Face Recognition Vendor Test (FRVT)

Performance of Automated Age Estimation Algorithms

NIST Interagency Report 7995

M. Ngan and P. Grother

Information Access Division

National Institute of Standards and Technology

NIST

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

March 20, 2014

AUTHENTICATED
U.S. GOVERNMENT
INFORMATION
GPO

Executive Summary

Introduction

Facial age estimation is an area of study new to the Face Recognition Vendor Test (FRVT) with Still Facial Images Track. While peripheral to automated face recognition, it has become a growing area of research, given its potential use in various applications. The motivation for age estimation systems has grown in the last few decades, given the rise of the digital age and the increase in human-computer interaction. Age-based access control and verification (e.g., age verification for alcohol/tobacco purchases), age estimation in crime and mass disaster investigation (e.g., age determination of unknown human bodies at a crime scene to help with victim identification), age-adaptive targeted marketing (e.g., displaying age-specific advertisements from digital signage), age-invariant person identification (e.g., identifying missing children), and age-based indexing of face images are potential applications of automated facial age estimation.

NIST performed a large scale empirical evaluation of facial age estimation algorithms, with participation from five commercial providers and one university, using three large operational datasets comprised of facial images from visas and law enforcement mughots, leveraging a combined corpus of over 7 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 algorithmic training. Core age estimation accuracy was baselined over a large homogeneous population, then assessed demographically by age group, gender, and ethnicity. The impact of input-driven variations, namely image quality and number of image samples per subject was captured, and assessments of age-verification accuracy and estimation accuracy in children were documented.

Key Results

Core Accuracy and Speed: Age estimation accuracy depends strongly on the provider of the core technology. Broadly, there is a twofold difference between the most accurate and the least accurate algorithm in terms of the percentage of images correctly classified to within five years and mean absolute error (MAE)[1]. Using the most accurate age estimation algorithm, (i.e., B31D from Cognitec), the chance of accurately estimating the age of a person within five years of their actual age over an ethnically-homogeneous database of 6 million images is 67%, with an MAE of 4.3 years. All algorithms can perform age estimation on a single image in less than 0.15 seconds with one server-class processor. The most accurate algorithm, on average, performs estimation in 0.125 seconds.

The main dataset used for overall accuracy assessment is comprised of 6 million ethnically-homogeneous images. Although image collection was subject to the guidelines published by the Department of State (DoS), 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, age estimation performance may improve, but errors will still likely exist due to ageing variation driven by intrinsic and extrinsic factors.

Impact of Demographic Data on Accuracy: For a heterogeneous dataset of 240 thousand images, it is empirically observed that age is more accurately estimated in males than females, with the tendency for adult females to be underestimated in age. A majority of the algorithms demonstrated lower accuracy and higher MAE on an ethnically-heterogeneous population than a homogeneous population, which suggests that ethnicity has an impact on age estimation. South Americans tend to be overestimated in age, and Asians tend to be understimated. A majority of the algorithms estimate age more accurately for the most operationally relevant age group, i.e., adults age 18-55. The adult age group is also where estimation accuracy is closest among the algorithms. The majority of algorithms exhibit the highest MAE in the senior age group, i.e., age 56-99.

These results state empirical observations for the particular dataset, but they do not determine cause. The impact of extrinsic factors potentially driving the observed results between gender and ethnicity, such as cosmetics and plastic surgery, are not studied in this report. Further research would be required to objectively verify these conjectures.

Age Verification Accuracy: For a system with an objective to verify that a person is at least 21 years old, a 17 year-old

[1] For more details on cumulative score and mean absolute error, see sections 2.4.2 and 2.4.1.

has a 29% chance of passing for 21, as achieved by the most accurate algorithm (i.e., B30D from Cognitec). This false-verification percentage increases as a person gets closer to age 21. To ensure that 98% of underage individuals are detected would falsely provide an underage result for 39% of people who are actually above age 21. The same Detection Error Tradeoff (DET) analysis can be done for other verification ages of interest to support specific applications.

Impact of Image Quality on Accuracy[2]: Comparing poor quality webcam photos and better quality mugshot images, the majority of the algorithms demonstrate lower accuracy with the webcam images, with varying rates in accuracy degradation between the algorithms. Observable biasing in the error distribution to the right is seen in poor quality webcam images, which suggests the occurrence of overestimation. There is no clear explanation for the overestimation seen, although it may anecdotally be related to the demographics of the subjects captured in the webcam images.

Impact of Number of Image Samples on Accuracy: The FRVT Application Programming Interface (API) [11] supports multiple still image input to the algorithm software for age estimation, which enables the analysis of age estimation 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 MAE monotonically decreasing as the number of image samples provided increased for all algorithms. There is an improvement in MAE of about one year between one and four input images. Age estimation times increase linearly with respect to the number of image samples, which is the expected behavior.

Comparison against Academic Methods: A performance evaluation was done with the commonly benchmarked FG-NET Aging Database [1] in an attempt to compare FRVT age estimation participants with published methods from the academic literature. A fair performance comparison could not be made due to observed fundamental differences in testing protocol employed by academia versus NIST. The published methods from academia were tested through a protocol which allowed the implementation to train on a subset of the images through every iteration of testing. NIST employed a lights-out, black-box testing methodology that did not involve any type of algorithm training during evaluation, as designed to model operational reality.

[2] Guidelines for quality aspects of facial images are documented in ANSI/NIST-ITL 1-2011 [4], Annex E. For the study documented in this report, image quality is broadly defined by image size, resolution, and contrast and subject illumination and pose.

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, the Office of Biometric Identity Management (OBIM) office in the Department of Homeland Security (DHS), and the Science and Technology (S&T) Directorate, also in 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 [22] package running under R, 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

Facial age estimation is an area of study new to the Face Recognition Vendor Test (FRVT) with Still Images Track. Automated facial age estimation is the calculation of an individual's age by computer software based on features derived from a person's face image. Age estimates are a soft biometric [16] and its characterization has become a growing area of study given its potential use in various applications. The progression of age brings changes to the appearance of the human face. A number of factors including the variation between ethnicities and gender, as well as dependence on external factors such as health conditions and lifestyle [9], have introduced challenges to age estimation.

The main goals of this evaluation are to:

- Provide an objective assessment of current automated age estimation technology.
- Leverage massive operational corpora. The availability of images from large populations (in the millions) supports statistical significance of the studies. The use of operational images brings greater operational relevance to the test results.
- Investigate age estimation accuracy across various factors, including age group, ethnicity, and gender.

1.2 Application Scenarios

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

Age-based access control and verification has long been a familiar concept where a person's age is verified (e.g., ID check) prior to physical access to a place or product being sold or virtual access to a website is granted. Examples include preventing minors from accessing adult websites and age verification for tobacco and alcohol purchases. In Japan, age-verification cameras are installed in a number of cigarette vending machines to estimate the patron's age prior to allowing their purchase.

Age estimation in crime and mass disaster investigation, for example, age determination of unknown human bodies is important in the setting of a crime investigation or a mass disaster, because the age can guide investigators to the correct identity among a large number of possible matches. Automated facial age estimation may offer a less invasive alternative for age determination as compared to some traditional techniques used in forensic sciences [5].

Age-adaptive human-computer interaction is on the rise given the popularity of digital signage and the opportunity for targeted digital marketing. Advertisements targeted for a certain age group can be displayed based on the age of the audience walking past a digital sign.

Age-based indexing of face images, that is, the use of age as criterion for indexing into large-scale biometric databases for faster retrieval has been discussed [21] and can also apply to automatic sorting and image retrieval from digital photo albums and the internet.

Age-invariant person identification is required in some face recognition applications where age compensation is required (e.g., identifying missing children over time), given a significant age difference may exist between the probe and gallery images.

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 6-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 age estimation 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 [11]. 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 age estimation, 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
NEC	E	E10D		E30D,E31D,E32D
Tsinghua University	F	F10D		F30D
MITRE	K	K10D		
Zhuhai-Yisheng	P			P30D
JunYu Tech.	Q	Q10D		

Table 1: FRVT Age Estimation Participants

2.3 Image Dataset

This report documents the use of the following datasets[3]:

- LEO: 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. The majority of images are traditional mugshot photos with a

[3] 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.

small subset of images captured with webcams.

- DoS/P: This dataset consists of facial images for visa applicants.
- DoS/Natural: This dataset consists of facial images for non-immigrant visa applicants.
- FG-NET: This is a public dataset composed of personal photographs that is widely used for benchmarking age estimation performance in academia.

Facial images from visa applications will henceforth be referred to as ”visa images” in this report.

The dataset properties are summarized in Table 2.

Property	LEO	DoS/P	DoS/Natural	FG-NET
Collection Environment	Law enforcement booking	Visa application process	Visa application process	Personal photos
Collection Era	˜1960s-2008	˜2006-2010	˜1996-2002	Unknown
Digital, Paper Scan	Digital, few paper	Mostly digital	Mostly digital	Unknown
Documentation	See NIST Special Database 32 Vol. 1			
Image size	Various, 480x600, 240x240, 768x960	Most 252x300	Most 252x300	Most ˜400x500
Compression	JPEG ˜20:1	JPEG, mean size: 16.2kB	JPEG, mean size: 9.2kB	JPEG, mean size: 44.2kB
Eye to eye distance	Mean = 108 pixels, SD = 40 pixels	Median = 71 pixels	Median = 71 pixels	
Frontal pose	Moderate control. Known profile images excluded.	Well controlled	Well controlled	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[4] Full Frontal requires.	Yes, in most cases. Faces are more cropped (i.e., smaller background) than ISO Full Frontal requires.	
Source	Operational data	Operational data	Operational data	Public dataset
Notable Population Characteristics	Predominantly mugshots with a small subset of webcam images	Various	Predominantly Mexican	Over 50% between age 0-13

Table 2: Image dataset descriptions.

The datasets are characterized by population sizes well in excess of all published age estimation tests (See Table 9a). The number of images are given in Table 3.

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

Quantity	LEO	DoS/P	DoS/Natural	FG-NET
Number of age labeled face images	2378635	243023	6249313	1002
Number of subjects	1802874	222862	5738141	82
Age	18-109	0-100	0-99	0-69

Table 3: Image dataset sizes.

2.4 Performance Metrics

The following performance measures will be reported in the assessment of age estimation:

2.4.1 Mean and Median Absolute Error

Mean Absolute Error (MAE) is defined as the average of the absolute errors between the estimated ages and the actual ages. i.e.,

$$\text{MAE} = \frac{\sum_{k=1}^{N} |\hat{a_k} - a_k|}{N}, \tag{1}$$

where $\hat{a_k}$ is the estimated age for the k-th test image, a_k is the corresponding ground-truthed age, and N is the number of test images.

Median Absolute Error is defined as the median of the absolute error values, i.e., in an ordered set of increasing values,

$$\text{Median Absolute Error} = (\frac{N+1}{2})^{th} \text{ value}, \ \frac{(\frac{N}{2})^{th} \text{ value} + (\frac{N}{2}+1)^{th} \text{ value}}{2} \text{ for } N_{odd} \text{ , } N_{even} \text{ respectively} \tag{2}$$

2.4.2 Cumulative Score (CS)

CS is defined as the percentage of test images such that the absolute error is not higher than a threshold, t, (in years). i.e., Given

$$\begin{gathered} H(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases}, \\ \text{CS}(t) = (1 - \frac{\sum_{k=1}^{N} H(|\hat{a_k} - a_k| - t)}{N}) \times 100, \end{gathered} \tag{3}$$

where $\hat{a_k}$ is the estimated age for the k-th test image, a_k is the corresponding ground-truthed age, and N is the number of test images. "Accuracy" is defined by CS(t), and both terms are used interchangeably in this report.

