm submissions to NIST was described in the FRVT Still Face Image and Video Concept, Evaluation Plan and Application Programming Interface (API) document [12]. Participants provided their submissions in the form of libraries compiled on a specified Linux kernel, which were linked against NIST's test harness to produce executables. NIST provided a validation package to participants to ensure that NIST's execution of submitted libraries produced the expected output on NIST's test machines.

FRVT had three submission phases where participants could submit algorithms to NIST. This report documents the results of all algorithms submitted in the final phase or the most recent submission for participants who only submitted in prior phases.

Table 1 lists the FRVT participants who submitted algorithms for gender classification, and the alphanumeric code associated with each of their submissions. For each participant, the algorithms are labeled numerically by chronological order of submission. The letter codes assigned to the participants are also located at the bottom of each page for reference.

Participant	Letter Code	Submissions		
		Aug. 2012	Mar. 2013	Oct. 2013
Cognitec	B	B10D	B20D	B30D,B31D
Neurotechnology	C			C30D
NEC	E	E10D		E30D,E31D,E32D
Tsinghua University	F	F10D		F30D
MITRE	K	K10D		
Zhuhai-Yisheng	P			P30D

Table 1: FRVT Gender Classification Participants

2.3 Image Dataset

This report documents the use of the following datasets[1]:

- Mugshot images: This dataset consists of facial images collected by various law enforcement (LEO) agencies and transmitted to the FBI as part of various criminal record checks.
- Visa images: This dataset consists of facial images for visa applicants.
- LFW: This is a public dataset composed of unconstrained facial images collected from the web [15].
- GROUPS: This is a public dataset, which is a collection of unconstrained images of groups of people from Flickr [10].
- FERET Sketch images: The FERET [20] database was collected in the 1990s and has been very widely studied. The City University of Hong Kong employed an artist to produce, for each person, "a face photo with lighting variation

[1]Operational datasets used in this study were shared with NIST only for use in biometric technology evaluations under agreements in which biometric samples were anonymously coded by the provider; code translations were never shared with NIST; and no personally identifiable information (PII) beyond the biometric sample was shared with NIST.

and a sketch with shape exaggeration drawn by an artist when viewing this photo", published as the CUHK Face Sketch FERET Database (CUFSF) [25].

The dataset properties are summarized in Table 2.

Property	MUGSHOT	VISA	LFW	GROUPS	CUFSF
Collection Environment	Law enforcement booking	Visa application process	Internet images	Internet images	Sketches
Collection Era	~1960s-2008	~1996-2010	Unknown	Unknown	
Digital, Paper Scan	Digital, few paper	Mostly digital	Unknown	Unknown	Paper Scan
Documentation	See NIST Special Database 32 [2]	[3]	See [15]	See [10]	See [25]
Image size	Various, 480x600, 768x960	Most 252x300	250x250		
Compression	JPEG ~20:1	JPEG, mean size: 16.2kB	JPEG	JPEG	JPEG
Eye to eye distance	Mean = 108 pixels, SD = 40 pixels	Median = 71 pixels		Median = 18.5 pixels	
Frontal pose	Moderate control. Known profile images excluded.	Well controlled	Uncontrolled	Uncontrolled	
Full frontal geometry	Mostly not. Varying amounts of the torso are visible.	Yes, in most cases. Faces are more cropped (i.e., smaller background) than ISO[2] Full Frontal requires.			
Source	Operational data	Operational data	Public dataset	Public dataset	Public dataset

Table 2: Image dataset descriptions.

The operational datasets are characterized by population sizes well in excess of all published gender classification tests. The number of images are given in Table 3.

Quantity	MUGSHOT	VISA	LFW	GROUPS	CUFSF
Number of gender labeled face images	587 680	364 086	13 233	28 231	856
Gender	Male: 293 840 Female: 293 840	Male: 185 164 Female: 178 922	Male: 10 256 Female: 2 977	Male: 13 672 Female: 14 559	Male: 494 Female: 363

Table 3: Image dataset sizes.

[2]The International Organization of Standardization, (ISO), is an international standard-setting body composed of representatives from various national standards organizations.

2.4 Performance Metrics

The following performance measures will be reported in the assessment of gender classification:

2.4.1 Classification Accuracy

Male Accuracy is defined as the number of correctly classified male images, TM, divided by the total number of male images, M. i.e.,

$$\text{Male accuracy} = \frac{TM}{M} \tag{1}$$

Female Accuracy is defined as the number of correctly classified female images, TF, divided by the total number of female images, F. i.e.,

$$\text{Female accuracy} = \frac{TF}{F} \tag{2}$$

Overall accuracy is defined as the sum of correctly classified male and female images divided by the total number of images. i.e.,

$$\text{Overall accuracy} = \frac{TM + TF}{M + F} \tag{3}$$

2.4.2 Gender Verification Error

Per the FRVT API [12], a real-valued measure of maleness-femaleness on [0,1] is provided by the gender classification algorithms, with a value of 0 indicating certainty that the subject is a male and 1 indicating certainty that the subject is a female. The maleness-femaleness values can be plotted as a Detection Error Tradeoff (DET) characteristic, where, for example, in a gender-based access control scenario that only allows female entry, the rate of males being granted access (i.e., the false female rate) is traded off against the rate of females being denied access (i.e., the false male rate).

For gender verification, the fundamental error rates are defined as:

$$\text{False male rate}(T) = \frac{\text{Number of females with a maleness-femaleness value } < \text{ threshold, } T}{\text{Number of females}} \tag{4}$$

$$\text{False female rate}(T) = \frac{\text{Number of males with a maleness-femaleness value } \geq \text{ threshold, } T}{\text{Number of males}} \tag{5}$$

3 Results

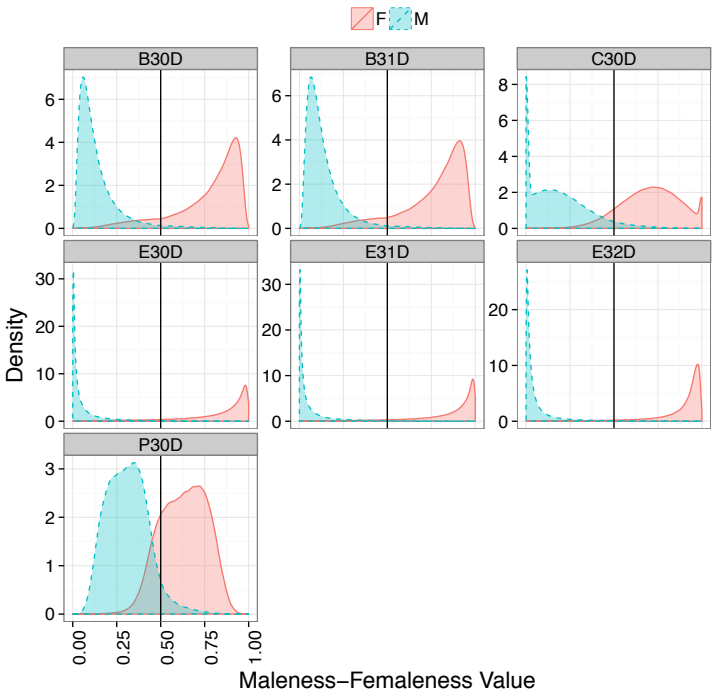
3.1 Gender Classification on Large Constrained Dataset

