TECHNICAL REPORT

ISO/IEC TR 22116

First edition
2021-06

Information technology — A study of the differential impact of demographic factors in biometric recognition system performance

Reference number
ISO/IEC TR 22116:2021(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Automated systems (including biometrics) are increasingly used in decision-making processes. In recent years, systemic performance differentials reflected in several automated decision systems have been reported and hotly debated. In the context of this report, an algorithm exhibiting performance differentials produces statistically different outcomes or decisions for different groups of individuals, for example, based on gender, age and race/ethnicity. In the context of biometric recognition, this means that probabilities of false positives and/or false negatives can differ among the demographic groups. The impacts of such performance differentials on the affected individuals can range from mere inconvenience in cooperative access control systems, to consequential harms such as varying arrest rates for certain demographic groups based on decisions produced by facial recognition systems.

Although such systems are almost certainly not designed to be explicitly differential against any group, implicit differences can occur independently of the intentions of the system designers. They can be exhibited and propagated at many stages of the decision-making pipeline, including but not limited to training data itself as well as the data processing. Due to the scalability of such systems, a higher quantity of erroneous or inaccurate decisions can be generated than in the typical, human-based processes. Consequently, in recent years, measuring and ensuring the fairness (i.e. lack of differential performance) of such systems has often been discussed in the media and political circles, with research and commercial interest increasing accordingly. With increasing deployments of the technology, it is important to consider whether it performs similarly for all users. This document helps to identify where recognition performance differences related to demographic factors can exist in biometric systems.

Information technology — A study of the differential impact of demographic factors in biometric recognition system performance

1 Scope

This document introduces the effects of population demographics on biometric functions. It:

— establishes terms and definitions relevant to the study of demographic factors in biometric recognition system performance;

— identifies areas where biometric systems can exhibit different performances based on different demographic factors of the individuals submitting the biometric samples;

— explains how different demographic factors can influence the biometric characteristics captured by different biometric modalities and how these influences can affect biometric performance measures;

— presents a case study on existing scientific material that explores the impact of demographic factors on biometric system performance. Only biometric modalities where quantitative information is available on the impact of demographic factors are considered.

Outside of the scope of this document are:

— effects of disease and injury on biometric performance; and

— how religious and cultural norms can affect biometric operations.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 2382-37, *Information technology — Vocabulary — Part 37: Biometrics*

3 Terms, definitions, symbols and abbreviated terms

For the purposes of this document, the terms, definitions, symbols and abbreviated terms given in ISO/IEC 2382-37 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
age
length of time an individual has lived

3.2
ageing
natural progression of an individual's characteristics over time

Note 1 to entry: The impact of ageing will vary at different ages.

3.3
categorical demographic variable
demographic characteristic that is nominally or ordinally described

EXAMPLE Gender categories consist of "Male", "Female", or "Other".

3.4
detection error trade-off
DET
relationship between false positive and false negative errors of a binary classification system as the discrimination threshold varies

3.5
differential performance
differences in system variables or system processing between different demographic groups

EXAMPLE Differences in comparison scores, feature-level fusion, and/or image-level fusion.

3.6
differential outcomes
difference in system results between different demographic groups

EXAMPLE Differences in match rate.

3.7
ethnicity
state of belonging to a group with a common origin, set of customs or traditions

Note 1 to entry: Scientifically, race can be defined as a group of humans that share biological features related to genetic ancestry. Practically, race is primarily a social construct, i.e. not related to biology but instead related to self-identity. Ethnicity is frequently used as a proxy for self-reported "race". In the context of this report, "ethnicity" and "race" are considered interchangeable and can both be taken to mean a social identity that was reported or assigned to a particular subject. These reported values can, but do not always, correlate with underlying genetic features.

3.8
failure-to-acquire
FTA
failure-to-accept for subsequent comparison the output of a biometric capture process, a biometric sample of the biometric characteristic of interest

3.9
failure-to-enrol
FTE
failure-to-create and store a biometric enrolment data record for an eligible biometric capture subject in accordance with a biometric enrolment policy

3.10
false accept rate
FAR
proportion of transactions with false biometric claims erroneously accepted

3.11
false match rate
FMR
proportion of the completed biometric non-mated comparison trials that result in a false match

3.12
false negative differential
tendency for mated biometric samples from subjects in one demographic group not to match relative to another demographic group

3.13
false non-match rate
FNMR
proportion of the completed biometric mated comparison trials that result in a false non-match

3.14
false positive differential
tendency for non-mated biometric samples from one demographic group to falsely match relative to another demographic group or a tendency for this effect to occur across demographic groups

3.15
false reject rate
FRR
proportion of biometric transactions with true biometric claims erroneously rejected

3.16
gender
classification as male, female or another category based on social, cultural, or behavioural factors

Note 1 to entry: Gender is generally determined through self-declaration or self-presentation and can change over time.

Note 2 to entry: Depending on jurisdiction recognition, it is possible that gender will or will not require assessment by a third party.

3.17
phenotypic demographic variable
observable demographic characteristic of an individual resulting from the interaction of that individual's genotype and the environment

EXAMPLE An individual's skin reflectance.

3.18
sex
state of being male or female as it relates to biological factors such as DNA, anatomy and physiology

4 Understanding demographic factors in biometric systems

4.1 Introduction

Demographic factors include any characteristics or attributes that apply to a specific group within a population[1]. There are potentially an infinite number of different demographic factors that can be considered in terms of how they impact biometric systems. In order to maintain a manageable scope for this document, the demographic factors considered are limited to those where at least some research is available to evaluate the impact on biometric comparison performance, quality score, failure-to-enrol rate or other performance metrics[1]. The state-of-the-art of biometric system performance can change rapidly, and performance can improve by an order of magnitude over the course of just a few years. Therefore, performance results observed in this report can be overcome or obviated by more recent studies. It is valuable to continuously monitor for publications that provide new insights into aspects of differential performance. Specifically, this document considers the following demographic factors that have been shown to impact the performance of biometric systems (see Clause 3 for definitions):

— Age and ageing

— Ethnicity

— Gender and sex

The concept of differential performance is closely related with the long-studied biometric concept of the biometric menagerie[2] which divides individuals into individual "animal" categories based on the profile of their mated and non-mated score distributions.