2.4.3 Age Verification Error

For age verification, the fundamental error rates for a particular verification age A is defined as:

$$\text{False negative rate}_A(T) = \frac{\text{Number of people equal or above age } A \text{ with an age estimate } < \text{ threshold, } T}{\text{Number of people equal or above age } A} \tag{4}$$

$$\text{False positive rate}_A(T) = \frac{\text{Number of people below age } A \text{ with an age estimate} \geq \text{threshold, } T}{\text{Number of people below age } A} \tag{5}$$

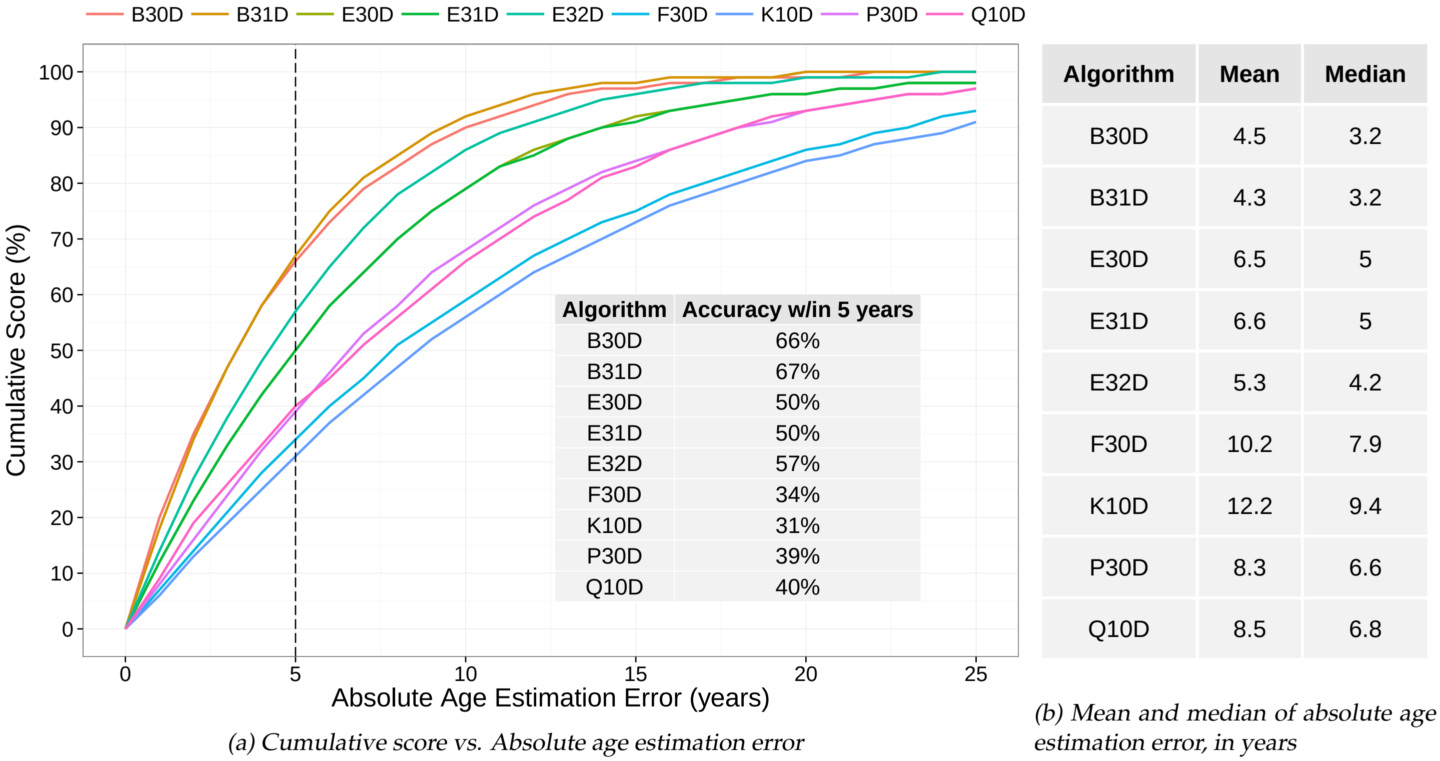
A false negative would occur when a person who is older than a certain verification age A is estimated as being younger than A. A false positive occurs when a person who is younger than A is estimated as being older than A. These error rates are plotted as a Detection Error Tradeoff (DET) characteristic, where, for example, in an age-based access control scenario, the rate of people younger than the minimum age requirement being granted access (i.e., the false positive rate) is traded off against the rate of people who meet the minimum age requirement being denied access (i.e., the false negative rate).

3 Results

3.1 Age Estimation in Large Homogeneous Population

3.1.1 Accuracy

The DoS/Natural dataset contains a subset of 6,172,395 images over an ethnically-homogeneous population spanning ages 0-99, both male and female. As such, statistically significant baseline age estimation performance results can be generated, as presented in Figure 1.



Algorithm	Accuracy w/in 5 years
B30D	66%
B31D	67%
E30D	50%
E31D	50%
E32D	57%
F30D	34%
K10D	31%
P30D	39%
Q10D	40%

(a) Cumulative score vs. Absolute age estimation error

Algorithm	Mean	Median
B30D	4.5	3.2
B31D	4.3	3.2
E30D	6.5	5
E31D	6.6	5
E32D	5.3	4.2
F30D	10.2	7.9
K10D	12.2	9.4
P30D	8.3	6.6
Q10D	8.5	6.8

(b) Mean and median of absolute age estimation error, in years

Figure 1: Line plot showing the accuracy of algorithms at absolute error levels and table of summary error statistics generated with 6,172,395 images over an ethnically-homogeneous population.

Results and notable observations:

- For the most accurate algorithm (i.e., B31D), 67% of estimates were accurate to within five years with an MAE of 4.3 years.
- It is clear from the results that participant B's algorithms outperform the rest. The other algorithms have substantially lower performance numbers, with the next most accurate, which is participant E, being 10% lower in accuracy within five years (i.e., E32D).

- For all of the algorithms, the mean absolute error is higher than the median absolute error, which is indicative of skewness of the error distribution and the existence of large error values that drive the mean upward. Depending on the algorithm, this could be a result of higher error driven by certain age groups, which is discussed in Section 3.2.
- The dataset used is comprised of visa images. Although image collection was subject to the guidelines published by the DoS, the images are highly compressed, which reduces the amount of detail in the images. With more detail available in less compressed images, age estimation accuracy may improve.

3.1.2 Speed

Speed could be an important performance factor in some age estimation 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 age as criterion for indexing into large-scale biometric databases would levy rapid speed requirements on age estimation algorithms to make it operationally viable.

Figure 2 presents the distribution of age estimation times for each algorithm. Age estimation time is the amount of time elapsed computing the age estimation 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.

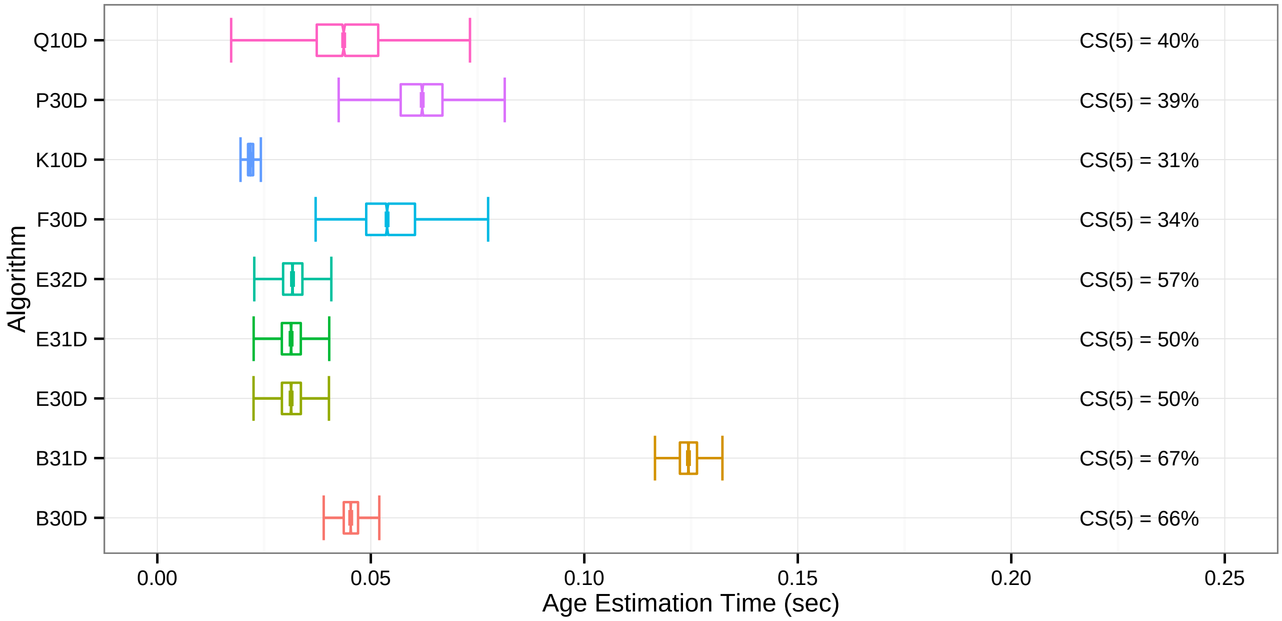


Figure 2: Boxplots of the distribution of age estimation times. Plots were generated over 5,000 age estimates. For reference, CS(5) against a population of 6,172,395 is reported on the right.

Results and notable observations:

- Age estimation time varies considerably from one participant to another. K10D can perform estimation in less than 0.025 seconds while B31D takes about five times longer than that, on average.
- While B31D has the highest estimation times, it has the highest accuracy, while K10D, with the fastest estimation speeds, exhibits the lowest accuracy. No clear speed-accuracy tradeoff exists between the other algorithms. B30D is among the most accurate, but not among the slowest.
- Participant E appears to have fixed age estimation times, while exhibiting significant variation in accuracy between its algorithms.

3.1.3 Failure to Compute Rate

The accuracy results presented above were computed for cases where age estimation did not fail. The error metrics do not include a penalty for cases where an algorithm failed to generate an age estimate. Per the FRVT API [11], a failure to compute occurs when an algorithm's code returns a non-zero return value from a call to its age estimation function, and hence fails to generate an age 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 an age estimate over 6,172,395 images.

Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
6172395	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.4787	0.0006	0.0003

Table 4: Table summarizing failure to compute ratio over 6,172,395 images.

Results and notable observations:

- 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 images collected under published DoS 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 exhibits a significantly high failure rate, failing on approximately half of the images, with the reason being involuntary failure to extract features from the image (as indicated in the FRVT API).

3.2 Age Groups

The facial ageing process drives different types of changes among different age groups. While facial ageing is mostly represented by craniofacial growth in younger age groups, it is mostly represented by relatively large texture changes in older age groups [9]. This introduces a challenge to age estimation algorithms given the ageing variation between age groups. Accuracy across three major age groups, which are youth (0-17), adult (18-55), senior (56-99), over an ethnically-homogeneous population is assessed and summarized in Figure 3 and Tables 5 and 6. Given that age group definition and composition are often application-specific, the CS(5) and MAE at each age is provided in Appendix A for reference.

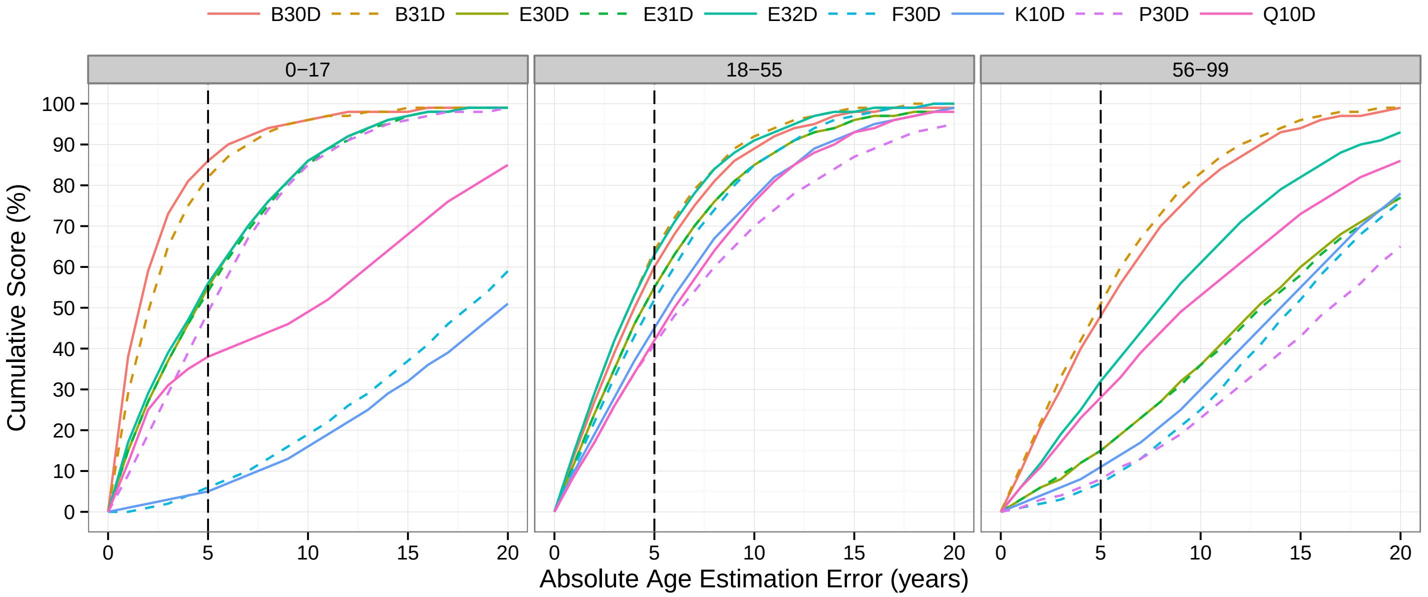


Figure 3: Line plots showing the accuracy of algorithms at absolute error levels by age group. Plots were generated over an ethnically-homogeneous population.