3.1.1 Accuracy

Gender classification accuracy was baselined against a large dataset composed of visa and mugshot images collected under well-controlled pose, illumination, and facial expression conditions. The dataset consisted of 951 766 images (472 762 females and 479 004 males) with a relatively balanced number of females and males. Gender classification performance for each algorithm is presented in Table 1a. Per the FRVT API [12], in addition to providing a binary male or female classification, the gender classification algorithms were asked to provide a real-valued measure of maleness-femaleness on [0,1], where a value of 0 indicates certainty that the subject is a male and 1 indicates certainty that the subject is a female. Figure 1b plots the distribution of maleness-femaleness values for the algorithms that provided continuous maleness-femaleness values.

	Female Accuracy (%)	Male Accuracy (%)	Overall Accuracy (%)
# Images	472762	479004	951766
B30D	88.8	97.6	93.2
B31D	88.7	97.9	93.3
C30D	88.7	95.0	91.9
E30D	91.0	97.3	94.2
E31D	91.9	97.2	94.5
E32D	95.6	97.5	96.5
F30D	87.7	89.5	88.6
K10D	88.6	93.0	90.8
P30D	81.9	94.4	88.1

(a) Classification accuracy



(b) Distribution of maleness-femaleness value

Figure 1: Table summarizing gender classification accuracy and density plots showing distribution of maleness-femaleness value. Table and plots were generated with 951 766 images.

Results and notable observations:

- The algorithm with the highest overall accuracy (E32D) can correctly classify gender 96.5% of the time over a dataset of 951 766 images.
- For males, classification performance is closer between the algorithms, with participants B and E being the top performers, achieving accuracies of 97.9% (B31D) and 97.5% (E32D) respectively. For females, the leading participant (E32D) achieves accuracy of 95.6%, with the next most accurate participant (B30D) being 6.8% lower in accuracy for the same gender class.
- Gender classification accuracy is empirically lower in females than males for all of the algorithms. Accuracy in females is 1.8% to 12.5% lower than males, depending on the algorithm. This could be a result of higher female misclassification rates observed in certain age groups, which is discussed in Section 3.2.
- The distribution of the maleness-femaleness value for most of the algorithms shows clear separation of male and female classification. The higher misclassification rate seen in females is also supported in the density plots where overlapping past the implied male threshold value of 0.5 or less is visible. Algorithms F30D and K10D did not provide continuous values for maleness-femaleness.

3.1.2 Speed

Speed could be an important performance factor in some gender classification applications where there exists a limited window of time for a decision based on the outcome, such as when a person walks past a digital sign. The use of gender as criterion for indexing into large-scale biometric databases would levy rapid speed requirements on gender classification algorithms to make it operationally viable.

Figure 2 presents the distribution of gender classification times for each algorithm. Gender classification time is the amount of time elapsed computing the gender from pixel data of a face image. It does not include any pre-processing steps

performed by the test software such as loading the image from disk or extracting image data from a compressed JPEG file. The timing machine was a server-class blade with a CPU running at 3.47 GHz. For more details on the testing environment, see Section 2.1.

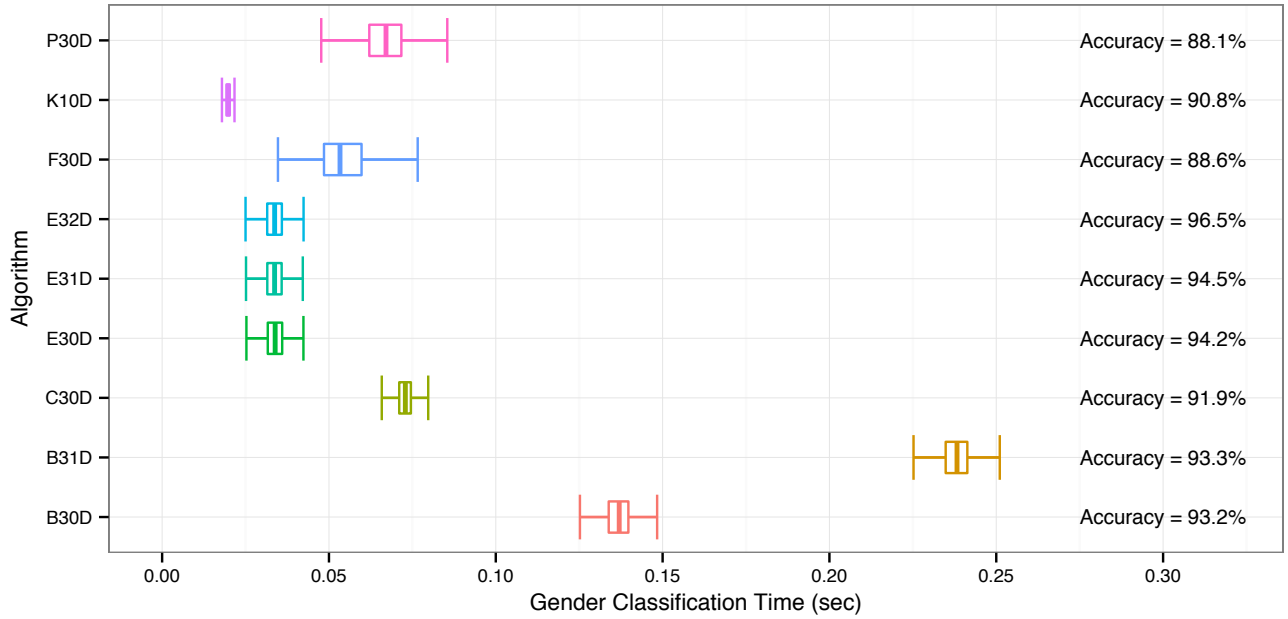


Figure 2: Boxplots of the distribution of gender classification times. Plots were generated over 5 000 gender estimates. For reference, overall classification accuracy against a population of 951 766 is reported on the right.

Results and notable observations:

- Gender classification time varies considerably from one participant to another. K10D can perform classification in less than 0.025 seconds while B31D takes about nine times longer than that, on average.
- No clear speed-accuracy tradeoff exists between the gender classification algorithms. While K10D has the fastest classification speeds, it is not the lowest in accuracy, and while participant B's algorithms are the slowest, they do not achieve the highest accuracy. Partipant E's algorithms are among the lowest in classification times, and they also attain the highest overall classification accuracy.
- Participant E appears to have fixed gender classification times, while exhibiting evident variance in accuracy between its algorithms. Participant B shows observable differences in classification times between its two algorithms with nominal changes in accuracy.