4.2 Biometric system components

The functions of the generalized components of a biometric system, applied in ISO/IEC 19795-1 and ISO/IEC 2382-37, can be affected by the demographics of the target population. Those functions are:

— **Acquisition:** Some applications detect, localize, and acquire biometric characteristics, for example counting faces in a crowded area. Acquisition accuracy can depend on demographics. Some systems can include biometric sample quality estimation in the acquisition process, which can also be affected by the demographics of the target population. A failure-to-acquire (FTA) can also impede downstream biometric processes.

— **Enrolment:** For certain individuals it can be difficult enrolling in a biometric system. This can be caused by sensor or system properties, or by innate characteristics of these individuals, including characteristics associated with demographics. In such cases, the failure-to-enrol (FTE) rate can vary by demographic group.

— **Verification:** In verification processes such as access control, accuracy (false rejection rates or false acceptance rates) can vary. One potential reason for such a variation are demographic effects. In addition, biometric performance can be affected by failure-to-acquire rates which can deviate for different demographic groups.

— **Identification:** In identification systems demographics can affect both false negative and false positive identification rates. Unlike verification applications, demographic effects in identification can consider the demographics of the probe and the gallery biometric samples. Again, biometric performance can be affected by failure-to-acquire rates influenced by demographic properties.

4.3 The influence of demographics on biometric recognition

4.3.1 The influence of sex and gender

Biometric comparison performance, enrolment success and other aspects of performance can be affected by sex and by gender. As explained in Clause 3, these two terms are distinct. For the purposes of statistical analysis when correlating demographic factors with biometric performance, gender is usually recorded based on the gender (recorded as "sex" on most government issued documents; male or female or undefined) listed on the identity documents belonging to an individual.

The broadly generalizable influence of gender on biometric performance can be difficult to characterize as practices associated with gender vary. For example, manual labour can result in friction ridge degradation over time that can impact all phases of biometric operations. However, these activities are associated with different gender identities in different geographic locations. Since manual labour can result in degradation of the quality of fingerprints and lower fingerprint comparison performance, this is an example of a demographic factor that is related to gender but can be more significant for males in some cultures and for females in others.

4.3.2 The influence of age and ageing

Biometric performance can vary substantially with a person's age. Physiological properties such as skin elasticity and bone structure change with age, meaning the process of ageing can change a subject's biometric characteristics, causing a general decline in comparison performance[4][5][6]. Behavioural aspects of very young or very old subjects can impact the comprehension of instructions, thereby affecting the usability of a biometric system[7][8].

Comparison performance depends on both age and ageing. For example, enrolment and subsequent verification of fingerprints can be much less accurate for children than adults[4][5][6].

4.3.3 The influence of race and ethnicity

The performance of some biometric modalities has been shown to vary with race/ethnicity. This is particularly true for modalities and algorithms where biometric features depend on anatomical traits formed under genetic expression. Most studies seem to use broad categories of ethnicity based on the country of origin or nationality of an individual, or sometimes based on their personal declaration of ethnic identity. In a few studies, images of the individuals have been examined and classified by a human into specific ethnic groups. Both self and group classifications are likely sub-optimal in terms of consistency and description of the underlying physiological or behavioural effect.

4.4 Measurement and analysis

Given sufficient amounts of subject-specific demographic data and associated biometric processing results (e.g. comparison scores, transaction times, or recognition decisions), an analysis can be performed to expose the effects of demographics on biometric processing outcomes. At a high level, in an investigation of demographic effects in biometric systems, it is important to consider answering the following questions, prior to execution:

1) Is this study investigating effects in the mated distribution (false negative differentials) or non-mated distribution (false positive differential)?

2) Is this study investigating differential performance (i.e. score level, feature level fusion, low or image level fusion, etc.) effects or differential outcome (i.e. FNMR, TIR, etc.) effects?

3) Is this study using self-reported, categorical demographic variables or more descriptive phenotypes?

4) Are there any uncontrolled demographic variables that are confounded with the demographic variable under study? For example, on average women are shorter than men. In a study comparing the biometric performance of men and women, could any observed effects in the study be due to the confounding variable (height) and not the studied variable (gender or sex)?

Once these questions have been answered, the appropriate study description and statistical techniques for determining the presence and significance of an effect can be selected.

5 Impact of demographic factors on facial recognition systems

5.1 Existing literature on demographic factors impacting facial recognition systems

5.1.1 General notes

There are publications available on the demographic factors impacting the performance of facial recognition systems, based on some databases containing facial images from passports, visa mugshots, and driver's licences. Some factors, such as plastic surgery, the wearing of glasses, use of face covering aids, and significant changes to the face due to makeup or hairstyle are known to affect the comparison performance of facial recognition systems. Such factors are not precisely demographic factors but are influenced by them.

The following eight references were selected to highlight some of the key findings in the existing literature showing the impact of demographic factors on the comparison performance of facial recognition systems.

5.1.2 Factors that influence algorithm performance in the Face Recognition Grand Challenge

Reference [9] looks at the comparison performance of three comparison algorithms on facial images from the NIST Facial Recognition Grand Challenge. The specific data set used consisted of enrolment images taken under controlled lighting and verification images taken under uncontrolled lighting for 351 individuals from the University of Notre Dame. Each individual had multiple enrolment and verification images. This study examined numerous effects, including the impact of age, gender and race.

The investigated algorithms showed a strong correlation between the age of the test subject and the likelihood of successful verification, suggesting that 1:1 facial recognition gets easier as the subject ages. The study also showed that there is a slightly higher probability of correct verification for males than for females.

The data set was also divided by race, but only Caucasians and East Asians formed groups of significant size. Between these groups in the study, the probability of correct verification for East Asian subjects was slightly higher than for Caucasian subjects.

Overall, this study[9] found that the magnitude of the effect of demographic covariates varied across algorithms.