Age Group	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
0−17	1605807	86	82	55	54	56	6	5	49	38
18−55	3781607	60	64	55	55	63	52	45	41	42
56−99	784981	48	51	15	15	32	7	11	8	28
Overall	6172395	66	67	50	50	57	34	31	39	40

Table 5: CS(5), in percentage, by age group and overall CS(5) from Figure 1a for reference.

Age Group	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
0−17	1605807	2.6	3	5.3	5.4	5.3	18.6	21	6.1	10.9
18−55	3781607	4.9	4.5	5.5	5.5	4.6	5.6	6.6	7.6	7
56−99	784981	6.2	5.7	13.8	13.9	8.9	14.7	14	16.7	10.9
Overall	6172395	4.5	4.3	6.5	6.6	5.3	10.2	12.2	8.3	8.5

Table 6: MAE, in years, by age group and overall MAE from Table 1b for reference.

Results and notable observations:

- All of the algorithms estimate age more accurately for a particular age group better than others, and no algorithm has consistent MAE across all age groups.
- For the most operationally relevant age group, i.e., adults (age 18-55), algorithms are closer in performance, with B31D and E32D being the top performers.
- Participant B's performance is superior in the youth and senior age groups, leading the next most accurate algorithm in 5-year accuracy by 30% and 16% respectively.

- Algorithms exhibit lower accuracy and higher MAE in the senior age group, which could be driven by large errors seen in the higher ages in this group.

3.3 Ethnicity

The way a person ages can depend on a number of intrinsic factors, ethnicity being an important one of them, driving genetically inherited ageing patterns as well as extrinsic characteristics such as habitation climate and cultural behavior choices. Methods to address age estimation across ethnicity have been published in academic studies [12, 17]. The DoS/P dataset 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 2,800 images were extracted and used in the analysis captured in Tables 7 and 8 and Figure 4.

	ARG	BRZL	CHIN	COL	DF	DOMR	ECUA	GUAT	IND	ISRL	KOR	PERU	PHIL	POL	RUS	TWAN	VENZ	Mean
B30D	70	67	45	63	68	58	65	68	61	72	46	69	55	63	57	37	67	61
B31D	70	66	41	61	70	55	65	67	63	72	42	71	52	63	57	32	67	60
E30D	51	47	42	47	50	41	43	45	50	52	50	50	47	46	42	50	48	47
E31D	52	49	44	48	50	40	44	44	50	53	50	48	47	46	43	51	47	47
E32D	63	60	52	57	65	51	56	57	61	64	58	65	54	58	48	54	58	58
F30D	36	35	24	28	35	28	25	25	41	34	27	37	31	31	26	33	27	31
K10D	29	29	29	25	32	26	24	23	38	30	24	26	34	31	31	30	27	29
P30D	44	39	22	33	40	25	33	37	35	41	30	47	22	36	31	18	32	33
Q10D	34	34	44	41	40	37	39	37	51	38	52	33	49	35	36	50	44	41

Table 7: CS(5), in percentage, by ethnic proxy group[5].

	ARG	BRZL	CHIN	COL	DF	DOMR	ECUA	GUAT	IND	ISRL	KOR	PERU	PHIL	POL	RUS	TWAN	VENZ	Mean
B30D	4.1	4.5	6.8	4.8	4.1	5.3	4.6	4.3	5	3.8	7.5	4.2	5.7	4.7	5.3	8.8	4.3	5.2
B31D	4.1	4.6	7.3	5	3.9	5.5	4.8	4.3	4.8	3.9	8.1	4	6.1	4.6	5.4	9.6	4.5	5.3
E30D	6.8	7.2	8	7.6	7.1	8.5	8.1	7.9	7.1	6.2	7.2	7	7.1	7.3	7.9	7.2	7.3	7.4
E31D	6.7	7	7.7	7.7	7.1	8.7	7.9	7.9	7.1	6	7	6.9	7.1	7.3	7.8	6.8	7.3	7.3
E32D	4.9	5.3	6	5.8	4.7	6.3	6	5.6	5.3	4.7	5.7	4.9	5.7	5.4	6.5	6.1	5.4	5.5
F30D	10	10.1	12.9	12	9.8	11.1	12.9	12.3	8.9	10.4	12.1	9.6	10.4	10.9	11.7	11.1	12.4	11.1
K10D	10.8	10.9	10.1	12.6	10.8	11.6	12.9	13.4	8.9	11.2	12	11.2	9.7	10.2	10.2	10.5	12	11.1
P30D	8	9.1	12.7	10.2	7.9	11.8	9.9	9.7	9.9	8.2	10.9	8.1	12.2	9.6	10.1	14	9.9	10.1
Q10D	10	9.4	7.6	9.2	10.3	9.5	9.9	10.3	8.2	9.3	6.5	9	6.9	10.1	10	6.8	9.2	9

Table 8: MAE, in years, by ethnic proxy group[5].

[5] DF is Mexico City, and DOMR is the Dominican Republic.

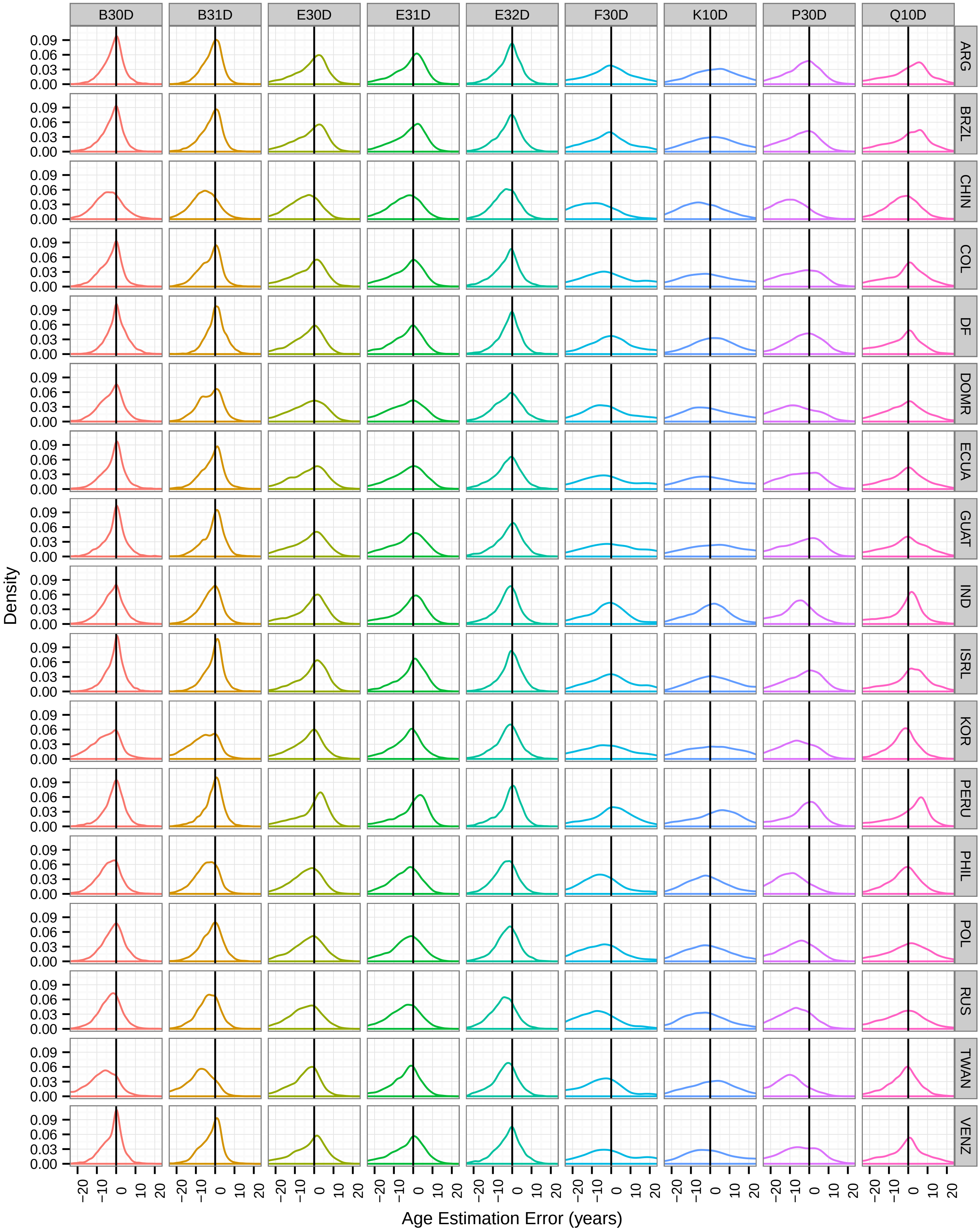


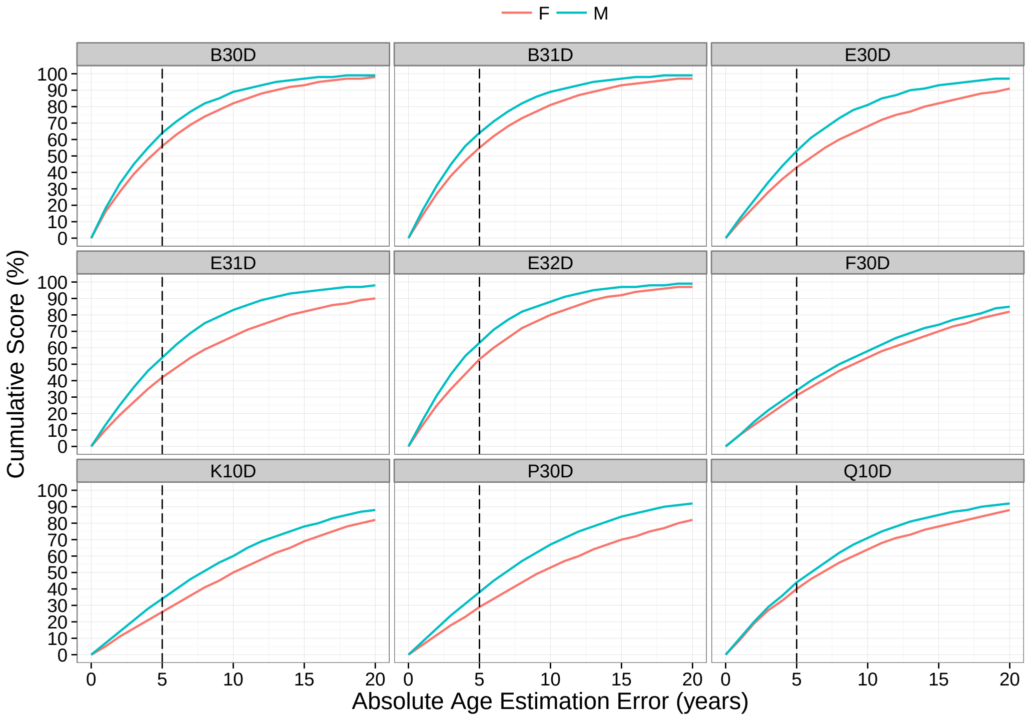
Figure 4: Density plots showing age estimation error across various ethnic proxies[5]*. Plots were generated for ethnic proxy groups with at least 2,800 images.*

Results and notable observations:

- The average of the CS(5) and MAE across all of the ethnic proxy groups is shown in the last column of Tables 7 and 8, respectively. Compared to the performance results over an ethnically homogeneous population (Figure 1), a majority of the algorithms demonstrate a lower mean accuracy and higher average MAE on this heterogeneous dataset, which suggests that ethnicity has an impact on age estimation. It is interesting to note that E32D and Q10D appear to have nearly consistent performance results between the two datasets.
- Observable bias and skew in error distribution can be seen across certain countries. Bias/skewing to the right, which is indicative of overestimation in age, is seen in South American countries like Argentina, Brazil, and Peru. Conversely, bias/skewing to the left indicates underestimation in age, which is observed in Asian countries such as China, Korea, Taiwan, and the Philippines.
- There could be extrinsic factors driving the deviations seen in the error distributions, such as increased sun exposure in the South American countries and the popularity of plastic surgery in the east Asian countries [2], although any explanation would be anecdotal at this point in time and would require further research to solidify.

3.4 Gender

Gender is one of a number of demographic traits that drives a person's ageing pattern, both inherently and extrinsically. Gender has been known to impact age estimation, and methods to address age estimation across gender have been published in academic studies [12, 13]. The DoS/P dataset contains a large number of gender-labeled visa images with a balanced number of males and females. Accuracy between males and females is assessed and summarized in Figure 5.