3.1.3 Failure to Compute Rate

The accuracy results presented above were computed for cases where gender classification did not fail. The error metrics do not include a penalty for cases where an algorithm failed to generate a gender estimate. Per the FRVT API [12], a failure to compute occurs when an algorithm's code returns a non-zero return value from a call to its gender classification function, and hence fails to generate an gender estimate. This can be a result of software issues (e.g., memory corruption), algorithmic limitations (e.g., failure to find eyes in small images), elective refusal to process the input (e.g., image is assessed to have insufficient quality), or specific vendor-defined failures. Table 4 presents the fraction of images for which algorithms failed to generate a gender estimate over 951 766 images.

Num Images	B30D	B31D	E30D	E31D	E32D	K10D	P30D
951766	0.0011	0.0011	0.0063	0.0063	0.0063	0.1554	0.0031

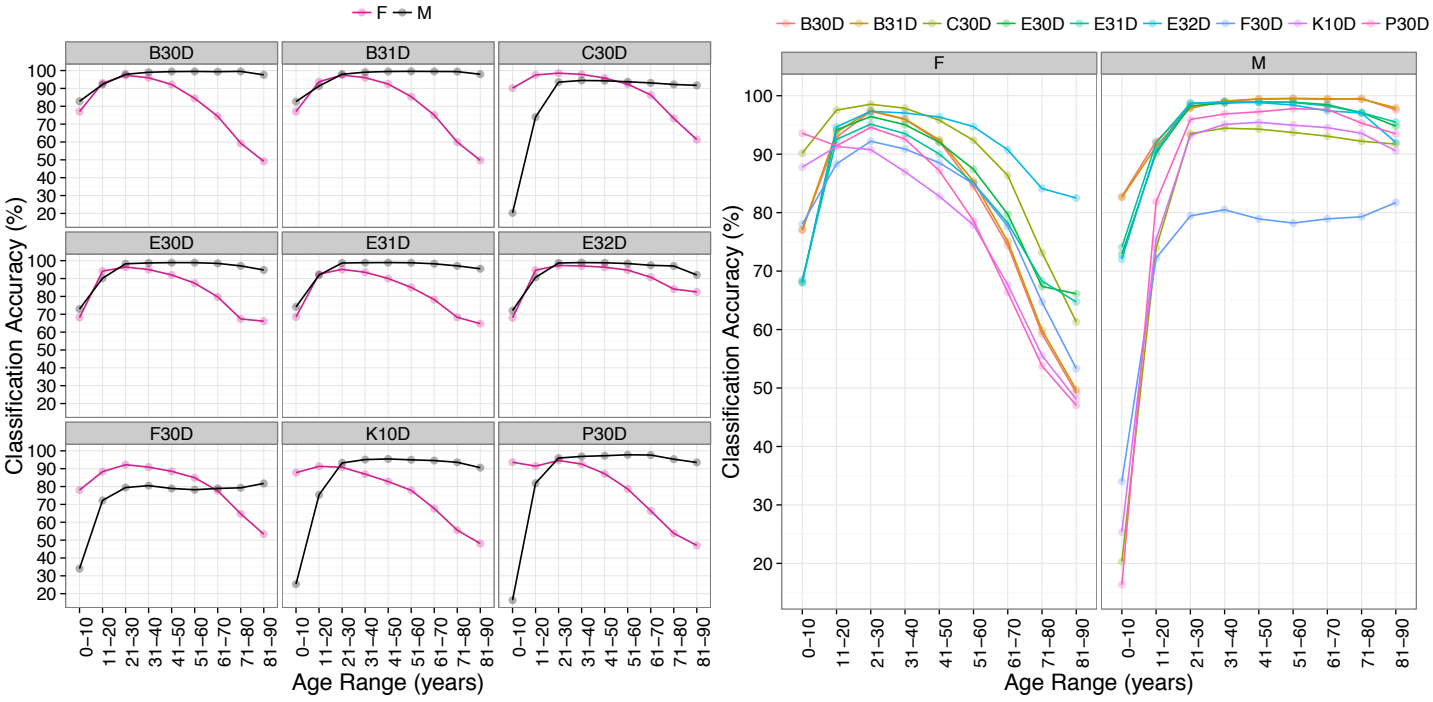
Table 4: Table summarizing failure to compute ratio over 951 766 images.

Results and notable observations:

- Algorithms C30D and F30D did not produce any failures to generate a gender estimate on the dataset of 951 766 images.
- Eight out of nine of the algorithms have insignificant failure to compute ratios over the massive number of images processed. While the dataset used is comprised of visa and mugshot images collected under published collection guidelines, the existence of a small number of bad images is inevitable given the operational nature of the data. Issues with images included occlusion, closed eyes, and pathological quality.
- Participant K fails on approximately 15% of the images, with the reason being "involuntary failure to extract features from the image" (as indicated in the FRVT API).

3.2 Age

Age is a demographic trait that can impact gender perception. Studies from the anthropological literature have indicated that there are a number of skeletal structure differences between the faces of male and female adults. In contrast, claims have been made that the faces of boys and girls are very similar in skeletal facial structure, posing a challenge to gender classification in children [5,8]. Studies have also examined the impact of a person's age on gender classification accuracy on limited amounts of data [13]. The dataset used in this experiment contains a large number of gender-labeled visa images collected under well-controlled pose, illumination, and facial expression conditions. The dataset contains a balanced number of males and females over each of the age ranges. Gender classification accuracy between males and females across age ranges is assessed and summarized in Figure 3 and Table 5.



(a) By algorithm *(b) By gender*

Figure 3: Line plots showing the classification accuracy over age ranges (a) by algorithm and (b) by gender. Plots were generated with 243 023 visa images.

Age Range	# Females	# Males	B30D	B31D	C30D	E30D	E31D	E32D	F30D	K10D	P30D
0–10	11141	11442	77.0 | 82.8	77.1 | 82.6	90.2 | 20.3	68.1 | 72.9	68.5 | 74.1	68.0 | 72.1	78.0 | 34.0	87.8 | 25.3	93.6 | 16.3
11–20	11067	10859	92.9 | 92.2	93.7 | 91.3	97.5 | 74.0	94.2 | 90.2	92.5 | 91.9	94.7 | 90.7	88.3 | 72.2	91.3 | 75.4	91.4 | 81.9
21–30	32966	36786	97.3 | 97.9	97.5 | 97.9	98.5 | 93.5	96.4 | 98.3	95.1 | 98.7	97.3 | 98.7	92.2 | 79.4	90.8 | 93.2	94.6 | 95.9
31–40	21848	27185	96.0 | 99.0	96.0 | 99.1	97.9 | 94.4	95.0 | 98.7	93.5 | 98.9	97.0 | 98.9	90.9 | 80.5	87.0 | 95.1	92.6 | 96.9
41–50	16330	18145	92.2 | 99.4	92.5 | 99.5	95.8 | 94.3	92.0 | 98.9	90.0 | 98.9	96.4 | 98.8	88.5 | 78.9	82.8 | 95.4	87.2 | 97.3
51–60	14376	12185	84.4 | 99.5	85.4 | 99.6	92.4 | 93.7	87.4 | 98.9	85.0 | 98.8	94.7 | 98.4	84.9 | 78.2	77.9 | 95.0	78.6 | 97.8
61–70	7320	5960	74.4 | 99.4	75.1 | 99.5	86.4 | 93.1	79.7 | 98.5	78.2 | 98.3	90.7 | 97.4	77.7 | 78.9	67.7 | 94.5	66.4 | 97.6
71–80	2579	1960	59.3 | 99.5	59.9 | 99.4	73.2 | 92.2	67.4 | 97.1	68.3 | 97.1	84.1 | 97.0	64.7 | 79.3	55.6 | 93.6	53.8 | 95.3
81–90	457	355	49.2 | 97.6	49.7 | 97.9	61.3 | 91.7	66.1 | 94.8	64.7 | 95.5	82.5 | 92.0	53.3 | 81.7	48.1 | 90.6	47.0 | 93.5