5.1.3 Face recognition performance: Role of demographic information

Reference [10] looks at the comparison performance of six algorithms on a large database of around one million mugshot images. Each mugshot came with complete demographic information on the subject, and the study extracted a specific subset of 102 942 images which were broken into cohorts by age, gender, and race (here: White, Black and Hispanic). These cohorts were selected so that all of the other demographic factors except the one being evaluated were approximately constant across the cohorts.

At a fixed FMR of 0,1 %, all investigated algorithms had a lower true accept rate (1 – FNMR) on the 18 - 30 age group than on the other two age groups (30 - 50, and 50 - 70). There was less difference between the 30 - 50 and the 50 - 70 age group. This observation suggests that older people can be easier to recognize, but that the decreasing difficulty with age levels off somewhere in middle age.

Considering race, the true accept rate at an FMR of 0,1 % was lowest for the Black cohort, higher for the White cohort and then again higher for the Hispanic cohort.

Finally, at an FMR of 0,1 %, the true accept rate for females was lower than for males. This observation suggests that there might be an intrinsic difference between males and females that makes females more difficult for facial recognition algorithms to recognize. Overall, this study[7] found the potential for false negative differential outcomes in three demographic cohorts; the young, African Americans, and females.

5.1.4 Issues related to face recognition accuracy varying based on race and skin tone

Reference [11] investigated the impact of sex and race on biometric performance of facial recognition systems. The study used the MORPH facial image database and the open-source VGGFace and ArcFace recognition algorithms.

The study[11] found that at a fixed, operationally relevant decision threshold, the benchmarked algorithms exhibit a lower biometric performance for females in general, a higher FMR for African-Americans, and a higher FNMR for Caucasians.

5.1.5 Report on the FG 2015 Video Person Recognition Evaluation

Reference [12] looked at facial recognition of subjects in videos acquired as part of the Point-and-Shoot Face Recognition Challenge Problem (PaSC). These videos were acquired over a seven-week period in Spring, 2011 on the campus of University of Notre Dame. They represent a very challenging problem for facial recognition algorithms since they were acquired in a range of environments, both indoors and outdoors, and by a variety of hand-held cameras. The challenge was further heightened because the

subjects were not looking directly at the cameras but were engaged in activities involving movement such as swinging a golf club or blowing bubbles. By design, there were few clear frontal views of the faces of the subjects. Each video featured one of the 265 test subjects performing one of seven actions in one of six locations and was captured by one of five hand-held cameras and by a higher resolution tripod mounted camera. This resulted in 1 401 tripod mounted videos and 1 401 handheld videos. Each comparison algorithm was given two videos and required to return a comparison score indicating the likelihood that the subject in both videos was the same. This resulted in 3 128 match pairs and 977 572 non-match pairs of videos for each of the handheld and tripod mounted data sets. Five algorithms were then used to evaluate comparison performance.

This study[12] is different from the previous studies because it operated in a very challenging part of the spectrum of facial recognition and because it operated by direct comparison of one entire video clip to another, rather than by comparison of single images to one another. Despite this, all five algorithms showed false negative differential outcomes where males were easier to recognize than females and Asians were easier to recognize than Caucasians.

5.1.6 Demographic effects on estimates of automatic face recognition performance

Reference [13] looked at data from the Facial Recognition Vendor Test (FRVT) 2006 conducted by the NIST in the US. In this study, the focus was on how the demographic distribution of the non-match pairs would impact the comparison performance. This is an interesting question and distinguishes this from the studies in the previous references.

Two data sets were selected from the FRVT 2006 data. Set 1 was captured with a 6 Megapixel camera under uncontrolled lighting conditions. Set 2 was captured with a 4 Megapixel camera under a mixture of controlled and uncontrolled lighting conditions. The comparison performance on each data set was measured by fusing together the comparison scores from the three highest performing algorithms tested in FRVT 2006. Then the data sets were broken down into three strata of image difficulty based on the comparison scores for each image pair.

In the study[13], the true accept rate at a specific FMR were lower when the non-match pairs were more closely compared because the comparison threshold had to increase in order to maintain a specific FMR as the non-match pairs became more similar.

The study then analysed the results in a number of ways, looking at changes in the percentages of a particular demographic in the non-match distribution and looking at cases where the non-match faces were of one race and the matching faces of a different race.

This research was limited to studies of one-to-one verification outcomes. The demographic effects reported there do not automatically apply to identification systems using the same underlying recognition algorithms. The general conclusion was that if the non-match faces are more similar to the genuine face, then the chance of a false match goes up at a fixed match threshold and so the choice of an appropriate threshold might have to be re-assessed periodically and adjusted as necessary.

The study suggested two practical lessons for operational systems. Firstly, in order to correctly estimate the comparison performance of a facial recognition system, it can be tested on a mixture of both genuine and imposter face pairs that correctly model the demographic composition of the individuals who will ultimately be using the system. Secondly, since most systems operate using a fixed threshold, an imposter can significantly increase their chance of being falsely matched by the system, if they attempt to match against an individual with their approximate demographic characteristics.

NOTE 1 The above research often stated demographic effects in terms of FNMR for specific groups at a specific FMR. This is seldom the correct approach because operationally the threshold is the fixed quantity and then FNMR and FMR both vary with demographics.