(a) Cumulative score vs. Absolute age estimation error

Gender	Female	Male
Num Images	118108	124894
B30D	5.7	4.7
B31D	5.9	4.7
E30D	8.5	6.1
E31D	8.6	5.8
E32D	6.2	4.8
F30D	11.2	10.2
K10D	11.9	9.9
P30D	11.6	8.5
Q10D	9.5	8.2

(b) MAE, in years

Figure 5: Line plot showing the accuracy of algorithms at absolute error levels and table summarizing MAE by gender. Plot and table were generated with 243,023 visa images.

Results and notable observations:

- All algorithms estimate age more accurately on males than females and exhibit higher MAE in females than males.

- Figure 6 shows the fraction of age estimates that are less than the actual age, broken out by age group and gender. A notable observation is that much of the difference in error between male and female occurs in the adult (18-55) age group, with females having higher understimation ratios. This could be attributed to the common use of cosmetics in young and middle-aged adult females, although any explanation would be anecdotal and requires further research to objectively solidify.

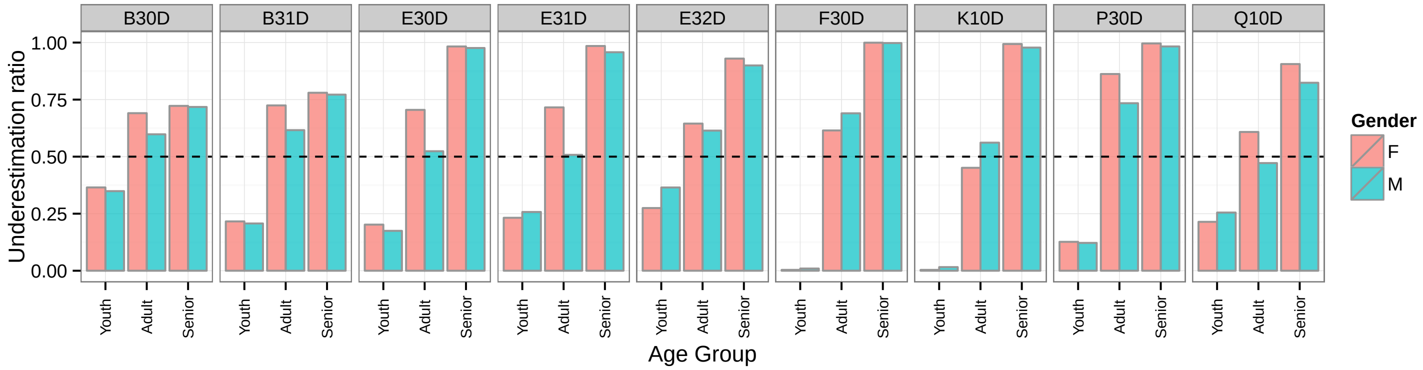
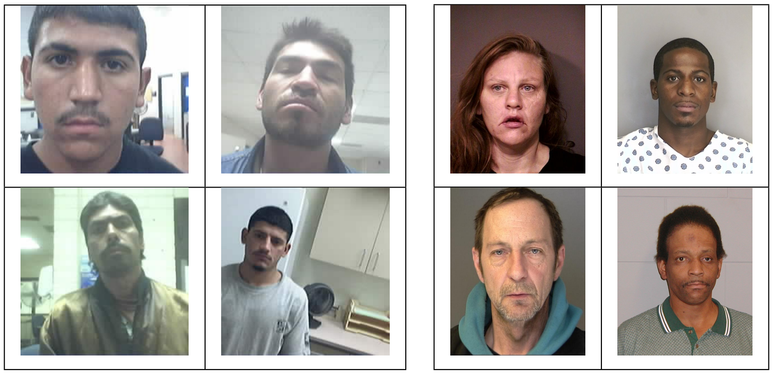


Figure 6: Bar plots showing the fraction of age estimates that are less than the actual age by gender between youths (age 0-17), adults (age 18-55), and seniors (age 56-99). Plots were generated with 243,023 images over a heterogeneous population.

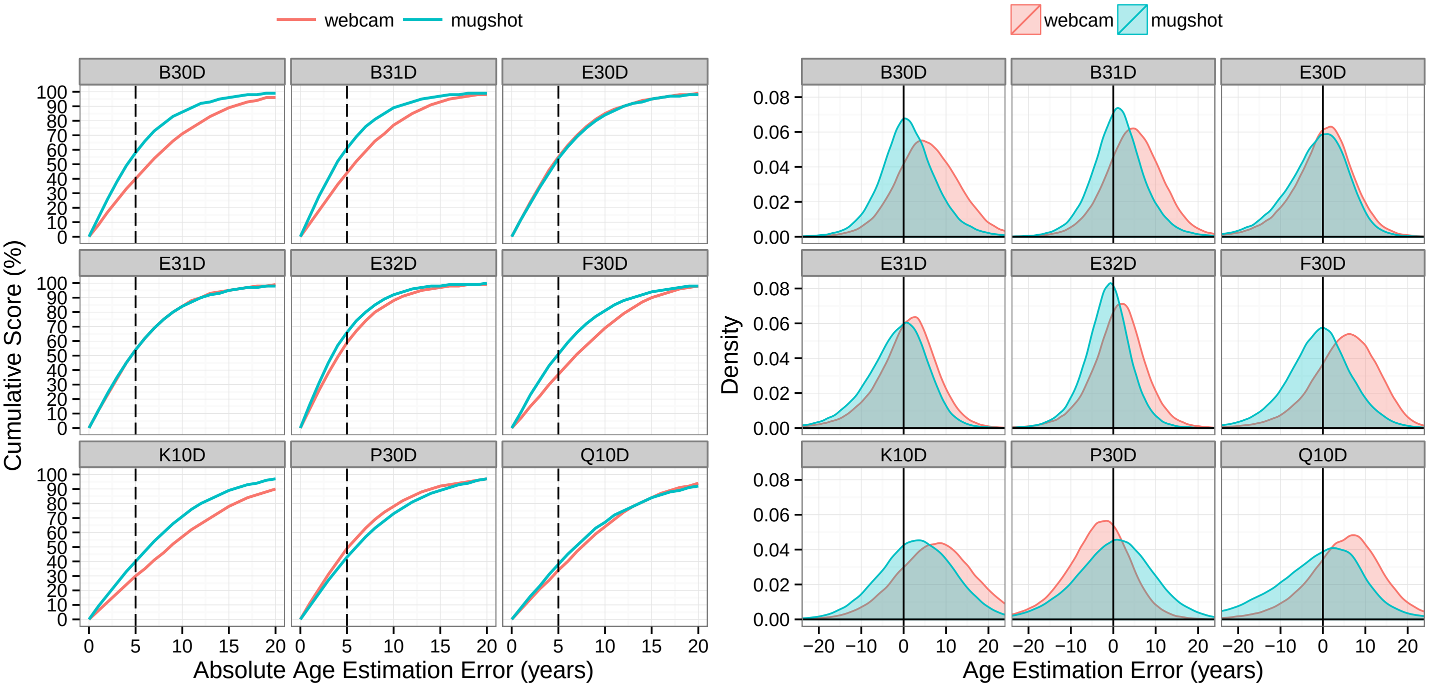
3.5 Image Quality

The LEO dataset is a large database composed of images collected during law enforcement booking. A large portion of the images are traditional mugshot photos captured under controlled lighting, background, and pose conditions. However, there is a subset of low resolution webcam images captured under moderately to less controlled environment, with large variations in illumination, contrast, and frontal pose. Webcams are commonly used in remote internet-based applications from a laptop or desktop computer, and webcam images may have unique value in virtual age verification for certain websites and potentially webcam-enabled banking ATMs. Guidelines for quality aspects of facial images are documented in ANSI/NIST-ITL 1-2011 [4], Annex E. For this particular study, image quality is broadly defined by image size, resolution, and contrast and subject illumination and pose. Leveraging the LEO database allows for the extraction of a fair number of images used in the analysis of age estimation accuracy versus image quality. 190,852 poor quality webcam images of size 240x240 and 157,906 better quality mugshot images of size 768x960 were used in this analysis. Table 7 contains sample photos extracted from the publicly available Multiple Encounter Deceased Subject (MEDS) Database [3], which is representative of image qualities from the LEO dataset. Figure 8 shows the accuracy and error distribution of the algorithms on the two different types of images.



(a) Webcam *(b) Mugshot*

Figure 7: Examples of webcam and mugshot images from the MEDS dataset.



(a) Cumulative score vs. Absolute age estimation error

(b) Density of age estimation error distribution

Figure 8: Line plot of accuracy and density plot of age estimation error between webcam and mugshot images.

Results and notable observations:

- Six out of nine algorithms demonstrate lower accuracy with poor quality webcam images.
- The impact of image quality on accuracy varies significantly between algorithms. B30D exhibits an 18% difference in 5-year accuracy while E30D and E31D show no discernible differences between webcam and mugshot images. P30D exhibits slightly higher accuracy in webcam over mugshot images.
- There is an observable bias in the error distribution to the right exhibited by algorithms on webcam images, which suggests that overestimation is occurring. There is no clear explanation for this biasing, although it may anecdotally be related to the demographics of the subjects captured in the webcam images.

3.6 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 age estimation 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 [11] supports multiple still image input to the algorithm software for age estimation, which enables the analysis of age estimation performance versus the number of image samples of the same person. The LEO 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 modelling of a scenario where age estimation implementations can exploit multiple images. 11,920 subjects with at least four contemporaneous mugshot images were extracted. The subjects' age ranged from 18-74, with a large distribution falling between age 18-42. Figure 9 shows the effects of the number of image samples on MAE. Note that some algorithms did not support processing $K > 1$ images and are not included in Figure 9.

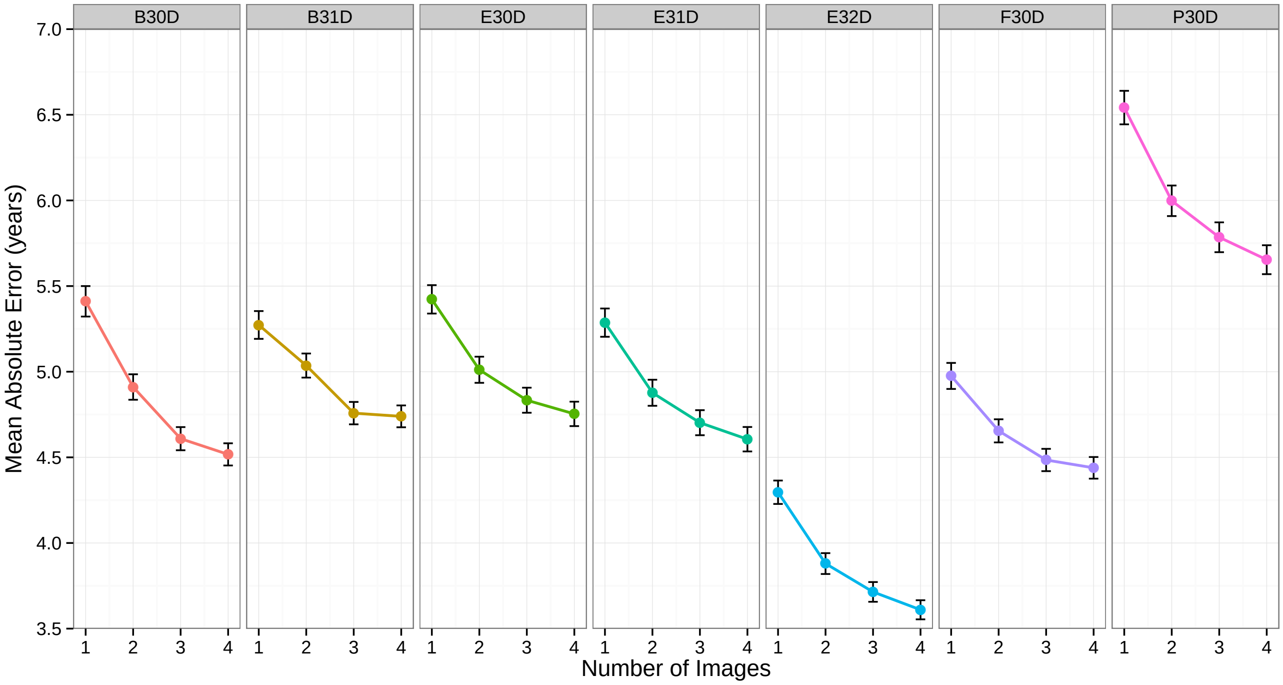


Figure 9: Line plots showing MAE vs. the number of image samples per subject and the bootstrap 95% confidence interval around the MAE. Plots were generated with 11,920 subjects.