Table 5: Gender classification accuracy, in percent, by age range and gender (Female | Male).

Results and notable observations:

- All of the algorithms show a significant decrease in gender classification accuracy as age increases for adult females. There is an empirically decreasing trend in accuracy seen in females past age 50, which implies algorithms are misclassifying females as males more often as age increases. Accuracy for males across increasing age groups does not demonstrate such evident decrease but remains comparatively stable. The correlation between age and gender classification error observed in older females presents an opportunity where gender classification could be used in conjunction with age estimation to mitigate the loss in accuracy in older women, for example, by moving the minimum "femaleness" threshold to a lower value if the person's estimated age is above a certain threshold. Figure 4 shows examples of misclassifications in females above age 50.
- For females, gender classification is the most accurate in young adults (ages 21-30) as seen across all algorithms.

Contrastingly, the highest accuracy is observed in older adult age groups (ages 31-60) for males. Gender classification is more accurate in adult males (ages 21-60) than young boys (ages 0-10) for all of the algorithms.

Figure 4: Examples of misclassifications in females above age 50 extracted from the FERET database, which are representative of image qualities from misclassifications observed in the operational dataset.

3.3 Ethnicity

While the effects of racial features on gender classification have been studied [9, 19] in the academic literature, the experimental results were derived using small datasets with a limited set of ethnicity groups. The visa dataset was used to investigate the effect of ethnicity on gender classification as it contains a respectable number of visa images across multiple ethnic proxies. The term ethnic proxy is used, because an individual could be a citizen of a country but not necessarily be of that country's ethnic descent. Ethnic proxy groups with a minimum of at least 1 800 images of subjects between ages 11-40 and a relatively balanced number of males and females were extracted and used in the analysis captured in Tables 6 and Figure 5.

	ARG	BRZL	CHIN	COL	DF	IND	ISRL	JPN	KOR	PERU	PHIL	POL	RUS	TWAN
B30D	95.0	97.7	96.4	96.4	99.0	97.2	97.8	90.0	94.4	97.2	97.9	98.1	97.6	94.1
B31D	95.3	97.8	96.7	96.7	99.0	97.7	98.0	90.1	94.3	97.6	98.1	98.2	97.7	94.0
C30D	93.3	95.0	89.2	94.6	96.7	98.1	96.4	82.9	83.6	95.4	92.8	97.3	96.8	81.5
E30D	91.2	96.6	96.4	95.0	97.2	97.8	96.6	95.1	95.6	96.8	97.7	97.5	96.9	96.1
E31D	89.8	97.0	94.4	94.7	96.1	96.9	95.8	96.7	94.9	96.9	97.6	97.7	96.6	96.8
E32D	93.9	97.2	96.6	95.5	97.6	98.4	97.3	95.8	95.8	97.5	98.0	98.1	97.6	97.1
F30D	80.9	86.4	80.8	86.2	83.6	82.5	87.1	84.3	81.7	86.2	78.6	89.8	87.8	76.1
K10D	86.8	91.2	89.6	87.0	88.1	94.6	94.5	85.5	86.0	90.1	90.9	92.6	91.4	88.4
P30D	90.8	95.0	93.2	93.8	96.1	96.0	95.3	89.2	90.5	95.0	95.3	94.5	94.8	88.9

Table 6: Overall gender classification accuracy, in percentage, by ethnic proxy group[3]. For each algorithm, the lowest accuracy number is highlighted in yellow, and the highest accuracy number is highlighted in green.

[3]DF is Mexico City.