5.1.7 US National Institute of Standards and Technology (NIST) FRVT test

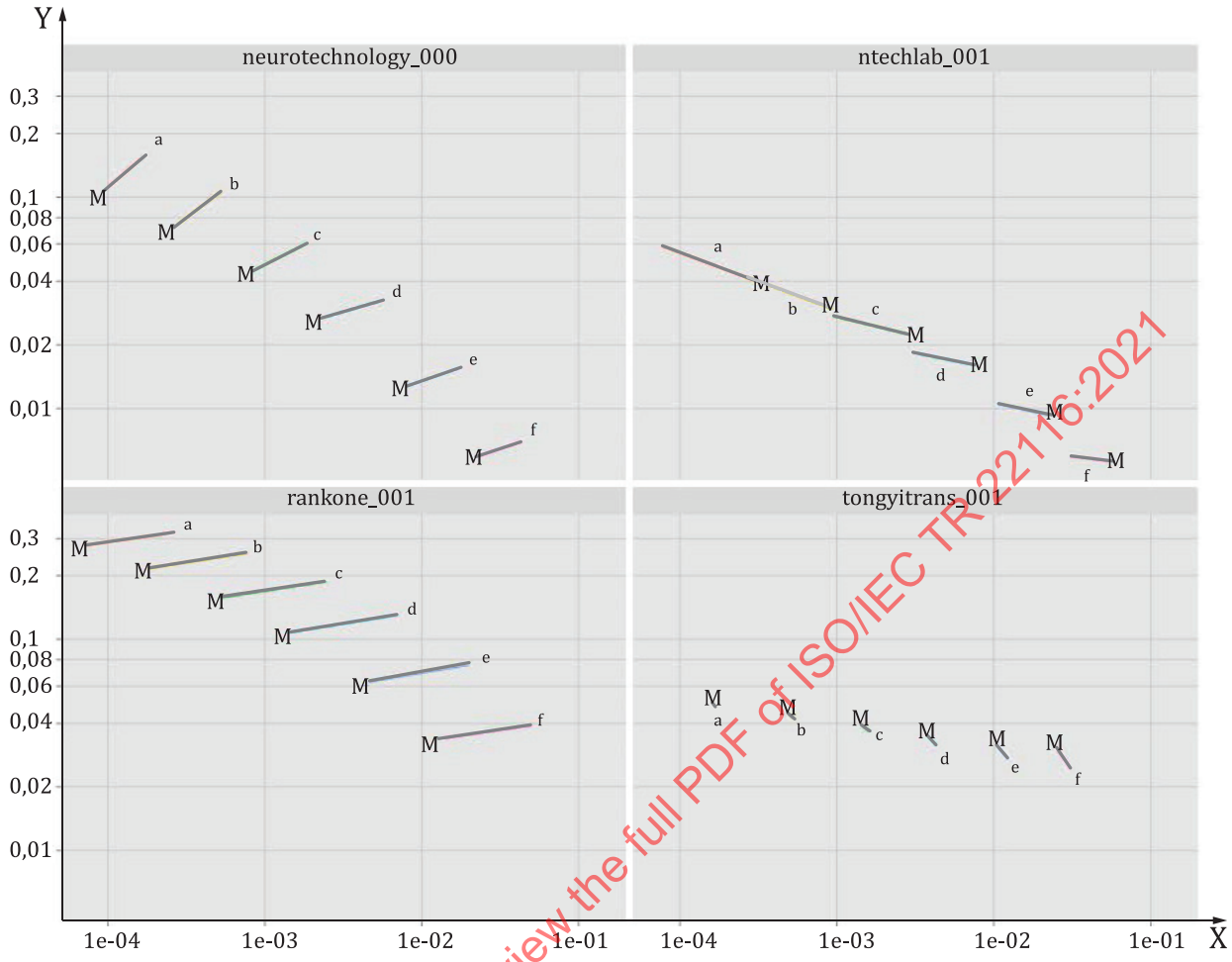
NIST's ongoing evaluation of face recognition algorithms [6][14] allows developers to submit algorithms at any time. These are tested using several databases, including a set of nearly 250 000 visa photographs.

Each photograph is accompanied by the subject's age, country-of-birth and gender. For each algorithm tested, NIST reports multiple metrics related to comparison performance, including the following:

— How FMR and FNMR differ by gender. Figure 1 shows, for four algorithms, the FNMR and FMR measured for each gender at six different possible operating thresholds. At each point a line is drawn between (FMR, FNMR) MALE and (FMR, FNMR) FEMALE showing which gender has lower FMR and/or FNMR. The "M" label denotes male, the other end of the line corresponds to female. The six operating thresholds are selected to give the nominal FMRs given in the legend, and are computed over all impostor pairs regardless of age, gender and place of birth. The plotted FMR values are broadly an order of magnitude larger than the nominal rates because FMR is computed over demographically-compared impostor pairs (i.e. individuals of the same gender, from the same geographic region and age group). The plots show that FMRs for men and women differ, with men giving lower FNMR for three of the four algorithms, and also lower FMR for three of the four algorithms.

— How FMR depends on the age of the impostor and the enrolee. For the algorithm neurotechnology_000 operating on visa images, the heatmap in Figure 2 shows FMRs observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0,001 over all ten billion (10^{10}) impostor comparisons. The text in each box gives the same quantity as that coded by the colour. Light colours present a security vulnerability to, for example, a positive verification system. Figure 2 shows, for age groups spanning infant to elderly, how FMR is larger for the very young and very old, and is very low when an impostor is of a different age to the enrolee. In particular, note that the FMR for children under four with matched gender and region imposters is $10^{-0.4}$ (see green square in Figure 2) or 40 % instead of the nominal 0,1 % that is measured for all imposters.

— How FMR depends on the country of the birth of the impostor and enrolee. Figure 3 shows the FMR expected when an impostor from one region of the world is compared to an enrolee from another. For the algorithm ntechlab_001 operating on visa images, the heatmap shows FMRs observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0,001 over all ten billion (10^{10}) impostor comparisons. If text appears in each box it gives the same quantity as that coded by the colour. Grey indicates that FMR is at the intended 0,001 level. Light red colours indicate FMR > 0,001 which represents a security vulnerability to, for example, a passport gate. FMR varies globally, with high FMR in East Asia, South Asia and sub-Saharan Africa. In the worst case, FMR for South Asians compared against other South Asians with the same gender and approximate age is $10^{-1,1}$ (see green square in Figure 3) or 8 % rather than the nominal 0,1 % for all imposters. The FMR is also high when comparing across certain regions, such as when imposters from the Caribbean are compared against enrollees from sub-Saharan Africa or vice versa. NIST's report[14], which includes analogous figures for all algorithms tested, shows that most algorithms exhibit wide variations in FMR geographically. This likely occurs as face structure is known to have a genetic component.