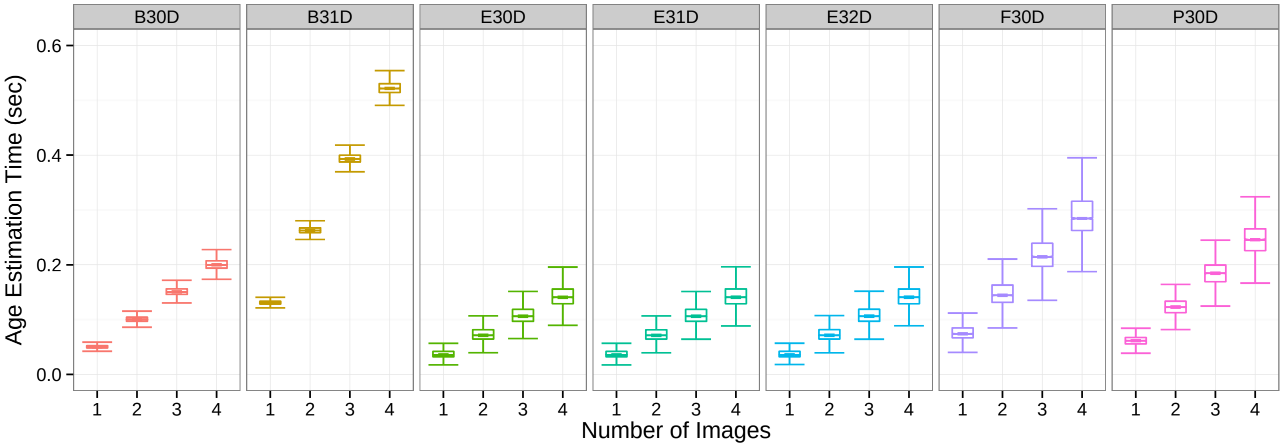


Figure 10: Boxplots summarizing age estimation time vs. the number of image samples per subject. Plots were generated with 11,920 subjects.

Results and notable observations:

- Figure 9 shows the MAE monotonically decreasing as the number of image samples increases for all algorithms. There is a decrease in MAE of 10-20% (˜1 year) between one and four images, depending on the algorithm. The separation between confidence intervals (not overlapping) between one, two, and three image samples demonstrates statistical significance and stability in the observed decrease in MAE. While the results documented are generated from images collected within a twelve month span, analysis with truly contemporaneous data (e.g., sequential frames from video) may further solidify the trends observed.
- It is also interesting to observe that the improvement in MAE starts to decrease as the number of image samples increases, and overlapping of the confidence intervals starts to occur between three and four images, which could be attributed to the law of dimishing returns.

- Figure 10 shows age estimation times increasing linearly with respect to the number of image samples, as expected.

3.7 FG-NET

The FG-NET Aging Database [1] is a publicly available dataset that is widely used in academia for age estimation performance benchmarking. It contains 1,002 images for 82 subjects ranging from age 0-69, although over 50% of the images are between age 0-13. Given published results from academia are publicly available for this dataset, it would be of interest to conduct a performance comparison against the FRVT participants, many of which are commercial algorithms. Table 9 tabulates the performance of the FRVT participants against the published methods for automatic age estimation from academic literature.

Publication	MAE (years)	CS(5)
Luu et al. [18]	4.1	73%
Chao et al. [7]	4.4	NA
Chang et al. [6]	4.5	74.7%
Han et al. [15]	4.6	74.8%
Choi et al. [8]	4.7	73%
Guo et al. [14]	4.8	47%
Wu et al. [23]	5.9	62%
Suo et al. [19]	6.0	55%
Thukral et al. [20]	6.2	NA
Geng et al. [9]	6.8	65%

(a) Published methods [15], using LOPO testing protocol

Algorithm	MAE (years)	CS(5)
B30D	7.7	59%
B31D	8.6	51%
E30D	6.9	48%
E31D	6.5	53%
E32D	5.8	56%
F30D	15.9	16%
K10D	23.8	8%
P30D	7.3	44%
Q10D	15.1	23%

(b) FRVT participants, using lights-out, black-box testing protocol

Table 9: Tables summarizing MAE and CS(5) on FG-NET.

Results and notable observations:

- Many of the academic methods performed better than the FRVT participants in terms of MAE and CS(5).
- An important point to be made is that all of the published methods in Table 9a were tested through a Leave-One-Person-Out (LOPO) protocol when running their algorithms on FG-NET, which allowed the implementation to train on a subset of the images through every iteration of testing. For the participant results documented in this report, 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. Given the fundamental differences in testing approaches, a fair benchmark comparison against academic methods cannot be performed for the FG-NET dataset in this report.

3.8 Specific Applications

3.8.1 Age Verification

Many age estimation systems strive to classify an individual's age or age range to verify that the person meets specific age requirements. Common examples in the United States include the purchasing of alcohol or admission into a casino, which has a minimum age requirement of 21. Figure 11 shows the probability of algorithms classifying a person at a particular age as equal or above age 21. It can be observed that a person at age 17 has anywhere from a 29% to 85% chance of passing for age 21 or above, depending on the algorithm. The most accurate algorithm (B30D) achieves the lowest false verification percentage of 29% at age 17. This percentage increases as a person gets closer to 21. System thresholds can be set based on the probabilities observed and the costs associated with errors in age verification.

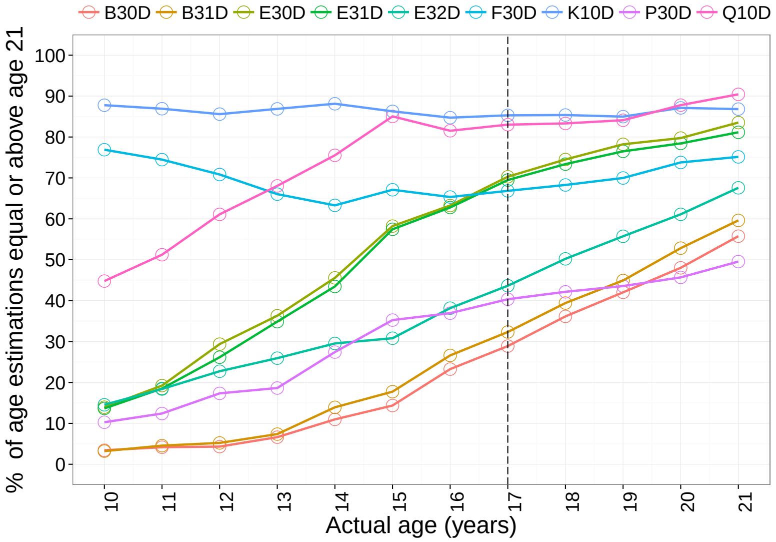
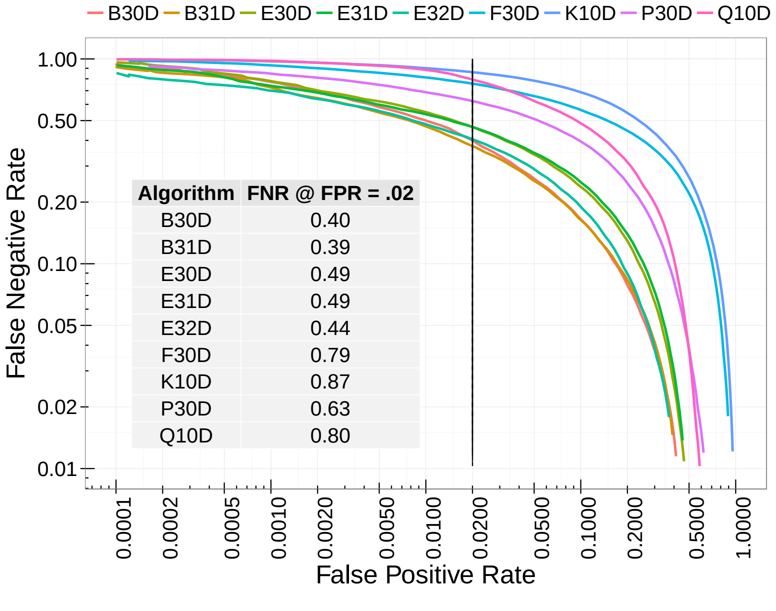


Figure 11: Line plot showing the probability of algorithms estimating that a person at a particular age is equal or above age 21. Plot was generated over a heterogeneous population of 243,023 images.



Algorithm	FNR @ FPR = .02
B30D	0.40
B31D	0.39
E30D	0.49
E31D	0.49
E32D	0.44
F30D	0.79
K10D	0.87
P30D	0.63
Q10D	0.80

Figure 12: DET curve plotting false negative rate against false positive rate for a verification age of 21. Plot was generated over a heterogeneous population of 243,023 images.

Consider an age estimation system that screens for underage individuals attempting to purchase alcohol. One might set the cost of a false negative, in this case, misclassifying someone over the age of 21 as under, to the inconvenience incurred by having to show proof of their age. The cost of a false positive, i.e., allowing a person under the age of 21 to purchase alcohol, could be the incurrence of a fine or even serving time in jail. Given it would be reasonable to argue that the costs are asymmetric in this scenario, i.e., the cost of a false positive is much greater than that of a false negative, tighter confidence levels could be set to minimize false positives. Figure 12 presents DET accuracy for a verification age of 21. A false positive rate of 0.02 would impose a false negative rate of 0.39 for the most accurate algorithm (B31D) at that threshold. If a system were to ensure that 98% of underage individuals are detected and not allowed to purchase alcohol or enter a casino, it would impose on 39% of people who are actually above 21 to be asked for proof of age, unnecessarily.

The same analysis can be done for other verification ages of interest. The legal age for purchasing alcohol and tobacco products in some places such as Hong Kong and the United Kingdom is 18. There are cigarette vending machines in Japan that have age-verification systems in place to verify that a person is at least 20 based on facial images captured prior to allowing them to make a purchase. Certain airlines have restrictions and extra fees for unaccompanied minors traveling under the age of 14, so the ability to detect minors under age 14 who appear to be traveling alone may be of interest to airports. The age verification probability and DET plots for these other scenarios are provided in Appendix A.2 for reference.

3.8.2 Children

There are age estimation applications that are specific to children. For example, age-invariant identity verification [10,24], which involves age regression or progression to predict how the subject looked like in the past or will look like in the future, certainly contains an age estimation component, and can support law enforcement in finding missing children. Missing children recovered from human trafficking or abduction may not know their own age, and automated age estimation could potentially aid in the investigation process. Tables 10 and 11 present the percentage of age estimates accurate to within one year and MAE for children ages 0-14 from an ethnically-homogeneous population extracted from the DoS/Natural dataset.

Age	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
0	133813	85	83	60	65	68	0	0	3	0
1	94610	74	34	29	29	33	0	0	5	62
2	82846	63	27	24	25	29	0	0	6	44
3	82872	50	33	21	21	27	0	0	7	17
4	84510	44	34	19	19	24	0	0	7	13
5	92982	39	33	17	17	20	0	0	8	14
6	92143	37	34	15	14	16	0	0	8	28
7	90537	36	29	12	11	12	0	0	8	25
8	89960	28	25	10	10	9	0	0	9	9
9	90182	23	21	9	8	7	0	0	10	6
10	98834	24	22	7	7	6	0	0	10	2
11	97996	24	23	6	6	4	0	0	10	2
12	94830	33	18	5	5	3	0	1	10	2
13	94545	38	12	4	5	3	0	1	10	1
14	97467	24	18	4	4	3	0	2	11	0

Table 10: CS(1), in percentage, for children by age.

Age	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
0	133813	2.4	2.0	1.8	1.7	1.6	28.3	33.7	10.9	3.3
1	94610	1.4	2.0	2.3	2.2	2.3	27.0	31.1	7.2	2.7
2	82846	1.5	2.1	3.0	3.0	2.9	25.3	29.4	6.7	4.3
3	82872	1.7	2.2	3.5	3.6	3.3	24.0	28.2	6.4	6.0
4	84510	1.8	2.2	4.1	4.3	3.8	22.8	27.3	6.2	7.4
5	92982	1.9	2.3	4.5	4.7	4.4	21.5	26.5	6.0	9.1
6	92143	2.1	2.5	5.1	5.3	5.1	20.4	25.1	5.9	10.4
7	90537	2.3	2.8	5.8	6.1	6.1	19.3	23.1	5.8	12.2
8	89960	2.6	3.1	6.3	6.6	6.7	18.4	21.4	5.6	13.8
9	90182	2.8	3.2	6.7	6.9	7.2	17.5	19.9	5.4	14.9
10	98834	2.9	3.3	7.0	7.3	7.6	16.6	18.6	5.3	15.7
11	97996	2.9	3.4	7.2	7.4	7.7	15.5	17.1	5.3	15.7
12	94830	2.9	3.6	7.2	7.3	7.6	14.3	15.8	5.3	15.4
13	94545	3.1	3.8	7.1	7.1	7.1	12.8	14.5	5.2	14.9
14	97467	3.6	4.1	6.8	6.7	6.5	11.4	13.2	5.0	14.1

Table 11: MAE, in years, for children by age.