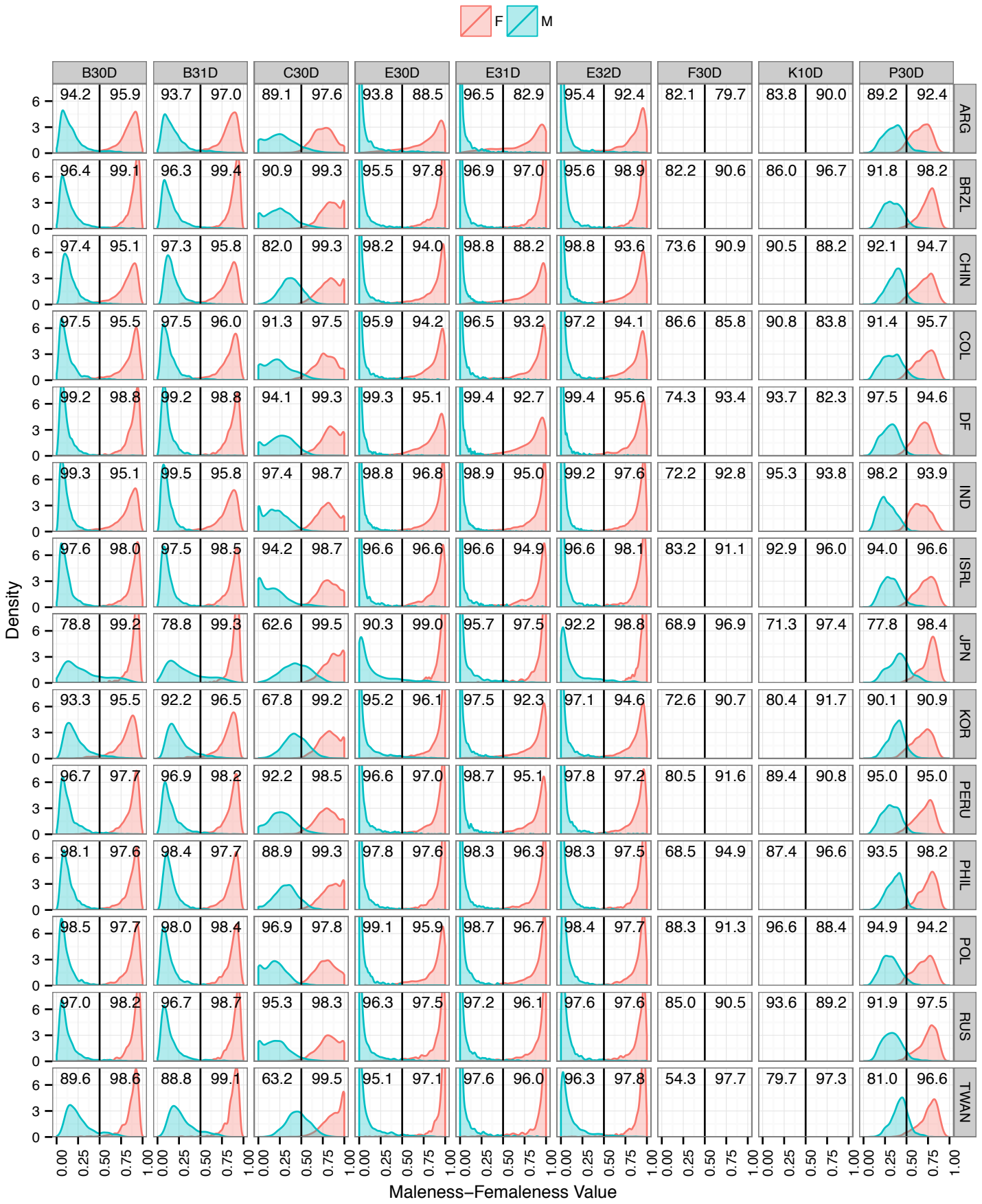


Figure 5: Density plots showing distribution of maleness-femaleness value across various ethnic proxies[3]. Plots were generated for ethnic proxy groups with at least 1 800 images of subjects between ages 11-40. Gender classification accuracy for male and female are shown in the upper left and right of each graph, respectively.

Results and notable observations:

- A majority of the algorithms show the lowest gender classification accuracy on subjects from Taiwan and Japan. It can be observed from Figure 5 that males are more often misclassified as females on people from these east Asian countries, which is a consistent trend between most of the algorithms. Algorithms F30D and K10D did not provide continuous values for maleness-femaleness.
- The higher misclassification rates observed in males from the Taiwan and Japan may anecdotally be related to slimmer and more defined face shapes seen in Asian men, facial hair being an uncommon trait, and the popularity of longer hairstyles, especially in Japanese men, but further research would be required to objectively verify these conjectures.

3.4 Multiple Image Samples

In certain applications, there are opportunities for multi-sampling of images, such as imagery being captured from video of people walking past a digital sign. For such scenarios, the question arises of whether accuracy improves if the gender classification implementation is provided multiple contemporaneous images of the same subject. This could drive whether a system, for example, used for targeted digital marketing, could set a minimum threshold on the number images of a person to process, based on some time-accuracy tradeoff, prior to making a decision on the type of advertisement to display.

The FRVT API [12] supports multiple still image input to the algorithm software for gender classification, which enables the analysis of gender classification performance versus the number of image samples of the same person. The mugshot dataset includes $K > 1$ contemporaneous images for some subjects, with contemporaneous, here, being defined as images of the same subject collected within a twelve month span. This allows for the modeling of a scenario where gender classification implementations can exploit multiple images. 11 920 subjects with at least four contemporaneous mugshot images were extracted. Figure 6 shows the effects of the number of image samples on classification accuracy for the algorithms that supported multiple image input.

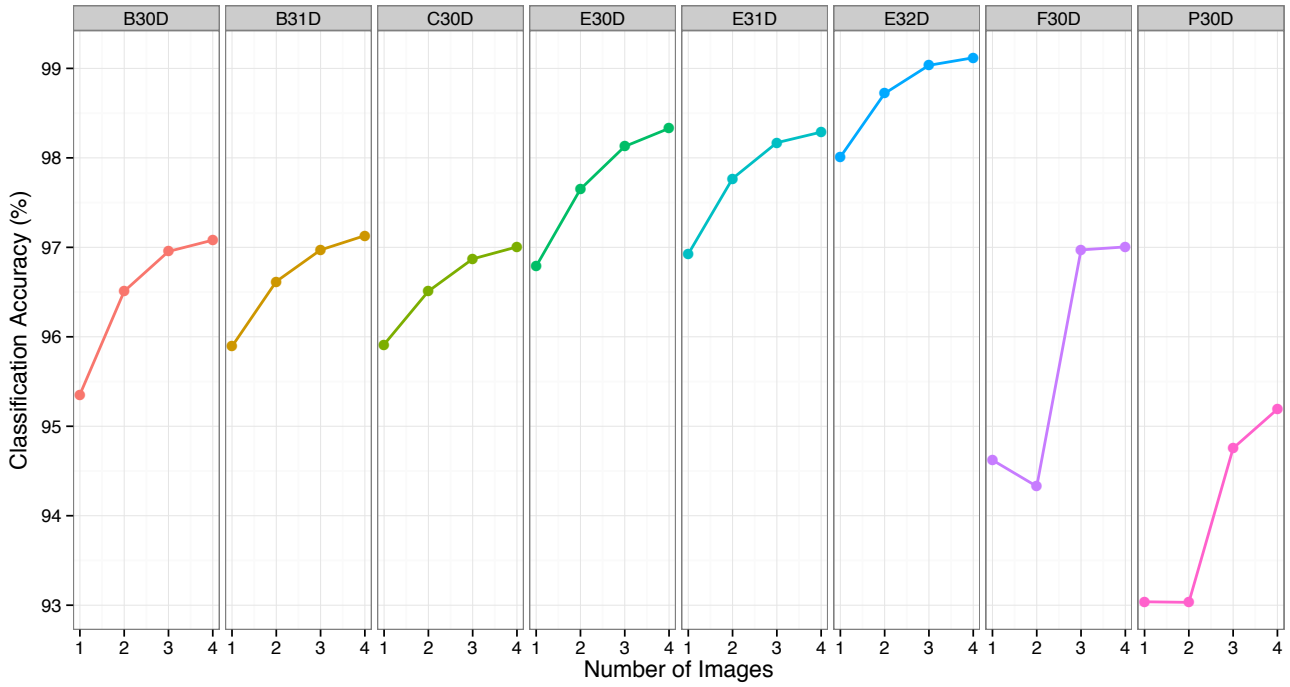


Figure 6: Line plots showing gender classification accuracy vs. the number of image samples per subject. Plots were generated with 11 920 subjects.