Key

X false match rate (FMR)

Y false non-match rate (FNMR)

a fmr_nominal 0,00001

b fmr_nominal 0,00003

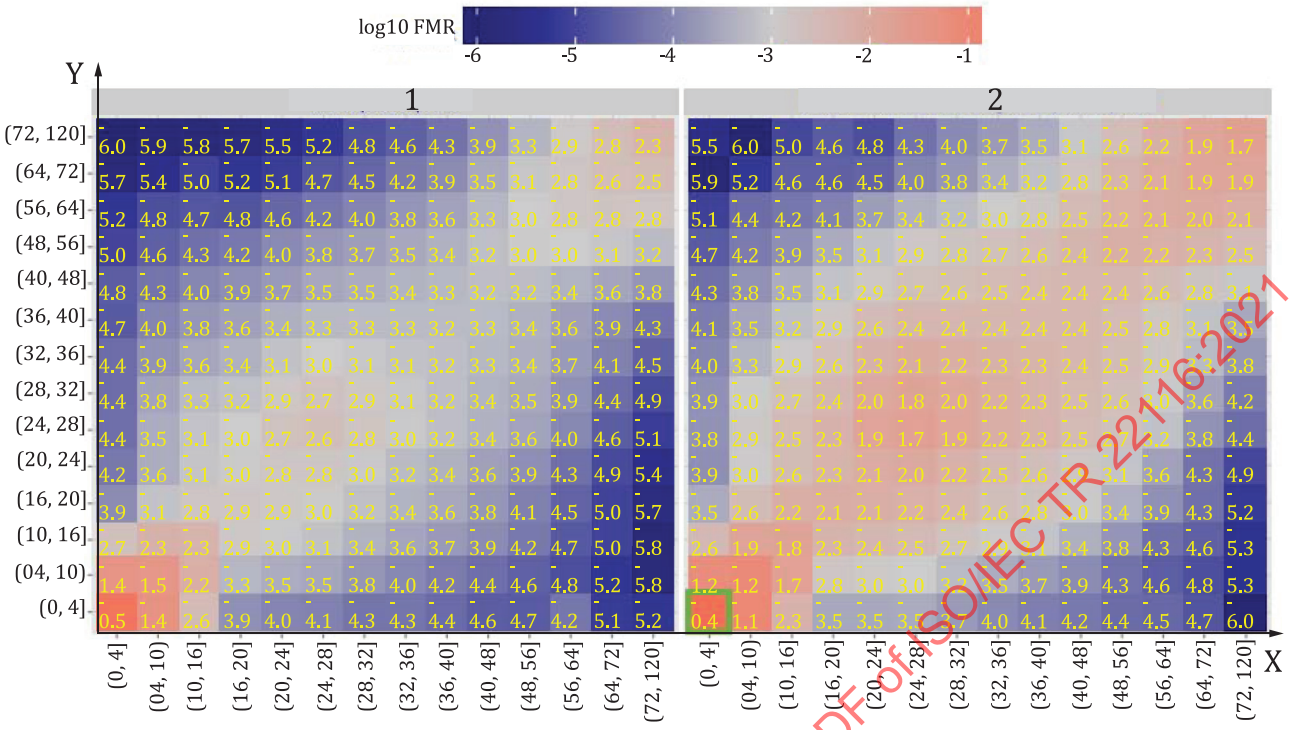
c fmr_nominal 0,00010

b fmr_nominal 0,00030

e fmr_nominal 0,00100

f fmr_nominal 0,00300

Figure 1 — FNMR and FMR performance for four different algorithms at six operating thresholds (M – "male", opposite line end – "female")