3.8.3 Time Interval Estimation

The ability to accurately determine the amount of time that has passed between two images of the same subject has potential application in law enforcement investigations. 473,923 pairs of same-subject images with an age delta between 0-4 years were used in the analysis of using age estimation technology in time interval determination. Figure 13 presents the age estimation error for the first image plotted against the error associated with the second image, over all pairs of images, which supports correlation analysis between the age estimation errors.

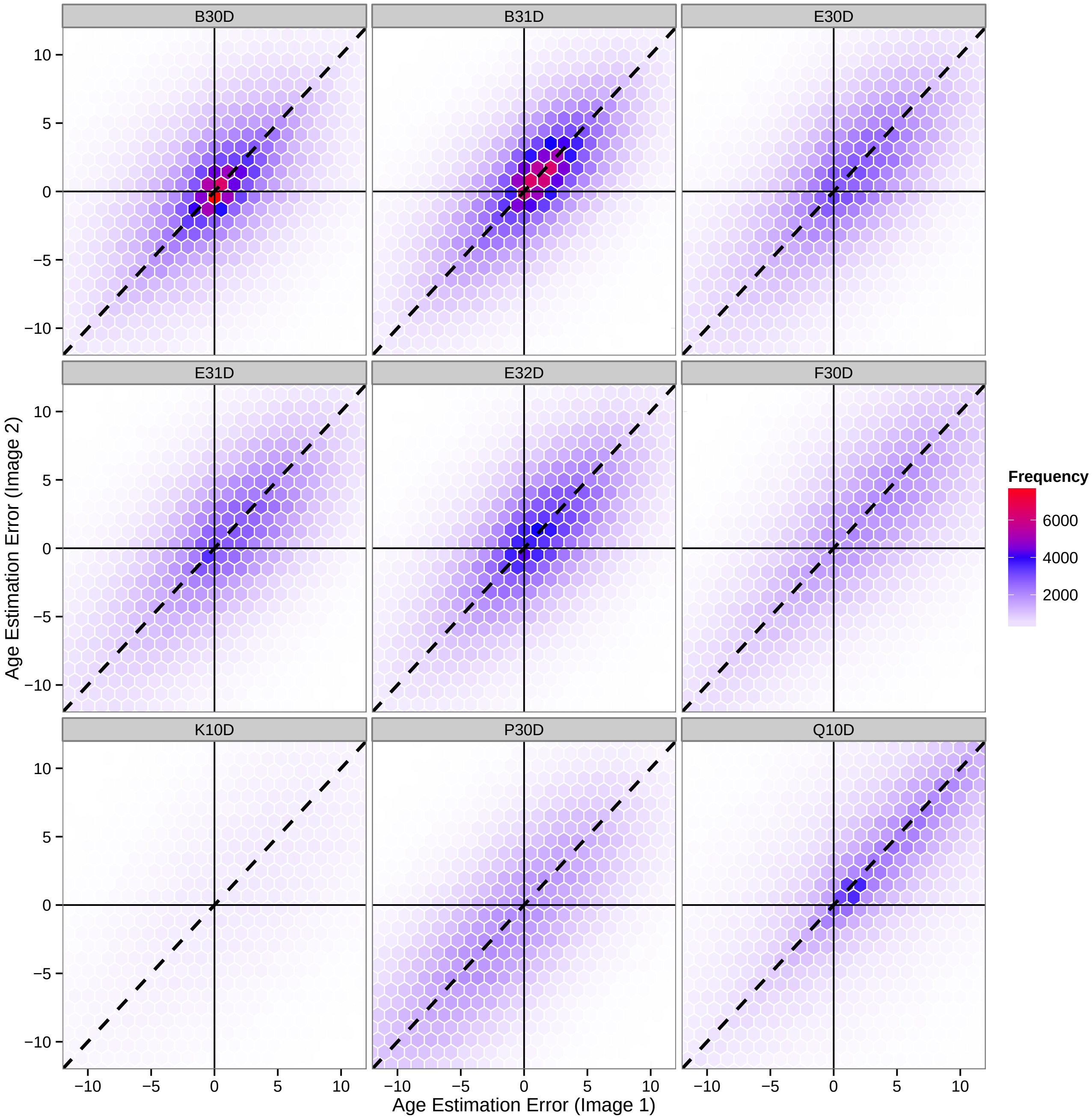


Figure 13: Hexbin plots showing the correlation between age estimation error for image 1 vs. image 2. Plots were generated with 473,923 pairs of same-subject images. The dashed line annotates where estimation error for image 1 and image 2 are equal.

Given the standard deviation of the age estimation error for image 1, σ_x, and image 2, σ_y, the correlation coefficient, ρ, is defined as the covariance, σ_{xy}, divided by the product of σ_x and σ_y. i.e.,

$$\begin{aligned} \rho &= \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \\ \sigma_{xy} &= \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{N}, \end{aligned} \tag{6}$$

where x_k is the age estimation error for the k-th first image, y_k is the error for the k-th second image, $\bar{x}$ is the mean error across all first images, $\bar{y}$ is the mean error across all second images, and N is the number of image pairs.

For the dataset used, the maximum time interval between any given image pair is less than four years, so for simplicity, $\sigma_x = \sigma_y = \sigma$. The standard deviation of the difference of x and y can be related to their individual standard deviations and the correlation coefficient between them. i.e.,

$$\sigma_{x-y} = \sqrt{2\sigma^2(1-\rho)} \tag{7}$$

Table 12 presents the standard deviation and correlation coefficient of the age estimation errors along with the standard deviation associated with time interval estimation error based on two independent age estimates of the same subject at different ages.

Algorithm	σ (years)	ρ	σ_{x-y} (years)
B30D	6.2	0.55	5.9
B31D	5.8	0.65	4.8
E30D	8.5	0.76	5.9
E31D	8.5	0.76	5.9
E32D	6.7	0.68	5.4
F30D	10.9	0.86	5.8
K10D	11.8	0.82	7.1
P30D	9.2	0.81	5.7
Q10D	10.1	0.71	7.7

Table 12: Table summarizing the standard deviation of time interval estimation error derived from the age estimation error and correlation coefficient between independent age estimates from 473,923 pairs of same-subject images.

As observed from Figure 13 and Table 12, there is a positive linear correlation between the error of two independent age estimates of the same subject at different ages, with varying degrees of correlation between the algorithms. The algorithm with the lowest time interval estimation error (i.e. B31D), produces a standard deviation in error of 4.8 years. Some of the algorithms that exhibit large age estimation error also demonstrate high correlation in error, which indicates consistency in overestimation or underestimation, resulting in relatively lower time interval estimation error.

References

[1] FG-NET Aging Database. http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html.

[2] International Society of Aesthetic Plastic Surgery (ISAPS) International Survey on Aesthetic/Cosmetic Procedures Performed in 2011. http://www.isaps.org/Media/Default/global-statistics/ISAPS-Results-Procedures-2011.pdf.

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

[4] NIST Special Publication 500-290, ANSI/NIST-ITL 1-2011, Data Format for the Interchange of Fingerprint, Facial and Other Biometric Information. http://www.nist.gov/itl/iad/ig/ansi_standard.cfm.

[5] K. Alkass, B. A. Buchholz, S. Ohtani, T. Yamamoto, H. Druid, and K. L. Spalding. Age estimation in forensic sciences: application of combined aspartic acid racemization and radiocarbon analysis. *Mol Cell Proteomics*, 9(5):1022–1030, 2010.

[6] K. Y. Chang, C. S. Chen, and Y. P. Hung. Ordinal hyperplans ranker with cost sensitivities for age estimation. In *Proc. IEEE CVPR*, pages 585–592, 2011.

[7] W. L. Chao, J. Z. Liu, and J. J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[8] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim. Age estimation using a hierarchical classifer based on global and local facial features. *Pattern Recognition*, 44(6):1262–1281, 2011.

[9] X. Geng, Z. H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12):2234–2240, 2007.

[10] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Proc. ICCV*, pages 1–8, 2013.

[11] 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.

[12] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG'10)*, 2010.

[13] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. S. Huang. A study on automatic age estimation using a large database. In *Proc. ICCV*, pages 1986–1991, 2009.

[14] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *In Proc. IEEE CVPR*, pages 112–119, 2009.

[15] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *Proc. of IAPR ICB*, pages 1–8, 2013.

[16] 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.

[17] K. Ricanek Jr., C. Chen Y. Wang, and S. J. Simmons. Generalized multi-ethnic face age-estimation. In *Third IEEE International Conference on Biometrics*, pages 127–132, 2009.

[18] K. Luu, K. Seshadri, M. Savvides, T. Bui, and C. Suen. Contourlet appearance model for facial age estimation. In *Proc. IJCB*, pages 1–8, 2011.

[19] J. Suo, S. C. Zhu, and X. Chen. A compositional and dynamic model for face aging. *IEEE Trans. PAMI*, 32(3):385–401, 2010.

[20] P. Thukral, K. Mitra, and R. Chellappa. A hierarchical approach for human age estimation. In *Proc. IEEE ICASSP*, pages 1529–1532, 2012.

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

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

[23] T. Wu, P. Turaga, and R. Chellappa. Age estimation and face verification across aging using landmarks. *IEEE Trans. IFS*, 7(6):1780–1788, 2012.

[24] F. J. Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen. Investigating age invariant face recognition based on periocular biometrics. In *Proc. IJCB*, pages 1–7, 2011.

Appendix A Additional Figures and Tables

Appendix A contains supplementary figures and tables for all age estimation algorithms.