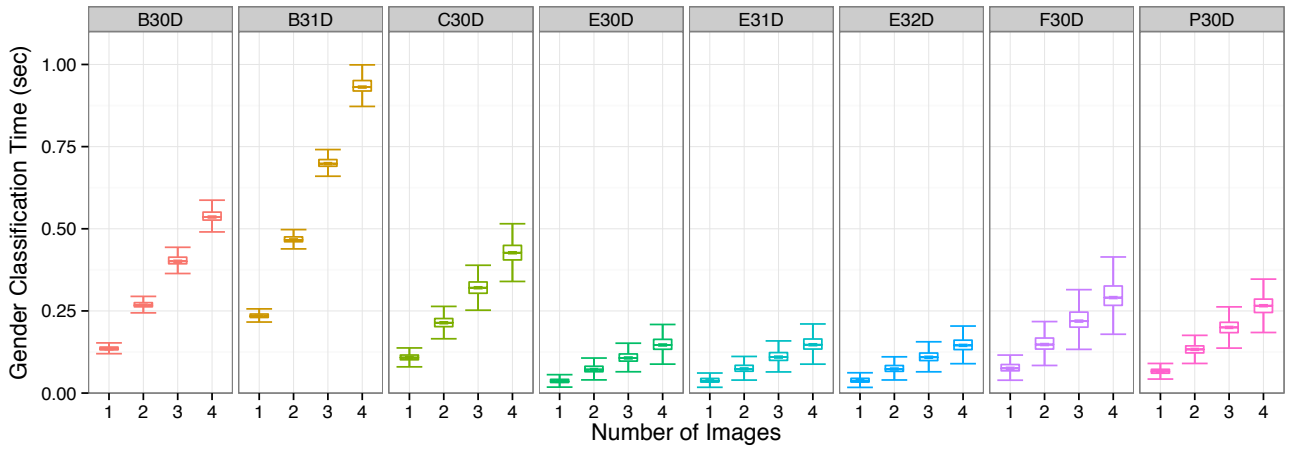


Figure 7: Boxplots summarizing gender classification time vs. the number of image samples per subject. Plots were generated with 11 920 subjects.

Results and notable observations:

- Figure 6 shows most of the algorithms demonstrate an increasing trend in gender classification accuracy as the number of image samples increases. Algorithms exhibit a decrease in misclassification rate of up to 50% between one and four images, depending on the provider. Algorithm K10D did not support multiple image input.
- Figure 7 shows gender classification durations increasing linearly with respect to the number of image samples, as expected.

3.5 In the Wild

In recent years, the study of gender classification on face databases acquired in the wild (i.e., unconstrained face images that show a large range of variation in factors such as pose, illumination, facial expression, and background) has gained popularity in the published literature. Among such unconstrained datasets are Images of Groups (GROUPS) [10] and Labeled Faces in the Wild (LFW) [15]. GROUPS consists of images of groups of people collected from Flickr, and in all, the dataset consists of 28 231 (14 559 females and 13 672 males) largely low resolution faces that are labeled with gender. The LFW dataset consists of 13 233 images (2 977 females and 10 256 males) of famous people collected over the Internet. See Section 2.3 for more details on these datasets.

Tables 7 and 8 tabulate the performance of FRVT participants on the GROUPS and LFW datasets, alongside results of methods published from the academic literature.

Algorithm	Overall Accuracy (%)
B30D	86.3
B31D	86.2
C30D	68.7
E30D	90.3
E31D	90.2
E32D	90.4
F30D	71.1
K10D	71.3
P30D	83.5

(a) FRVT participants, using lights-out, black-box testing protocol over entire dataset (28 231 images)

Publication	Overall Accuracy (%)	# Images
Gallagher and Chen (2009) [10]	74.1	25 099
Shan (2010) [22]	77.4	12 080
Dago-Casas et al. (2011) [6]	86.6	14 760
Han and Jain (2014) [14]	87.1	28 231
Bekios-Calfa et al. (2014) [4]	80.5	Faces of children $\leq$ age 12 were excluded.

(b) Published methods, using various testing protocols and images

Table 7: Tables summarizing overall gender classification accuracy on the GROUPS dataset.

Algorithm	Overall Accuracy (%)
B30D	94.4
B31D	94.4
C30D	84.5
E30D	95.0
E31D	95.2
E32D	94.9
F30D	79.1
K10D	76.2
P30D	93.9

(a) FRVT participants, using lights-out, black-box testing protocol over entire dataset (13 233 images)

Publication	Overall Accuracy (%)	# Images
Dago-Casas et al. (2011) [6]	94.0	13 233
Shan (2012) [23]	94.8	7 443
Tapia and Perez (2013) [24]	98.0	7 443
Bekios-Calfa et al. (2014) [4]	79.5	13 233
Shaefey et al. (2014) [7]	94.6	13 233
Ren and Li (2014) [21]	98.0	6 840

(b) Published methods, using various testing protocols and images

Table 8: Tables summarizing overall gender classification accuracy on the LFW dataset.

Results and notable observations:

- For both datasets, the top performing FRVT participants achieved comparable or higher overall gender classification accuracy when compared to published methods that tested their algorithms with the same number of images.
- While standard evaluation protocols for gender classification have been proposed for these datasets [1], straight comparison between performance results remains to be a challenge given some methods are tested on only a subset of the images (e.g. excluding non-frontal images or children) and others using single versus cross-database validation, which in all cases, allowed the implementation to train on a set of images through every iteration of testing. For the FRVT participant results documented, all images from the datasets were used for testing, and NIST employed a lights-out, black-box testing methodology that did not involve any type of algorithm training during evaluation. This is designed to model operational reality where software is shipped and used "as-is" without subsequent algorithmic training.

3.6 Constrained versus In the Wild

The performance of gender classification over constrained and in the wild datasets are presented in Sections 3.1 and 3.5, respectively. Figure 8a contains sample photos extracted from the publicly available Multiple Encounter Dataset (MEDS) [2], which are representative of image qualities from the constrained dataset. Figure 8b contains sample photos from the public in the wild datasets used in this study, which contain images that exhibit a large range of variation in pose, illumination, facial expression, and background. Comparative analysis of the performance on the two different types of images would reveal whether gender classification is affected by factors such as pose, illumination, facial expression, and background. Figure 9 compares the distribution of the maleness-femaleness value, and Table 9 presents the average maleness-femaleness value, by gender for constrained versus in the wild images. The constrained dataset discussed in Section 3.1, and the aggregate of the two in the wild datasets discussed in Section 3.5 were used in this study.

(a) Constrained *(b) In the Wild*

Figure 8: Examples of constrained images extracted from the MEDS dataset and in the wild images extracted from LFW and GROUPS.