Key

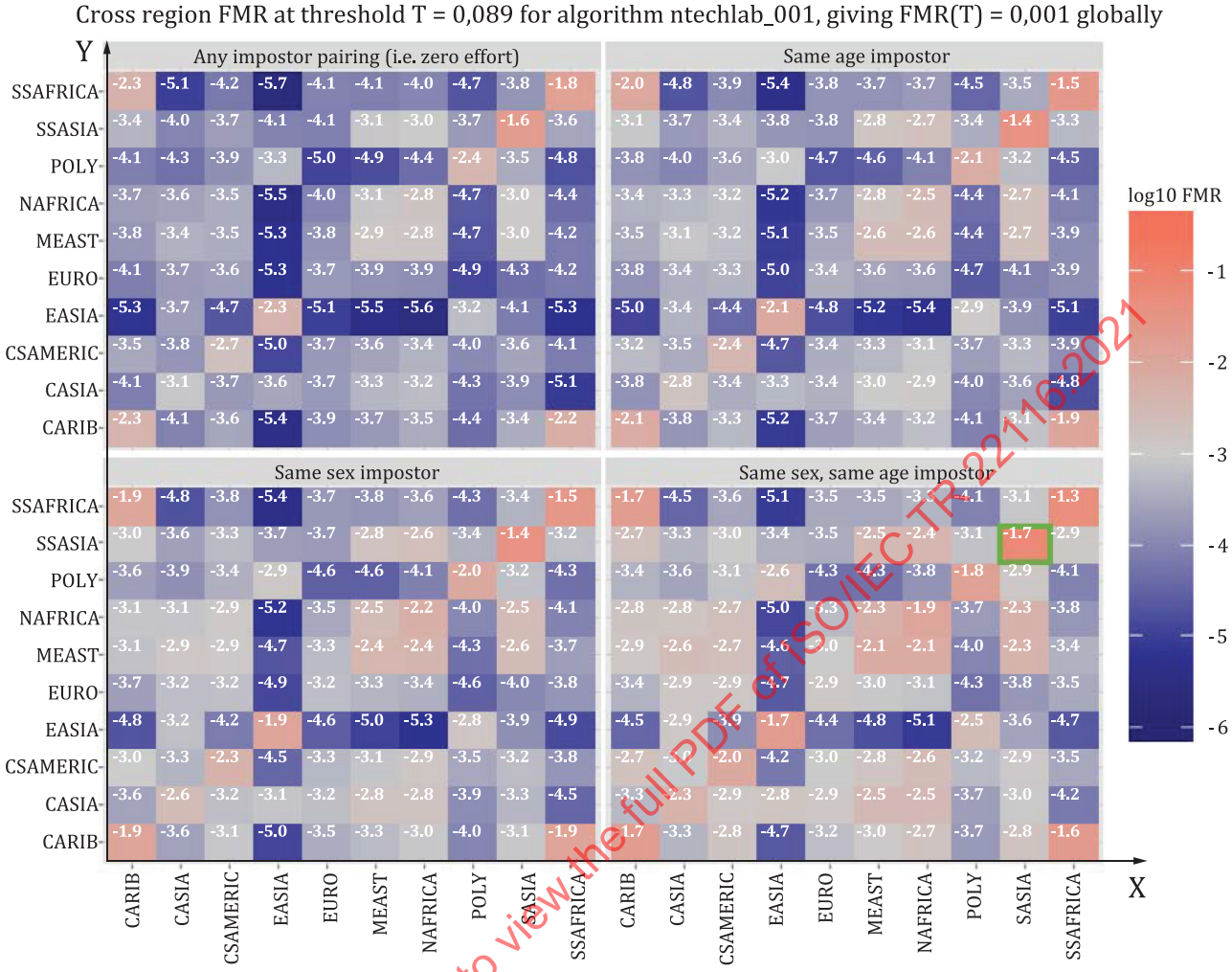
X age of enrollee

Y age of imposter

1 all imposter pairs

2 same sex and same region imposter pairs

Figure 2 — Heatmap of false match over the face impostor comparisons from different individuals with the given age pair



Key

X region of birth of enrollee

Y region of birth of impostor

Figure 3 — Heatmap of false match over the face impostor comparisons from different individuals born in the given region pair

5.1.8 Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems

Reference [15] looked at data from the 2018 U.S. Department of Homeland Security Biometric Technology Rally. In this study, a population of 363 individuals were collected using 11 different face acquisition systems. These samples were then compared to a gallery of historic images of those same subjects from 1 - 4 years prior. The authors also created an automated means of estimating the skin reflectance of the subjects in their population. The results showed that both the efficiency and effectiveness of acquisition systems was impacted by multiple demographic covariates, including gender, age and skin reflectance. Skin reflectance had the strongest net linear effect on performance. Linear modelling showed that lower (darker) skin reflectance was associated with lower efficiency (higher transaction times) and accuracy (lower mated comparison scores). Skin reflectance was also a statistically better predictor of these effects than self-identified race labels. Unlike other covariates, the degree to which skin reflectance altered accuracy varied between systems. Reference [15] also documented an effect of gender, showing that mated comparison score for males was on average higher than for females, but only when compared to historical gallery images. When probe images were compared to existing same-day images, there was no impact of gender on mated comparison score.

In conclusion, this study[15] showed that the skin reflectance effect was inversely related to the overall accuracy of the system such that the effect was almost negligible for the system with the highest overall accuracy. This study also showed that gender effects can be strongly linked to behaviour and not underlying physiological differences between genders. These results suggest that, in evaluations of biometric accuracy, the magnitude of measured demographic effects depends on image acquisition.

5.1.9 The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance

Reference [16] is one of the few studies that looked at false positive differentials. The authors used non-mated face pairs collected as part of the 2018 U.S. Department of Homeland Security Biometric Technology Rally and showed that demographic variables were highly correlated with false-positive outcomes in a diverse population. FMR at a fixed threshold increased >400-fold for broadly homogeneous groups (individuals of the same age, same gender and same race) relative to heterogeneous groups. However, for specific demographic homogeneity, the highest FMR was observed for older males that self-identified as White and the lowest for older males that self-identified as Black or African American. The magnitude of FMR differentials between specific homogeneous groups (<3-fold) was modest in comparison with the FMR increase associated with broad demographic homogeneity.

5.2 Summary of demographic impact on facial recognition systems

There is a strong correlation between the age of the subject and the ease of facial recognition. In 1:1 comparisons, at a fixed match threshold, FNMR becomes lower as the age of the subject increases. This is very significant for those under 18, but becomes less significant after middle age. All algorithms seem to exhibit this correlation. The NIST FRVT study also shows that for 1:1 comparisons, FMR can become very high as the age of the subject decreases. In particular, at a fixed threshold that provided an average FMR for all subjects of 0,1 %, children under four had an FMR with other children under four of 30 %. Most facial recognition algorithms have lower false rejection rates for males versus females; although Reference [15] suggests that this difference is driven by behavioural, not physiological, differences.

Some facial recognition algorithms appear to be better at recognizing certain ethnic groups. In one study, the difference in FNMR at a fixed FMR between ethnic groups was 10 % or more, which means that facial recognition systems can experience higher error rates for some groups. On other algorithms and capture systems, this effect does not exist, showing the importance of testing facial recognition systems on diverse subject populations that mimic the populations expected to interact with the system in operations.

More importantly, however, the NIST FRVT and Reference [16] studies show false positive differentials in facial recognition systems when the subjects being compared are of the same ethnicity and even more if the subjects are of same gender, same ethnicity and similar age. The NIST study shows that, for some ethnic groups, this can increase the FMR at a fixed threshold from 0,1 % to 8 %.

If the demographics of the subjects being recognized by a facial recognition system change significantly, then the biometric performance of the system will change, so it is important to test systems using a group with similar demographics to the subjects that will be part of the deployed system. An algorithm that can maintain a stable FMR for different demographic groups can help to mitigate any security risk.

5.3 Recommendations for facial recognition systems

Facial recognition relies on a biometric characteristic that is heavily influenced by genetic factors. As such, error rates of facial recognition systems will be influenced by certain demographic factors. Broadly speaking, everyone is more likely to match individuals that share their demographic covariates.

Facial recognition also relies on sampling an inherently noisy biometric characteristic. The face is subject to physiological, behavioural and environmental changes. Some of these factors are also influenced by demographics. Some studies have found facial recognition has lower false non-match

rates amongst men and older people, perhaps reflecting behavioural correlations with demographics. In consideration of these effects, the following recommendations are made for facial recognition systems.