A.1 Age Estimation Performance by Age

Age	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
0	133690	93 / 2.4	95 / 2	92 / 1.8	93 / 1.7	93 / 1.6	0 / 28.3	0 / 33.7	20 / 10.9	92 / 3.3
1	94406	97 / 1.4	97 / 2	90 / 2.3	90 / 2.2	89 / 2.3	0 / 27	0 / 31.1	35 / 7.2	90 / 2.7
2	82556	97 / 1.5	96 / 2.1	84 / 3	84 / 3	85 / 2.9	0 / 25.3	0 / 29.4	38 / 6.7	85 / 4.3
3	82481	97 / 1.7	95 / 2.2	79 / 3.5	78 / 3.6	81 / 3.3	0 / 24	0 / 28.2	41 / 6.4	78 / 6
4	84111	97 / 1.8	94 / 2.2	73 / 4.1	72 / 4.3	76 / 3.8	0 / 22.8	0 / 27.3	43 / 6.2	71 / 7.3
5	92524	96 / 1.9	91 / 2.3	65 / 4.5	64 / 4.7	69 / 4.4	0 / 21.5	0 / 26.4	46 / 6	62 / 9.1
6	91634	94 / 2.1	89 / 2.5	60 / 5.1	58 / 5.3	61 / 5.1	0 / 20.4	0 / 25.1	47 / 5.9	52 / 10.4
7	90004	91 / 2.3	86 / 2.8	53 / 5.8	50 / 6.1	51 / 6.1	0 / 19.3	0 / 23.1	47 / 5.8	40 / 12.2
8	89419	90 / 2.6	83 / 3.1	47 / 6.3	45 / 6.6	44 / 6.7	0 / 18.4	1 / 21.4	49 / 5.6	31 / 13.8
9	89652	90 / 2.8	79 / 3.2	43 / 6.7	41 / 6.9	38 / 7.2	0 / 17.5	1 / 19.8	52 / 5.4	22 / 14.9
10	98327	88 / 2.9	82 / 3.3	38 / 7	36 / 7.3	32 / 7.6	0 / 16.6	2 / 18.6	53 / 5.3	15 / 15.7
11	97440	84 / 2.9	82 / 3.4	35 / 7.2	33 / 7.4	29 / 7.7	2 / 15.5	3 / 17.1	53 / 5.3	10 / 15.7
12	94317	80 / 2.9	77 / 3.6	33 / 7.2	32 / 7.3	27 / 7.6	4 / 14.3	5 / 15.8	53 / 5.3	7 / 15.4
13	94070	77 / 3.1	71 / 3.8	32 / 7.1	32 / 7.1	30 / 7.1	8 / 12.8	8 / 14.5	54 / 5.2	6 / 14.9
14	96940	73 / 3.6	66 / 4.1	33 / 6.8	34 / 6.7	36 / 6.5	13 / 11.4	11 / 13.2	56 / 5.1	5 / 14.1
15	97522	71 / 3.9	64 / 4.3	37 / 6.5	37 / 6.3	44 / 5.8	19 / 10.2	15 / 12.1	60 / 4.7	6 / 13.3
16	96440	71 / 4	65 / 4.4	42 / 6.1	43 / 5.9	51 / 5.3	24 / 9.3	19 / 11.2	64 / 4.4	7 / 12.5
17	99469	72 / 4	65 / 4.3	48 / 5.7	49 / 5.5	58 / 4.8	29 / 8.6	22 / 10.4	69 / 4	9 / 11.7
18	98734	73 / 3.9	68 / 4.1	53 / 5.3	55 / 5.1	63 / 4.5	32 / 8.1	23 / 10.1	74 / 3.6	11 / 10.7
19	105770	69 / 3.8	72 / 3.9	58 / 4.9	60 / 4.7	68 / 4.1	35 / 7.6	25 / 9.6	77 / 3.4	13 / 10
20	118645	69 / 3.8	71 / 3.8	63 / 4.6	64 / 4.4	72 / 3.8	38 / 7.2	29 / 9	77 / 3.4	16 / 9.5
21	124101	70 / 3.8	71 / 3.8	66 / 4.3	67 / 4.2	75 / 3.5	40 / 6.9	32 / 8.4	77 / 3.4	19 / 8.9
22	128841	71 / 3.9	72 / 3.7	69 / 4.1	69 / 4.1	77 / 3.4	42 / 6.5	35 / 7.9	74 / 3.6	27 / 8.4
23	133521	72 / 4	72 / 3.7	70 / 4	70 / 4	78 / 3.4	45 / 6.2	38 / 7.4	70 / 3.9	35 / 8
24	138148	70 / 4	73 / 3.7	71 / 4	70 / 4	78 / 3.4	48 / 5.9	41 / 7	65 / 4.3	40 / 7.5
25	139873	68 / 4.2	72 / 3.8	70 / 4	69 / 4.1	76 / 3.5	49 / 5.7	45 / 6.6	59 / 4.7	43 / 7.1
26	141798	65 / 4.3	69 / 3.9	69 / 4.1	68 / 4.2	75 / 3.7	51 / 5.5	48 / 6.2	54 / 5.1	46 / 6.7
27	138453	63 / 4.5	68 / 4.1	67 / 4.3	66 / 4.3	72 / 3.8	53 / 5.2	50 / 5.9	48 / 5.6	47 / 6.4
28	133628	60 / 4.7	65 / 4.2	64 / 4.4	64 / 4.5	68 / 4.1	55 / 5	52 / 5.6	43 / 6	49 / 6.2
29	129021	59 / 4.9	62 / 4.4	61 / 4.6	61 / 4.6	65 / 4.3	57 / 4.9	54 / 5.5	39 / 6.5	45 / 6.1
30	123436	57 / 5.1	61 / 4.5	59 / 4.8	59 / 4.8	62 / 4.5	59 / 4.7	56 / 5.3	36 / 7	48 / 5.9

Table 13: CS(5), in percentage / MAE, in years, for age 0-30. Table was generated over an ethnically-homogeneous population.

Age	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
31	118117	56 / 5.3	61 / 4.7	56 / 5	57 / 4.9	60 / 4.7	62 / 4.5	57 / 5.1	33 / 7.4	50 / 5.9
32	113140	54 / 5.4	60 / 4.8	54 / 5.1	56 / 5.1	57 / 4.8	64 / 4.4	57 / 5.1	31 / 7.8	49 / 5.8
33	109142	53 / 5.5	59 / 4.9	53 / 5.3	54 / 5.2	56 / 5	66 / 4.2	58 / 5	30 / 8.2	47 / 5.8
34	105909	53 / 5.6	58 / 5	51 / 5.4	53 / 5.3	54 / 5.1	68 / 4.2	58 / 4.9	28 / 8.5	46 / 5.7
35	103135	54 / 5.6	57 / 5	51 / 5.5	52 / 5.4	54 / 5.2	69 / 4.1	58 / 4.9	27 / 8.9	50 / 5.6
36	99884	54 / 5.5	58 / 5	50 / 5.6	52 / 5.5	54 / 5.2	69 / 4	58 / 4.9	27 / 9.2	50 / 5.5
37	97533	53 / 5.5	57 / 5	50 / 5.7	51 / 5.6	53 / 5.3	69 / 4.1	57 / 5	26 / 9.4	51 / 5.5
38	93916	53 / 5.5	57 / 5.1	50 / 5.8	51 / 5.7	54 / 5.3	68 / 4.1	57 / 5.1	25 / 9.8	54 / 5.5
39	91463	54 / 5.5	58 / 5	50 / 5.9	51 / 5.8	54 / 5.3	66 / 4.2	56 / 5.2	25 / 10	61 / 5.4
40	94042	55 / 5.5	58 / 5	51 / 5.9	52 / 5.8	55 / 5.3	65 / 4.3	54 / 5.4	26 / 10	62 / 5.3
41	90172	57 / 5.5	59 / 5	51 / 6	51 / 6	55 / 5.3	63 / 4.5	52 / 5.6	25 / 10.3	57 / 5.4
42	86397	58 / 5.5	59 / 5.1	50 / 6.2	50 / 6.2	56 / 5.3	61 / 4.7	50 / 5.8	25 / 10.5	57 / 5.5
43	83247	57 / 5.5	60 / 5	49 / 6.3	49 / 6.3	56 / 5.3	58 / 4.9	48 / 6	25 / 10.7	56 / 5.7
44	80605	55 / 5.4	59 / 5	48 / 6.5	48 / 6.6	57 / 5.3	54 / 5.2	46 / 6.3	24 / 10.9	55 / 5.9
45	77477	55 / 5.4	59 / 5	48 / 6.7	47 / 6.8	57 / 5.2	51 / 5.4	44 / 6.5	24 / 11.1	52 / 6.2
46	74183	56 / 5.3	59 / 4.9	46 / 6.9	45 / 7.1	58 / 5.2	48 / 5.7	41 / 6.9	24 / 11.3	50 / 6.5
47	70689	57 / 5.4	59 / 4.8	45 / 7.2	42 / 7.3	58 / 5.2	46 / 6	39 / 7.1	23 / 11.4	48 / 6.8
48	67881	57 / 5.4	59 / 4.8	43 / 7.4	40 / 7.6	59 / 5.2	43 / 6.3	37 / 7.5	23 / 11.6	44 / 7
49	65641	56 / 5.4	61 / 4.8	41 / 7.7	38 / 7.9	58 / 5.2	41 / 6.6	35 / 7.8	23 / 11.8	37 / 7.3
50	64889	56 / 5.4	61 / 4.8	39 / 8	36 / 8.3	59 / 5.3	39 / 6.9	32 / 8.2	23 / 11.9	34 / 7.6
51	63424	56 / 5.4	62 / 4.8	36 / 8.3	34 / 8.6	58 / 5.3	36 / 7.3	31 / 8.5	22 / 12.1	37 / 7.9
52	60969	54 / 5.4	61 / 4.8	34 / 8.7	31 / 9	57 / 5.4	34 / 7.7	28 / 8.9	22 / 12.2	33 / 8.3
53	59545	56 / 5.4	60 / 4.8	32 / 9.1	30 / 9.3	57 / 5.5	31 / 8.1	26 / 9.3	22 / 12.5	32 / 8.6
54	56893	57 / 5.4	62 / 4.7	30 / 9.4	28 / 9.6	56 / 5.6	28 / 8.5	24 / 9.7	21 / 12.7	32 / 8.9
55	55271	56 / 5.4	62 / 4.8	28 / 9.7	27 / 10	54 / 5.7	26 / 9	22 / 10.1	20 / 12.9	27 / 9.2
56	52583	56 / 5.4	62 / 4.8	26 / 10.1	25 / 10.4	53 / 5.9	23 / 9.4	20 / 10.6	19 / 13.1	26 / 9.6
57	49828	55 / 5.4	61 / 4.9	24 / 10.5	24 / 10.7	51 / 6.1	21 / 9.9	19 / 11	18 / 13.3	26 / 9.9
58	48544	55 / 5.5	60 / 4.9	23 / 10.9	23 / 11.1	48 / 6.4	17 / 10.5	17 / 11.4	17 / 13.6	25 / 10
59	46032	55 / 5.7	57 / 5	22 / 11.2	22 / 11.4	46 / 6.6	15 / 11	16 / 11.7	15 / 13.9	25 / 10.3
60	46064	53 / 5.8	57 / 5.2	20 / 11.7	21 / 11.8	44 / 6.8	12 / 11.6	15 / 12.2	14 / 14.2	24 / 10.6
61	43832	50 / 5.9	54 / 5.3	19 / 12.1	20 / 12.2	41 / 7.2	8 / 12.2	13 / 12.5	12 / 14.5	24 / 10.7
62	41420	48 / 6.1	53 / 5.5	18 / 12.5	19 / 12.6	38 / 7.5	6 / 12.8	13 / 12.9	10 / 15	26 / 10.9
63	39435	45 / 6.2	51 / 5.7	16 / 13	17 / 13.1	35 / 7.9	4 / 13.5	12 / 13.5	8 / 15.5	26 / 11.1
64	37404	44 / 6.3	48 / 5.8	15 / 13.4	16 / 13.5	32 / 8.3	2 / 14.2	11 / 13.9	7 / 15.9	24 / 11.1
65	34833	44 / 6.5	46 / 5.9	14 / 13.8	15 / 13.9	29 / 8.7	1 / 14.8	10 / 14.2	5 / 16.4	25 / 11.1
66	32009	43 / 6.6	46 / 6.1	13 / 14.2	14 / 14.2	26 / 9.1	1 / 15.5	9 / 14.6	4 / 16.8	27 / 11.1

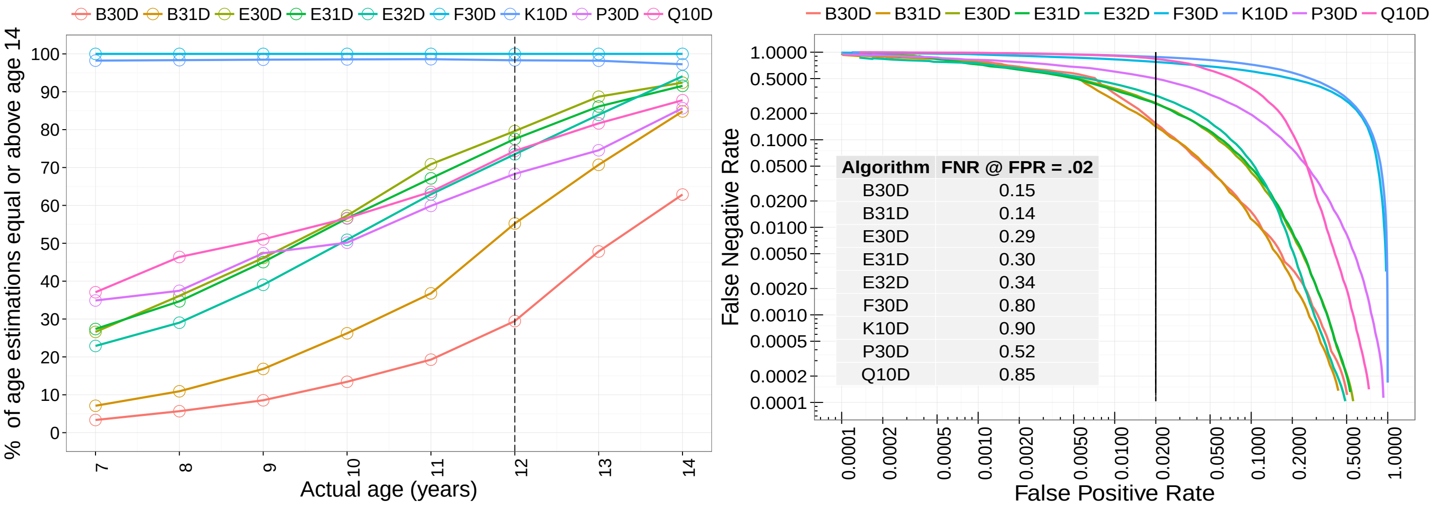
Table 14: CS(5), in percentage / MAE, in years, for age 31-66. Table was generated over an ethnically-homogeneous population.