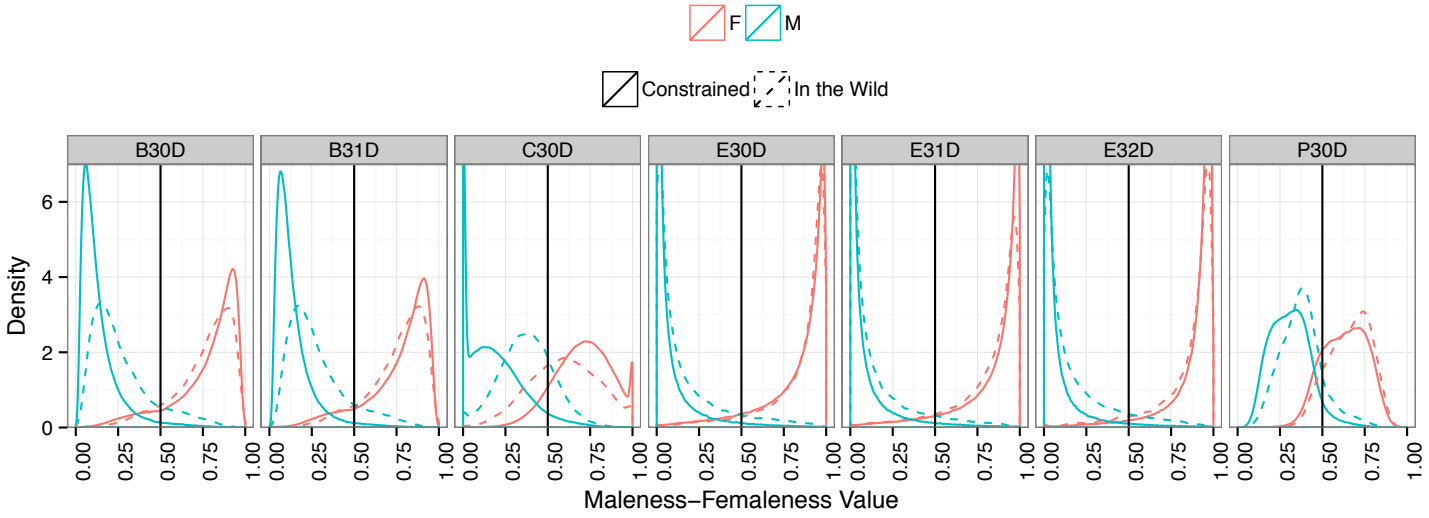


Figure 9: Density plots showing distribution of maleness-femaleness value, by gender and image type (constrained versus in the wild images).

Algorithm	Male		Female	
	Constrained	In the wild	Constrained	In the wild
B30D	0.14	0.27	0.77	0.76
B31D	0.14	0.28	0.76	0.75
C30D	0.18	0.37	0.71	0.60
E30D	0.07	0.16	0.83	0.84
E31D	0.07	0.15	0.84	0.82
E32D	0.07	0.17	0.88	0.85
P30D	0.31	0.38	0.63	0.66

Table 9: Average maleness-femaleness value, by gender and image type (constrained versus in the wild).

Results and notable observations:

- A majority of the algorithms show visible shifting of the distribution of maleness-femaleness value towards the middle and, the increase in overlap between male and female values is indicative that the qualities of the in the wild images are causing a decrease in algorithm certainty of gender classification. This observation is further supported by the comparison of average maleness-femaleness value where for all of the algorithms, the average maleness value is lower in constrained images (more certainty of maleness) and for a majority of the algorithms, the average femaleness value is higher in constrained images (more certainty of femaleness).
- Based on the average maleness-femaleness value, the impact of the qualities of images collected in the wild appear to have a bigger impact on males than females, given the larger differences observed in the average values for males.

3.7 Specific Applications

3.7.1 Gender Verification

Gender-targeted surveillance can assist with monitoring gender-restricted areas and potentially aid in filtering subjects of interest - for example, in the event where elevated threat levels may be associated with a specific gender. Consider a surveillance system with an automated gender classification system that has the ability to filter and highlight subjects based on gender. In a scenario where areas restricted to females (e.g., entrance to female restrooms or locker rooms) are being monitored, an operator may be alerted when a male is detected in view. In this case, one might set the cost of falsely classifying a female as a male (i.e., the false male rate) to the inconvenience of having to review the alert. The cost of falsely classifying a male as a female (i.e., the false female rate) could result in allowing suspicious or threatening activity to be conducted. Given it would be reasonable to argue that the costs are asymmetric in this scenario, i.e., the cost of a false female is greater than that of a false male, tighter confidence levels could be set to minimize false females. Figure 10 presents DET accuracy for gender classification for algorithms that supplied continuous maleness-femaleness values (F30D and K10D did not provide continuous maleness-femaleness values). A false female rate of 0.01 would impose a false male rate of 0.10 for the most accurate algorithm (E32D) at that threshold. In other words, if a system were to ensure that 99% of monitored subjects entering a female locker room were indeed female, it would mean that 10% of the time, the operator is falsely alerted of a male being detected in view. The same Detection Error Tradeoff (DET) analysis can be done by gender and for different verification thresholds to support specific applications. While the results documented are generated from single image input, analysis with multiple image input (i.e., sequential frames from video) may further solidify the results observed and be more representative of surveillance systems running in video mode.

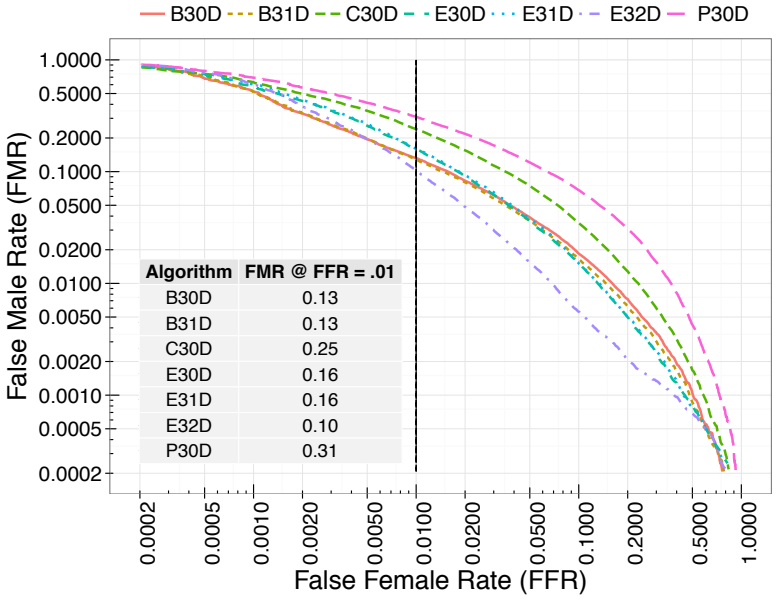


Figure 10: DET curve plotting false male rate against false female rate for gender classification. Plot was generated over 218,525 images of subjects between age 11-90.

3.7.2 Sketches

Sketches have long been used in criminal investigations. Historically the most common occurrence is for a forensic artist to interview an eye-witness and, iteratively, produce a likeness of the individual recollected by the witness. That sketch could then be used to try and match against photographs resident in a mugshot database. Given gender can be utilized

as an indexing technique to reduce the search space for automatic face recognition, the accuracy of gender classification algorithms on sketches has an interesting niche application.