— Performance of a facial recognition system can be tested using subjects with a similar demographic mix to those that will eventually use the system. Specific security issues can be tested with specific subsets of this demographic mix.

— Any performance test for a facial recognition system can report the gender, age and race or country of birth (as a proxy for ethnicity) of the test subjects.

— Security settings for a facial recognition system might take into account the demographics of the subjects and make appropriate adjustments. An algorithm which is designed to maintain a stable FMR across multiple demographic factors could also be useful if available.

— Any performance test for a facial recognition system can evaluate the stability of FMR for a given threshold with respect to changes in gender, age and race or country of birth (as a proxy for ethnicity) of the test subjects.

6 Impact of demographic factors on fingerprint systems

6.1 Existing literature on demographic factors impacting fingerprint systems

6.1.1 General notes

There is a long history of using fingerprints as part of criminal investigations. Most of the early development of large-scale automated biometric identification systems was to support fingerprint identification for criminal investigations. Several factors that affect the performance of fingerprint recognition systems have therefore been known for many years. Some of these factors can be correlated with demographics. These include the humidity of the skin, the level of wear to the friction ridges caused by daily activities of the subject (e.g. hard physical labour or working with caustic chemicals) and the ability of the subject to physically present their finger in a stable and flat position on the fingerprint scanner.

In many cultures, hard physical labour is more likely to be performed by certain genders. Manual labour can result in friction ridge degradation over time. This means that worn fingerprints are also affected by gender, but differently in different locations. Ability to properly present a fingerprint is also highly correlated to age, with very young children and elderly people lacking the motor skills to make a good presentation.

There is also some highly relevant literature which investigates the impact of demographic factors on fingerprint recognition, as indicated in the following subclauses.

6.1.2 IDENT/IAFIS image quality study

Reference [17] is a review of the Image Quality Study conducted by MitreTek Systems in 2000 to examine the use of the Federal Bureau of Investigation's Integrated Automated Fingerprint Identification System (IAFIS) to be used as part of large-scale visa processing. The testing used four different databases with a mixture of slap and rolled fingerprints and investigated both image quality and comparison performance. Although their results are specific to the quality and comparison algorithms that were in use by the IAFIS at that time, several of the findings can have general applicability.

Firstly, the researchers found that there was poor correlation between the image quality score and the comparison performance for an individual fingerprint image, except for those fingerprints with very low-quality scores which were unlikely to be successfully matched. Secondly, they found that female fingerprints are likely to have lower quality scores than male fingerprints. Specifically, 6,7 % of the fingerprint images of females were classified as "Very Poor" and thus unsuitable for use with the IAFIS, as opposed to only 2,4 % of male fingerprints. Reference [17] states that ridge flow and classification quality were notably worse for females and were the expected reason for decreased fingerprint image

quality. Specifically, they found that females had smaller ridge diameter, smaller ridge frequency and shallower ridges than men, which made it more difficult to determine ridge flow. This led to the conclusion that "Clearly, performance and throughput will be engineering challenges for systems with large female populations."

NOTE While historically important to mention, the research outlined in Reference [17] is over 20 years old and thus does not necessarily represent the current state of algorithms used by the IAFIS, now referred to as the Next Generation Identification (NGI) system.

6.1.3 Impact of gender on fingerprint recognition systems

Reference [18] collected data from 244 subjects, capturing three images of each of their right index, left index, right middle and left middle fingers with each of an optical and a capacitive fingerprint sensor. A commercially available image quality software was used to measure the quality of every image on a scale from 0 to 100. The mean quality score for males was 71,8 for the optical sensor and 71,3 for the capacitive sensor, whereas for females it was 63,1 for the optical sensor and 59,7 for the capacitive sensor. This was a statistically significant difference and is similar to what the IAFIS quality algorithm found in Reference [17]. Next, the researchers broke the data into four subsets, males on optical sensor, males on capacitive sensor, females on optical sensor and females on capacitive sensor. Each subset was then compared against itself and a DET curve was generated.

For both the capacitive and the optical sensor, the comparison performance for the female subset was better than for the male subset, in that the FNMR at every value of the FMR below 0,1 % was lower for the female subset. This is a surprising finding as the image quality score might be at least somewhat correlated to comparison performance. In this case, however, the image quality score for female fingerprints was lower even though the comparison performance for female fingerprints was better. The degree of accuracy differential for Reference [18] was not discussed; however, it recommends further statistically relevant testing to determine such.

6.1.4 Impact of gender on image quality, Henry classification and performance on a fingerprint recognition system

Reference [19] was a continuation of the work started in Reference [18]. In this case a group of 115 males and 81 females was used. Their fingerprint images were captured using a different optical sensor than in the previous experiment and the image quality and fingerprint comparison software were updated since the previous experiment. They also captured more fingers from each test subject, specifically three images of each of the left and right index, middle, ring and little fingers. Despite these differences, the results were very similar, in that the male fingerprints showed higher image quality scores but the females fingerprints performed better. In this case, the equal error rate (where FMR = FNMR) was 0,68 % for males and 0,42 % for females. Reference [19] recommends further testing in regards to gender and Henry classification.

6.1.5 Impact of age and ageing on sample quality and performance in fingerprint recognition systems

Reference [4] analysed a database of around 400 000 fingerprint images from around 250 000 different fingers acquired under real operational conditions (passport issuance). The database contains fingerprints from subjects aged between 0 and 98 years. The study analysed the sample quality scores using three different algorithms: the open source NFIQ 1.0 and NFIQ 2.0, as well as the commercial VERIQ. The main findings of the study were:

1) The elderly age group (65+ years) is the most challenging one in terms of image quality. In the age range 65 - 90, a linear decrease of sample quality can be observed, with around 1 % quality loss per year.

2) The sample quality for children is low, but it increases rapidly with age until the age of 12. From that age onwards, the quality stabilizes and is similar to that of young adults (18 - 25).