Age	Num Images	B30D	B31D	E30D	E31D	E32D	F30D	K10D	P30D	Q10D
67	28903	43 / 6.6	45 / 6.2	12 / 14.7	13 / 14.8	23 / 9.5	0 / 16.2	8 / 15.1	3 / 17.3	29 / 11
68	26617	41 / 6.7	46 / 6.3	11 / 15.2	11 / 15.3	21 / 10	0 / 16.9	7 / 15.7	2 / 17.9	32 / 11
69	24863	40 / 6.8	44 / 6.4	10 / 15.7	10 / 15.7	18 / 10.5	0 / 17.7	6 / 16.1	1 / 18.5	36 / 10.9
70	23395	41 / 6.7	44 / 6.4	9 / 16.2	9 / 16.2	16 / 11	0 / 18.4	5 / 16.4	1 / 19.2	44 / 10.8
71	20728	41 / 6.8	44 / 6.5	8 / 16.7	8 / 16.7	15 / 11.6	0 / 19.1	4 / 17.1	0 / 19.7	45 / 10.9
72	18417	41 / 6.8	45 / 6.4	7 / 17.4	7 / 17.4	13 / 12.1	0 / 20	4 / 17.5	0 / 20.5	45 / 10.8
73	16212	44 / 6.8	45 / 6.5	6 / 18	6 / 18	12 / 12.6	0 / 20.8	4 / 18	0 / 21.3	45 / 10.9
74	14849	59 / 6.6	56 / 6.4	5 / 18.6	5 / 18.6	10 / 13.1	0 / 21.6	3 / 18.7	0 / 21.8	43 / 11.1
75	13184	57 / 6.5	55 / 6.4	4 / 19.1	4 / 19.1	9 / 13.8	0 / 22.3	3 / 19.1	0 / 22.5	42 / 11.4
76	11889	55 / 6.4	54 / 6.4	4 / 19.5	4 / 19.5	8 / 14.3	0 / 23	2 / 19.4	0 / 23.1	40 / 11.8
77	10836	53 / 6.5	52 / 6.5	3 / 20.5	3 / 20.4	7 / 15.1	0 / 23.9	2 / 20.3	0 / 24	36 / 12.5
78	9351	54 / 6.3	50 / 6.4	2 / 20.9	2 / 20.9	6 / 15.7	0 / 24.7	1 / 20.7	0 / 24.7	31 / 13
79	7551	54 / 6.4	51 / 6.5	2 / 21.7	2 / 21.8	5 / 16.3	0 / 25.5	1 / 21.4	0 / 25.4	21 / 13.9
80	6248	54 / 6.8	51 / 6.9	1 / 22.6	1 / 22.6	4 / 17.1	0 / 26.5	1 / 22.1	0 / 26.5	0 / 14.4
81	5231	53 / 7.4	49 / 7.4	1 / 23.3	1 / 23.3	4 / 17.7	0 / 27.2	1 / 22.5	0 / 27.3	0 / 15.4
82	3953	51 / 8	45 / 8.1	1 / 23.8	1 / 23.7	3 / 18.3	0 / 28.1	0 / 22.8	0 / 27.8	0 / 15.5
83	3273	46 / 8.7	42 / 8.6	0 / 24.7	0 / 24.8	2 / 19	0 / 29.1	1 / 23.4	0 / 28.7	0 / 17
84	2617	0 / 9.4	0 / 9.4	0 / 25.3	0 / 25.5	2 / 19.9	0 / 29.9	0 / 24.2	0 / 29.6	0 / 17.7
85	2201	0 / 9.9	0 / 9.9	0 / 26.5	0 / 26.6	1 / 20.7	0 / 30.8	0 / 24.9	0 / 30.6	0 / 18.7
86	1834	0 / 10.8	0 / 10.7	0 / 27.2	0 / 27.2	0 / 21.5	0 / 31.5	0 / 25.6	0 / 31.3	0 / 19.7
87	1399	0 / 11.7	0 / 11.5	0 / 27.8	0 / 27.9	0 / 22.1	0 / 32.5	0 / 25.9	0 / 32.4	0 / 20
88	1251	0 / 12.3	0 / 12	0 / 28.5	0 / 28.5	0 / 23	0 / 33.5	0 / 26.5	0 / 33.5	0 / 21.8
89	909	0 / 13	0 / 12.7	0 / 29	0 / 29.2	0 / 23.6	0 / 34.1	0 / 26.5	0 / 34.1	0 / 22.2
90	668	0 / 14.5	0 / 14.3	0 / 30.5	0 / 31	0 / 24.9	0 / 35.5	0 / 28.8	0 / 35.6	0 / 23.5
91	514	0 / 14.7	0 / 14.6	0 / 31.9	0 / 31.9	0 / 25.6	0 / 36.2	0 / 29.6	0 / 36.5	0 / 25.4
92	322	0 / 15.6	0 / 15.5	0 / 32.4	0 / 32.8	0 / 26.5	0 / 36.8	0 / 29.8	0 / 36.9	0 / 25.4
93	235	0 / 17.2	0 / 16.8	0 / 33	0 / 33.7	0 / 27.6	0 / 38	0 / 30.3	0 / 38.6	0 / 26.3
94	156	0 / 17.7	0 / 17.4	0 / 34.2	0 / 34.6	0 / 28.4	0 / 38.9	0 / 31.5	0 / 39.1	0 / 28.8
95	117	0 / 20.5	0 / 20	0 / 36.3	0 / 36.9	0 / 30.7	0 / 40.6	0 / 32.4	0 / 41.1	0 / 30.3
96	62	0 / 20.4	0 / 19.7	0 / 34.7	0 / 35	0 / 29.9	0 / 41.6	0 / 31.3	0 / 41.6	0 / 29.2
97	43	0 / 20.1	0 / 20.3	0 / 35.6	0 / 37.7	0 / 31.2	0 / 42	0 / 30.3	0 / 42.3	0 / 28.3
98	35	0 / 23.2	0 / 22.7	0 / 38.2	0 / 39.8	0 / 32.8	0 / 42.4	0 / 34.5	0 / 41.7	0 / 30.6
99	15	0 / 23.2	0 / 23.2	0 / 37.8	0 / 38	0 / 34.4	0 / 47	0 / 44.1	0 / 47.8	0 / 38.8

Table 15: CS(5), in percentage / MAE, in years, for age 67-99. Table was generated over an ethnically-homogeneous population.

A.2 Age Verification Accuracy

A.2.1 Age 14



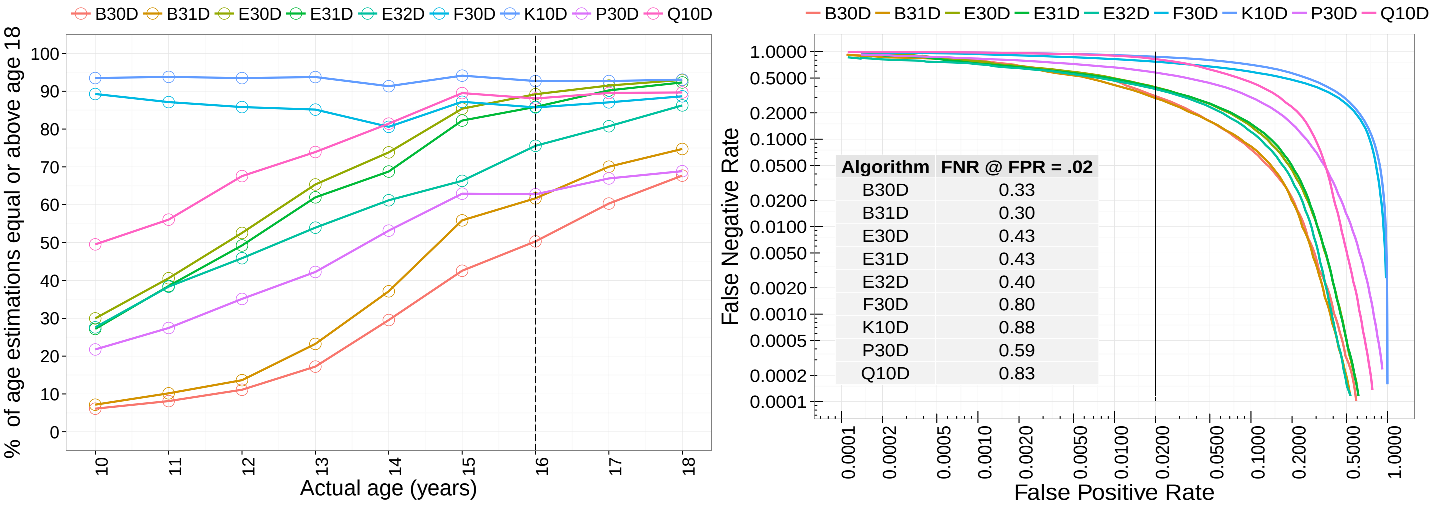
Algorithm	FNR @ FPR = .02
B30D	0.15
B31D	0.14
E30D	0.29
E31D	0.30
E32D	0.34
F30D	0.80
K10D	0.90
P30D	0.52
Q10D	0.85

(a) Probability that age estimate is equal or above age 14

(b) DET for verification age of 14

Figure 14: Line plot showing the probability of algorithms estimating that a person at a particular age is equal or above age 14 and DET curve plotting false negative rate against false positive rate for verification age of 14. The dotted line in (a) highlights the probability that a person at age 12 is estimated as equal or above age 14. Plots were generated over a heterogeneous population of 243,023 images.

A.2.2 Age 18



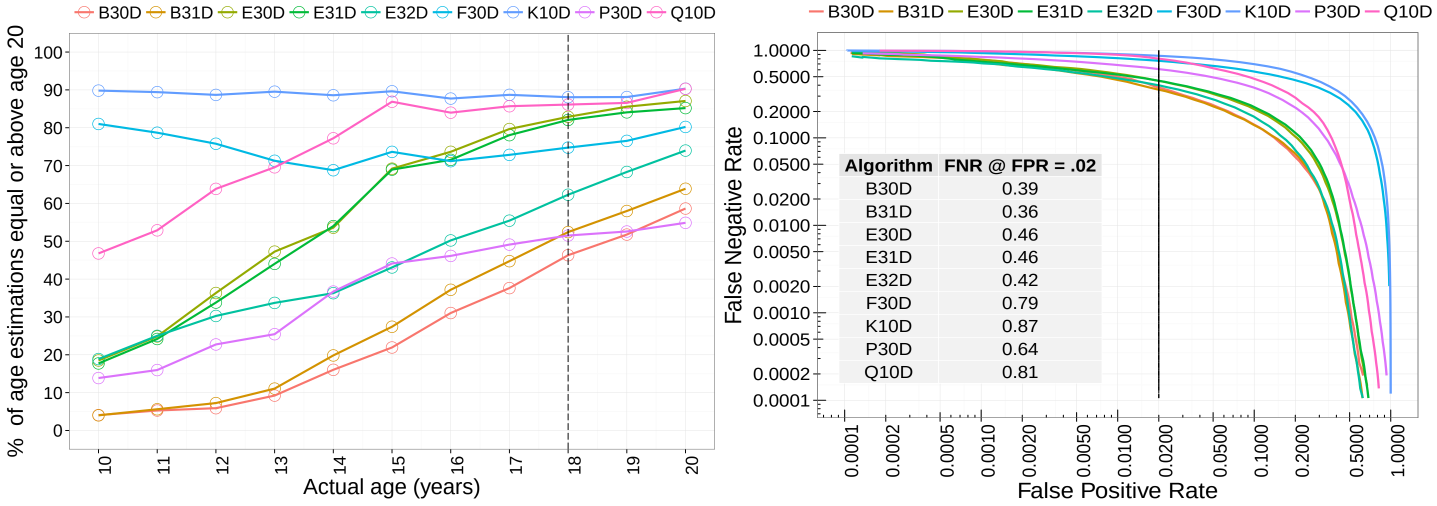
Algorithm	FNR @ FPR = .02
B30D	0.33
B31D	0.30
E30D	0.43
E31D	0.43
E32D	0.40
F30D	0.80
K10D	0.88
P30D	0.59
Q10D	0.83

(a) Probability that age estimate is equal or above age 18

(b) DET for verification age of 18

Figure 15: Line plot showing the probability of algorithms estimating that a person at a particular age is equal or above age 18 and DET curve plotting false negative rate against false positive rate for verification age of 18. The dotted line in (a) highlights the probability that a person at age 16 is estimated as equal or above age 18. Plots were generated over a heterogeneous population of 243,023 images.

A.2.3 Age 20



Algorithm	FNR @ FPR = .02
B30D	0.39
B31D	0.36
E30D	0.46
E31D	0.46
E32D	0.42
F30D	0.79
K10D	0.87
P30D	0.64
Q10D	0.81

(a) Probability that age estimate is equal or above age 20

(b) DET for verification age of 20

Figure 16: Line plot showing the probability of algorithms estimating that a person at a particular age is equal or above age 20 and DET curve plotting false negative rate against false positive rate for verification age of 20. The dotted line in (a) highlights the probability that a person at age 18 is estimated as equal or above age 20. Plots were generated over a heterogeneous population of 243,023 images.