The City University of Hong Kong published the CUHK Face Sketch FERET Database (CUFSF), which is composed of sketch images of subjects from the FERET dataset. A set of 856 sketches from CUFSF was used for this study. Figure 11 shows examples of the sketches, alongside the original FERET image, and Table 10 presents the gender classification accuracy of algorithms on sketches, alongside the performance on the corresponding photographs from the FERET database.

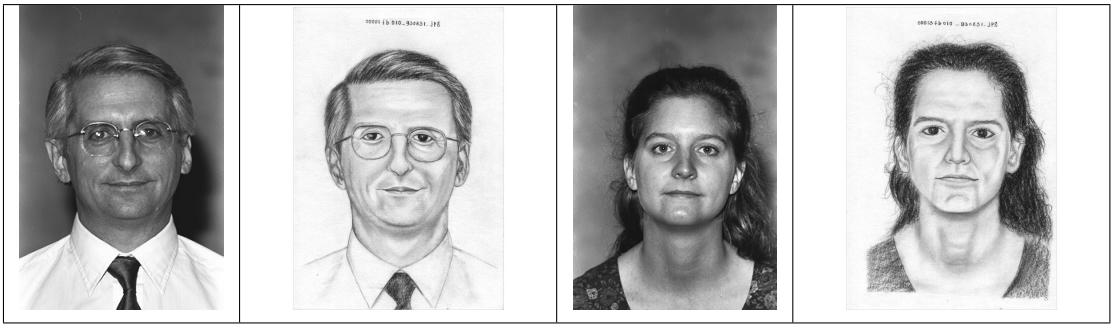


Figure 11: Images and sketches of subjects from the FERET database. This production of a sketch is atypical operationally given it is unusual for an artist to have access to an image of the individual.

Algorithm	Accuracy on sketches (%)	Failure to compute on sketches (%)	Accuracy on photos (%)
B30D	90.9	0.0	97.5
B31D	93.8	0.0	98.0
C30D	85.6	0.0	96.3
E30D	82.2	87.5	96.3
E31D	85.0	87.5	96.1
E32D	84.1	87.5	96.5
F30D	68.8	0.0	85.5
K10D	N/A	100.0	87.3
P30D	85.1	0.1	92.2

Table 10: Table summarizing overall gender classification accuracy on sketches, failure to compute percentages on sketches, and accuracy on original FERET photos.

As gender classification on sketches was never declared to be a part of the study, the algorithms are being used in a manner not expressly intended by the providers. Operationally though, automated face recognition algorithms are being used to recognize sketches in many police departments [11], and gender classification on sketches may help reduce the search space for face recognition in this scenario.

The use of this image set probably does reveal differences in algorithmic capability. Variation will in part depend on what facial information is represented. Participant E and K's algorithms exhibit very high failure to compute percentages on the sketch images, with the reason being "involuntary failure to extract features from the image" (as indicated in the FRVT API). All of the algorithms perform better on the FERET photographs than their corresponding sketches, which is a likely indication that there is information in the photographs used for gender classification that may not be represented in the sketch images. While many of the algorithms give very high classification accuracy on the FERET photographs, the most accurate algorithm on sketches (B31D) classifies gender correctly 93.8% of the time, which is better than some algorithms even on the good quality photographs.

References

[1] BeFIT - Benchmarking Facial Image Analysis Technologies. http://fipa.cs.kit.edu/412.php.

[2] NIST Special Database 32 - Multiple Encounter Dataset 2 (MEDS-II), NISTIR 7807. http://www.nist.gov/itl/iad/ig/sd32.cfm.

[3] U.S. Department of State, Bureau of Consular Affairs, Visa Photo Requirements. http://travel.state.gov/content/visas/english/general/photos.html.

[4] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela. Robust gender recognition by exploiting facial attributes dependencies. *Pattern Recogn. Lett.*, 36:228–234, January 2014.

[5] Y. D. Cheng, A. J. O'Toole, and H. Abdi. Classifying adults' and children's faces by sex: computational investigations of subcategorical feature encoding. In *Cognitive Science*, number 25, pages 819–838, September 2001.

[6] P. Dago-Casas, D. González-Jiménez, L. Long-Yu, and José Luis Alba Castro. Single- and cross- database benchmarks for gender classification under unconstrained settings. In *ICCV'11 -BeFIT 2011*, November 2011.

[7] Laurent El Shafey, Elie Khoury, and Sébastien Marcel. Audio-visual gender recognition in uncontrolled environment using variability modeling techniques. In *International Joint Conference on Biometrics*, 2014.

[8] D. Enlow. *Handbook of facial growth*. W. B. Saunders, Philadelphia, PA, 1982.

[9] G. Farinella and J. Dugelay. Demographic classification: Do gender and ethnicity affect each other? In *Informatics, Electronics Vision (ICIEV), 2012 International Conference on*, pages 383–390, May 2012.

[10] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.

[11] P. Grother and M. Ngan. Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms. NIST Interagency Report 8009, National Institute of Standards and Technology, 2014. http://biometrics.nist.gov/cs_links/face/frvt/frvt2013/NIST_8009.pdf.

[12] P. Grother, G. W. Quinn, and M. Ngan. FRVT Still Face Image and Video Concept, Evaluation Plan and API Version 1.4, 2013. http://www.nist.gov/itl/iad/ig/frvt-2012.cfm.

[13] G. Guo, C. R. Dyer, Y. F., and T. S. Huang. Is gender recognition affected by age. In *In ICCV (Workshops)*, 2009.

[14] H. Han and A. K. Jain. Age, gender and race estimation from unconstrained face images. MSU Technical Report MSU-CSE-14-5, Michigan State University, 2014.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[16] A. Jain and J. Huang. Integrating independent components and linear discriminant analysis for gender classification. *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR04)*, pages 159–163, 2004.

[17] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proc. Intl Conf. Biometric Authentication (ICBA 04) LNCS 3072*, pages 731–738, 2004.

[18] S. A. Khan, M. Ahmad, M. Nazir, and N. Riaz. A comparative analysis of gender classification techniques. *International Journal of Bio-Science and Bio-Technology*, 5(4):223–244, 2013.

[19] O. Ozbudak, M. Kirc, Y. Cakir, and E.O. Gunes. Effects of the facial and racial features on gender classification. In *MELECON 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*, pages 26–29, April 2010.

[20] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, October 2000.

[21] Haoyu Ren and Ze-Nian Li. Gender recognition using complexity-aware local features. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 2389–2394, Aug 2014.

[22] C. Shan. Learning local features for age estimation on real-life faces. In *Proc. ACM Workshop on MPVA*, 2010.

[23] Caifeng Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431 – 437, 2012. Intelligent Multimedia Interactivity.

[24] J.E. Tapia and C.A. Perez. Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape. *Information Forensics and Security, IEEE Transactions on*, 8(3):488–499, March 2013.

[25] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):1955–1967, Nov 2009.

[26] J. L. Wayman. Large-scale civilian biometric systems - issues and feasibility. In *Card Tech / Secur Tech ID*, 1997.

[27] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, 2009. ISBN 978-0-387-98141-3.