3) The sample quality for young adults (18 - 25) is the highest from the tested groups and appears to be quite stable. It can be noted that the results in the age range of 26 - 64 years have been extrapolated and can be verified experimentally in a follow-up study. The modelling used in the study predicts a relatively stable sample quality until the age of around 40 - 45 years, where after a linear decrease in quality (see point 1) is expected to happen.

Based on those findings, Reference [4] identifies the following quality zones for the fingerprint images depending on the subjects' age:

— 0 - 4: low quality

— 4 - 12: medium quality

— 12 - 70: high quality

— 70 +: low quality

Reference [5] analysed a database of 183 data subjects with fingerprint impressions gathered over a timespan of approximately 40 years. An evaluation in verification mode was conducted to ascertain and quantify the effects of ageing on the comparison score distributions and biometric performance. For all the tested algorithms, a decrease in biometric performance has been observed when comparing samples acquired at different times. For samples with a time delta of 10 or more years, the FRR (at $0 < FAR < 0,1$) has roughly doubled.

6.2 Summary of demographic impact on fingerprint systems

Fingerprint image quality algorithms tend to give lower scores to female fingerprints, but the comparison performance of female fingerprints is often higher than that of male fingerprints. This suggests that quality algorithms can be improved to better normalize some factors linked to gender like ridge flow and ridge density.

In regards to age, fingerprint image quality tends to be lower for children. Around 12 years of age, fingerprint image quality becomes more stable and of higher quality, and remains so until approximately age 45, with a gradual linear descent in quality after. In regards to ageing, fingerprint recognition performance decreases when samples are acquired at different times, with FRR doubling after a time delta of 10 years.

6.3 Recommendations for fingerprint systems

All fingerprint quality algorithms can be tested on separate sets of fingerprints from individuals of each of the male and female sexes, as well as individuals from different age bins, and the comparison performance on each set can be compared to the quality scores for each set. It can be a design goal of quality algorithms to achieve similar correlation between performance and quality for both male and female fingerprints, as well as members of varying age. Additionally, integrators of fingerprint systems can consider ageing when defining gallery policies, such that time deltas between gallery and probe images does not exceed 10 years, when possible for the application. This might not be possible for all applications. For example, renewal requirements would not be possible for most black-list or police investigation systems.

7 Impact of demographic factors on iris recognition systems

7.1 Existing literature on demographic factors impacting iris recognition systems

7.1.1 General notes

There is less information available in this context for iris recognition than for fingerprint or facial recognition. There are, however, some factors that are strongly correlated with demographics which are known to affect the performance of iris recognition systems. Most iris recognition cameras expect

that the individual whose iris images are being captured are following the directions given by the system. They prompt the user to look in the right direction and move their head into the relatively small focal volume of the camera. Young children can have difficulty following these directions due to level of education, mental acuity and previous exposure to other technical devices. Experience from the Unique Identification Authority of India (UIDAI) suggests that children under 4 are unable to properly follow the directions for iris capture[20]. This is also a problem for people who have poor eye-sight, as visual cues are typically used to help position the eyes in the focal volume of the iris camera.

As the population ages and visual capability tends to degrade, there can be an impact on usability. Individuals can be unable to perceive instructions or cooperate effectively with the capture system as the average age of the target population increases. Additionally, References [21] and [22] show that under objective metrics, the sample quality of iris images (both near-infrared and visible wavelength spectrums) is deteriorated by the presence of eyeglasses. As a consequence, the biometric performance of such systems can decrease. In this context, the explosive increase of myopia prevalence in recent years means that the number of people wearing eyeglasses is high and growing[23]. Finally, the use of textured contact lenses, which change the colour or look of the wearer's eyes in the visible spectrum, can have a significant impact on iris recognition performance. It is generally recommended that the use of such cosmetic contact lenses not be permitted with iris recognition systems.

7.1.2 The Canadian NEXUS system

In 2004, the Canada Border Services Agency (CBSA) began deployment of a system to authenticate pre-registered travellers in airports. This system, called NEXUS-Air, consisted of kiosks using iris recognition to authenticate the identity of travellers entering Canada or entering the US at a pre-clearance site located in a Canadian airport. A total of 69 iris recognition kiosks were installed at eight Canadian airports. After being approved for participation in the programme, each traveller attempts to enrol both their left and right eye at an enrolment office where they are given instruction on how to use the system. If one or both of their eyes are successfully enrolled, then these are stored independently in the enrolment database. When a traveller uses a kiosk, they do not make a claim of identity. Instead, they present their eye to the kiosk camera and an iris template is generated and compared with all stored left and right iris templates. If any template matches the presented iris template beyond a specific threshold, then the traveller's identity is confirmed as the identity associated with that iris template in the database.

References [7] and [8] examined iris imagery from the OPS-XING database created by the CBSA. This database was developed from transaction logs of the NEXUS system for 705 553 distinct NEXUS participants who used the system to cross the border at airports. The enrolment data goes back to 2003. Between 2007 and 2015, 467 314 travellers had encountered the system, with some using the system over 60 times. OPS-XING includes quality metrics, Hamming distance, and other ancillary data, but the only demographic information provided is the age of the traveller. Limited portions of OPS-XING have been provided to NIST and to some academic researchers.

The age range is very wide as there is no age restriction on participation in NEXUS. In one case, a child was enrolled at the age of 8 months and verified at a kiosk at 9 months, 12 months and 14 months. In another case, a 99 year-old traveller was able to use the system. This study showed some very interesting data related to the impact of age on biometric performance.

During enrolment, if only a single iris meets the quality threshold, then the traveller will be enrolled with only a single eye rather than two eyes. The number of travellers who enrolled in the system and the percentage of them who were only able to enrol a single eye are shown below in Figure 4. Travellers who could not enrol either eye are not included in the OPS-XING data as they would never have used a NEXUS kiosk. It is clear from this figure that successful enrolment of both eyes is much more difficult for children under 14 and for adults over 60. This means that iris recognition might not be a recommended modality for populations with large percentages of young children or elderly people.